remote sensing

Article

Towards Robust Semantic Segmentation of Land Covers in Foggy Conditions

Weipeng Shi [1], Wenhu Qin [1,*] and Allshine Chen [2]

[1] School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China
[2] Health Sciences Center, University of Oklahoma, Oklahoma City, OK 73106, USA
* Correspondence: qinwenhu@seu.edu.cn

Abstract: When conducting land cover classification, it is inevitable to encounter foggy conditions, which degrades the performance by a large margin. Robustness may be reduced by a number of factors, such as aerial images of low quality and ineffective fusion of multimodal representations. Hence, it is crucial to establish a reliable framework that can robustly understand remote sensing image scenes. Based on multimodal fusion and attention mechanisms, we leverage HRNet to extract underlying features, followed by the Spectral and Spatial Representation Learning Module to extract spectral-spatial representations. A Multimodal Representation Fusion Module is proposed to bridge the gap between heterogeneous modalities which can be fused in a complementary manner. A comprehensive evaluation study of the fog-corrupted Potsdam and Vaihingen test sets demonstrates that the proposed method achieves a mean $F1_{score}$ exceeding 73%, indicating a promising performance compared to State-Of-The-Art methods in terms of robustness.

Keywords: semantic segmentation; attention mechanism; robust deep learning; remote sensing; data fusion

check for updates

Citation: Shi, W.; Qin, W.; Chen, A. Towards Robust Semantic Segmentation of Land Covers in Foggy Conditions. *Remote Sens.* **2022**, *14*, 4551. https://doi.org/10.3390/rs14184551

Academic Editor: Gwanggil Jeon

Received: 5 August 2022
Accepted: 8 September 2022
Published: 12 September 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Computer vision has emerged as a powerful and labor-saving tool for automatic scene parsing of remote sensing images (RSIs). Land cover classification (LCC) of aerial imagery, which is also known as semantic labeling or segmentation, assigns a class label to each pixel in RSIs. As an integral part of computer vision, semantic segmentation plays a pivotal role in remote sensing for rapid and accurate detection. A considerable amount of literature has been published concerning applications like landslide extraction [1], road extraction [2], collapsed building detection [3], and so on. Semantic segmentation models can be summarized into three categories [4], namely FCNets [5] which yield a coarse feature map directly from the low-resolution representation, UNets [6] which perform high-resolution recovery from the downsampled representation as well as HRNets [4] which retain the high-resolution representation during all procedures. Despite plenty of research exploring semantic segmentation based on natural scene images (NSIs), there is still a lack of scientific literature specifically focusing on robust remote sensing in foggy conditions.

RSIs with fog are distinguished from NSIs by a number of challenging characteristics that may result in decreased classification robustness. It is believed that there exist two major challenges when it comes to LCC with fog. First, model robustness is susceptible to fog corrupted RSIs, which refer to a series of issues [7], including intra-class heterogeneity, inter-class homogeneity, geometric size diversity, and so on. In terms of intra-class heterogeneity, the models tend to classify objects with distinctive appearances as disparate species; they may belong to the identical yet [8]. For instance, various materials and structures may lead to different appearances and textures. Whereas, fog-covered objects affiliated with diverse species frequently exhibit close characteristics when they are made up of the same material. The concrete building and the impervious surface in Figure 1 appear similar. Models assume they belong to the same class by mistake accidentally, which is inter-class homogeneity. Additionally, objects under fog exhibit geometric size diversity and only a robust model can capture multi-scale attributes in RSIs. Overall, RSIs with low quality can exert negative impacts on classification robustness. Second, when dealing with dense fog, a single sensor is not always effective. It is difficult for a single optical camera to classify objects robustly when they are partially obscured. Figure 1 illustrates the failing cases, such as the shadow of buildings, fog coverage, and cars parked under the trees. The optical input is rich in semantic details, while Digital Surface Model (DSM) provides discriminative height information. It is imperative to excavate informative cues from multimodal inputs.

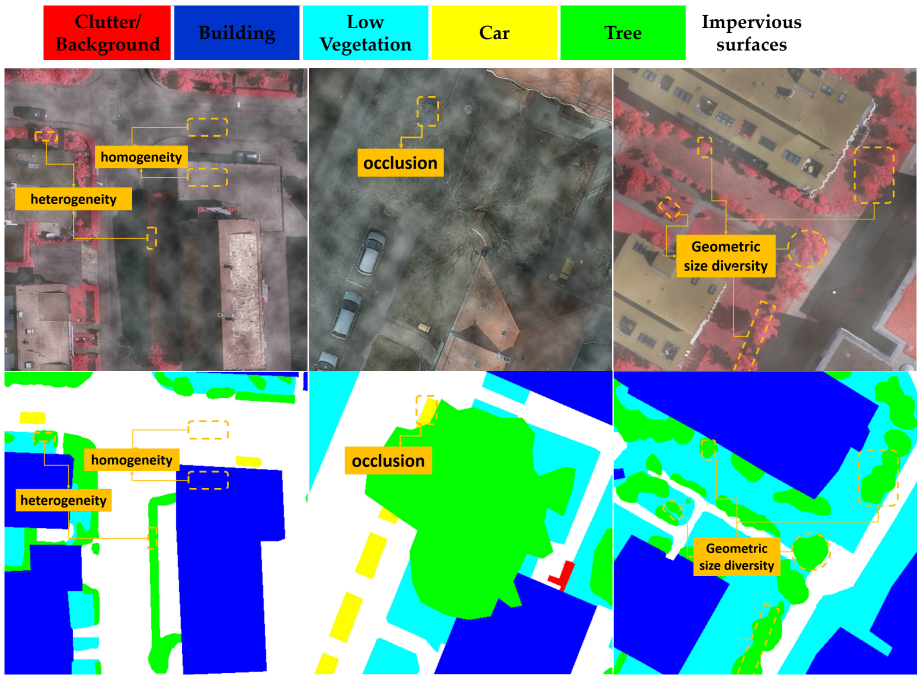

Figure 1. Challenges associated with robust LCC in the areas covered with fog. The first row is the fog corrupted image and the second row is the corresponding ground truth (GT).

To address the listed challenges of LCC in foggy conditions, we propose a framework with superior robustness based on attention mechanisms and multimodal fusion. We adopt HRNet as the backbone. Through compiling representations from all the high-to-low resolution streams in parallel, HRNet is robust to intra-class heterogeneity and geometric size diversity. The proposed Spectral and Spatial Representation Learning (SSRL) module probes into the relationship between spectral channels and spatial locations to improve robustness to intra-class heterogeneity. Thus, the output representation is gifted with semantic information and spatial accuracy. The introduced Multimodal Representation Fusion Module (MRFM) investigates the fusion of multimodal remote sensing data to learn the boundary connectivity and contour closure in RSIs to cope well with object occlusion and inter-class homogeneity issues. In summary, the main contributions are as follows:

- Based on multimodal fusion and attention mechanisms, we propose a robust end-to-end model that can fuse the optical and DSM input for LCC.
- Adopting HRNet as the backbone, we propose and incorporate SSRL and MRFM into the framework. To enhance the semantic information, a lightweight SSRL is inserted to capture the long-range dependencies and explore the interactions between various spectral channels. MRFM is employed for the effective fusion of multimodal remote sensing data. All the components cooperate and contribute to the classification robustness.
- We conduct an ablation study to evaluate the effectiveness of the proposed framework, including functions of different modules and modal inputs.

- We compare our model with SOTA methods to demonstrate the robustness against natural noise.

2. Related Work

There is a large volume of published studies describing how to conduct LCC. Most publications concentrate on accuracy instead of robustness, which is also vital in daily application. Thus far, several studies investigating robustness are predominantly associated with NSIs, such as scenes in ImageNet, Cityscape, BDD100k, and so on. Nonetheless, there is still a lack of relevant research focusing on LCC robustness. RSIs captured in foggy conditions are characterized by low quality, which poses challenges to robustness. This paper aims to design a robust model which can improve the classification performance in harsh environments. Our work refers to LCC, model robustness, and multimodal fusion. This section discusses the related work from these three perspectives.

2.1. Land Cover Classification

LCC actually refers to semantic segmentation of land covers using computer vision. There has been a great deal of research into semantic segmentation focusing on classification accuracy. Conventional segmentation algorithms are normally put forward on the premise of basic image attributes, e.g., grey-scale mutations are utilized to detect edges. Comparable grey scale values are partitioned into several regions according to the predefined criteria. However, it is extremely complex to detect boundaries when there exist substantial grey-scale changes. Considerable evidence has accumulated to show that deep learning-based models are more suitable for the semantic segmentation of NSIs. These models can usually be divided into three groups [4], namely, FCNet [5] type, UNet [6,9] type, and HRNet [4] type. FCNets learn the representation from high to low resolution in series to extract coarse feature maps. This group includes models like Deeplab [10], DenseASPP [11], PSPNet [12], and so on. UNets learn the encoded low-resolution representation and then recover to the high-resolution representation. Analogous models are DeepLabV3+ [13], SegNet [14], and so on. Moreover, ref. [9] inserts a cascaded dilated convolution in UNet to capture objects of diverse shapes, which is an effective approach to enhance robustness to multi-scale issues. Different from [9], we alleviate the influence of diverse shapes by adopting HRNet as the backbone because it retains a high-resolution representation throughout the process [15].

2.2. Model Robustness

Although neural networks are highly accurate for classification, they are not always as robust as human beings while actually applied [16]. Ref. [16] suggests that building multimodal and multitasking systems based on multi-sensor fusion is indispensable for robust decisions. Noises are categorized into three main groups [17], mainly adversarial noises, systematic noises, and natural noises. By comparing the robustness of three types of models, ref. [17] shows that CNN is more robust under natural noise and systematic noise, while Transformer is more robust against adversarial noise. Adversarial noise is the result of ambiguous decision-making at boundaries due to the limited training dataset and the inability to cover the whole sampling space. A small perturbation will always lead to completely distinct results. Ref. [18] constructed ImageNet-P and ImageNet-C datasets on top of ImageNet to facilitate researchers to evaluate and test the corruption and perturbation robustness. Based on this work, ref. [19] investigates the robustness of semantic segmentation. Researchers find that the Atrous Spatial Pyramid Pooling module significantly improves robustness, while the generalization performance depends heavily on the corruption degrees. Furthermore, there are some recent works on the adversarial noise robustness of Visual Transformers (ViTs). Ref. [20] found that shallow features in ViTs enable it to possess a better generalization than CNN, thus, better coping with adversarial noise. Meanwhile, the ensemble operation of CNN and ViTs can also improve the model robustness [21].

2.3. Attention Mechanism

A large and growing body of literature has investigated the role of attention mechanism [22] in deep visual models. Ref. [23] proposes an efficient channel module to explore the cross-channel interactions without dimension reduction. Ref. [24] designs a global context module to model the long-range dependencies with significantly less computation. There is a consensus among researchers that multi-head attention in ViTs[25] has acquired SOTA due to the uniform representation. Swin transformer [26] is capable of modeling input of different scales flexibly and the complexity is linear with input sizes. SETR [27] presents a multi-level feature fusion module to classify each pixel at a fine-grained level. Segformer [28] removes the complex position encoding binding a lightweight multi-layer perceptron to output feature maps of various sizes. Volo [29] introduces a novel outlook attention to grasp both coarse and fine-grained representations.

2.4. Multimodal Fusion

Data collected by a single sensor are often flawed and multi-sensor fusion is fundamental for accurate and robust decisions. Ref. [30] put forward a top-down pyramid fusion architecture for multimodal fusion. It is lightweight and can extract complementary features from multi-sources. Ref. [31] proposes a new and lightweight depth fusion transformer network for LCC, with different backbones extracting features from various inputs. Ref. [32] explores the pros and cons of early fusion along with late fusion to show that they all can utilize the complementarity of multimodal inputs. To calibrate features of the current modality from spatial and channel dimensions, ref. [33] has developed a Cross-Modal Feature Rectification Module for feature extraction. A multimodal fusion module was proposed in [15] to explore complementary features of heterogeneous inputs. In contrast, our method exploits the discriminate representation of each input from the perspective of the channel and spatial location before multimodal fusion, which greatly enhances the robustness of low-quality RSIs.

Overall, most studies remain narrow in focusing only on NSIs instead of RSIs. RSIs are characteristic of challenges of high resolution, multi-scales, class imbalance, occlusion, and so on. The attention mechanism is capable of capturing long-range dependencies. We extend ideas from deep learning and RGB-D semantic segmentation as well as an attention mechanism to establish a practical framework that can robustly classify land covers in foggy conditions.

3. Core of the Framework

3.1. Overview

The overall framework is illustrated in Figure 2. HRNetV2-W48 is adopted as the backbone for feature extraction. UperHead from [34] serves as the decoder. Each batch combines an Optical image ($\mathbf{X}_{\mathbf{RGB}} \in \mathbb{R}^{C\times H\times W}$) with the corresponding DSM ($\mathbf{X}_{\mathbf{DSM}} \in \mathbb{R}^{1\times H\times W}$), which contains the height information of land covers. H and W denote the height and width of RSIs, respectively. There is a considerable amount of noise in low-quality RSIs. We intend to design SSRL in such a manner that it would extract useful and discriminate representation efficiently without incurring excessive computational costs. Conventional methods simply aggregate two modalities without obtaining complementary features effectively. By contrast, MRFM, which is dedicated to multimodal fusion, exploits the complementarity between heterogeneous data. The backbone and decoder can be replaced by the other models. We will dig into the detailed design of HRNet, SSRL, and MRFM in Sections 3.2–3.4, respectively.

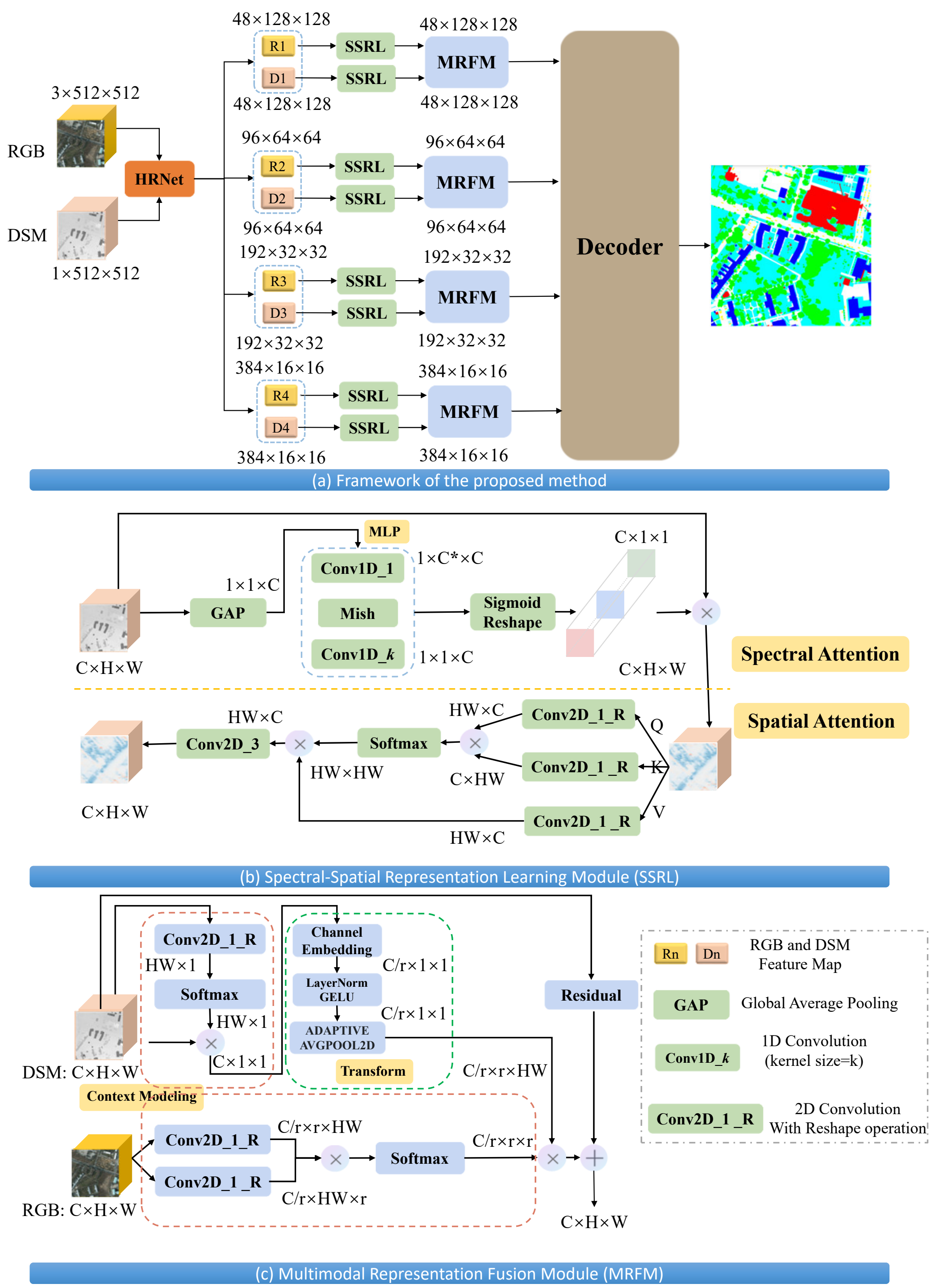

Figure 2. (**a**) Illustration of the proposed framework for LCC in foggy conditions. The input consists of a pair of the optical image and DSM. UperHead is selected as the decoder. (**b**) Detailed design of SSRL, which is composed of Spectral Attention and Spatial Attention modules. C and C^* are different channel numbers. (**c**) The framework of MRFM is to explore the complementarity of heterogeneous inputs.

3.2. Backbone

Low-quality RSIs collected under fog suffer from the issues of intra-class heterogeneity and diverse geometric shapes, which impairs classification robustness. LCC is actually a dense pixel prediction task that requires a strong backbone with powerful modeling capability. Semantic segmentation networks are often constructed based on encoder-decoder architecture, where ResNet [35] is usually applied as the encoder. It is characterized by the accurate prediction of the spatial location at the low-level stage but is limited by a small receptive field that lacks consistent semantic information. This could lead to blurry classification. At the high-level stage, the network possesses a larger receptive field to make fine semantic predictions, but is deficient in the global representation. Consequently, conventional CNNs are susceptible to the above issues which are attributed to a loss of spatial details with the degradation of resolution.

HRNet is composed of four parallel branches pertaining to different resolutions. They constantly exchange information across multiple scales. High-resolution representations contain more spatial details, while the low-resolution representations are more capable of fine-grained classification. By maintaining a high-resolution representation, HRNet can generate spatially accurate feature maps that contain semantic information abundantly. Accordingly, HRNet-W48 is selected as our backbone for robust semantic segmentation. It can learn discriminative and distinct representations efficiently. Furthermore, HRNet is capable of integrating both local and global features with multiple scales, thereby increasing robustness in the presence of fog corruptions.

The hierarchical structure of HRNet is illustrated in Figure 3, which consists of 4 multi-resolution branches, each with resolutions of $1/4, 1/8, 1/16, 1/32$. Each branch can be partitioned into 4 stages, and the output channel number of each branch is $C, 2C, 4C, 8C$. Between each stage, there are blocks of multi-resolution fusion which consist of a 3×3 stride convolution integrating with a 1×1 upsampling layer, represented by the crossed lines. The fusion module serves as a mechanism for transferring feature information between branches of different resolutions. HRNet can be applied to semantic segmentation by accessing a 1×1 convolution for mixing and merging the representations from four branch outputs to align channel numbers. Detailed convolution parameters are shown in Table 1 where parameters for each stage are in the form $[a \times a, nC] \times b \times c$. '[]' represents the residual connection unit. Parameters a,b represent kernel size and duplication times of the residual unit separately. c means to repeat entire modularized part c times. Four basic blocks like $\begin{bmatrix} 3 \times 3, & nC \\ 3 \times 3, & nC \end{bmatrix}$ accompanying with fusion modules constitute each branch of HRNet.

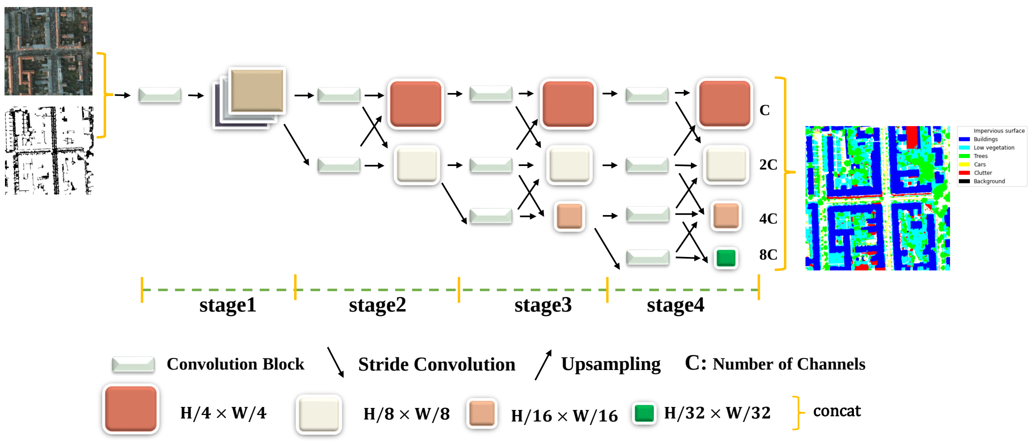

Figure 3. Overview of the backbone HRNet. HRNet maintains high-resolution representations and exchanges information throughout branches by means of 1×1 and 3×3 convolution. It can cope well with intra-class heterogeneity and multi-scale issues in RSIs with fog.

Table 1. Detailed HRNET specifications about every stage and channels.

Downsp. Rate	Stage1	Stage2	Stage3	Stage4
4×	$\begin{bmatrix} 1\times1,64 \\ 3\times3,64 \\ 1\times1,256 \end{bmatrix}\times4\times1$	$\begin{bmatrix} 3\times3,C \\ 3\times3,C \end{bmatrix}\times4\times1$	$\begin{bmatrix} 3\times3,C \\ 3\times3,C \end{bmatrix}\times4\times1$	$\begin{bmatrix} 3\times3,C \\ 3\times3,C \end{bmatrix}\times4\times3$
8×	$\begin{bmatrix} 3\times3,2C \\ 3\times3,2C \end{bmatrix}\times4\times1$	$\begin{bmatrix} 3\times3,2C \\ 3\times3,2C \end{bmatrix}\times4\times4$	$\begin{bmatrix} 3\times3,2C \\ 3\times3,2C \end{bmatrix}\times4\times3$	
16 ×		$\begin{bmatrix} 3\times3,4C \\ 3\times3,4C \end{bmatrix}\times4\times4$	$\begin{bmatrix} 3\times3,4C \\ 3\times3,4C \end{bmatrix}\times4\times3$	
32 ×				$\begin{bmatrix} 3\times3,8C \\ 3\times3,8C \end{bmatrix}\times4\times3$

3.3. Spectral and Spatial Representation Learning

Generally, RSIs captured in foggy conditions are of low quality, with a significant amount of noise existing in heterogeneous inputs. Merely incorporating encoded representations into the decoder will bring excess redundancy to the network, which reduces the classification robustness. There have been some proposals to enhance semantic features through an attention mechanism. The non-local operation proposed by [22] can obtain the attention map corresponding to a specific query tensor for modeling the global context. After rigorous experiments in [24], researchers argue that the gap between attention activation maps corresponding to different query locations is narrow, illustrating the non-necessity of query weights. In addition, ViT [25] can also enhance the semantic information through multi-head attention. It is featured with numerous parameters, high complexity as well as overfitting. This computationally intensive approach, however, ignores the correlation between various spectral channels. Additionally, the channel attention mechanism [23,36] has also been proposed to explore the interaction between different channels. Nevertheless, simply regarding 2D images as 1D disrupts the dependencies between different positions, which reduces the robustness in capturing long-range relationships. Inspired by CBAM [37], we propose SSRL which is composed of Spectral Attention and Spatial Attention. Spectral attention is to explore interdependencies between different spectral channels, thereby improving semantic representations. Spatial attention is to capture long-range dependencies. Thus, SSRL can generate a global context and acquire correlations between various pixels and spectral channels to improve the robustness of intra-class heterogeneity.

Spectral Attention is illustrated in the upper half of Figure 2b. It is constructed for acquiring the spectral-level representation weight $\Re_{spe}(\Re_{spe} \in \mathbb{R}^{C\times1\times1})$. SSRL firstly transforms the dimension of $\mathbf{X}_{RGB}$ or $\mathbf{X}_{DSM}$ ($\mathbf{X}_{RGB} \in \mathbb{R}^{C\times H\times W}$ or $\mathbf{X}_{DSM} \in \mathbb{R}^{C\times H\times W}$) to $1\times1\times C$ ($\boldsymbol{C}$ is the channel number) using a global average pooling layer (GAP). The compressed tensor is fed into MLP ($\boldsymbol{W_{mlp}}$) to compute the interaction between $\boldsymbol{k}$ adjacent channels. The MLP consists of two one-dimensional convolution layers with kernel size 1 and k, denoted as $Conv1D\_1(\mathbf{W}_{1D\_1})$ and $Conv1D\_k(\mathbf{W}_{1D\_k})$, respectively. $Conv1D\_1$ is applied for the dimension reduction, converting the channel number from $\boldsymbol{C}$ to $\boldsymbol{C}^*$. The same learnable weight values are shared among channels, where the efficiency is significantly improved because only $\boldsymbol{k}$ values are noted. Detailed formulas about Spectral Attention are as the following:

$$\Re_{spe} = \sigma\left(\mathbf{W}_{mlp}\frac{1}{WH}\sum_{i=1,j=1}^{W,H}\mathcal{X}_{ij}\right) \tag{1}$$

$$\mathbf{W}_{mlp}(x) = \mathbf{W}_{1D\_k}[F_{Mish}(\mathbf{W}_{1D\_1}(x))] \tag{2}$$

$$F_{Mish}(y) = x\cdot\tanh(\ln(1+e^{y})) \tag{3}$$

In Equation (1), σ is the sigmoid activation function. Local cross-channel attention interaction coverage is adaptively and dynamically adjusted according to the input channel numbers. There is a nonlinear mapping between the total number of channels $\boldsymbol{C}$ and the kernel size $\boldsymbol{k}$ of $Conv1D$. $\boldsymbol{k}$ increases with the number of channels. $\boldsymbol{odd}$ means to select the nearest odd number. $\mathbf{F}_{spe}$ means the feature map acquired through the Spectral Attention. The specific correspondence is as follows:

$$k = \psi(C) = \left| \frac{\log 2(C)}{2} + \frac{1}{2} \cdot \frac{1}{1+e^{-C}} + \frac{1}{2} \right|_{odd} \quad (4)$$

$$\mathbf{F}_{spe} = \Re_{spe} \otimes \mathrm{X} \quad (5)$$

Spatial Attention is illustrated in the lower panel of Figure 2b. This part is constructed for acquiring the spatial-level weight $\Re_{spe}$ ($\Re_{spe} \in \mathbb{R}^{HW \times HW}$). Through the linear transformation of $\Re_{spa}$, we can get three weight matrixes W_q, W_k, W_v. By calculating the interaction between positions (softmax operation, Equation (6)), the long range dependencies $\Re_{spa}$ can be captured. Finally, the feature map $\mathbf{F}_{spa}$ after Spectral Attention and Spatial Attention is obtained through matrix multiplication.

$$\Re_{spa} = \frac{\exp\left(\langle W_q \mathbf{F}_{spe}, W_k \mathbf{F}_{spe} \rangle\right)}{\sum\limits_{m} \exp\left(\langle W_q \mathbf{F}_{spe}, W_k \mathbf{F}_{spe} \rangle\right)} \quad (6)$$

$$\mathbf{F}_{spa} = W_c[\Re_{spa} \otimes (W_v \mathbf{F}_{spe})] \quad (7)$$

3.4. Multimodal Representation Fusion Module

RSIs are collected by satellites or drones far from the ground, which will inevitably cause pixel loss owing to the atmosphere and clouds. Meanwhile, the terrestrial environment is a three-dimensional space and there exist complicated interactions between land covers. The quality of RSIs with fog is low as some of the land covers will be obscured by fog, resulting in a single optical sensor failing. This phenomenon reduced the classification robustness by a large margin. DSM built with lidar will not suffer from this. We fuse multiple inputs by designing an effective MRFM to explore the respective characteristics of each modality. Different from the early and late fusion strategy in [32], we exploit semantic representation of different modalities through the cross attention mechanism. This can improve robustness to inter-class homogeneity and object occlusion issues.

The structure of MRFM is illustrated in Figure 2c. Red and green dashed boxes in Figure 2c function as context modeling and transform respectively. We extract the coarse representation $\Re_{DSM}$ from DSM since it contains the height information of each land cover. Firstly, utilizing 1×1 convolution (W_v) with softmax in Equation (8) to obtain the global semantic key weight from DSM in the batch, this step is to obtain the coarse correlation feature maps between different locations. N_p signifies the number of all pixels. $\boldsymbol{j}$ is for pixel indexing. Then, the computational cost is reduced by bottleneck. $\boldsymbol{r}$ is the reduction coefficient, which is set to 16 by default. Layer Normalization and GELU activation functions are integrated which enable a faster convergence as well as a stable training process. This step plays a role of transform in exploring channel-wise features while Channel Embedding (W_{CE}) can enhance the nonlinearity as well as reduce the dimension. Finally, the linked adaptive average pooling layer ($\boldsymbol{AAP}$) is utilized for the late fusion. $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$ in F_{LN} are trainable vectors, which is for affine transformation and ϵ is for numerical stability. $\boldsymbol{E}$ and $\boldsymbol{Var}$ mean expectation and standard deviation separately. We extract the fine-grained feature map from RGB input as a result of the abundant semantic features in this modality. Like non-local operation [22], we obtain the RGB feature weight $\Re_{RGB}$ through linear transformation and softmax. To exploit the complementary representation, we fuse both modalities using matrix multiplication accompanied by the residual connection. We likewise incorporate the residual structure into MRFM to make it easier for information to flow between layers, including providing feature reuse during forwarding propagation and mitigating the gradient vanishing phenomenon during backward propagation. Adding the original DSM input, dependencies between different positions obtained from DSM are fused with the exhaustive global information obtained from the optical so that each coarse position in DSM has an element-wise corresponding response generated.

$$\alpha(j) = \frac{e^{W_v X_{Dj}}}{\sum_{m=1}^{N_p} e^{W_v X_{Dm}}} \tag{8}$$

$$F_{GELU}(x) = 0.5x\left(1 + \tanh\left[\sqrt{2/\pi}\left(x + 0.044715x^3\right)\right]\right) \tag{9}$$

$$F_{LN}(x) = \frac{x - \mathrm{E}[x]}{\sqrt{\mathrm{Var}[x] + \epsilon}} * \gamma + \delta \tag{10}$$

$$\Re_{DSM} = AAP\langle F_{LN}\{F_{Gelu}[W_{CE}\sum_{j=1}^{N_p}\alpha(j)X_{Dj}]\}\rangle \tag{11}$$

$$\Re_{RGB} = \frac{\exp(\langle W_q\mathbf{X_{RGB}}, W_k\mathbf{X_{RGB}}\rangle)}{\sum_m \exp(\langle W_q\mathbf{X_{RGB}}, W_k\mathbf{X_{RGB}}\rangle)} \tag{12}$$

$$\mathbf{F}_{MRFM} = \Re_{RGB} \otimes \Re_{DSM} + X_{DSM} \tag{13}$$

3.5. Loss Function

From the analysis in Section 4.1, we can observe that the class imbalance issue exists in the ISPRS dataset. The proportion of impervious surface is sixteen times higher than the vehicle which is the tail class. Many deep models are heavily biased towards the dominant class during the training process and fail to classify the tail classes instead. Drawing on an extensive range of sources, we propose a unified loss function for our framework following a series of studies like [38,39]. Three elements constitute the unified loss. It is generalized for RSIs with long-tail class imbalance. FC, FT, and CE in Equation (14) stand for the modified focal loss [40], focal Tversky loss [41], and cross-entropy loss [42], respectively. $\boldsymbol{f}$ refers to the final loss originating from [38], whose input is the prediction result of MRFM. $\boldsymbol{aux}$ is the auxiliary loss employed to supervise the coarse object area estimation of SSRL output. $\boldsymbol{x}$ stands for the input data and $\boldsymbol{y_{gt}}$ is the ground truth. $\boldsymbol{y_{pred}}$ is the prediction output values. $\boldsymbol{back}$ denotes the background class. $\boldsymbol{gt}$, $\boldsymbol{pred}$, and $\boldsymbol{coar}$ are ground truth, prediction, and coarse feature map, respectively.

$$\mathscr{L}_{unified} = \alpha\mathscr{L}_{\widetilde{FC}}^{f} + (1-\alpha)\mathscr{L}_{\widetilde{FT}}^{f} + \lambda\sum_{x\in\{rgb,d\}}\mathscr{L}_{CE}^{aux}\left(y_{pred}^{coar}, y_{gt}\right) \tag{14}$$

$$\mathscr{L}_{\widetilde{FC}}^{f} = \frac{1}{N_c+1}[\sum_{j=1}^{N_c}\delta\mathscr{L}_{CE} + (1-\delta)(1-y_{pred}^{back})^{\gamma_1}\mathscr{L}_{CE}] \tag{15}$$

$$\mathscr{L}_{\widetilde{FT}}^{f} = \frac{1}{N_c+1}[\sum_{j=1}^{N_c}(1-DSC(x^j))^{1-\gamma_2} + (1-DSC(x^{back}))] \tag{16}$$

$$DSC(x) = \frac{TP+\varepsilon}{TP+\delta FN+(1-\delta)FP+\varepsilon} \tag{17}$$

$$\mathscr{L}_{CE} = -\sum_{j=1}^{N_c}[y_{gt}log(y_{pred}) + (1-y_{gt})log(1-y_{pred})] \tag{18}$$

α (e.g., 0.5) is designed to balance the relative weights of the final loss while λ is the weight for auxiliary bootstrap loss. $\boldsymbol{N_c}$ is the number of classes. $\boldsymbol{\delta}$ is the threshold parameter (e.g., 0.7) related to the proportion of positive and negative samples. γ controls the degree of down-weighting of easy samples while enhancing the rare. $\gamma_1|\gamma_2$ are 2 and 0.75 by default. $\boldsymbol{\varepsilon}$ is the small number for numerical stability in $\boldsymbol{DSC}$, which acts similarly to the Tversky index to control the optimizing for output imbalance.

4. Experiment

4.1. Dataset Overview

ISPRS provided true orthophotos (TOP) of Potsdam and Vaihingen for model training and validation, whose resolutions are 6000 × 6000 and average 2500 × 2500, respectively. Ground sampling distance (GSD) is 5 cm and 9 cm. We need to eliminate the Image Index 7–10 in Potsdam dataset due to the error in GT. Each TOP has six classes: car, low vegetation, tree, building, impervious surface, and cluttered background. We split both datasets into the training, test, and validation sets according to the ratios of 0.8, 0.1, and 0.1. DSM, which is acquired through lidars, contains the height information for land covers in 32-bit floating type. In addition to DSM, we use RGB and IRRG optical bands of Potsdam and Vaihingen for training and inference. nDSM is the normalized DSM which supplies the height of each pixel accompanied by ground elevation subtracted.

In order to intuitively observe the composition of different classes, we pull out all pixels from GT for an exploratory data analysis. Class proportions are shown in Figure 4. In both datasets, the ratios of each class are similar, with the proportion of cars being the smallest, less than 2%. The volume of buildings along with impervious surfaces, which are difficult to distinguish, is between 26% and 28%. Potsdam has a proportion of low vegetation that is 8.92% higher than the number of trees, whereas the difference between the two classes in Vaihingen is only 2.35%.

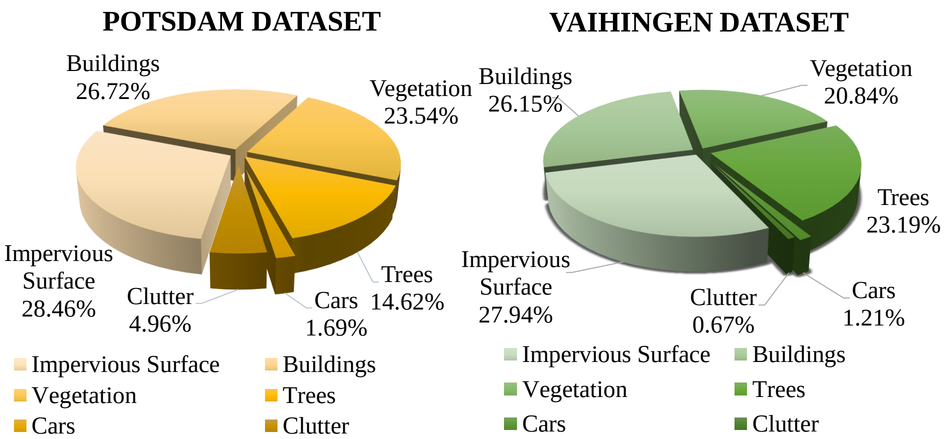

Figure 4. Illustrations about the composition of each class in ISPRS datasets.

4.2. Implementation Details

ISPRS datasets consist of high-resolution RSIs. Limited by GPU memory, we crop the optical image into 512 × 512 for model training and inference. In view of reducing overfitting, we augment the training set. Each pair of the image and the corresponding GT are rotated in arbitrary directions. Basic attributes such as contrast, brightness, saturation, and so on are randomly set for the augmentation. Reflection padding is conducted at the edges after cropping, this adjustment is particularly effective for urban complexes like buildings. Details are inevitably lost when cropping randomly, hence the symmetry of the buildings is well preserved by adopting reflection padding. Figure 5 illustrates samples for two datasets.

The hardware and software environment is listed in Table A1. We choose HRNet-W48 as the backbone, whose four branches yield feature maps ($\boldsymbol{R1, D1} \sim \boldsymbol{R4, D4}$) with 1/4, 1/8, 1/16, 1/32 of the original size. UPerHead [34] is selected as the decoder. The learning rate is set to 0.00006 with the AdamW optimizer. The weight decay is 2×10^{-2} and the power of poly optimization strategy is 1. A mixed precision scheme is employed to reduce memory usage. The model is trained for 40k iterations loaded with weight pretrained on the ImageNet.

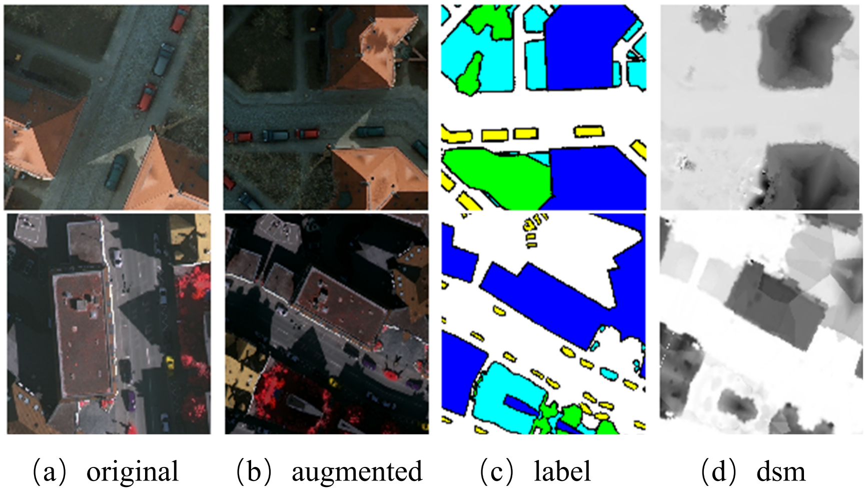

Figure 5. Illustrations about samples of Potsdam and Vaihingen dataset. *Augmented* means the image after the augmentation operation.

4.3. Test Set Transformation

The original dataset was captured in normal weather conditions. It is necessary to augment the test set for robustness evaluation. Inspired by the generation of corrupted ImageNet data sets in [18], we also render the corrupted RSIs test set with different degrees of fog and average the classification results for judging the robustness performance. We model the fog corruption through a diamond square algorithm which is to create a weighted heat map blended with the clean image. Thus, we can acquire corrupted test sets of Potsdam and Vaihingen, which are employed to measure the classification robustness in clean and foggy conditions. Corrupted samples are displayed in Figure 6, with fog generated corresponding to five levels of severity. This facilitates the robustness evaluation in various foggy conditions. Evaluation values are averaged over all five severity levels. As can be observed from Figure 6, the fog-covered region is indistinguishable from the actual scenario.

Figure 6. Illustrations of five severity levels of fog rendered the ISPRS dataset. First row: Potsdam. Second row: Vaihingen

4.4. Metrics

Metrics like overall accuracy (OA), $F1_{score}$ ($F1_{score}$) are selected to evaluate the classification accuracy. Following the robustness evaluation in [18,19], we take Corruption degradation (CD) and relative corruption degradation (rCD) into consideration for measuring LCC robustness. Specifically, metrics for accuracy evaluation are defined in Equation (19) to Equation (21). TP, FP, TN, FN represent true positive, false positive, true negative, and false negative classifications, respectively. Higher $F1_{score}$ and OA indicate a better classification accuracy.

$$Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN} \tag{19}$$

$$OA = \frac{TP + TN}{TP + FP + TN + FN} \tag{20}$$

$$F1_{score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{21}$$

From Equation (22) to Equation (24), ref stands for the baseline which is regarded as a reference model in the ablation and comparison experiments. D refers to the degree of degradation. S signifies obtaining the mean value across different corruption degrees. f is the selected model. $Clean$ and F represents clean and corrupted datasets. $\widetilde{D^{F}_{s,f}}$ means that we acquire the average degradation value using model f on RSIs across different degrees of fog corruption.

$$D = 1 - F1_{score} \tag{22}$$

$$CD^{f}_{F} = \frac{\widetilde{D^{f}_{s,F}}}{\widetilde{D^{ref}_{s,F}}} \times 100\% \tag{23}$$

$$rCD^{f}_{F} = \frac{\widetilde{D^{f}_{s,F}} - D^{f}_{clean}}{\widetilde{D^{ref}_{s,F}} - D^{ref}_{clean}} \times 100\% \tag{24}$$

CD is a measure of absolute robustness. CD greater than 100% indicates a decline in robustness compared to the reference. The part over 100% represents the degradation in performance. To evaluate the relative robustness, rCD takes the performance on the clean dataset into account. Based on the reference, it is a proportional measure of the degradation in robustness relative to the clean data. When $rCD < 100\%$, it indicates that the performance degradation in foggy conditions is less than that of the corresponding reference value compared with the clean. When CD or $rCD > 100\%$, it means that model is not as robust as the reference. Robustness is better when both values are lower.

5. Result

5.1. Impact of Different Modules

To verify the effectiveness of each component, we conduct an ablation study on Vaihingen by removing or replacing the original part. From the qualitative visualization in Figure 7, it can be noted that when SSRL is added alone, the model can enhance capturing the global context and semantic features for classifying land cover edges. However, owing to the lack of height information, it is tough to grasp the correlation in the vertical space precisely (e.g., car in the box in Figure 7d). When MRFM is incorporated, the model can obtain the complementary information of multiple modalities, yet boundary labeling is coarse due to the absence of semantic representation details. The integration of both increases the model robustness and allows it to classify various land covers in foggy conditions more accurately.

In Tables 2–7, ✓ indicates the element that we incorporate on top of the baseline. Values in the *fog* column refer to the average value across five severity levels of fog corruption. We first illustrate the effectiveness of each constituent. To demonstrate the robustness and accuracy of HRNet-W48, ResNet50 is selected for comparison in view of the comparable size. H and R in Tables 2–4 correspond to HRNet and ResNet respectively. When only SSRL is integrated, we directly transfer the Optical and DSM feature maps following convolutional layers into the decoder. When only MRFM is available, we exclude SSRL and transfer both modalities into MRFM. U and C in *loss* column represent the Unified loss and Cross-entropy loss. $ImpSurf^{*}$ and $LowVeg^{*}$ signify the impervious surface and low vegetation.

Table 2. Ablation study results with different components integrated based on the clean and fog corrupted variants of the Vaihingen test set. The $F1_{score}$ is averaged across five severity levels. The best results are marked in bold.

Method				Per-Class $F1_{score}$ (%)										Mean F1 (%)		OA (%)	
				Imp Surf*		Building		Low Veg*		Tree		Car					
Baseline	SSRL	MRFM	Loss	Clean	Fog	Clean	Fog	Clean	Fog	Clean	Fog	Clean	Fog	Clean	Fog	Clean	Fog
H	✓		U	84.93	72.18	88.71	76.71	81.70	68.32	82.51	70.93	83.46	63.82	84.26	70.39	85.92	73.12
H		✓	U	88.13	75.10	90.97	80.21	82.66	70.38	84.36	71.32	85.79	65.92	86.38	72.58	86.91	75.16
R	✓	✓	U	85.94	75.86	88.85	77.04	82.96	65.71	83.83	69.01	84.09	63.65	85.13	70.25	86.63	72.53
H	✓	✓	C	91.74	80.70	92.05	80.08	81.86	71.47	86.97	73.74	84.21	64.08	87.36	74.01	88.55	76.97
H	✓	✓	U	**92.98**	**80.85**	**93.19**	**81.54**	**83.39**	**73.50**	**87.91**	**75.28**	**92.12**	**72.67**	**89.92**	**76.77**	**90.49**	**79.15**

Table 3. *CD* in ablation study with different components integrated based on clean and fog corrupted variants of the Vaihingen test set. Our framework is regarded as the reference and values above 100% denote a decrease in robustness compared with the reference. The highest *CD* is bold.

Method				*CD* for Per-Class $F1_{score}$ (%)										Mean *CD* (%)		*CD* for OA (%)	
				Imp Surf*		Building		Low Veg*		Tree		Car					
Baseline	SSRL	MRFM	Loss	Clean	Fog	Clean	Fog	Clean	Fog	Clean	Fog	Clean	Fog	Clean	Fog	Clean	Fog
H	✓		U	**214.50**	**145.23**	**165.73**	**126.15**	**110.16**	119.53	**144.69**	117.61	**209.78**	132.41	**156.06**	127.44	**148.08**	128.94
H		✓	U	169.01	130.01	132.58	107.19	104.39	111.78	129.40	116.04	180.21	124.74	135.05	118.01	137.69	119.14
R	✓	✓	U	200.23	126.05	163.72	124.36	102.58	**129.40**	133.78	**125.37**	201.80	**133.03**	147.44	**128.04**	140.65	**131.75**
H	✓	✓	C	117.66	100.77	116.64	107.90	109.22	107.66	107.81	106.24	200.27	131.45	125.30	111.86	120.36	110.48
H	✓	✓	U	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table 4. The relative *CD* in ablation study with different components integrated based on clean and fog corrupted variants of the Vaihingen test set. Lower *rCD* suggests an improvement of robustness in the presence of fog corruption. The highest *rCD* is bold.

Method				*rCD* for Per-Class $F1_{score}$ (%)					Mean *rCD* (%)	*rCD* for OA (%)
Baseline	SSRL	MRFM	Loss	Imp Surf*	Building	Low Veg*	Tree	Car		
H	✓		**U**	105.12	**103.01**	135.26	91.70	101.03	105.49	112.89
H		✓	**U**	**107.42**	92.35	124.19	103.25	102.24	104.94	103.59
R	✓	✓	**U**	83.10	101.36	**174.43**	**117.31**	**105.13**	**113.17**	**124.28**
H	✓	✓	**C**	90.99	102.78	105.05	104.74	103.54	101.55	102.19
H	✓	✓	**U**	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table 5. Quantitative evaluation for clean and fog corrupted variants of the Vaihingen test set about using different input data. Every $F1_{score}$ is averaged over all five severity levels. The best results are marked in bold.

Method		Per-Class $F1_{score}$ (%)										Mean F1 (%)		OA (%)	
Optical	DSM	Imp Surf*		Building		Low Veg*		Tree		Car					
		Clean	Fog	Clean	Fog	Clean	Fog	Clean	Fog	Clean	Fog	Clean	Fog	Clean	Fog
✓		90.70	77.05	91.58	83.49	82.32	70.91	85.32	73.56	90.63	65.87	88.11	74.17	88.69	73.25
	✓	85.50	69.28	88.66	79.17	76.21	59.70	83.59	65.12	83.89	55.85	83.57	65.82	80.18	65.21
✓	✓	**92.98**	**80.85**	**93.19**	**81.54**	**83.39**	**73.50**	**87.91**	**75.28**	**92.12**	**72.67**	**89.92**	**76.77**	**90.49**	**79.15**

Table 6. Quantitative evaluation of *CD* on clean and fog corrupted variants of the Vaihingen test set about using different input data. Our framework is regarded as the reference and values above 100% denote a decrease in robustness compared with the reference. The highest *CD* is bold.

Method		*CD* for Per-Class $F1_{score}$ (%)										Mean *CD* (%)		*CD* for OA (%)	
Optical	DSM	Imp Surf*		Building		Low Veg*		Tree		Car					
		Clean	Fog	Clean	Fog	Clean	Fog	Clean	Fog	Clean	Fog	Clean	Fog	Clean	Fog
✓		132.42	119.84	123.59	89.42	106.41	109.76	121.47	106.96	118.83	124.91	117.91	111.16	118.93	128.32
	✓	**206.37**	**160.39**	**166.46**	**112.84**	**143.21**	**152.07**	**135.73**	**141.11**	**204.35**	**161.56**	**162.92**	**147.11**	**208.43**	**166.88**
✓	✓	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table 7. Quantitative evaluation of *rCD* on clean and fog corrupted variants of the Vaihingen test set in ablation study about using different input data evaluated. Our framework is regarded as the reference and lower *rCD* indicates an improvement of robustness in the presence of fog corruption. The highest *rCD* is bold.

Method		*rCD* for Per-Class $F1_{score}$ (%)					Mean *rCD* (%)	*rCD* for OA (%)
Optical	DSM	Imp Surf*	Building	Low Veg*	Tree	Car		
✓		112.56	69.44	115.40	93.08	127.37	105.98	136.18
	✓	**133.76**	**81.49**	**166.94**	**146.25**	**144.21**	**134.98**	**132.04**
✓	✓	100.00	100.00	100.00	100.00	100.00	100.00	100.00

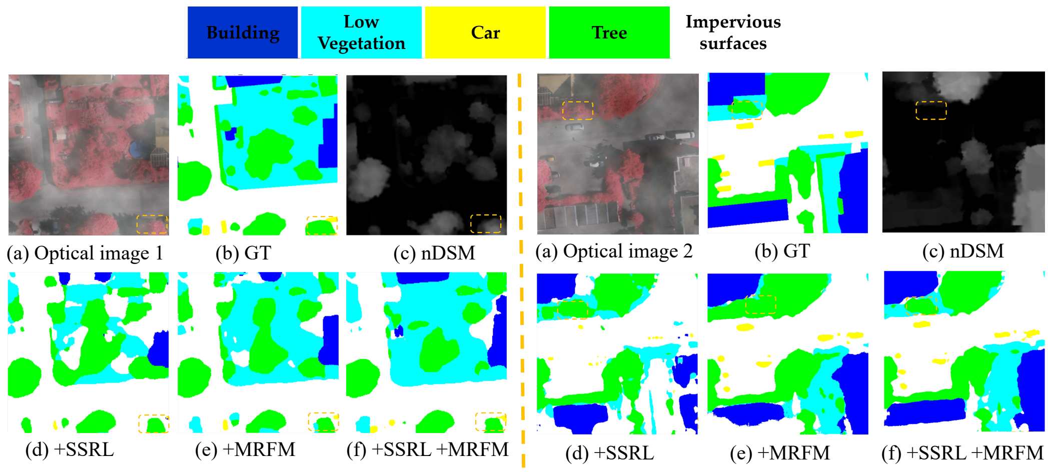

Figure 7. Illustrations about some ablation results of each component in the framework. In this case, optical images are corrupted by fog, which belongs to severity level 2.

Combining Tables 2–4 with Figure 8, we observe that the incorporation of SSRL alone could improve the accuracy, but for impervious surface (*CD* = 145.23%, *rCD* = 105.12%) and low vegetation (*CD* = 119.53%, *rCD* = 135.26%), values of both exceed 100% and robustness is still inferior to the others. In the absence of effective multimodal fusion, SSRL is more biased towards the regular shape and small objects in Vaihingen. Moreover, the addition of MRFM is conducive to the improvement of robustness. Misclassification result of tail-end distributed cars ($F1_{score}$ = 64.08%, *CD* = 131.45%) manifests if just cross-entropy loss function is utilized. When compared with the Unified loss, $F1_{score}$ is reduced by 8.59%, while *CD* is increased by 31.45%, indicating that UFL improves the robustness of imbalanced distributed objects ($F1_{score}$ = 72.67%).

5.2. Impact of Multimodal Fusion

We also perform an ablation study with different inputs to investigate the improvement of robustness under multimodal fusion. ✓ represents the input modality. In the case of single modal input, the original multimodality is replaced by the identical modal input. From Tables 5–7, we can conclude that utilizing a single modality alone is less effective than multimodal fusion. When using DSM alone, the model performs poorly because DSM contains fewer semantic features compared with the optical. There is an 8.35% and 35.95% performance loss compared to the corresponding result of the optical input. Fusing multimodalities improves accuracy and robustness in foggy environments. Compared to the single optical modal input, the accuracy is 5.9% higher and *rCD* is 36.18% lower. From Figure 9, we can observe that our model is capable of classifying edges of the cars robustly in dense fog.

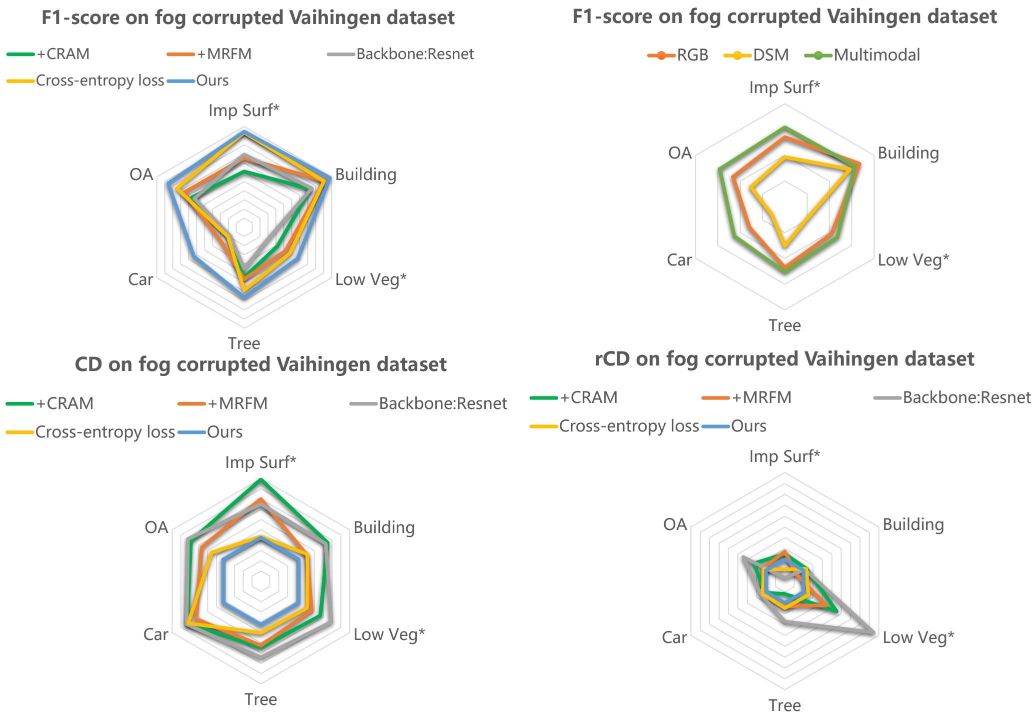

Figure 8. Radar plot visualization of the ablation study results. The first in top two is the classification result when different modules are integrated into the backbone. The second is about different inputs. Based on the radar plots regarding *CD* and *rCD*, the smaller envelope range is indicative of a model that is more robust on a fog-corrupted test set.

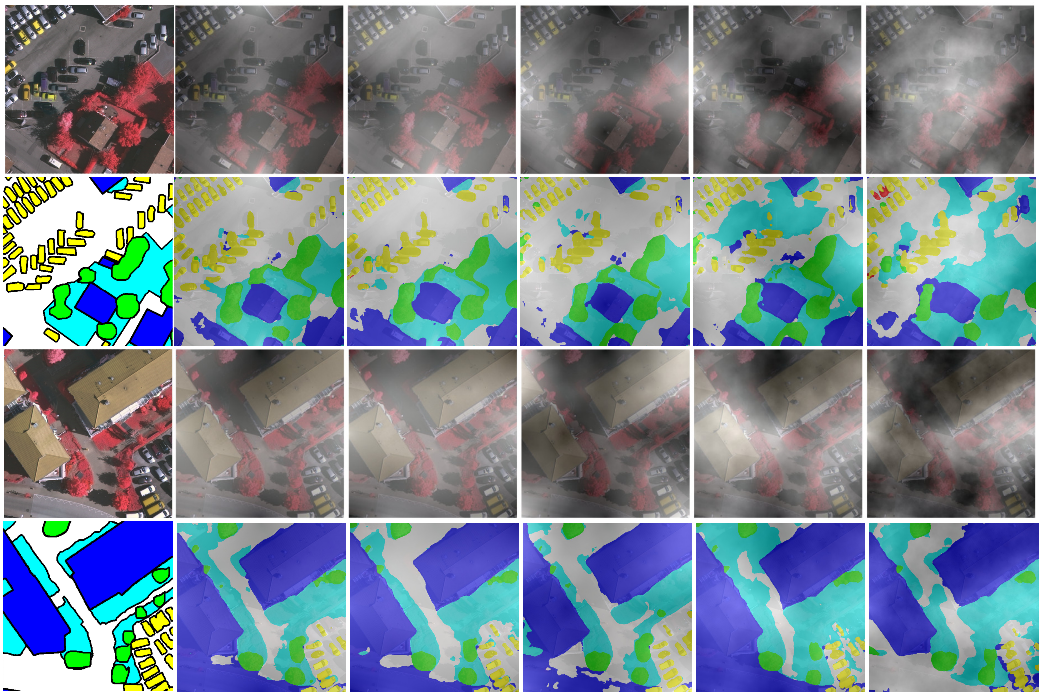

Figure 9. Comparing the LCC results across different corruption levels. 1st column displays the clean image and ground truth. Others are images with fog of various severity levels, accompanied by the corresponding semantic labeling result. From left to right, the fog intensity increases gradually.

6. Discussion

To further elucidate the model robustness, we select some existing SOTA methods to conduct a comparison experiment on the Potsdam dataset. Models in the experiment can be grouped into CNNs and Transformers. Specifically, CNN-based models contain FCN [5], UNet [6], PSPNet [12], DeepLabV3+[13], CCNet [43], OCRNet [44], TRM [15], and Transformer-based models include SETR [27], Segmenter [45], and Segformer [28]. To ensure that models in the experiment have comparable parameters, we adopt ResNet101 [35] as the backbone for FCN, UNet, PSPNet, DeepLabV3+, and CCNet. TRM, OCRNet, and our framework are built on top of HRNet-W48 [4]. Encoder backbones selected for SETR, Segmenter, and Segformer are DeiT-B [46], DeiT-B, and MiT-B5 [28], respectively. Thus, the parameter of each model is around 70–90M in size. To utilize multimodal data, we stack the optical and DSM inputs in the channel dimension for all models excluding ours.

Results obtained from Figures 10 and 11 and Table 8 show that Transformers perform better compared to CNNs on the clean test set with comparable sizes. However, the robustness of most ViTs is significantly reduced on the fog corrupted test set, with the exception of Segformer. There is a coarse classification of cars and edges in the box regions. Multimodal fusion allows our model to precisely learn the hierarchical features of inter-modal and the relationship between neighboring objects and the global. In this way, edges and interiors can be accurately classified. Compared to our previously proposed algorithm TRM, the accuracy and robustness LCC have been improved as a result of SSRL and MRFM, which enhance the ability to capture semantic information in low-quality images with a more effective data fusion approach. Specifically speaking, the $F1_{score}$ on the corrupted dataset improved by 1.3%, while there is a reduction of over 3% on both *CD* and *rCD*.

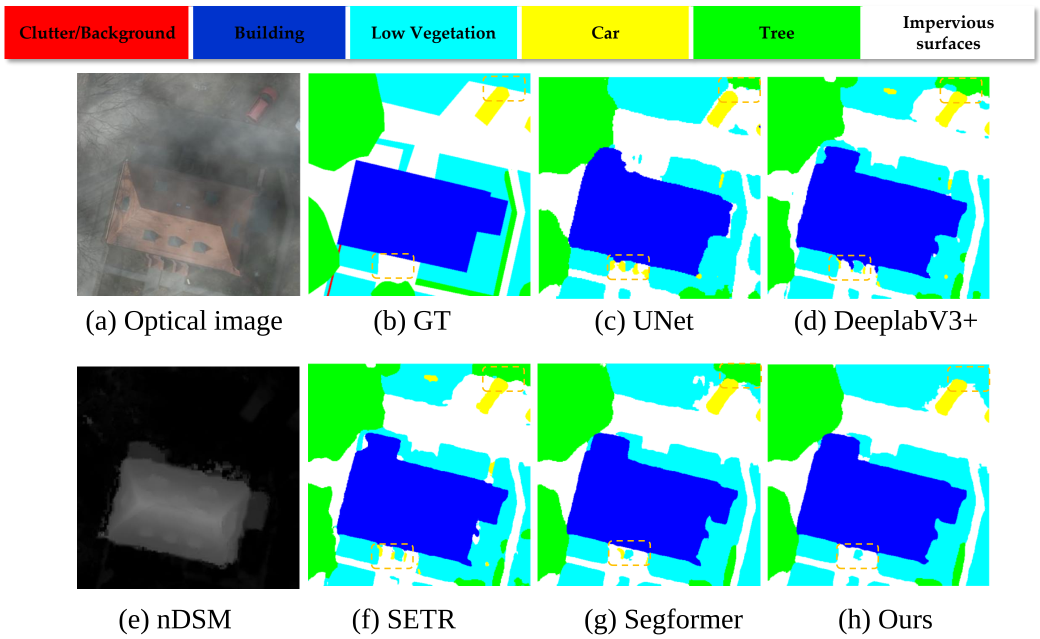

Figure 10. Qualitative comparisons between different methods applied to semantic segmentation of RSIs. The optical image is corrupted by the third severity level of fog.

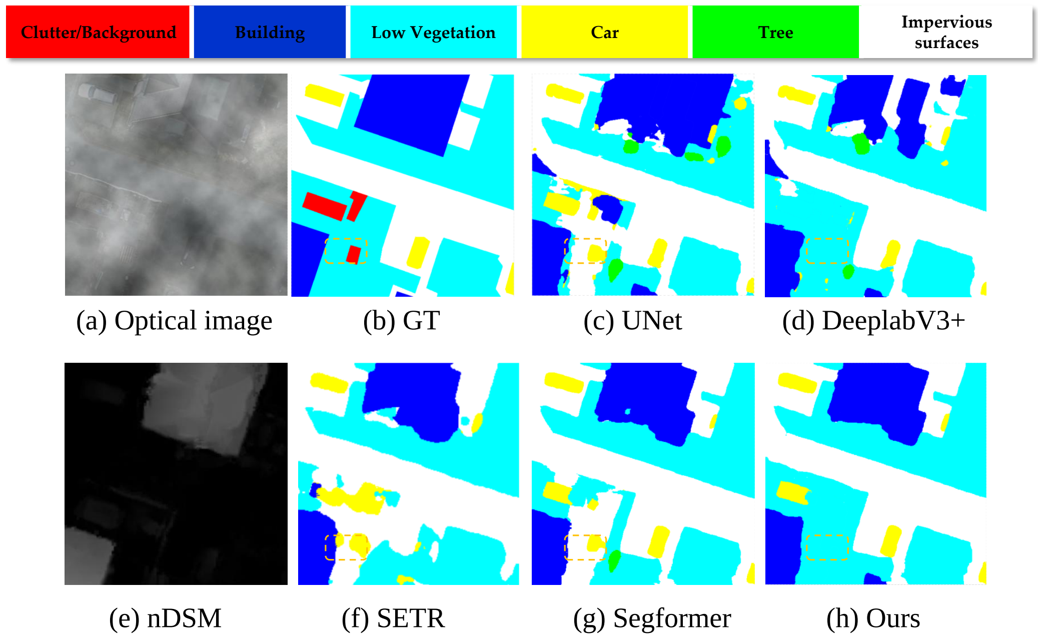

Figure 11. Qualitative comparisons between different methods applied to semantic segmentation of RSIs. The optical image is corrupted by the fourth severity level of fog.

It can be concluded from Tables 8–10 that performance on the clean and fog corrupted test set varies significantly with regard to different models, whereas *CD* and *rCD* results are generally stable. ViTs perform better than CNNs on the clean test set (Segformer is 0.31% better on the Mean $F1_{score}$ and 0.71% better on OA), which can be attributed to the capability of capturing global information. However, in terms of robustness, CNNs represented by OCRNet are stronger than the best-performing ViTs (OCRNet's *CD* decreased by 0.36% and *rCD* decreased by 2.23% in the mean $F1_{score}$ compared to Segformer), which is attributable to the fact that ViTs require more sophisticated training strategies with data augmentation. We are able to achieve a balance between accuracy and robustness compared to several SOTAs, as shown in Figure 12, where our model encircles a relatively smaller area (*CD* decreases by 3.96% and *rCD* decreases by 2.87% on OA compared to OCRNet). The balanced classification result of each class also reflects the robustness. This illustrates the effectiveness of the proposed framework in generalizing and handling with class imbalance.

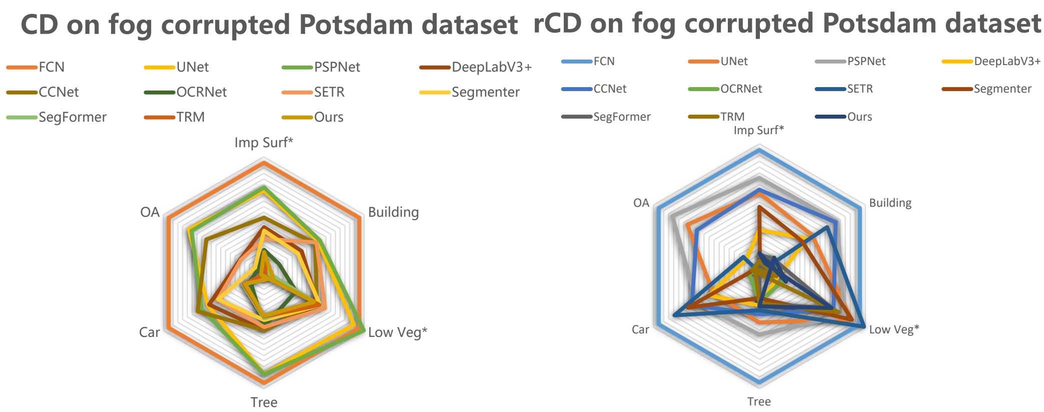

Figure 12. Radar plot for the robust performance of several SOTAs on the Potsdam test set. The envelope area of a robust model should be small and balanced. Although ViTs can boost the performance, CNNs manifest stronger robustness compared with ViTs.

Table 8. Quantitative comparison with SOTA methods on the clean and fog corrupted variants of the Potsdam test set. Each $F1_{score}$ is averaged over all severity levels. The best results are marked in bold.

Method	Per-Class $F1_{score}$ (%)										Mean F1 (%)		OA (%)	
	Imp Surf*		Building		Low Veg*		Tree		Car					
	Clean	Fog	Clean	Fog	Clean	Fog	Clean	Fog	Clean	Fog	Clean	Fog	Clean	Fog
FCN [5]	85.15	63.22	86.08	67.12	79.02	63.15	81.72	63.58	78.24	55.61	82.04	62.54	85.86	58.33
UNet [6]	85.61	66.93	88.17	72.79	79.31	63.99	79.16	64.77	80.61	62.89	82.57	66.27	86.25	61.85
PSPNet [12]	86.34	66.52	89.79	72.55	77.34	62.46	79.88	64.73	80.95	61.74	82.86	65.60	88.34	62.31
DeepLabV3+[13]	87.73	71.83	90.06	75.08	80.27	69.59	83.89	70.56	81.23	63.12	84.64	70.04	88.76	70.86
CCNet [43]	89.49	70.55	90.37	73.24	83.91	69.71	84.47	70.59	80.63	60.98	85.77	69.01	88.27	64.91
OCRNet [44]	88.77	74.86	90.69	78.19	84.02	**72.95**	84.11	71.02	85.33	**70.79**	86.58	73.56	89.31	74.92
SETR [27]	86.09	73.38	89.53	73.04	84.57	68.44	84.61	70.98	**85.93**	64.71	86.15	70.11	89.26	70.99
Segmenter [45]	89.94	72.30	90.22	75.61	**84.58**	69.21	84.77	71.90	84.85	64.83	86.87	70.77	89.88	73.42
SegFormer [28]	**89.53**	75.52	92.40	80.02	84.07	69.99	85.38	72.13	83.05	69.48	**86.89**	73.43	90.01	74.91
TRM [15]	89.11	**75.57**	91.45	80.61	83.83	69.29	85.56	72.37	83.58	69.62	86.71	73.49	89.92	75.27
Ours	89.31	**75.16**	**92.49**	**80.70**	84.23	70.15	**85.83**	**72.45**	82.08	69.83	86.79	**73.66**	**90.17**	**76.57**

Table 9. Quantitative evaluation of *CD* on clean and fog corrupted variants of Potsdam test set about different SOTA methods. FCN is regarded as the reference and values lower than 100% represent an improvement in the robust performance compared with the reference. The minimum in each ***Fog*** column is bold.

Method	*CD* for Per-Class $F1_{score}$ (%)										*CD* for Mean F1 (%)		*CD* for OA (%)	
	Imp Surf*		Building		Low Veg*		Tree		Car					
	Clean	Fog	Clean	Fog	Clean	Fog	Clean	Fog	Clean	Fog	Clean	Fog	Clean	Fog
FCN [5]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
UNet [6]	96.90	89.91	84.99	82.76	98.62	97.72	114.00	96.73	89.11	83.60	97.05	90.02	97.24	91.55
PSPNet [12]	91.99	91.03	73.35	83.49	108.01	101.87	110.07	96.84	87.55	86.19	95.44	91.82	82.46	90.45
DeepLabV3+[13]	82.63	76.59	71.41	75.79	94.04	82.52	88.13	80.83	86.26	83.08	85.56	79.98	79.49	69.93
CCNet [43]	70.77	80.07	69.18	81.39	76.69	82.20	84.96	80.75	89.02	87.90	79.22	82.71	82.96	84.21
OCRNet [44]	75.62	68.35	66.88	66.33	76.17	73.41	86.93	79.57	67.42	65.80	74.71	70.57	75.60	60.19
SETR [27]	93.67	72.38	75.22	82.00	73.55	85.64	84.19	79.68	64.66	79.50	77.15	79.78	75.95	69.62
Segmenter [45]	67.74	75.31	70.26	74.18	73.50	83.55	83.32	77.16	69.62	79.23	73.10	78.02	71.57	63.79
SegFormer [28]	70.51	66.56	54.60	60.77	75.93	81.44	79.98	76.52	77.90	68.75	73.03	70.93	70.65	60.21
TRM [15]	73.33	**66.42**	61.42	58.97	77.07	83.34	78.99	75.86	75.46	68.44	74.03	70.76	71.29	59.35
Ours	71.99	67.54	53.95	**58.70**	75.17	**81.00**	77.52	**75.65**	82.35	**67.97**	73.57	**70.31**	69.52	**56.23**

Table 10. Quantitative evaluation of *rCD* on clean and fog corrupted variants of Potsdam test set about different SOTA methods. FCN is regarded as the reference and values lower than 100% represent an improvement in the robust performance compared to the reference. The minimum in each column is in bold.

Method	*rCD* for Per-Class $F1_{score}$ (%)					Mean *rCD* (%)	*rCD* for OA (%)
	Imp Surf*	Building	Low Veg*	Tree	Car		
FCN [5]	100.00	100.00	100.00	100.00	100.00	100.00	100.00
UNet [6]	85.18	81.12	96.53	79.33	78.30	83.55	88.63
PSPNet [12]	90.38	90.93	93.76	83.52	84.89	88.49	94.55
DeepLabV3+ [13]	72.50	79.01	**67.30**	73.48	80.03	74.85	65.02
CCNet [43]	86.37	90.35	89.48	76.52	86.83	85.92	84.85
OCRNet [44]	63.43	65.93	69.75	72.16	64.25	**66.76**	52.27
SETR [27]	57.96	86.97	101.64	75.14	93.77	82.21	66.36
Segmenter [45]	80.44	77.06	96.85	**70.95**	88.47	82.55	59.79
SegFormer [28]	63.89	65.30	88.72	73.04	59.96	68.99	54.85
TRM [15]	**61.74**	**57.17**	91.62	72.71	61.69	67.74	53.21
Ours	64.52	**62.18**	88.72	73.76	**54.13**	67.31	**49.40**

7. Conclusions

This study set out to design a robust model for LCC. The framework utilizes multimodal fusion and attention mechanisms to achieve a robust segmentation of RSIs in foggy conditions. We transfer heterogeneous data into HRNet, which serves as the backbone to maintain the high-resolution representation. Incorporating MRFM into the framework can exploit cross-modal complementary fusion. SSRL is deployed for exploring the correlations between different channels and positions. Unified loss helps to mitigate class imbalance issues. Multiple experiment analyses reveal that the proposed model has superior robustness on the fog-corrupted Potsdam and Vaihingen test sets. In addition, this study has also confirmed that in terms of robustness, ViTs are often inferior to CNNs in the presence of natural noises. Overall, this study highlights the importance of multimodal fusion and attention mechanisms for enhancing segmentation robustness. Our future research plans to investigate this topic further by combining ViTs with more fundamental attributes of RSIs.

Author Contributions: Conceptualization, W.S.; methodology, W.S.; validation, W.S.; writing—review and editing, W.S. and A.C.; project administration, W.Q. All authors have read and agreed to the published version of the manuscript.

Funding: 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.

Data Availability Statement: No new data were created or analyzed in this study.

Acknowledgments: A publicly available GitHub repository fastai was adapted for the experiment.

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

Abbreviations

The following abbreviations are used in this manuscript:

RSIs	Remote Sensing Images
NSIs	Natural Scene Images
LCC	Land Cover Classification
SSRL	Spectral and Spatial Representation Learning
MRFM	Multimodal Representation Fusion Module
DSM	Digital Surface Model
SOTA	State Of The Art
CNN	Convolution Neural Network
ViT	Visual Transformer
TOP	True Ortho Photos
GSD	Ground Sampling Distance

OA	Overall Accuracy
CD	Corruption Degradation
rCD	Relative Corruption Degradation
GT	Ground Truth
ISPRS	International Society for Photogrammetry and Remote Sensing

Appendix A

Table A1. Experiment Environment.

Software	Software Version	Hardware	Hardware Version
CUDA	10.2	CPU	i7-5930K CPU @ 3.50 GHz
cuDNN	7.6	GPU	2 × Titan XP(12G)
Pytorch	1.7	RAM	64 GB
Fast.ai	2.2.2	HARD DISK	Toshiba SSD 2T
Wandb	0.1.20	SYSTEM	Ubuntu 18.0.4

References

1. He, H.; Li, C.; Yang, R.; Zeng, H.; Li, L.; Zhu, Y. Multisource Data Fusion and Adversarial Nets for Landslide Extraction from UAV-Photogrammetry-Derived Data. *Remote Sens.* **2022**, *14*, 3059. [CrossRef]
2. Shao, S.; Xiao, L.; Lin, L.; Ren, C.; Tian, J. Road Extraction Convolutional Neural Network with Embedded Attention Mechanism for Remote Sensing Imagery. *Remote Sens.* **2022**, *14*, 2061. [CrossRef]
3. Ding, J.; Zhang, J.; Zhan, Z.; Tang, X.; Wang, X. A Precision Efficient Method for Collapsed Building Detection in Post-Earthquake UAV Images Based on the Improved NMS Algorithm and Faster R-CNN. *Remote Sens.* **2022**, *14*, 663. [CrossRef]
4. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep High-Resolution Representation Learning for Visual Recognition. *arXiv* **2020**, arXiv:1908.07919.
5. Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef]
6. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* **2015**, arXiv:1505.04597.
7. Xu, Q.; Yuan, X.; jun Ouyang, C.; Zeng, Y. Attention-Based Pyramid Network for Segmentation and Classification of High-Resolution and Hyperspectral Remote Sensing Images. *Remote Sens.* **2020**, *12*, 3501. [CrossRef]
8. Zhang, G.; Lei, T.; Cui, Y.; Jiang, P. A dual-path and lightweight convolutional neural network for high-resolution aerial image segmentation. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 582. [CrossRef]
9. Li, X.; Jiang, Y.; Peng, H.; Yin, S. An aerial image segmentation approach based on enhanced multi-scale convolutional neural network. In Proceedings of the 2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS), Taipei, Taiwan, 6–9 May 2019; pp. 47–52.
10. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *arXiv* **2017**, arXiv:1606.00915.
11. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. DenseASPP for Semantic Segmentation in Street Scenes. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; IEEE: Salt Lake City, UT, USA, 2018; pp. 3684–3692. [CrossRef]
12. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. *arXiv* **2017**, arXiv:1612.01105.
13. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *arXiv* **2018**, arXiv:1802.02611.
14. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *arXiv* **2016**, arXiv:1511.00561.
15. Shi, W.; Qin, W.; Yun, Z.; Chen, A.; Huang, K.Y.; Zhao, T. Land Cover Classification in Foggy Conditions: Toward Robust Models. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
16. Zhang, X.Y.; Liu, C.L.; Suen, C.Y. Towards Robust Pattern Recognition: A Review. *Proc. IEEE* **2020**, *108*, 894–922. [CrossRef]
17. Tang, S.; Gong, R.; Wang, Y.; Liu, A.; Wang, J.; Chen, X.; Yu, F.; Liu, X.; Song, D.; Yuille, A.; et al. RobustART: Benchmarking Robustness on Architecture Design and Training Techniques. *arXiv* **2021**, arXiv:2109.05211.
18. Hendrycks, D.; Dietterich, T. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *arXiv* **2019**, arXiv:1903.12261.
19. Kamann, C.; Rother, C. Benchmarking the Robustness of Semantic Segmentation Models. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 8825–8835. [CrossRef]
20. Shao, R.; Shi, Z.; Yi, J.; Chen, P.Y.; Hsieh, C.J. On the Adversarial Robustness of Visual Transformers. *arXiv* **2021**, arXiv:2103.15670.

21. Mahmood, K.; Mahmood, R.; van Dijk, M. On the Robustness of Vision Transformers to Adversarial Examples. *arXiv* **2021**, arXiv:2104.02610.
22. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-Local Neural Networks. *arXiv* **2018**, arXiv:1711.07971.
23. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. *arXiv* **2020**, arXiv:1910.03151.
24. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. GCNet: Non-local Networks Meet Squeeze-Excitation Networks and Beyond. *arXiv* **2019**, arXiv:1904.11492.
25. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929.
26. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. *arXiv* **2021**, arXiv:2103.14030.
27. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.S.; et al. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. *arXiv* **2021**, arXiv:2012.15840.
28. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *arXiv* **2021**, arXiv:2105.15203.
29. Yuan, L.; Hou, Q.; Jiang, Z.; Feng, J.; Yan, S. VOLO: Vision Outlooker for Visual Recognition. *arXiv* **2021**, arXiv:2106.13112.
30. Gu, Y.; Hao, J.; Chen, B.; Deng, H. Top-Down Pyramid Fusion Network for High-Resolution Remote Sensing Semantic Segmentation. *Remote Sens.* **2021**, *13*, 4159. [CrossRef]
31. Yan, L.; Huang, J.; Xie, H.; Wei, P.; Gao, Z. Efficient Depth Fusion Transformer for Aerial Image Semantic Segmentation. *Remote Sens.* **2022**, *14*, 1294. [CrossRef]
32. Audebert, N.; Le Saux, B.; Lefèvre, S. Beyond RGB: Very High Resolution Urban Remote Sensing with Multimodal Deep Networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *140*, 20–32. [CrossRef]
33. Liu, H.; Zhang, J.; Yang, K.; Hu, X.; Stiefelhagen, R. CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation with Transformers. *arXiv* **2022**, arXiv:2203.04838.
34. Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; Sun, J. Unified Perceptual Parsing for Scene Understanding. *arXiv* **2018**, arXiv:1807.10221.
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
36. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *arXiv* **2019**, arXiv:1709.01507.
37. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. *arXiv* **2018**, arXiv:1807.06521.
38. Yeung, M.; Sala, E.; Schönlieb, C.B.; Rundo, L. Unified Focal Loss: Generalising Dice and Cross Entropy-Based Losses to Handle Class Imbalanced Medical Image Segmentation. *Comput. Med Imaging Graph.* **2022**, *95*, 102026. [CrossRef] [PubMed]
39. Ma, J.; Chen, J.; Ng, M.; Huang, R.; Li, Y.; Li, C.; Yang, X.; Martel, A.L. Loss Odyssey in Medical Image Segmentation. *Med. Image Anal.* **2021**, *71*, 102035. [CrossRef]
40. Lin, T.Y.; Goyal, P.; Girshick, R.B.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef] [PubMed]
41. Abraham, N.; Khan, N.M. A Novel Focal Tversky Loss Function With Improved Attention U-Net for Lesion Segmentation. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019; pp. 683–687.
42. Zhang, Z.; Sabuncu, M.R. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. In Proceedings of the NeurIPS, Montreal, QC, Canada, 3–8 December 2018.
43. Huang, Z.; Wang, X.; Wei, Y.; Huang, L.; Shi, H.; Liu, W.; Huang, T.S. CCNet: Criss-Cross Attention for Semantic Segmentation. *arXiv* **2020**, arXiv:1811.11721.
44. Yuan, Y.; Chen, X.; Chen, X.; Wang, J. Segmentation Transformer: Object-Contextual Representations for Semantic Segmentation. *arXiv* **2021**, arXiv:1909.11065.
45. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for Semantic Segmentation. *arXiv* **2021**, arXiv:2105.05633.
46. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training Data-Efficient Image Transformers & Distillation through Attention. *arXiv* **2021**, arXiv:2012.12877.