remote sensing

MDPI

Article

OSNet: An Edge Enhancement Network for a Joint Application of SAR and Optical Images

Keyu Ma [1], Kai Hu [1,2,*], Junyu Chen [1], Ming Jiang [3], Yao Xu [4], Min Xia [1,2] and Liguo Weng [1,2]

[1] School of Automation, Nanjing University of Information Science and Technology, Nanjing 210044, China; 202212220002@nuist.edu.cn (K.M.); 202212490018@nuist.edu.cn (J.C.); xiamin@nuist.edu.cn (M.X.); 002311@nuist.edu.cn (L.W.)
[2] Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET), Nanjing University of Information Science and Technology, Nanjing 210044, China
[3] Faculty of Geosciences, Utrecht University, 3584 CS Utrecht, The Netherlands; m.jiang2@students.uu.nl
[4] Department of Computer Science, University of Reading, Whiteknights, Reading RG6 6DH, UK; hz841881@student.reading.ac.uk
* Correspondence: 001600@nuist.edu.cn; Tel.: +86-13770569871

Abstract: The combined use of synthetic aperture radar (SAR) and optical images for surface observation is gaining increasing attention. Optical images, with their distinct edge features, can accurately classify different objects, while SAR images reveal deeper internal variations. To address the challenge of differing feature distributions in multi-source images, we propose an edge enhancement network, OSNet (network for optical and SAR images), designed to jointly extract features from optical and SAR images and enhance edge feature representation. OSNet consists of three core modules: a dual-branch backbone, a synergistic attention integration module, and a global-guided local fusion module. These modules, respectively, handle modality-independent feature extraction, feature sharing, and global-local feature fusion. In the backbone module, we introduce a differentiable Lee filter and a Laplacian edge detection operator in the SAR branch to suppress noise and enhance edge features. Additionally, we designed a multi-source attention fusion module to facilitate cross-modal information exchange between the two branches. We validated OSNet's performance on segmentation tasks (WHU-OPT-SAR) and regression tasks (SNOW-OPT-SAR). The results show that OSNet improved PA and MIoU by 2.31% and 2.58%, respectively, in the segmentation task, and reduced MAE and RMSE by 3.14% and 4.22%, respectively, in the regression task.

Keywords: multimodal neural networks; multi-source fusion; attention mechanism

Academic Editor: Giuseppe Casula

Received: 31 October 2024
Revised: 29 January 2025
Accepted: 30 January 2025
Published: 31 January 2025

Citation: Ma, K.; Hu, K.; Chen, J.; Jiang, M.; Xu, Y.; Xia, M.; Weng, L. OSNet: An Edge Enhancement Network for a Joint Application of SAR and Optical Images. *Remote Sens.* **2025**, *17*, 505. https://doi.org/10.3390/rs17030505

1. Introduction

Optical and synthetic aperture radar (SAR) images are widely used in remote sensing, serving critical roles in environmental monitoring, natural disaster assessment, and resource management [1]. Optical images, with their rich color, boundary, and texture information, can effectively distinguish different objects and edges. However, optical images are susceptible to weather conditions such as clouds and fog. In contrast, SAR images, with their excellent penetration capability, can capture internal information of objects even in adverse weather conditions. Therefore, researchers have increasingly focused on integrating optical and SAR images to leverage their complementary advantages, providing a more comprehensive and accurate land surface analysis [2].

With the rapid development of deep learning techniques, the joint application of optical and SAR images has become a mainstream trend to improve the accuracy of remote

sensing tasks [3]. Li et al. [4] first proposed the land cover segmentation dataset based on optical and SAR images, which served as a foundational dataset for deep learning models. Since then, researchers have conducted extensive work around efficient feature extraction and fusion methods, significantly improving the accuracy of land cover classification. Additionally, optical and SAR images have also achieved groundbreaking applications in snow depth estimation. Daudt et al. [5] were the first to combine SAR and optical images for snow depth retrieval by employing neural networks to explore the nonlinear relationships between optical images, SAR images, and snow depth.

However, while existing studies have designed detailed models to decouple the overall features of the two modalities, they often overlook crucial edge information [6,7]. Edge information is crucial in both of these tasks. In land segmentation, edges are essential for identifying the boundaries between different land cover types, such as forests, water, and urban areas [8]. By detecting these edges, segmentation models can more accurately classify distinct land features and reduce misclassification. In snow depth estimation, optical edges clearly mark the boundaries between snow-covered and non-snow-covered areas [9], assisting SAR in reducing the misclassification of snow-free regions as deep snow. Regions such as ridges and bare ground often have little snow, and accurately identifying these areas can help reduce redundant information in SAR images [10].

Obtaining accurate boundary information through optical and SAR images still faces challenges. Despite SAR images' ability to penetrate surfaces and reveal internal edges, they often suffer from blurred edges due to scattering noise [11]. To address the above issues, this study made the following contributions:

(1) This study introduces OSNet (network for optical and SAR images), a bidirectional feature exchange network that leverages the strengths of both optical and SAR images to achieve complementary edge information fusion.
(2) This study introduces a Laplacian convolution designed for neural networks, incorporating a differentiable Lee filter and a Laplacian edge detection operator. This approach effectively suppresses SAR noise while enhancing edge features.
(3) We construct the SNOW-OPT-SAR dataset, which integrates optical and SAR images for snow depth inversion. This dataset combines regional optical images (RGB and near-infrared bands) with SAR images to perform regression predictions of snow depth at central location, using station measurements as the ground truth.

The rest of this paper is organized as follows: Section 2 introduces related work of land segmentation and snow depth estimation. Section 3 introduces land cover classification dataset called WHU-OPT-SAR and snow depth inversion dataset, which we created and called SNOW-OPT-SAR. Section 4 presents the structure of OSNet. Section 5 outlines a comparison and ablation experiments that we performed on WHU-OPT-SAR [4] for a quantitative analysis to validate the effectiveness of OSNet. In Section 6, we conduct experiments on SNOW-OPT-SAR to verify the performance of OSNet in a regression task. Section 7 summarizes our paper's contributions and limitations. The flowchart of the methodology for OSNet validation is shown in Figure 1.

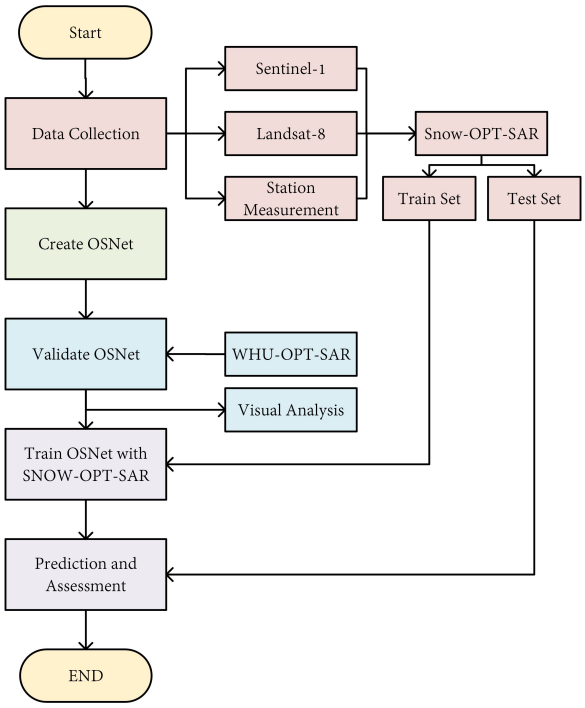

Figure 1. Flowchart of the methodology for OSNet's validation.

2. Related Work

2.1. Land Segmentation

Currently, research on land segmentation using the joint images of optical and SAR is rapidly developing [12]. Optical sensors capture electromagnetic radiation in the visible spectrum, are highly sensitive to color information, and can directly reflect land objects. However, they are susceptible to weather and lighting conditions. In contrast, SAR collects radar signals reflected by land surface to generate backscatter images, effectively identifying specific land objects (such as water and snow) even under adverse weather or dim lighting conditions. Therefore, extracting information from optical and SAR images to improve land classification holds great research value.

Li et al. [4] established the first deep learning-based optical and SAR joint land classification dataset, WHU-OPT-SAR, and proposed MCANet, used for feature extraction from multi-source data with Siamese networks. It demonstrates that the complementarity between optical and SAR images can significantly enhance the accuracy of land classification. After this, various fusion strategies have been explored in recent studies, which can be broadly classified into pixel-level fusion and feature-level fusion. Pixel-level fusion, as exemplified by methods like PSCNN [6], directly combines the two modalities by stacking them at the pixel level. However, this approach tends to be sensitive to noise and suffers from instability during training. In contrast, feature-level fusion has become the dominant approach in recent studies, with researchers also investigating multi-scale feature fusion techniques [4,13–15]. Hu et al. [7] proves that feature-level fusion is more effective than pixel-level fusion in terms of performance and robustness. Accordingly, in this work, we adopt a feature-level fusion strategy to construct a dual-branch neural network.

2.2. Snow Depth Estimation

The retrieval of snow depth mainly depends on the ability of synthetic aperture radar to penetrate the snow surface and obtain the internal information of the snow [16]. Snow is a poor conductor of heat and acts as an insulator, providing thermal insulation to the ground [17]. During the transition from fall to winter, the cooling of the near-surface

air increases the temperature gradient between the ground and the air, resulting in more longwave radiation from the ground and a decrease in ground temperature [18]. Changes in ground temperature affect moisture and liquid water, which change the ground dielectric constant, which is closely related to the backscatter coefficient in SAR imagery. The insulating effect of snow is proportional to its depth; deeper snow provides better insulation, resulting in less energy loss, a smaller decrease in ground temperature, and a lower moisture and dielectric constant [19]. This results in a decrease in the backscatter coefficient in SAR images. By establishing a quantitative relationship between the backscatter coefficient and snow depth, snow depth can be estimated [20–22]. Daudt et al. [5] are the first to combine SAR and optical images for snow depth inversion by employing neural networks to explore the nonlinear relationships between optical and SAR images, and snow depth.

Although optical remote sensing cannot penetrate snow, it can accurately distinguish between snow-covered and snow-free areas. This capability can assist SAR in reducing misclassification of snow-free and shallow snow areas as deep snow. Additionally, there is a correlation between snow cover extent and snow depth; a larger snow cover extent often indicates a greater average snow depth [23]. Therefore, Zhao et al. [23] combined regional optical and SAR images to retrieval the snow depth at a center single point, with optical images providing information on snow cover location, color, and texture to assist SAR in prediction.

3. Datasets

In this study, we use the WHU-OPT-SAR dataset [4] for the semantic segmentation task and create the SNOW-OPT-SAR dataset, which combines optical and SAR images, for the snow depth regression task. For both datasets, we allocate 80% of the data for training and 20% for testing.

3.1. WHU-OPT-SAR Dataset

WHU-OPT-SAR dataset: WHU-OPT-SAR is a publicly available dataset that matches optical images with SAR images to train and test the effectiveness of models in the task of semantic segmentation, available at [24]. The dataset originates from Hubei Province, China, covering a wide range of remote sensing images with varying terrains and vegetation. The dataset is annotated with seven main categories: farmland (brown), urban (red), village (yellow), water (blue), forest (green), road (cyan), and others (white).

In this study, 100 sets of image from the dataset are cropped to a size of 256 $\times$ 256. We carefully examine the cropped dataset, removing samples with only a single category, resulting in a total of 22,409 datasets.

3.2. SNOW-OPT-SAR Dataset

3.2.1. Study Area and Dataset

The study area of SNOW-OPT-SAR is located in the Tibetan Plateau (ranging from 26°00′N to 39°47′N in latitude and from 73°19′E to 104°47′E in longitude). The Tibetan Plateau is the highest region in the mid-latitudes of the Northern Hemisphere and the area with the most extensive snow cover. Figure 2 presents a topographic map of the Tibetan Plateau in China, along with the distribution of meteorological stations used in this paper.

We create SNOW-OPT-SAR by merging data from three sources: VV-polarized SAR images from Sentinel-1, optical images from Landsat-8 covering four bands (RGB, NIR), and daily snow depth observations from the National Meteorological Information Center. Since the station's data span from 2014 to 2017, we select Landsat-8 for optical images instead of Sentinel-2.

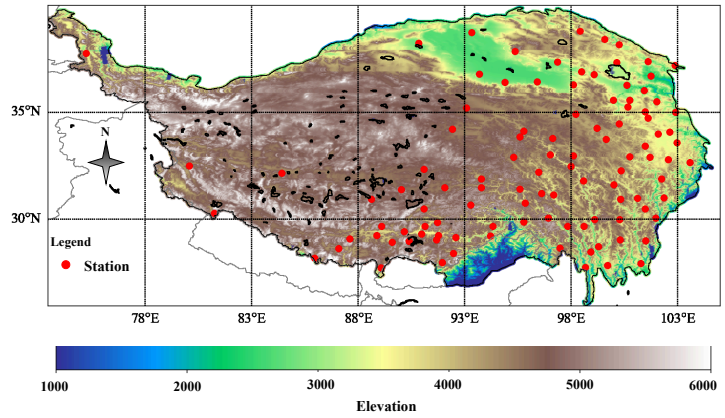

Figure 2. A topographic map of the Tibetan Plateau and the distribution of stations used in this paper.

3.2.2. Sentinel-1

C-band SAR images obtained by Sentinel-1 are robust to variations in solar illumination, cloud cover, and other meteorological events, which makes them a reliable, timely tool for observing the Earth at regular intervals.

We preprocess Sentinel-1 VV-polarization images by ESA's SNAP toolbox, including orbit correction, thermal noise removal, radiometric calibration, terrain correction and conversion to decibel scale.

3.2.3. Landsat-8

Landsat-8 provides seasonal coverage of the global landmass at a spatial resolution of 30 meters. We select only optical images taken in clear, cloud-free weather conditions. The RGB and NIR bands are then selected for layer stacking and resampled to 10 meters to match the resolution of SAR images. All these processes are implemented in ENVI.

3.2.4. Ground Observation

The daily snow depth observation data are obtained from the National meteorological Information Center, available at [25]. We selected stations that are displayed in Figure 2, covering the period from 2014 to 2017, specifically from November to March of the following year. We selected the data in winter to minimize the impact of snow melting. The site observation values are used as labels to annotate the optical and SAR images. The snow depth values range from 0 to 42 cm, and the data distribution is displayed in the Figure 3.

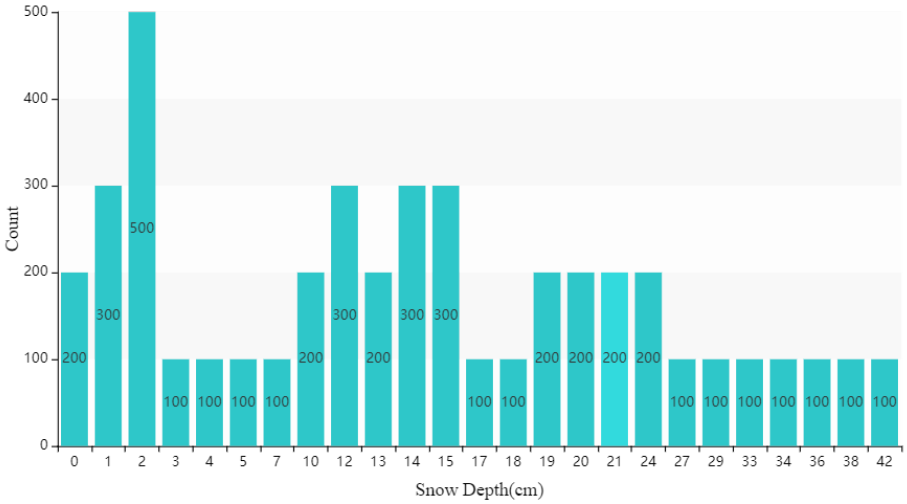

Figure 3. Data distribution of SNOW-OPT-SAR. The snow depth range from 0 to 42cm.

3.2.5. Creation Process

To effectively combine these diverse data sources and create a cohesive dataset, we follow a detailed preprocessing and alignment procedure.

Firstly, we obtain cloud-free optical and SAR images based on the time and location of ground observations. To ensure data accuracy, we select remote sensing images acquired on the same day as the measurements. We collect the observational data during winter to minimize changes in snow cover, thereby reducing the impact of temporal displacement. Then, we preprocess the image pairs, as mentioned previously, and align them using the registration tool in ENVI. Subsequently, we crop a 64 × 64 pixel area centered on the site location and annotate it with ground observation data. Finally, we expand the original 4000 samples to 8000 sample sets by applying random rotations and flipping.

This approach involves using optical and SAR images of 640 m × 640 m as inputs to determine snow depth at the central location. Unlike point-to-point snow estimation, this method leverages the spatial distribution and edge information of snow in the optical images to distinguish between snow-covered and snow-free areas, aiding the SAR analysis. Furthermore, edges in snow images, such as ridgelines, typically indicate snow-free zones and can be ignored in SAR backscatter analysis, thereby eliminating redundant information and improving snow depth prediction accuracy in SAR images.

Figure 4 displays a group of images pairs in SNOW-OPT-SAR.

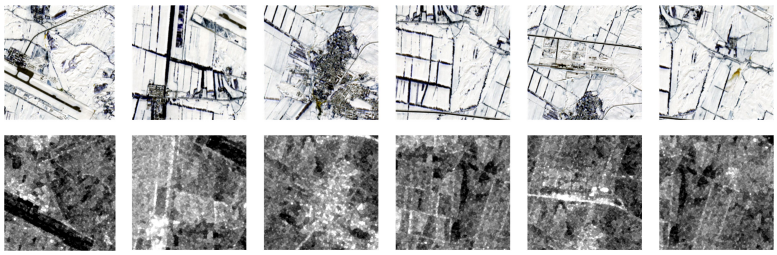

Figure 4. Partial dataset display: Each column represents a set of images for which alignment has been completed. The first row corresponds to optical images, and the second row corresponds to the VV-polarization SAR image.

4. Methodology

4.1. Architecture of OSNet

We proposed a joint framework called OSNet (network for optical and SAR images) in Figure 5, which can be divided into Backbone, synergistic attention integration module (SAIM), and global-guided local fusion module (GLFM).

The entire network can be viewed as an encoder-decoder structure. The backbone and SAIM serve as the encoder to extract features, while the GLFM acts as the decoder to transform these features into the desired output format. OSNet takes a pair of optical and SAR images as input. Depending on the task type, we can switch output by simply modifying the final part of the GLFM. For segmentation tasks, we use upsampling to restore the features to their original dimensions. For regression tasks, we flatten the features and use a fully connected network to output the predicted snow depth. In following subsections, we will introduce the details of each module.

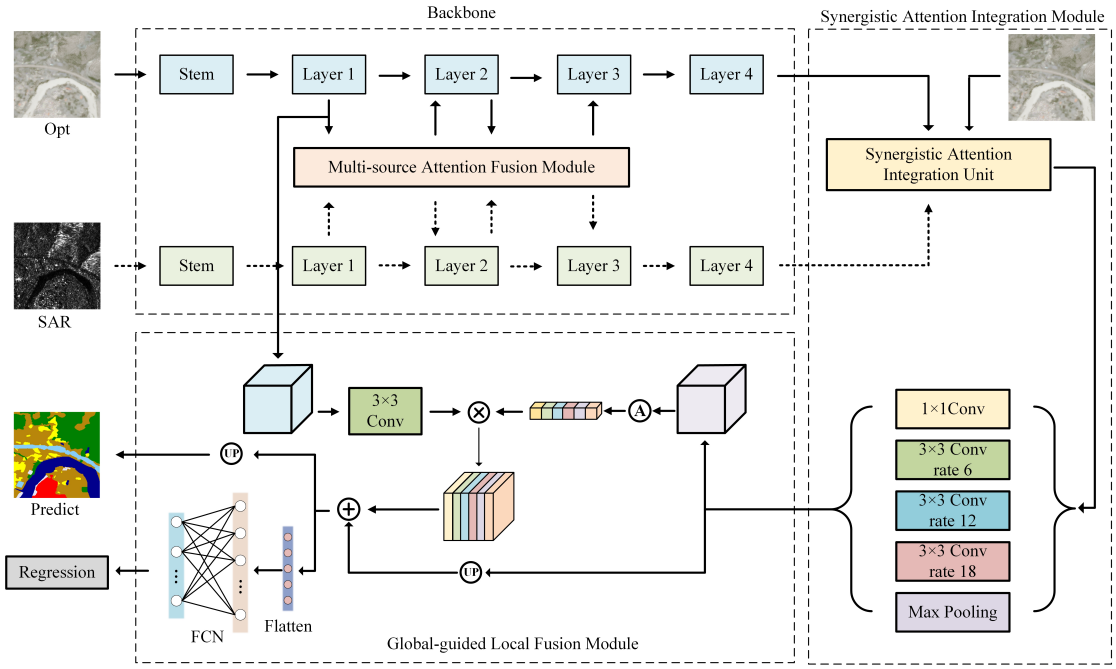

Figure 5. Framework of OSNet.

4.2. Backbone

We proposed an asymmetric dual-branch backbone and each branch is modified based on ResNet [26], which is a classical model in deep learning for extracting features at multiple scales. ResNet introduces residual structures to address the issues of gradient explosion and vanishing gradients in deep networks [8]. ResNet-50 is divided into five layers, with the first layer being the Stem and the remaining four layers consisting of 3, 4, 6, and 3 residual modules, respectively, as described below:

$$x_i' = Conv_i^3\left\{\sigma\left[Conv_i^1(x_i)\right]\right\} \tag{1}$$

$$x_{i+1} = x_i + \sigma\left\{Conv_{i+1}^1[\sigma(x_i')]\right\} \tag{2}$$

where x_i is the matrix input, x_i' is the output of the intermediate block, and x_{i+1} is the output of the entire residual module. $Conv_i^j$ denotes the convolution with kernel size j of the ith residual module. $\sigma(\cdot)$ denotes the BatchNorm and the activation function ReLU.

Table 1 displays the design of each branch. Specifically, we remove the pooling layers in L3 and L4 of the original ResNet50, resulting in the final feature map size being reduced to only 1/8 of the input size. Since the original ResNet is designed for classification tasks, this change is more suitable for segmentation tasks [27]. Additionally, given that SNOW-OPT-SAR has a small input size of 64 × 64, using an 8-folder down-sampling will not result in significant feature loss. In the SAR branch, we introduce a Laplacian convolution (detailed in Section 3.2.1) to suppress noise and strengthen edge information. The multi-source attention fusion module (MAFM) is introduced at L2 and L3 (in Section 3.2.2) to enhance the mutual representation of features between the two branches.

Table 1. Comparison between original ResNet50 and backbone of OSNet. The left is original ResNet50 and the right is the modified two-branch backbone we proposed. Size indicates the scale change compared to the original input size. (n × n, m) represents m convolutional kernels of size n. Max denotes the max pooling.

	Original		**Modified**		
Layer	**50-Layer**	**Size**	**Opt-branch**	**SAR-branch**	**Size**
Stem	7×7, stride=2	1/2	7×7, stride=2	Laplacian-Conv 7×7, stride=2	1/2
L1	3×3, Max, stride=2 $\begin{pmatrix} 1 \times 1, & 64 \\ 3 \times 3, & 64 \\ 1 \times 1, & 64 \end{pmatrix} \times 3$	1/4	3×3, Max, stride=2 $\begin{pmatrix} 1 \times 1, & 64 \\ 3 \times 3, & 64 \\ 1 \times 1, & 64 \end{pmatrix} \times 3$	3×3, Max, stride=2 $\begin{pmatrix} 1 \times 1, & 64 \\ 3 \times 3, & 64 \\ 1 \times 1, & 64 \end{pmatrix} \times 3$	1/4
L2	$\begin{pmatrix} 1 \times 1, & 128 \\ 3 \times 3, & 128 \\ 1 \times 1, & 512 \end{pmatrix} \times 4$	1/8	MAFM → Exchange ← MAFM $\begin{pmatrix} 1 \times 1, & 128 \\ 3 \times 3, & 128 \\ 1 \times 1, & 512 \end{pmatrix} \times 4$	$\begin{pmatrix} 1 \times 1, & 128 \\ 3 \times 3, & 128 \\ 1 \times 1, & 512 \end{pmatrix} \times 4$	1/8
L3	$\begin{pmatrix} 1 \times 1, & 256 \\ 3 \times 3, & 256 \\ 1 \times 1, & 1024 \end{pmatrix} \times 6$	1/16	MAFM → Exchange ← MAFM $\begin{pmatrix} 1 \times 1, & 256 \\ 3 \times 3, & 256 \\ 1 \times 1, & 1024 \end{pmatrix} \times 6$	$\begin{pmatrix} 1 \times 1, & 256 \\ 3 \times 3, & 256 \\ 1 \times 1, & 1024 \end{pmatrix} \times 6$	1/8
L4	$\begin{pmatrix} 1 \times 1, & 512 \\ 3 \times 3, & 512 \\ 1 \times 1, & 2048 \end{pmatrix} \times 3$	1/32	$\begin{pmatrix} 1 \times 1, & 512 \\ 3 \times 3, & 512 \\ 1 \times 1, & 2048 \end{pmatrix} \times 3$	$\begin{pmatrix} 1 \times 1, & 512 \\ 3 \times 3, & 512 \\ 1 \times 1, & 2048 \end{pmatrix} \times 3$	1/8

4.2.1. Laplacian Convolution

We propose a Laplacian convolution composed of a differentiable Lee filter and a Laplacian edge detection operator, which can be fully integrated into neural networks for enhanced performance.

The Lee filter is a commonly used method for suppressing speckle noise in SAR images [28]. Its basic principle involves using local statistical characteristics to estimate the filtered pixel values. The standard Lee filter formula involves nonlinear operations, which are not fully differentiable. We approximate these nonlinear operations with differentiable functions, making the Lee filter more suitable for integration into neural networks.

The differentiable Lee filter is used as follows:

Initially, the image is divided into multiple 3×3 windows. For each window, we calculate the mean $\hat{m}(x, y)$ and variance $s^2(x, y)$ of the pixels. The mean variance of all windows $\hat{\sigma}^2$ is used to estimate the noise variance, reflecting the noise level in the image. These calculations can be efficiently implemented using Pytorch's 2D convolution as follows:

$$\hat{m}(x, y) = F.conv2d(I, mean, pad) \tag{3}$$

$$s^2(x, y) = F.conv2d(I^2, mean, pad) - \hat{m}^2(x, y) \tag{4}$$

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} s^2(x_i, y_i) \tag{5}$$

where $F.conv2d$ is the 2D convolution function in PyTorch. I denotes the input image, and pad is short for padding indicating window size. x and y represent the position of the window in the image, and N is the number of windows.

We then calculate the signal-to-noise ratio (SNR) for each window as

$$SNR = \frac{\hat{m}(x, y)}{\sqrt{\hat{\sigma}^2}} \tag{6}$$

Traditionally, a manually set threshold T is used for a SNR comparison. If the $SNR < T$, filtering is needed; otherwise, the original pixel values are retained. We replace T

with a trainable parameter adaptively adjusted through backpropagation. Since traditional comparison is not differentiable, we use the Sigmoid function σ with a trainable parameter b, which adapts the slope of the Sigmoid function to output the filtering coefficient $c(x, y)$:

$$c(x, y) = \sigma(b \times (SNR - T)) \tag{7}$$

When the SNR exceeds T, the filtering coefficient approaches 1; otherwise, it nears 0. The final filtered pixel value, $L(x, y)$, is calculated using Equation (8). This equation integrates the local mean and the original pixel value within each window, weighted by the filtering coefficient.

$$L(x, y) = \hat{m}(x, y) + c(x, y) \times (I(x, y) - \hat{m}(x, y)) \tag{8}$$

In SAR images, boundaries are identified by rapid changes in grayscale values [29]. To detect these edges, we apply the Laplacian operator, following the differentiable Lee filter.

Common edge detection operators, such as Roberts, Prewitt, and Sobel, detect edges primarily in horizontal and vertical directions [30]. However, remote sensing images often feature irregular edges. To address this, we use the Laplacian operator, a second-order gradient operator that identifies edges by calculating pixel curvature in all directions.

Xu et al. [31] uses Laplacian edge detection tools integrated in MATLAB to extract image edges before feeding them into a deep learning network. This method does not allow for end-to-end optimization and training. Zhang et al. [32] implement the Laplacian operator by entirely adjustable convolution kernels, without any preset parameters. This method relies entirely on network training, which may deviate from the intended effect, especially in the presence of significant noise. We implement the Laplacian operator with a fixed 2D convolution kernel via PyTorch:

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & -8 & 1 \\ 1 & 1 & 1 \end{bmatrix} \tag{9}$$

Finally, we enhance SAR images by adding edge features extracted by the Laplacian operator to the filtered images using a residual connection.

Since most parameters are preset and not involved in training, it introduces only a few trainable parameters. We listed the FLOPs in Table 2, showing that this module has a low computational cost but delivers good performance.

Table 2. Statistics of ablation experiment results on WHU-OPT-SAR. (a) ResNet-50; (b) ResNet-50 + SAIM; (c) Laplacian + SAIM; (d) Laplacian + MAFM + SAIM; (e) Laplacian + MAFM + SAIM + GLFM. Bold and underlined indicates the best result.

Method	PA (%)	MIoU (%)	Params (M)	FLOPs (G)	P (%)						
					Farmland	City	Village	Water	Forest	Road	Others
a	77.27	51.83	**52.7**	**40.1**	79.88	56.23	48.07	68.46	78.77	27.87	17.21
b	79.25	53.43	71.9	57.29	81.73	59.86	53.52	69.75	80.51	28.28	21.54
c	80.31	54.61	71.9	59.52	82.59	60.02	53.24	69.89	82.17	29.02	22.94
d	80.71	55.06	81.7	71.80	83.61	62.18	54.50	74.24	83.15	29.71	22.81
e	**81.32**	**55.70**	85.1	78.54	**84.02**	**62.21**	**55.07**	**76.30**	**86.91**	**30.16**	**23.00**

4.2.2. Multi-Source Attention Fusion Module

In the multi-level features extracted by the backbone network, there exists both complementary information and redundant information. Before feature fusion, it is essential to identify valuable feature channels and filter out redundant ones. SENet [33] proposed a channel attention mechanism that squeezes features along the channel dimension and employs a multi-layer perceptron (MLP) to model the importance of each channel. The obtained channel weights will be weighted channel-by-channel to the original features.

However, SENet is limited to a single-modal mode. We designed the multi-source attention fusion module (MAFM) in Figure 6 to overcome this limitation. The MAFM enriches the feature representation by expanding the receptive field and increasing the feature squeeze methods. In the end, a cross-fusion of the weighted channel features is needed to better utilize the complementary information between modalities.

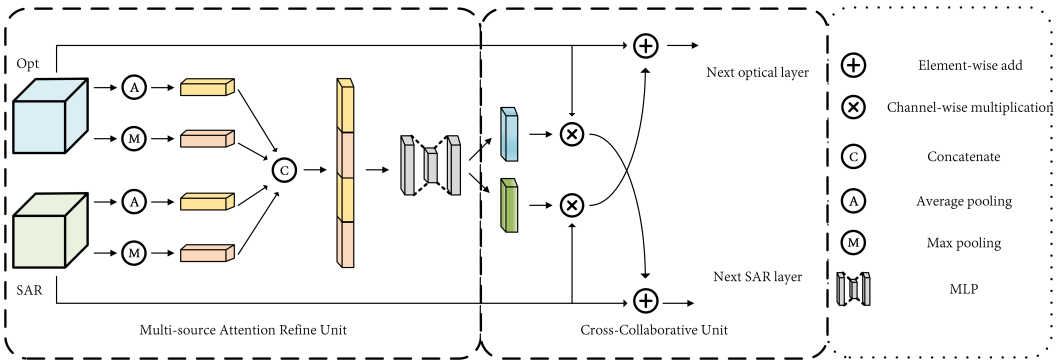

Figure 6. Structure of the multi-source attention fusion module. The architecture can be divided into two main components: the multi-source attention refine unit and the cross-collaboration unit.

Compared to the single squeeze method in SENet, the MAFM utilizes both global average pooling (GAP) and global maximum pooling (GMP) for each modality. GAP performs an averaging operation on the feature map of each channel, extracting overall trends and global information. GMP, on the other hand, extracts the most salient features from each channel's feature map. The mathematical expressions for the two squeeze methods are as follows:

$$GAP_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{(i,j,c)} \tag{10}$$

where H and W are the height and width of the feature map, respectively, and $X_{(i,j,c)}$ denotes the value of the cth channel of the feature map at position (i, j).

$$GMP_c = max_c X_{(i,j,c)} \tag{11}$$

where $max_c(\cdot)$ indicates that the maximum value is taken at all positions of channel c.

GAP and GMP generate two one-dimensional vectors of size $c \times 1 \times 1$ for each modality. These four one-dimensional vectors are then concatenated along the channel dimension to form a vector of $4c \times 1 \times 1$, representing the global distribution of feature channels and providing a comprehensive global receptive field of two modalities. Then, this concatenated vector is fed into a multi-layer perceptron (MLP), described by the following expression:

$$z^{(l)} = \sigma(\omega^{(l)} z^{(l-1)} + b^{(l)}) \tag{12}$$

where l denotes the position of layers of the current multi-layer perceptron, $\sigma(\cdot)$ denotes the activation function, ω denotes the weight coefficients, and b denotes the bias coefficients. The MLP has three layers with output dimensions of $2c$, c, and $2c$, respectively. This design

explicitly models the importance of feature channels, resulting in a channel attention weight of size $2c \times 1 \times 1$.

Finally, we split this channel attention weight into two vectors of $c \times 1 \times 1$ for each modality, applying them to the original features for channel-wise weighting. The weighted features are then cross-fused into the other modality, enabling effective inter-modal information sharing. The process can be expressed as

$$Out_{opt} = (Input_{sar} \otimes W_{sar}) + Input_{opt} \tag{13}$$

$$Out_{sar} = (Input_{opt} \otimes W_{opt}) + Input_{opt} \tag{14}$$

where $Input_{opt}$ and $Input_{sar}$ represent the original input features. W_{opt} and W_{sar} denote the previously obtained channel attention weights, and $\otimes$ signifies channel-wise multiplication.

Both the first layers of L2 and L3 in the backbone introduce the MAFM module. This facilitates the multiscale enhancement and cross-modal transmission of domain-shared features, achieving complementary features at different scales.

4.3. Synergistic Attention Integration Module

Attentional mechanisms are extensively employed in computer vision tasks to exploit correlations between features [27,34–36]. Optical images provide rich information, such as color, boundaries, and texture, which can be used to classify different objects. These features are distributed across various feature channels. In contrast, synthetic aperture radar (SAR) has the capability to penetrate objects and capture internal edge information. Based on these characteristics, we designed the synergistic attention integration module, illustrated in Figure 7. The module employs a channel attention mechanism to integrate land cover classification information from optical images into SAR features, enhancing SAR's ability to model the relationship between backscatter and various target types. Simultaneously, it utilizes a positional attention mechanism to incorporate SAR's internal edge information into optical features. This dual approach improves the model's capability to differentiate between objects with similar colors and textures.

In the optical branch, we take a three 1×1 convolution with shared weights to extract the optical input into three feature vectors Q , K, and V of shape (C, H, W), and we then reshape Q and K into $(C, H \times W)$ and $(H \times W, C)$, respectively, denoted as A and B. Finally, a Softmax operation is performed on the result of multiplication of A and B to obtain the channel attention weights, which can be expressed as follows:

$$A = reshape[Conv_1(opt), (C, H \times W)] \tag{15}$$

$$B = reshape[Conv_1(opt), (H \times W, C)] \tag{16}$$

$$F_{channel} = Softmax(A \otimes B) \tag{17}$$

where $reshape(a, b)$ denotes the deformation operation, a denotes the deformed object, and b denotes the target shape. $Conv_i()$ denotes the convolution operation with kernel i, and $\otimes$ is the matrix inner product operation.

In the SAR branch, K and V are reshaped to (H×W, C) and (C, H×W), respectively, to calculate the positional attention weights. The remaining operations are similar to those in the optical branch. The entire process can be expressed as follows:

$$C = reshape[Conv_1(SAR), (H \times W, C)] \tag{18}$$

$$D = reshape[Conv_1(SAR), (C, H \times W)] \tag{19}$$

$$F_{pos} = Softmax(C \otimes D) \tag{20}$$

However, the positional attention mechanism may encounter difficulties when dealing with irregular objects (such as forests, villages, etc.). For example, the transition between trees and bare land is abrupt rather than gradual. This discontinuity makes it challenging for positional attention to find consistent similarity or distance metrics to identify these features effectively. To address this limitation, we incorporate the edge detected by the Laplacian operator from the optical images into the positional attention feature maps to assist in object localization. This operation can be expressed as follows:

$$F_{Laplacian} = Laplacian(GAP(opt_{raw})) \tag{21}$$

$$F_{Lap+pos} = Softmax(F_{pos} \oplus flatten(F_{Laplacian})) \tag{22}$$

where $Laplacian(\cdot)$ denotes Laplacian edge detection, $GAP(\cdot)$ denotes global average pooling operation, $\oplus$ denotes element-by-element summation, and $flatten(\cdot)$ denotes spreading the matrix.

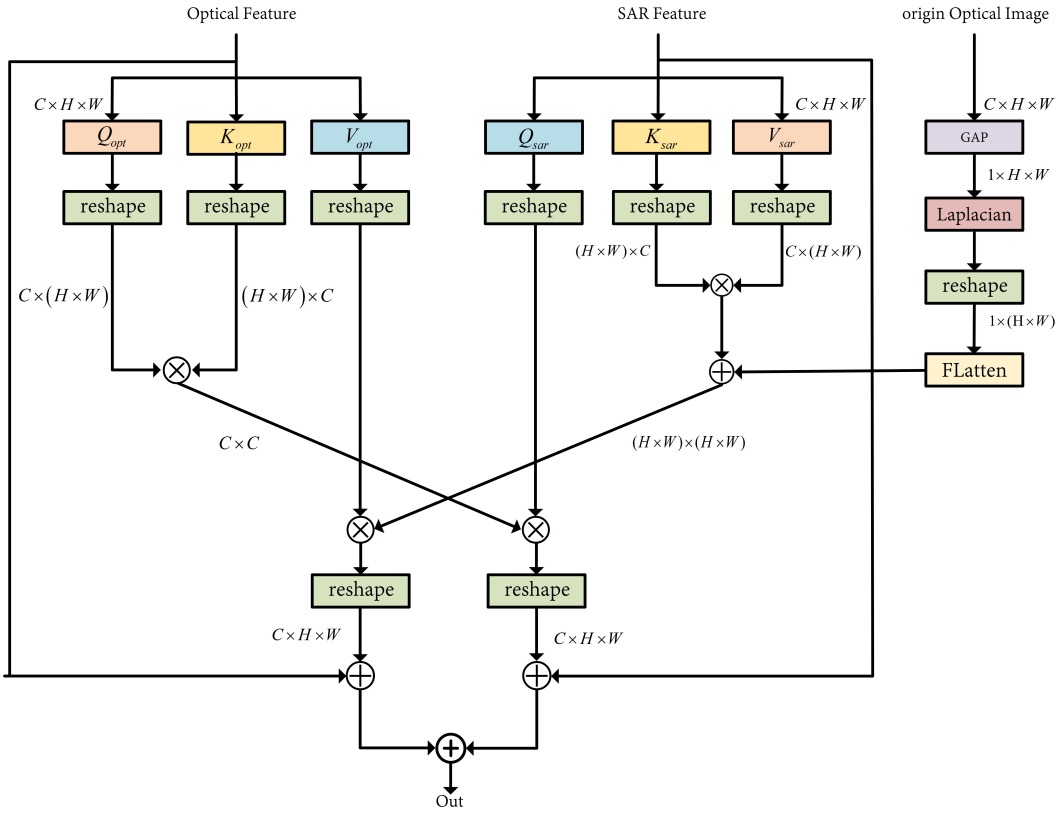

Figure 7. Synergistic attention integration unit, where $\otimes$ denotes tensor multiplication operation and $\oplus$ is element-wise addition operation.

Subsequently, the attention weights from each branch are cross-weighted, and residual connections are used to retain the original information in each branch. The expression for this operation is as follows:

$$Out_{sar} = SAR + F_{channel} \otimes \{reshape[Q_{sar}, (C, H \times W)]\} \tag{23}$$

$$Out_{opt} = Opt + \{reshape[V_{opt}, (C, H \times W)]\} \otimes F_{lap+pos} \tag{24}$$

The outputs from the dual branches are fused into a pyramid structure. Through dilated convolutions with varying dilation rates, the convolutional receptive fields are altered to integrate features from different scales [37,38].

4.4. Global-Guided Local Fusion Module

An upsampling module global-guided local fusion module (GLFM) (Figure 8) is designed to fully utilize deep and low-level feature information.

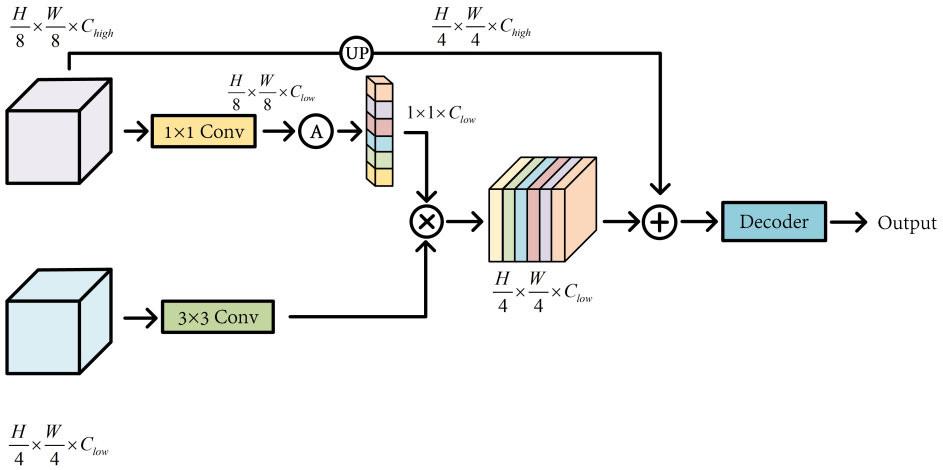

Figure 8. Structure of the global-guided local fusion module. "up" represents the upsampling operation, ⊗ denotes the tensor multiplication operation, and ⊕ signifies the element-wise summation operation.

Using both scales would lead to feature redundancy [39], increasing computational complexity. Therefore, we need to select appropriate scales that represent both shallow and deep features while considering the computational efficiency. Taking a 256 × 256 image as an example, Figure 9 illustrates the extracted feature information at different scales:

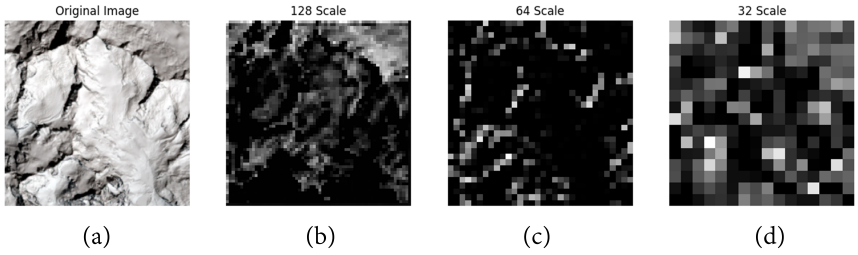

Figure 9. Feature visualization results at different scales. Part (**a**) is the original image, and parts (**b**), (**c**), (**d**) are the results of down-sampling 2, 4, and 8 times, respectively.

Features at the 1/2 scale exhibit a high degree of similarity with those at the 1/4 scale. However, shallower-level features increase the model's sensitivity, which fails to enhance the model's generalization ability [40]. The 1/4 scale features capture the high-frequency details of the image, reflecting its fundamental structural information. Additionally, the 1/4 scale is more computationally efficient than the 1/2 scale.

The features at a 1/8 scale are more abstract, and the deeper semantic information is less sensitive to small changes [41], enhancing the model's generalization capabilities when faced with new data [42].

Therefore, features that are down-sampled by factors of 4 and 8 are selected for multiscale fusion. We discard shallow SAR features, although we designed modules to

suppress SAR noise, but it is necessary to ensure that noise impact is isolated in the fusion close to the output layer.

The deep features are first processed through a 1×1 convolution, creating a new feature map that retains the same number of channels as the shallow features. This is followed by global average pooling to generate channel attention weights with a size of $1 \times 1 \times C$. These weights are then applied channel-wise to the shallow features, which were processed by a 3×3 convolution, thereby enabling a global guiding effect [43]. Furthermore, deep features are upsampled by a factor of two to align with the size of the shallow features. After the element-wise addition of multi-scale features, different decoders are selected depending on the specific downstream tasks; for instance, a fully connected network is used to flatten the features for final output when estimating snow parameters.

5. Experiments on WHU-OPT-SAR

To verify the effectiveness of OSNet in the segmentation task, experiments were conducted on the publicly available WHU-OPT-SAR dataset [4]. The final fully connected layer of the model was replaced with an upsampling layer to adapt to the semantic segmentation task.

5.1. Experimental parameter setting

All experiments are conducted using an Intel Core i5-12400F CPU (2.50 GHz) sourced from Intel Corporation, Santa Clara, California, USA, and an NVIDIA RTX 3080 GPU sourced from NVIDIA Corporation, Santa Clara, California, USA. The deep learning framework employed is PyTorch (version 1.10.0), and the optimizer used is Adaptive Moment Estimation (Adam).

CrossEntropyLoss is used as the loss function, and precision (P), pixel accuracy (PA), and mean intersection over union (MIoU) are used as the evaluation indexes. The formula for each evaluation index is as follows:

$$P = \frac{TP}{TP + FP} \tag{25}$$

where TP, true positive, represents the number of pixels correctly predicted as positive class; FP, false positive, represents the number of pixels incorrectly predicted as positive class.

$$PA = \frac{\sum_{i=0}^{k} p_{(i,i)}}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{(i,j)}} \tag{26}$$

where $p_{(i,i)}$, diagonal elements, represents the number of pixels correctly predicted as class i; $p_{(i,j)}$, non-diagonal elements, represents the number of pixels belonging to class i but predicted as class j; k is number of classes (excluding the background).

$$MIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{\sum_{j=0}^{k} p_{(i,j)}}{\sum_{j=0}^{k} p_{(i,j)} + \sum_{j=0}^{k} p_{(j,i)} - p_{(i,i)}} \tag{27}$$

where $p_{(i,j)}$ represents the number of pixels correctly predicted as class i and actually belonging to class j; $p_{(j,i)}$ represents the number of pixels correctly predicted as class j and actually belonging to class i; $p_{(i,i)}$ represents the number of pixels correctly predicted as class i; and k is number of classes (excluding the background).

5.2. Ablation Experiments on WHU-OPT-SAR Dataset

The proposed modules (modified backbone (Laplacian convolution + MAFM), SAIM, and GLFM) are integrated into the model step by step, and network performance is

evaluated using MIoU, PA, and Params, with the computation cost presented in Table 2. Params refer to the model's total number of trainable parameters used to measure the model's complexity and scale. FLOPs refer to the floating-point operations required for a single forward pass, representing the model's computational complexity.

Each module improves the model's accuracy but inevitably increases the computational cost. Among them, SAIM and MAFM are the main contributors to the increased computing cost, with an increase of 17.19 G and 12.28 G in FLOPs, respectively. The reason is that these two modules use the attention mechanism in the high-dimensional feature space, and many dot product operations lead to increased computing costs. However, these two modules significantly improve model performance: SAIM increases accuracy by 5.45% in the village category, and MAFM increases accuracy by 4.35% in the water category. Laplacian convolution has low computational costs due to preset parameters and simple calculations, but it improves the feature quality of SAR branches and shows good improvement in various categories. In the decoding stage, GLFM integrates multi-scale features with fewer parameters and computational costs, among which low-scale features include shallow information, such as edge and surface color, effectively improving the accuracy of forest classification.

Through heat maps in Figure 10, we can visualize the effect of each module more intuitively. The heat map demonstrates how much attention the model pays to different categories of regions. The intensity of the red region indicates the model's primary focus, followed by the yellow-green area, with the blue representing areas of lower attention.

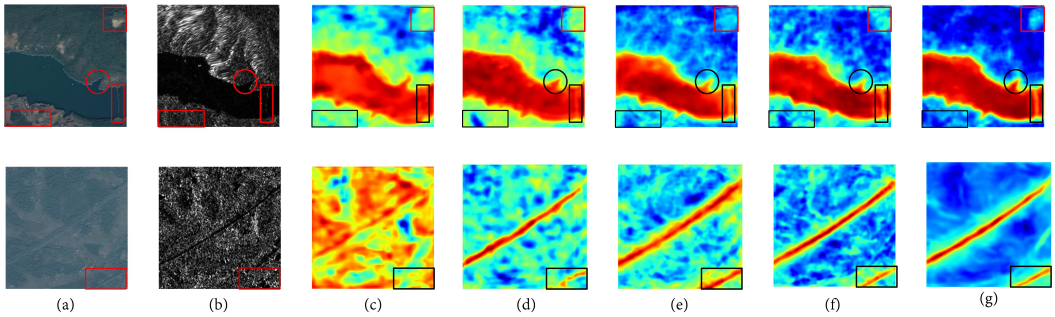

Figure 10. Heat maps of ablation experiments. (**a**) Optical image. (**b**) SAR image. (**c**) ResNet50. (**d**) ResNet-50 + SAIM. (**e**) Laplacian + SAIM. (**f**) Laplacian + MAFM + SAIM. (**g**) Laplacian + MAFM + SAIM + GLFM.

Heat maps of attention for water and roads are purposely selected and presented in rows one and two of the figure. Water and roads usually have a variety of scales, shapes, and texture variations, as well as complex boundaries and features similar to those of their surroundings. Therefore, accurately detecting and identifying watersheds and roads is challenging.

Ablation of SAIM:Compared to ResNet, SAIM corrects the wrong area of the water (rectangle in the bottom left corner of the first row). However, its outline shape is slightly rough (round in the first row, square in the bottom right corner of the second row). Furthermore, too many yellow and green areas almost cover the entire image, which means that the model still focuses on too much redundant information.

Ablation of Laplacian Convolution: We find that in the optical image of the second row, the color of the road looks similar to the surroundings, and in the SAR image, the location of the road has an obvious edge over the surroundings. Laplacian convolution helps the model to capture the location of the feature boundaries better, so Column e has a sharper outline of the region of interest compared to the previous heat map. The black box of the

first row in the lower right corner shows that the Laplacian operator's introduction makes the model notice the slender bridge.

Ablation of MAFM: The MAFM module eliminates unnecessary regions of interest by dynamically adjusting the weights between different modal channels. It selectively emphasizes feature channels that are critical to the task while suppressing responses from task-irrelevant or noisy channels. Column f clearly shows the decrease in the yellow–green area.

Ablation of GLFM: The deep features of the model encapsulate the overall structure and semantic information of the entire image, whereas the shallow features focus on local details. The GLFM enables the model to comprehend local information from a global perspective. As depicted in the figure, the multi-scale module effectively eliminates unnecessary regions of interest, preserving and optimizing the details of the relevant regions.

Figure 11 visualizes the impact of the introduction of different modules on the segmentation results.

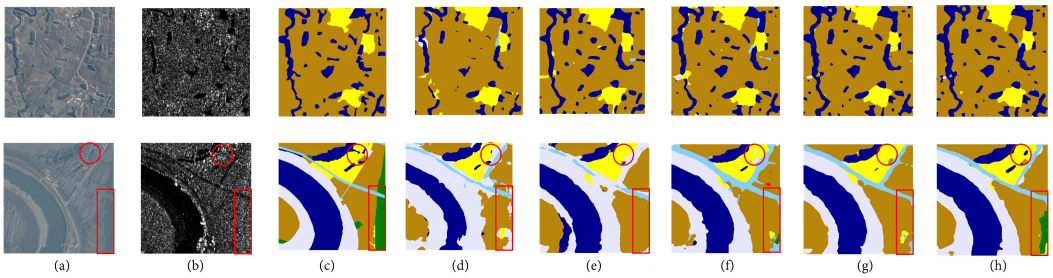

Figure 11. Segmentation result for ablation experiments. (**a**) Optical image. (**b**) SAR image. (**c**) Labels. (**d**) ResNet-50. (**e**) ResNet-50 + SAIM. (**f**) Laplacian + SAIM. (**g**) Laplacian + MAFM + SAIM. (**h**) Laplacian + MAFM + SAIM + GLFM. The dataset is annotated with seven main categories: farmland (brown), urban (red), village (yellow), water (blue), forest (green), road (cyan), and others (white).

In the first set of images, the lakes are fragmented and scattered, with complex contour shapes. The SAR images are almost covered by scattering noise, with only part of the waters showing clear contours. The second set of images shows a typical complex scene of a river running through a village. The river and the road are clearly distinguishable in the SAR image.

The segmentation results of ResNet-50 in Column d have the problem of blurred edges. The scattered water area in the middle of the image in the first row is not correctly identified, and the water area in the second row is much smaller than the label, resulting in many misidentifications. After the introduction of SAIM, the accuracy of the water area is improved. However, there is still much misidentification, indicating that only the fusion of deep features cannot significantly improve the model performance. After introducing Laplacian convolution in Column f, the road recognition in the second-row image is improved, and the contour information is more precise. Column g introduces the MAFM module, which allows multi-scale features in the backbone network to be fused, improving the accuracy of roads and waters. It can also be seen from the heat map that the MAFM module effectively weakens the influence of redundant information, so the false recognition area is reduced in the segmented image. There is a forest area in the lower right corner of the second line image, darker than the surrounding color in the optical image but similar to the surrounding backscattering in the SAR image. The GLFM module combines deep semantic and shallow information to distinguish this area from the surrounding land and improve the recognition rate of forest area. Although some areas are still not identified, this reflects GLFM's necessity.

However, the model exhibits an excessive association between low backscattering and classification as the water category. In the first row of Figure 11, several scattered farmland

areas are misclassified as water, typically corresponding to regions with low backscatter. In the upper right corner of the second row, indicated by the red circle, the model accurately identifies only the low backscattering area as water.

5.3. Comparison Test of WHU-OPT-SAR Dataset

MCANet [4], ACNet [44], RDFNet [45], V-FuseNet [46], CMGFNet [47], and DeepLab v3+ [48] were selected for comparative experiments. Figure 12 shows four groups of segmentation results.

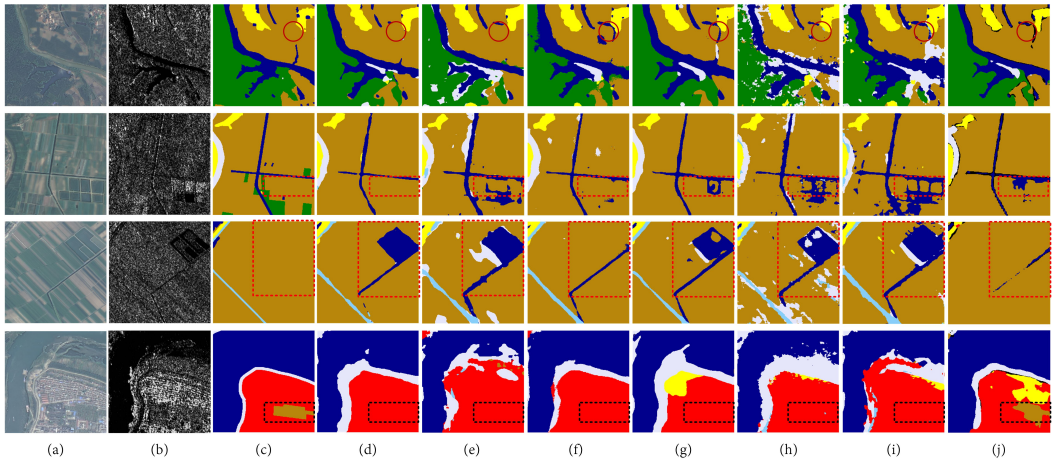

Figure 12. Comparison results on the WHU-OPT-SAR dataset. (**a**) Optical image. (**b**) SAR image. (**c**) Labels. (**d**) OSNet. (**e**) MCANet. (**f**) ACNet. (**g**) RDFNet. (**h**) V-FuseNet. (**i**) CMGFNet. (**j**) DeepLab v3+. The dataset is annotated with seven main categories: farmland (brown), urban (red), village (yellow), water (blue), forest (green), road (cyan), and others (white).

It is evident that Columns e, h, and i are seriously affected by noise, and many error pixels are scattered in the picture, demonstrating that the noise problem cannot be ignored.

In the red circle of the first row, the optical images show an obvious outline. However, due to the low backscattering coefficient of the water area, the SAR image can accurately determine that this region is not water. In the lower right corner of the second line, the SAR image shows a low backscatter, but the optical image can determine that the area is farmland. Through a comprehensive analysis of optical images and SAR images, the combination makes the judgment more accurate. However, other models produce many misjudgments in these two regions, proving that our model effectively utilizes the difference between the two modalities. Notably, the failure of forest areas in the second line to be correctly identified is a common problem for all models. There is no obvious outline in the optical image and SAR image. In this case, distinguishing forest and farmland is still a problem that needs to be solved in the future.

In the third row, the backscattering in the red-boxed area is significantly lower than in the surrounding areas, and OSNet misclassifies it as water.

In the WHU-OPT-SAR dataset, water labels constitute 38% of the total, leading to class imbalance that affects model training. We addressed this by setting the cross-entropy loss weights based on the proportion of each class. Despite the model correctly identifying some road areas (e.g., the upper left corner of the third row in Figure 12 and the roads in the second row of Figure 11), the red-boxed area is still misclassified as water. This suggests that weight adjustment alone may not completely resolve the issue of class imbalance.

We analyze the distribution of three low-backscattering categories in the labels (water, road, and farmland), which are 51.35%, 1.36%, and 47.29%, respectively. For these pixels, the predicted proportions are 53.8%, 1.17%, and 43.2%. Compared to the true values,

the proportion of predicted water pixels increases, while the other two categories decrease. This indicates that the model has a tendency to classify low-backscattering areas as water, leading to misclassification of some farmland and road pixels.

To understand the model's misclassification tendencies, we analyze the pixel values representing backscattering intensity in the SAR images. The average values for water, road, and farmland in the true labels are 27.87, 44.26, and 57.46, respectively; in the predictions, they increase to 28.38, 52.07, and 59.54. The significant rise in the road category's value, nearing that of farmland, suggests that some road and farmland areas with lower backscattering intensity (potentially due to surface cover or soil type) are misclassified as water. Consequently, roads and farmlands in the predictions appear only when the backscattering intensity is relatively high, highlighting the model's bias towards classifying low-backscattering regions as water.

In addition, although our model performs well on contour detection, small areas within large outlines are not accurately captured. The small edges in the farmland area in the city (last line) are challenging to distinguish for current models.

A quantitative evaluation is performed to compare the effectiveness of these methods presented in Table 3.

Table 3. Comparison statistics on the WHU-OPT-SAR dataset. Bold and underlined represent the best results.

	PA	MIoU	Params (M)	FLOPs (G)	P						
					Farmland	City	Village	Water	Forest	Road	Others
MCANet	0.7901	0.5312	95.92	102.38	0.8111	**0.624**	0.5235	0.7299	0.8573	0.2293	**0.2585**
ACNet	0.7820	0.5252	116.60	53.10	0.8256	0.6115	0.4935	0.7155	0.8559	0.2451	0.2126
RDFNet	0.7891	0.5291	443.86	178.70	0.8102	0.5843	0.5261	0.7126	**0.8708**	0.2764	0.2337
V-FuseNet	0.7487	0.4850	**58.90**	81.48	0.7829	0.5510	0.4695	0.7027	0.8263	0.1796	0.0672
CMGFNet	0.7693	0.5055	85.21	**38.45**	0.8107	0.5687	0.4784	0.7250	0.8326	0.2323	0.1886
DeepLab v3+	0.7760	0.5194	59.34	40.80	0.8087	0.5709	0.4786	0.7179	0.8629	0.2563	0.1783
OSNet	**0.8132**	**0.5570**	85.1	78.54	**0.8402**	0.6221	**0.5507**	**0.7630**	0.8691	**0.3016**	0.2300

In Table 3, we compare the computational cost of the MCANet model (proposed by the authors of WHU-OPT-SAR) and its baseline Deeplab V3+ model. It can be observed that the FLOPs of MCANet have nearly doubled compared to the baseline model, with MIoU and PA metrics increasing by 1.5%. In contrast, our model, OSNet, is more computationally efficient than MCANet, achieving a % improvement compared to the baseline model.

OSNet achieved the highest pixel accuracy (PA) of 81.32% and the highest MIoU of 55.7% with a moderate computational cost. Compared to other methods, our model significantly outperforms in accuracy for farmland, village, water and road, attaining 84.02%, 55.07%, 76.3%, and 30.16%, respectively. In the city and forest metrics, OSNet is slightly lower than the best model by 0.19% and 0.17%, respectively.

6. Experiments on SNOW-OPT-SAR

In this section, we will first explore the contribution of different modalities to snow depth prediction in the SNOW-OPT-SAR dataset. Subsequently, we will verify the performance of OSNet on the regression task through comparative experiments and ablation studies.

6.1. Experimental Parameter Settings

The experimental configuration is the same as in the previous section. The experiments utilized the Huber loss function, which behaves like Mean Squared Error (MSE) for small errors and resembles Mean Absolute Error (MAE) for large errors. This loss function is

more robust to data outliers. The experiment involved training for 150 epochs, with 80% of the data allocated for training and 20% for testing. The experiment is repeated ten times and the average was taken as the final result.

Mean absolute error (MAE) and root mean squared error (RMSE) were used as evaluation indexes, and the mathematical expressions are as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{28}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{29}$$

where y_i denotes the ith predicted value, $\hat{y}_i$ denotes the ith labeled value, and n is the amount of data. The smaller the two parameter indicators, the higher the accuracy.

6.2. Ablation Experiments on Data Combination

In this section, we conduct data ablation experiments with the ResNet-50 network. We split the optical images into four combinations: RGB, RGB + NIR (RGBN), SAR, and RGBN + SAR. This approach allows us to discuss the contribution of different band at various snow depth levels and demonstrate that data fusion improves prediction accuracy.

Table 4 shows results of data ablation. It demonstrates the sensitivity of the sensors involved in retrieving snow depth. Due to the limited penetration ability of optical images, it cannot obtain internal information of snow, making the overall performance of RGB and RGBN inferior to SAR. The optical band can extract the snow cover range, and the snow cover and color information are more suitable for retrieving very shallow snow (0–10 cm), which is difficult to distinguish further on in deeper snow. It can be seen that only in the prediction of 0–10 cm can the optical images reach an MAE below 0.3. Compared with RGB, adding NIR slightly improves the accuracy, but the improvement is limited. SAR is more advantageous than the optical band in acquiring the internal information of the snow layer. It is significantly better than the optical combination in different snow depth levels, especially in the 20–30 cm prediction, where it achieves 26% lower MAE than the optical band.

Table 4. Ablation experiment of data combination (bold and underlined represent the best result).

Data Combination	Metric	0–10 cm	10–20 cm	20–30 cm	>30 cm
RGB	MAE	0.2694	0.3446	0.4947	0.4198
	RMSE	0.3608	0.4399	0.5570	0.4993
RGB + NIR	MAE	0.2753	0.3368	0.4572	0.3759
	RMSE	0.3661	0.3902	0.5360	0.4953
SAR	MAE	0.1811	0.2262	0.2338	0.2771
	RMSE	0.2633	0.2747	0.3534	0.3752
RGB + NIR + SAR	MAE	**0.1425**	**0.1780**	**0.1959**	**0.2185**
	RMSE	**0.2080**	**0.2427**	**0.2971**	**0.3303**

The combination of SAR and optical images achieved optimal prediction results. SAR plays a primary role in snow depth prediction, and the integration of optical images significantly enhances the accuracy across various snow depth levels. Optical images clearly delineate the boundaries between snow-covered and snow-free areas, helping SAR reduce misclassifications in snow-free regions. As a result, incorporating optical images improves the mean absolute error (MAE) in shallow snow layers by 3.86% compared to using SAR alone.

Ridges segment the snow-covered areas into discrete snow patches, and we believe these regions exhibit different pattern characteristics. Optical images accurately identify these scattered snow areas, aiding SAR in learning features from different patterns. The ridges typically have shallow or no snow, helping SAR to focus on deep snow regions. This combined approach significantly enhances the accuracy of deep snow area predictions.

6.3. Comparative Experiments on SNOW-OPT-SAR

In the comparative experiment, we tested the prediction accuracy of each model on the same training and testing datasets. VGG [49], ResNet [26], ResNeXt [50], Inception [51], Xception [52], and SENet [33] are selected as the comparison models. Since all these models are single-modal networks, we adopt the stacking of optical and SAR images as input.

Table 5 presents the performance of various classical deep learning models on the SNOW-OPT-SAR dataset. After conducting ten rounds of cross-validation, the metrics' average, best (upper), and worst (lower) values indicate that our proposed model demonstrates higher accuracy and excellent robustness. The mean MAE and RMSE of our model are higher by 3.14% and 4.22%, respectively, compared to the second-best model. The difference between the best and worst values is 1.65% for MAE and 2.6% for RMSE, indicating that our model demonstrates stronger robustness compared to other models.

Table 5. Performance comparison of each model in snow depth estimation with SNOW-OPT-SAR (bold and underlined represent the best result).

Model	MAE			RMSE		
	Average	**Upper**	**Lower**	**Average**	**Upper**	**Lower**
Vgg19	0.2940	0.2832	0.3132	0.5555	0.5215	0.5992
ResNet50	0.1710	0.1609	0.1854	0.2612	0.2544	0.2778
ResNeXt	0.1563	0.1511	0.1620	0.1659	0.1625	0.1697
Inception	0.2091	0.1957	0.2281	0.3565	0.3417	0.3933
Xception	0.3612	0.3569	0.3643	0.4109	0.3916	0.4241
SENet	0.1510	0.1462	0.1635	0.1757	0.1559	0.2002
Ours	**0.1196**	**0.1116**	**0.1281**	**0.1335**	**0.1179**	**0.1439**

Figure 13 visualizes the predictions made by each model in comparison to the station observations, providing a clear visual assessment of their accuracy.

VGG's simple network structure design fails to fully comprehend modal features, resulting in poor performance. Inception and Xception, despite having complex architectures, lack sufficient parameters to ensure thorough decoupling of the modalities. ResNet and ResNeXt produce large errors in some samples. SENet, incorporating an attention mechanism, achieves relatively good results, demonstrating that attention mechanisms are rational for modal fusion tasks. They highlight important features and reduce the focus on irrelevant information. Our model surpasses others in prediction accuracy, with fewer gross errors. The following section will further discuss the model through ablation experiments, discussing how each module in our model enhances snow depth retrieval accuracy.

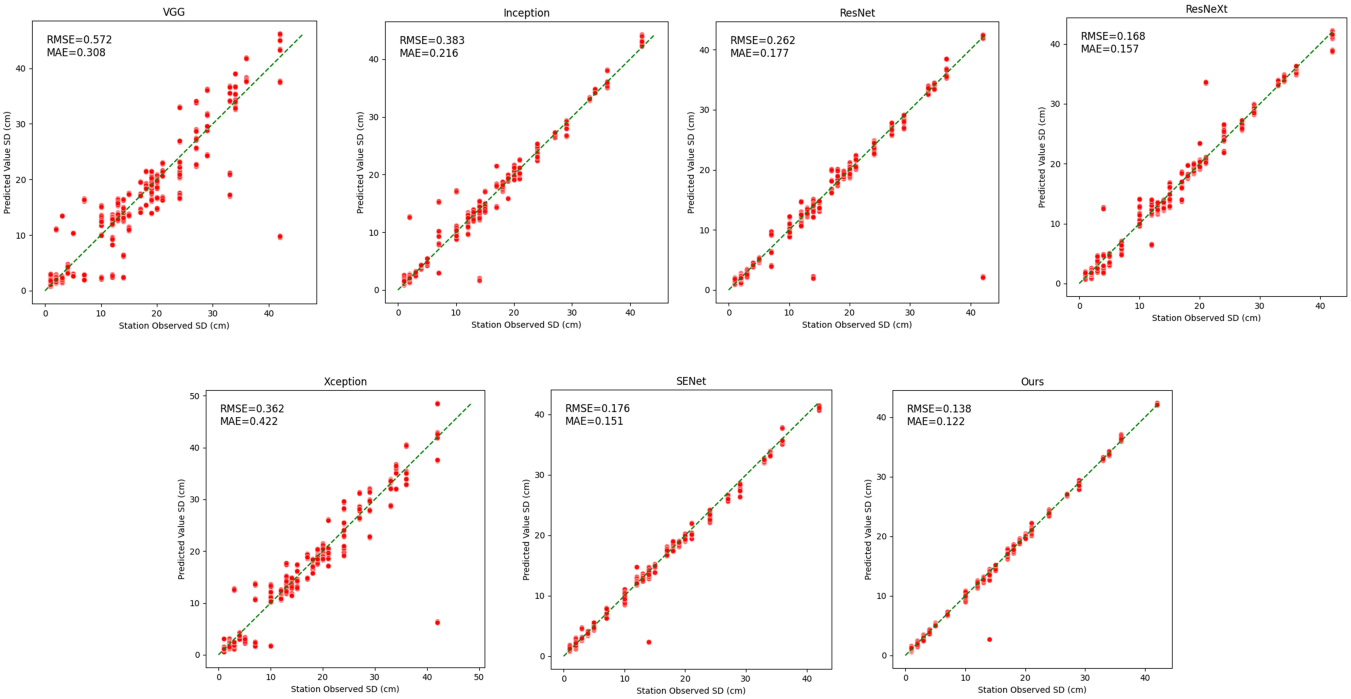

Figure 13. A scatter diagram of station measurements versus predictions by each network. The green line is the y = x line.

6.4. Ablation Experiments on SNOW-OPT-SAR Dataset

We conducted a series of ablation experiments to investigate the impact of each module on snow depth estimation. We will discuss the effect of the modified backbone (Laplacian convolution and MAFM) compared to ResNet50, followed by the incremental addition of SAIM and GLFM modules. Table 6 shows the impact on metrics at different snow depths by introducing each module. Figure 14 shows the density distribution of the model's prediction results on the test set by each module. Brighter colors indicate higher data point density in that area. A target line is provided in the figure to help assess the deviation between prediction and true value. The closer the data points are to this line, the more accurate the predictions are.

Table 6. Ablation experiments of modules of OSNet on SNOW-OPT-SAR (bold and underlined represent the best results).

Model	Metric	0–10 cm	10–20 cm	20–30 cm	>30 cm
ResNet50	MAE	0.1425	0.1780	0.1959	0.2185
	RMSE	0.2080	0.2427	0.2971	0.3303
Modified-Backbone	MAE	0.1219	0.1653	0.1399	0.1486
	RMSE	0.1321	0.2057	0.1425	0.1418
Modified-Backbone + SAIM	MAE	0.1130	0.1549	0.1307	0.1256
	RMSE	0.1291	0.1780	0.1404	0.1337
Modified-Backbone + SAIM + GLFM	MAE	**0.1093**	**0.1505**	**0.1260**	**0.1127**
	RMSE	**0.1218**	**0.1750**	**0.1297**	**0.1264**

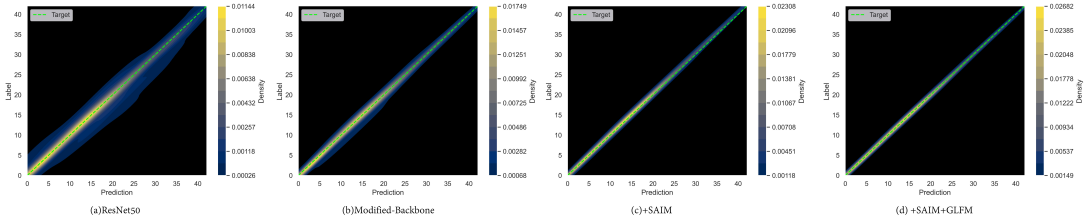

Figure 14. Two-dimensional histograms for reference data versus predicted snow depth values for each ablation experiment.

The predictions from ResNet are relatively scattered, with density regions not being concentrated. The results show a downward trend in the 10–35 cm range, indicating that the model often underestimates snow depth in this range. The introduction of the modified backbone significantly improves the prediction accuracy for depths above 20 cm, resulting in an overall improvement. This improvement is due to modified backbone's dual-branch design that allows the model not to fuse the features from the input but to weight channels after extracting corresponding features in their respective branches via MAFM, emphasizing features that are more effective for snow depth prediction and then fusing them.

Introducing SAIM at the end of the backbone enhances the semantic information in the optical branch and the positional information in the SAR branch. The deep-scale fusion moves the prediction distribution closer to the target line, especially in the 0–20 cm range. For deep snow, SAR images mainly provide effective features, so the improvement through fusion is moderate. For shallow snow, the complementarity of optical and SAR images is utilized, improving accuracy through fusion. Finally, GLFM introduces low-scale and deep-scale fusion during decoding, correcting the data distribution and aligning predictions more closely with actual values.

The weaker prediction results for snow depths in the 10-30 cm range may be due to the model's inability to extract sufficient features in this range. Since the dataset's SAR images only use VV polarization, we suggest that adding more polarization modes in the future, along with incorporating traditional methods, could help improve predictions in this range.

Additionally, in the design of SAIM, we propose using channel attention mechanisms and positional attention mechanisms on the optical and SAR branches, respectively. To verify the optimality of this combination, we conducted comparative experiments under the same experimental conditions to evaluate different combinations. The results are shown in Table 7. "Exchange" in the table indicates whether cross-weighting calculations were performed.

Table 7. Comparison of different combination of attention mechanism in SAIM. Bold and underlined represent the best combination. The checkmark symbol represents the category of attention mechanism used for each scheme.

OPT		SAR		Exchange	MAE	RMSE
Channel	Position	Channel	Position			
✓		✓		✓	0.1363	0.1559
✓			✓	✓	**<u>0.1196</u>**	**<u>0.1336</u>**
	✓	✓		✓	0.1334	0.1449
	✓		✓	✓	0.1381	0.1534
✓		✓			0.1510	0.1603
✓			✓		0.1288	0.1333
	✓	✓			0.1485	0.1895
	✓		✓		0.1357	0.1405

We find that the positional attention mechanism is essential for optimal performance. The worst results are observed when only channel attention is used without cross-weighting, yielding an MAE of 15.1%. This outcome occurs because our regression task requires predicting snow depth in center from its surrounding region. Without attention to the relevant area, the substantial amount of irrelevant information in the region can distort the estimation. Furthermore, the cross-weighting method proves effective in most combinations, indicating that exchanging weighted features is advantageous.

6.5. Visual Analysis

We present the heat maps before and after the application of the SAIM module in Figure 15. The redder the color, the more the model focuses; the darker the color, the less the model focuses. The SAIM module is located at the end of the encoder, where it fuses the boundary features of optical and SAR images before passing the output to the decoder. Consequently, the heat maps output by the SAIM module show the model's final focus. Parts (a) and (b) are the reference optical and SAR images, respectively. Part (c) shows the heat map before the application of SAIM, while part (d) shows the heat map after the application of SAIM.

We primarily rely on SAR images to detect snow depth. Optical images accurately distinguish between shallow or snow-free areas and deep snow regions, often revealing clear edges. In the SAIM module, we integrate optical features into SAR images. As shown in Figure 15, with the assistance of optical images, SAR's focus on snow-free and shallow snow areas is reduced, while attention to deep snow regions is increased. This approach decreases the likelihood of misclassifying shallow or snow-free areas as deep snow and filters out redundant information for SAR.

The images show that in areas with prominent edges, such as ridges and roads, the attention of SAR is diminished. These regions appear to be free of snow, the integration of optical images reduces the misclassification of snow depth in snow-free areas. Additionally, the successful detection of scattered black areas demonstrates that the color information from optical images helps SAR eliminate redundant information from snow-free regions. These edges divide the snow into several sections, each potentially influenced by the same weather pattern, leading to higher correlation in backscatter within these sections. Identifying these individual sections through optical images also helps SAR locate backscatter regions more relevant to the predicted locations, enhancing the model's utilization of useful information.

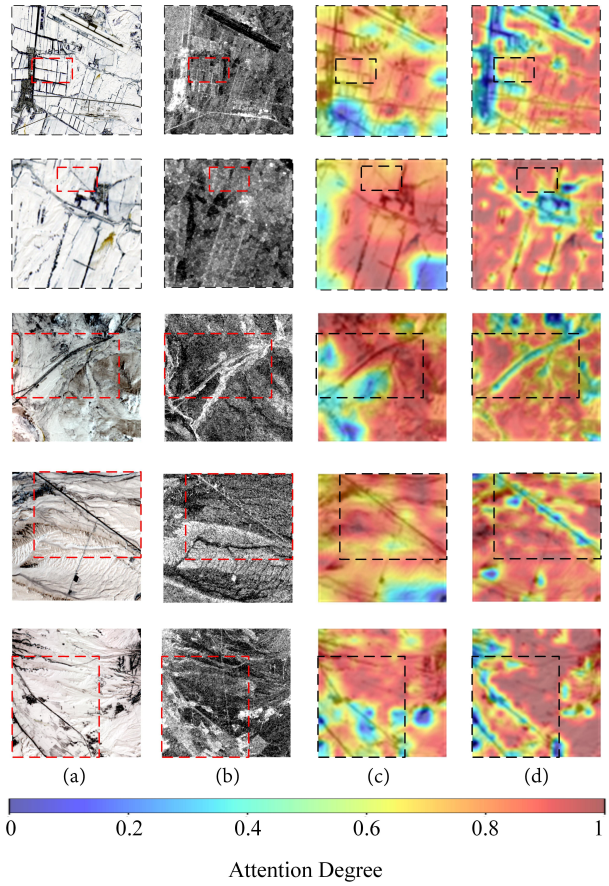

Figure 15. Heat maps before and after SAIM. Parts (**a**,**b**) are the reference optical and SAR images. Parts (**c**,**d**) are the heat maps before and after the application of SAIM. The first two rows are typical urban areas, while the remaining rows depict typical mountainous landscapes.

We find that in regions where both optical and SAR images exhibit distinct edges, our model can accurately identify these areas. Examples in the third and fourth rows demonstrate the effectiveness of the module in cross-fusing edge information from different modalities. However, the primary information for determining the boundaries between snow-covered and snow-free areas still comes from the optical images. In the red box at the bottom left of the last row, we can see that in areas where the SAR edges are blurry, the model relies solely on the optical edges to make an accurate judgment. In the middle section of the third row, the SAR image shows an additional branch compared to the optical image, which the model does not focus on. The optical image reveals that this extra branch is still snow-covered and should not be ignored. This highlights the importance of optical images in assisting SAR in predicting snow depth.

Figure 16 shows the snow depth map predicted by our model through optical and SAR images. We cut the optical and SAR into 64-by-64-sized samples with a step size of one pixel. The model predicts the snow depth at the center of each sample and then rearranges these predictions into a snow depth map.

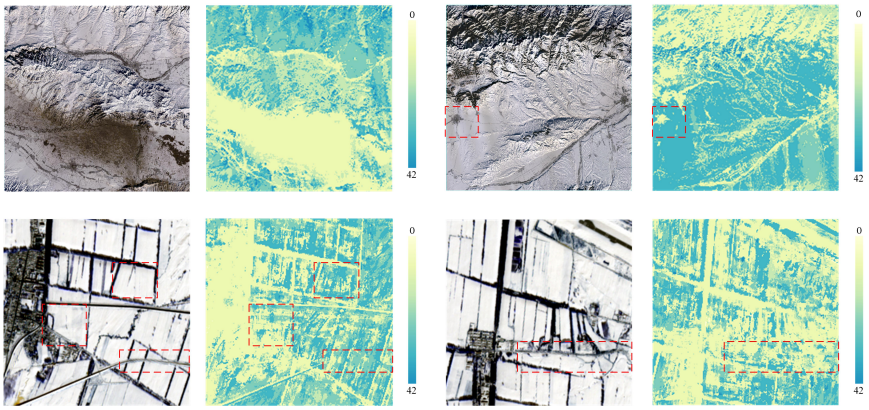

Figure 16. Snow depth mapping by OSNet. The first row displays the snow depth prediction map for mountainous areas, while the second row shows the prediction for urban areas.

The first row are snow depth prediction maps in mountainous areas. Ridges and bare ground are more accurately detected, and the predictions have distinct edges [53]. The continuity and texture of the predicted snow depths show the advantages of combining optical and SAR images to predict snow depths. This demonstrates that our method of using edge features to identify snow-free and snow-covered areas significantly enhances the model's performance, especially in shallow snow layers and snow-free regions. The red box in first row shows that the information of the optical image helps the model to effectively determine the presence or absence of snow, which is beneficial in shallow snow prediction.

The second row of images shows a typical urban scene, where the prediction results indicate that most snow-covered areas are predicted with reasonable accuracy, but the performance is relatively poorer compared to mountainous regions, as reflected in the following aspects: snow patches on roofs are scattered, and some snow-free roads are incorrectly predicted to being snow-covered.

By analyzing the heat map, we observe that in densely populated urban areas with multiple roads and buildings, the model's performance in recognizing fine edges is suboptimal. Unlike in mountainous regions, urban edges are more complex and discontinuous, making it more challenging to extract dense edges from optical images. Furthermore, SAR struggles to provide effective edge information, as shown by the red boxes in the first and second rows in Figure 15, where SAR images reveal almost no edge information.

We attribute these issues to several factors. Firstly, in cities, frequent human activities and higher temperatures cause snow to melt more readily. As the water content in the snow increases, the reflectivity of wet snow rises significantly, resulting in brighter backscatter and making it harder to distinguish snow from snow-free edges such as roads. This leads to incorrect estimates of snow depth on some roads.

Additionally, the relationship between snow depth and backscatter becomes more complex in urban areas. One reason we select winter data is to reduce the uncertainty caused by melting snow and ice. However, unavoidable human activities and temperature factors in cities accelerate the melting rate. The increase in wet snow means that both humidity and depth affect SAR backscatter, adding to the uncertainty in predictions. Moreover, urban buildings typically have highly reflective surfaces. Their complex structures, such as walls and roofs, cause multiple reflections and scattering of electromagnetic waves, resulting in multipath effects. The combined electromagnetic waves lead to signal distortion, interfering with prediction results. This can be seen in the SAR images in Figure 15, where finding regular patterns is almost impossible.

7. Conclusions

Heterogeneity and complementarity are always critical issues in the application of multi-source data. In this study, we proposed the OSNet model to extract modality features and achieve effective fusion. By utilizing different attention mechanisms, we explored the heterogeneity between optical and SAR images. Additionally, we strengthened the complementary edge features through edge enhancement methods. The results show that OSNet can obtain high-quality features from optical and SAR images, learn their recognition patterns and objects' edges, and achieve higher accuracy than existing models in both segmentation and regression tasks, demonstrating a good generalization ability.

We also identified some limitations in our study, which future research should focus on:

1. In both segmentation and regression tasks, OSNet can effectively identify edges in images. However, the model often treats objects within enclosed edges as a single class. When small objects of different types are present within the same class, the recognition accuracy is relatively low. Future research should focus on improving the model's ability to detect heterogeneous objects within same-class regions.
2. All optical images used in this study are in RGB and NIR bands, while the SAR images are all VV-polarized. In future studies, it is essential to expand the dataset to include more bands. Particularly in snow depth studies, the 1.57–1.65 μm shortwave infrared band is sensitive to snow. At the same time, other polarizations in SAR can provide more informative representations. Adding more spectral features will further deepen the study of multi-source data fusion. Additionally, our current snow depth dataset is relatively limited to the 0–42 cm range. In the future, we plan to expand the depth range of the dataset to explore the performance of C-band SAR under deeper snow conditions.
3. OSNet has shown high accuracy in both segmentation and regression tasks, but we suppose that its advantages in regression tasks are not as significant as in segmentation tasks. Segmentation tasks require classifying each pixel in the image and accurately detecting boundaries. In regression tasks, while reliable feature extraction is also necessary, the need for boundary detection is less critical, thus reducing the dependency on OSNet's design of enhanced edge representation. Therefore, future research should focus on adjusting the network structure and feature extraction methods for regression tasks to better adapt to the different tasks.
4. The dual-branch network effectively extracts features from both modalities, but this comes at the cost of higher computational demands [54,55]. Therefore, in our future work, we plan to focus more on improving the model's efficiency and computational speed. Lastly, ss OSNet's design primarily focuses on the encoder, enhancing the efficiency of the decoder could be a promising direction for reducing the overall computational cost.

Author Contributions: Conceptualization, K.M., K.H. and X.M.; methodology, K.H., K.M. and X.M.; software, K.M., J.C. and M.J.; formal analysis, K.H and K.M.; investigation, K.M. and J.C.; writing—original draft preparation, K.M. and Y.X.; writing—review, K.M., K.H., Y.X. and M.X.; editing, K.M., K.H. and L.W.; visualization, K.M. and L.W.; supervision, K.H., M.X. and L.W.; project administration, K.H. and L.W.; funding acquisition, K.H. and L.W. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Natural Science Foundation of China under Grant 42075130.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and code that support the findings of this study are available from the first author upon reasonable request.

Acknowledgments: During the preparation of this work, the authors used ChatGPT in order to improve this study's language and readability. After using this tool/service, the authors reviewed and edited the content of this publication as needed and take full responsibility for it.

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

References

1. Letsoin, S.M.A.; Herak, D.; Purwestri, R.C. Evaluation land use cover changes over 29 years in papua province of indonesia using remote sensing data. In *IOP Conference Series: Earth and Environmental Science*; IOP Publishing: Bristol, UK, 2022; Volume 1034, p. 012013.
2. Ye, Y.; Zhang, J.; Zhou, L.; Li, J.; Ren, X.; Fan, J. Optical and SAR image fusion based on complementary feature decomposition and visual saliency features. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–15.
3. Gómez-Chova, L.; Tuia, D.; Moser, G.; Camps-Valls, G. Multimodal classification of remote sensing images: A review and future directions. *Proc. IEEE* **2015**, *103*, 1560–1584.
4. Li, X.; Zhang, G.; Cui, H.; Hou, S.; Wang, S.; Li, X.; Chen, Y.; Li, Z.; Zhang, L. MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *106*, 102638. https://doi.org/10.1016/j.jag.2021.102638.
5. Daudt, R.C.; Wulf, H.; Hafner, E.D.; Bühler, Y.; Schindler, K.; Wegner, J.D. Snow depth estimation at country-scale with high spatial and temporal resolution. *ISPRS J. Photogramm. Remote Sens.* **2023**, *197*, 105–121.
6. Mou, L.; Schmitt, M.; Wang, Y.; Zhu, X.X. Identifying corresponding patches in SAR and optical imagery with a convolutional neural network. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 5482–5485.
7. Hu, W.; Wang, X.; Zhan, F.; Cao, L.; Liu, Y.; Yang, W.; Ji, M.; Meng, L.; Guo, P.; Yang, Z.; et al. OPT-SAR-MS2Net: A Multi-Source Multi-Scale Siamese Network for Land Object Classification Using Remote Sensing Images. *Remote Sens.* **2024**, *16*, 1850.
8. Chen, J.; Xia, M.; Wang, D.; Lin, H. Double Branch Parallel Network for Segmentation of Buildings and Waters in Remote Sensing Images. *Remote Sens.* **2023**, *15*, 1536. https://doi.org/10.3390/rs15061536.
9. Dumont, M.; Gascoin, S. Optical remote sensing of snow cover. In *Land Surface Remote Sensing in Continental Hydrology*; Elsevier: Amsterdam, The Netherlands, 2016; pp. 115–137.
10. Litaor, M.; Williams, M.; Seastedt, T. Topographic controls on snow distribution, soil moisture, and species diversity of herbaceous alpine vegetation, Niwot Ridge, Colorado. *J. Geophys. Res. Biogeosciences* **2008**, *113*. https://doi.org/10.1029/2007JG000419.
11. Besic, N.; Vasile, G.; Dedieu, J.P.; Chanussot, J.; Stankovic, S. Stochastic approach in wet snow detection using multitemporal SAR data. *IEEE Geosci. Remote Sens. Lett.* **2014**, *12*, 244–248.
12. Liu, S.; Qi, Z.; Li, X.; Yeh, A.G.O. Integration of convolutional neural networks and object-based post-classification refinement for land use and land cover mapping with optical and SAR data. *Remote Sens.* **2019**, *11*, 690.
13. Zhang, H.; Yu, A.; Gao, K.; Lu, X.; Cao, X.; Guo, W.; Lian, W. M2Caps: Learning multi-modal capsules of optical and SAR images for land cover classification. *Int. J. Digit. Earth* **2025**, *18*, 2447347.
14. Yu, K.; Wang, F. A Dual Attention Fusion Network for SAR-Optical Land Use Classification Based on Semantic Balance. In Proceedings of the 2024 7th International Conference on Pattern Recognition and Artificial Intelligence (PRAI), Hangzhou, China, 15–17 August 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 707–712.
15. Li, X.; Zhang, G.; Cui, H.; Hou, S.; Chen, Y.; Li, Z.; Li, H.; Wang, H. Progressive fusion learning: A multimodal joint segmentation framework for building extraction from optical and SAR images. *ISPRS J. Photogramm. Remote Sens.* **2023**, *195*, 178–191.
16. Awasthi, S.; Varade, D. Recent advances in the remote sensing of alpine snow: A review. *GISci. Remote Sens.* **2021**, *58*, 852–888.
17. Zhang, T. Influence of the seasonal snow cover on the ground thermal regime: An overview. *Rev. Geophys.* **2005**, *43*. https://doi.org/10.1029/2004RG000157.
18. Cook, B.I.; Bonan, G.B.; Levis, S.; Epstein, H.E. The thermoinsulation effect of snow cover within a climate model. *Clim. Dyn.* **2008**, *31*, 107–124.
19. Patil, A.; Singh, G.; Rüdiger, C. Retrieval of snow depth and snow water equivalent using dual polarization SAR data. *Remote Sens.* **2020**, *12*, 1183.
20. Bernier, M.; Fortin, J.P. The potential of times series of C-band SAR data to monitor dry and shallow snow cover. *IEEE Trans. Geosci. Remote Sens.* **1998**, *36*, 226–243.

21. Bernier, M.; Fortin, J.P.; Gauthier, Y.; Gauthier, R.; Roy, R.; Vincent, P. Determination of snow water equivalent using RADARSAT SAR data in eastern Canada. *Hydrol. Process.* **1999**, *13*, 3041–3051.
22. Chokmani, K.; Bernier, M.; Gauthier, Y. Uncertainty analysis of EQeau, a remote sensing based model for snow water equivalent estimation. *Int. J. Remote Sens.* **2006**, *27*, 4337–4346.
23. Zhao, L.; Chen, J.; Shahzad, M.; Xia, M.; Lin, H. MFPANet: Multi-Scale Feature Perception and Aggregation Network for High-Resolution Snow Depth Estimation. *Remote Sens.* **2024**, *16*, 2087.
24. AmberHen. WHU-OPT-SAR-Dataset. 2023. Available lone: https://github.com/AmberHen/WHU-OPT-SAR-dataset (accessed on January 31, 2025.).
25. The National Meteorological Information Center. 2024. Available lone: https://data.cma.cn/site (accessed on January 31, 2025.).
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
27. Dai, X.; Xia, M.; Weng, L.; Hu, K.; Lin, H.; Qian, M. Multiscale location attention network for building and water segmentation of remote sensing image. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–19.
28. Lee, J.S. Digital image enhancement and noise filtering by use of local statistics. *IEEE Trans. Pattern Anal. Mach. Intell.* **1980**, *2*, 165–168. https://doi.org/10.1109/tpami.1980.4766994.
29. Chen, S.C.; Chiu, C.C. Texture Construction Edge Detection Algorithm. *Appl. Sci.* **2019**, *9*, 897. https://doi.org/10.3390/app9050897.
30. Cherri, A.K.; Karim, M.A. Optical symbolic substitution: Edge detection using Prewitt, Sobel, and Roberts operators. *Appl. Opt.* **1989**, *28*, 4644–4648. https://doi.org/10.1364/ao.28.004644.
31. Xu, D.; Zhao, Y.; Jiang, Y.; Zhang, C.; Sun, B.; He, X. Using improved edge detection method to detect mining-induced ground fissures identified by unmanned aerial vehicle remote sensing. *Remote Sens.* **2021**, *13*, 3652.
32. Zhang, W.; Jiao, L.; Liu, F.; Liu, J.; Cui, Z. LHNet: Laplacian convolutional block for remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13.
33. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *arXiv* **2019**, arXiv:1709.01507.
34. Hu, K.; Li, Y.; Zhang, S.; Wu, J.; Gong, S.; Jiang, S.; Weng, L. FedMMD: A Federated weighting algorithm considering Non-IID and Local Model Deviation. *Expert Syst. Appl.* **2024**, *237*, 121463. https://doi.org/https://doi.org/10.1016/j.eswa.2023.121463.
35. Ji, H.; Xia, M.; Zhang, D.; Lin, H. Multi-Supervised Feature Fusion Attention Network for Clouds and Shadows Detection. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 247. https://doi.org/10.3390/ijgi12060247.
36. Hu, K.; Shen, C.; Wang, T.; Shen, S.; Cai, C.; Huang, H.; Xia, M. Action Recognition Based on Multi-Level Topological Channel Attention of Human Skeleton. *Sensors* **2023**, *23*, 9738.
37. Wang, Z.; Gu, G.; Xia, M.; Weng, L.; Hu, K. Bitemporal Attention Sharing Network for Remote Sensing Image Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 10368–10379.
38. Dai, X.; Chen, K.; Xia, M.; Weng, L.; Lin, H. LPMSNet: Location Pooling Multi-Scale Network for Cloud and Cloud Shadow Segmentation. *Remote Sens.* **2023**, *15*, 6005. https://doi.org/10.3390/rs15164005.
39. Jiang, S.; Lin, H.; Ren, H.; Hu, Z.; Weng, L.; Xia, M. MDANet: A High-Resolution City Change Detection Network Based on Difference and Attention Mechanisms under Multi-Scale Feature Fusion. *Remote Sens.* **2024**, *16*, 1387. https://doi.org/10.3390/rs16081387.
40. Hu, K.; Feng, X.; Zhang, Q.; Shao, P.; Liu, Z.; Xu, Y.; Wang, S.; Wang, Y.; Wang, H.; Di, L.; et al. Review of Satellite Remote Sensing of Carbon Dioxide Inversion and Assimilation. *Remote Sens.* **2024**, *16*, 3394. https://doi.org/10.3390/rs16183394.
41. Miao, S.; Xia, M.; Qian, M.; Zhang, Y.; Liu, J.; Lin, H. Cloud/shadow segmentation based on multi-level feature enhanced network for remote sensing imagery. *Int. J. Remote Sens.* **2022**, *43*, 5940–5960. https://doi.org/10.1080/01431161.2021.2014077.
42. Lu, C.; Xia, M.; Lin, H. Multi-scale strip pooling feature aggregation network for cloud and cloud shadow segmentation. *Neural Comput. Appl.* **2022**, *34*, 6149–6162. https://doi.org/10.1007/s00521-021-06802-0.
43. Song, L.; Xia, M.; Jin, J.; Qian, M.; Zhang, Y. SUACDNet: Attentional change detection network based on siamese U-shaped structure. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *105*, 102597. https://doi.org/10.1016/j.jag.2021.102597.
44. Hu, X.; Yang, K.; Fei, L.; Wang, K. ACNet: Attention Based Network to Exploit Complementary Features for RGBD Semantic Segmentation. *arXiv* **2019**, arXiv:cs.CV/1905.10089.
45. Lee, S.; Park, S.J.; Hong, K.S. RDFNet: RGB-D Multi-level Residual Feature Fusion for Indoor Semantic Segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
46. Audebert, N.; Le Saux, B.; Lefevre, S. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *140*, 20–32. https://doi.org/10.1016/j.isprsjprs.2017.11.011.
47. Hosseinpour, H.; Samadzadegan, F.; Javan, F.D. CMGFNet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 96–115. https://doi.org/10.1016/j.isprsjprs.2021.12.007.
48. 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.

49. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
50. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. *arXiv* **2017**, arXiv:1611.05431.
51. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. *arXiv* **2014**, arXiv:1409.4842.
52. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. *arXiv* **2017**, arXiv:1610.02357.
53. Lievens, H.; Demuzere, M.; Marshall, H.P.; Reichle, R.H.; Brucker, L.; Brangers, I.; de Rosnay, P.; Dumont, M.; Girotto, M.; Immerzeel, W.W.; et al. Snow depth variability in the Northern Hemisphere mountains observed from space. *Nat. Commun.* **2019**, *10*, 4629.
54. Hu, K.; Li, M.; Song, Z.; Xu, K.; Xia, Q.; Sun, N.; Zhou, P.; Xia, M. A review of research on reinforcement learning algorithms for multi-agents. *Neurocomputing* **2024**, *599*, 128068.
55. Hu, K.; Xu, K.; Xia, Q.; Li, M.; Song, Z.; Song, L.; Sun, N. An overview: Attention mechanisms in multi-agent reinforcement learning. *Neurocomputing* **2024**, *598*, 128015.