remote sensing

MDPI

Article

ABNet: An Aggregated Backbone Network Architecture for Fine Landcover Classification

Bo Si [1], Zhennan Wang [2], Zhoulu Yu [1,*] and Ke Wang [1]

[1] College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China; bo_si@zju.edu.cn (B.S.); kwang@zju.edu.cn (K.W.)
[2] Zhejiang Shuzhi Space Planning and Design Co., Ltd., Hangzhou 310030, China; 1108130121001@stu.bucea.edu.cn
* Correspondence: yuzl@zju.edu.cn; Tel.: +86-13958123139

Abstract: High-precision landcover classification is a fundamental prerequisite for resource and environmental monitoring and land-use status surveys. Imbued with intricate spatial information and texture features, very high spatial resolution remote sensing images accentuate the divergence between features within the same category, thereby amplifying the complexity of landcover classification. Consequently, semantic segmentation models leveraging deep backbone networks have emerged as stalwarts in landcover classification tasks owing to their adeptness in feature representation. However, the classification efficacy of a solitary backbone network model fluctuates across diverse scenarios and datasets, posing a persistent challenge in the construction or selection of an appropriate backbone network for distinct classification tasks. To elevate the classification performance and bolster the generalization of semantic segmentation models, we propose a novel semantic segmentation network architecture, named the aggregated backbone network (ABNet), for the meticulous landcover classification. ABNet aggregates three prevailing backbone networks (ResNet, HRNet, and VoVNet), distinguished by significant structural disparities, using a same-stage fusion approach. Subsequently, it amalgamates these networks with the Deeplabv3+ head after integrating the convolutional block attention mechanism (CBAM). Notably, this amalgamation harmonizes distinct scale features extracted by the three backbone networks, thus enriching the model's spatial contextual comprehension and expanding its receptive field, thereby facilitating more effective semantic feature extraction across different stages. The convolutional block attention mechanism primarily orchestrates channel adjustments and curtails redundant information within the aggregated feature layers. Ablation experiments demonstrate an enhancement of no less than 3% in the mean intersection over union (mIoU) of ABNet on both the LoveDA and GID15 datasets when compared with a single backbone network model. Furthermore, in contrast to seven classical or state-of-the-art models (UNet, FPN, PSPNet, DANet, CBNet, CCNet, and UPerNet), ABNet evinces excellent segmentation performance across the aforementioned datasets, underscoring the efficiency and robust generalization capabilities of the proposed approach.

Keywords: backbone network; landcover classification; semantic segmentation; aggregated feature

Citation: Si, B.; Wang, Z.; Yu, Z.; Wang, K. ABNet: An Aggregated Backbone Network Architecture for Fine Landcover Classification. *Remote Sens.* **2024**, *16*, 1725. https://doi.org/10.3390/rs16101725

Academic Editors: Sawaid Abbas, Janet E. Nichol, Faisal M. Qamer and Jianchu Xu

Received: 31 March 2024
Revised: 7 May 2024
Accepted: 10 May 2024
Published: 13 May 2024

1. Introduction

Landcover classification is the process of distinguishing different types of naturally occurring landcover through remote sensing image data processing, which provides important geospatial data for decision making and monitoring in the fields of environmental monitoring, urban planning, precision agriculture, resource management, and so on [1,2]. With the rapid development of remote sensing imaging technology, the rapid acquisition of meter-scale remote sensing images provides the possibility of fine landcover classification and mapping [3]. However, the feature differences within high-precision landcover categories become significantly larger, making it more difficult to distinguish surface covers

with complex compositional elements [4]. Misidentification of landcover categories will cause a huge waste of remote sensing data resources, and it is difficult to produce effective application value. Therefore, how to realize the fine classification of landcover has become a hot issue to be solved [5].

In recent years, deep learning models for landcover classification based on convolutional neural networks (CNNs) have been widely studied and have shown powerful performance over traditional methods that rely on manual feature extraction and shallow machine learning algorithms [6]. Michael Kampffmeyer et al. [7] first introduced an end-to-end convolutional neural network model for remote sensing imagery landcover mapping. After that, the semantic segmentation model for remote sensing images gradually formed an encoder–decoder architecture containing a backbone network, a neck, and a segmentation head [8–10]. Therefore, the semantic segmentation model based on the CNN as the backbone network has become a mainstream method for studying the high-precision landcover classification problem. However, different backbone networks have different abilities to extract features due to their different convolutional structures, which leads to different classification results [11]. Therefore, how to efficiently apply different backbone networks for landcover classification still deserves further exploration.

To compare the performance of semantic segmentation models based on different backbone networks, scholars have carried out related studies on backbone networks and provided some new inspirations. Marc et al. [12] used two model structures, UNet and Deeplabv3+, in combination with three backbone networks, namely, MobileNet-V3, ResNet-50, and EfficientNet-B4, to test their abilities to segment flooded areas under different environmental conditions and data availability. Results showed that the MobileNet-V3-based UNet model performed the best. This implies that the appropriate backbone network needs to be selected in advance under different datasets and environmental conditions. In a pioneering study, Liu et al. [13] generated a more powerful backbone network (CBNet) by combining ResNet and ResNeXt through a stage-by-stage composite connection strategy of different backbone networks, and the experimental results demonstrated that CBNet possessed a stronger generalization ability than a single backbone network on different target detection datasets. Liang et al. [14] further improved CBNet by a dense connectivity strategy of backbones and assisted supervised training to achieve an increase in classification accuracy, which provides a novel idea for model architecture improvement.

In the pursuit of leveraging a backbone network fusion strategy for the development of a remote sensing semantic segmentation model, this study meticulously selects the following three pre-eminent backbone networks for fusion: ResNet, HRNet, and VoVNet. Notably, ResNet facilitates feature extraction by serially connecting residual convolutional blocks at each network stage, thereby enabling profound feature extraction. In contrast, HRNet fosters feature extraction through parallel connections at each stage of the convolutional blocks, adeptly amalgamating features from multi-scale receptive fields. Furthermore, VoVNet strategically aggregates all preceding convolutional blocks exclusively at the final layer of each stage, thereby amplifying the receptive fields and enhancing feature extraction efficiency. After the multi-scale fusion of the backbone network, feature optimization is meticulously achieved through the integration of a convolutional attention mechanism. The amalgamation with the Deeplabv3+ network architecture gives rise to the aggregated backbone network (ABNet), tailored specifically for the fine-grained classification of landcover.

The main contributions of this paper include the following: (1) An efficient feature extraction strategy using multiple existing backbone networks is proposed. (2) An efficient and effective aggregated backbone network (ABNet) to enhance multi-scale feature extraction and fusion capabilities for remote sensing image semantic segmentation work is proposed. ABNet leverages the Deeplabv3+ head and reconstructs the backbone network through the same-stage fusion of ResNet-50, HRNet-48, and VoVNet-39. (3) Experimental results demonstrate that ABNet achieves comparable performance to the state-of-the-art methods on the LoveDA and GID datasets.

The rest of this paper is organized as follows. Section 2 introduces related research work, mainly reviewing the development of representative backbones and ensemble methods of semantic segmentation in the field of fine landcover classification in recent years. Section 3 elaborates on the structure of the proposed ABNet and details the design ideas of the aggregated backbone network. Section 4 gives the experimental details and results on the LoveDA dataset and Gaofen Image Dataset (GID). The advantages, limitations, and potential improvements of our proposed method will be elaborated on in Section 5. In Section 6, conclusions and future works are provided.

2. Related Work

2.1. Backbones for Remote Sensing Image Semantic Segmentation Work

The pivotal role of backbone networks in feature extraction remains integral to the fabric of semantic segmentation models [15], which is particularly evident in the realm of remote sensing classification, as shown in Table 1. Since the inception of AlexNet in 2012, progressive enhancements in backbone networks have predominantly revolved around augmenting both network depth and width. VGG-16 [16] constitutes a prototypical deep network by increasing the number of network layers. However, the deepening of network layers often precipitates the potential vanishing or explosion of computational gradients. Addressing this concern, He et al. [17] ingeniously introduced the residual connection module to alleviate the convergence challenges encountered by deep networks. Furthermore, Inceptionv3 [18] astutely observed that augmenting the model width effectively bolsters feature extraction through the incorporation of branches with multi-scale convolutional kernels. Beyond the model architecture, the effective interconnection between feature layers emerges as a focal avenue for enhancing backbone networks. In this context, DenseNet [19] forges dense connections between preceding and subsequent layers, harnessing multi-layer features to optimize segmentation performance. In a bid to pare down the parameter count, VoVNet [20] boosts computational efficiency by densely linking previous layer features exclusively to the last layer, drawing inspiration from DenseNet. HRNet [21], on the other hand, intertwines the outputs of multiple resolutions in parallel, engendering a richer, high-resolution representation. Leveraging the neural structure search technique, EfficientNet [22] delves into an exploration of network resolution, depth, and width parameters, rationalizing their configuration to bolster model enhancement efficiency. Meanwhile, Res2Net [23] delineates multi-scale features at a granular level by constructing hierarchical residual connections within the same residual block, sans an expansion in the network depth. These contemporary mainstream backbone networks exhibit distinct model structures and generalization proficiencies tailored to their specific developmental tasks.

Table 1. Representative backbones used in remote sensing classification.

Backbone	First Publication	Features	Application in Remote Sensing
VGG-16 [16]	2014	Deep convolutional network	Topography and geomorphology classification [24]
ResNet-50 [17]	2016	Residual connection	Multispectral image classification [25]
Inceptionv3 [18]	2016	Multi-scale convolutional kernel	Sheltered Vessels classification [26]
DenseNet [19]	2017	Dense connection of all layers	Road extraction [27]
VoVNet [20]	2019	Dense connection of the final layer	Open-pit mine extraction [28]

Table 1. *Cont.*

Backbone	First Publication	Features	Application in Remote Sensing
HRNet [21]	2019	Parallel connection of multi-resolution layers	Landcover classification [29]
EfficientNet [22]	2019	Neural structure search	Mars scene recognition [30]
Res2Net [23]	2021	Hierarchical residual connection	Remote sensing scene recognition [31]

2.2. Research on Fine Landcover Classification with Semantic Segmentation Network

Since the full convolutional neural network (FCN) [32] realizes the pixel-by-pixel semantic segmentation of images, many classical semantic segmentation models based on convolutional neural networks have been developed, such as UNet [33], Deeplabv3+ [34], FPN [35], PSPNet [36], DANet [37], UPerNet [38], and CCNet [39], extensively embraced in remote sensing semantic segmentation applications [40–43]. Remote sensing semantic segmentation models dedicated to landcover classification chiefly fall into distinct categories, including the pixel-based CNN, object-based CNN, graph-based CNN, siamese CNN, and ensembled CNN, with the ensembled CNN model emerging as adept at efficiently addressing complex landcover scenarios [44].

In the pursuit of heightened accuracy, Bigdeli et al. [45] achieved superior classification by optimally amalgamating the classification results of CNN models constructed from diverse sensor datasets. Faced with a dearth of labeled samples, Fan et al. [46] crafted an integrated teacher model to autonomously pick samples from an extensive pool of unlabeled data, culminating in dataset generation. A pivotal notion entails the integration of results generated by multiple backbone network models in a channel-by-channel superposition fashion. Complementing inputs or outputs, scholars have sought to consolidate deep learning models during their training process. Notably, Cao et al. [47] bolstered the classification accuracy of the ISPRS 2-D dataset by harnessing complementarities between diverse models, coalescing to jointly learn the solution domains of multiple semantic segmentation models. Additionally, Ekim et al. [48], premised on the training process of three loss optimization methods, honed the integration of distinct model outputs to fortify the robustness of the landcover classification model.

Cumulatively, these investigations meld the computational attributes of divergent networks to yield refined semantic segmentation results. Yet, the variance in classification outcomes stems from disparities in feature layer outputs across dissimilar model structures, disregarding the potential integration of feature layers encapsulating distinct semantic information. This oversight of structural dissimilarities may precipitate the omission of essential extraction outcomes.

3. Methods

3.1. Aggregated Backbone Network

Deep learning semantic segmentation models based on features extracted from backbone networks have been widely used in the fine landcover classification of remote sensing images. However, for remote sensing datasets with complex feature distributions, it is still challenging to select an appropriate backbone network. In addition, a single backbone network is usually unable to extract effective and relevant features from complex scenes, especially on datasets with diverse classes and large intra-class variations. To overcome this problem, this study selects three backbone networks to construct an aggregated backbone network (ABNet), aggregates the feature layers of different backbone networks by same-stage fusion, then optimizes the merged features by a convolution block attention mechanism, and finally outputs the classification results by the Deeplabv3+ head [34] (as

shown in Figure 1). Specifically, ResNet, HRNet, and VoVNet are used in this study and described in Section 3.2. Different backbone networks extract the semantic features of the classified objects in different ways, and the aggregation backbone network utilizes the characteristics of each backbone network and enhances the feature extraction ability by using the same-stage fusion method at different scale stages of feature extraction, which will be described in detail in Section 3.3. The CBAM attention mechanism extracts the features of the different backbone networks aggregated at different stages adaptively and optimally using the channel attention module and the spatial attention module, which will be described in Section 3.4.

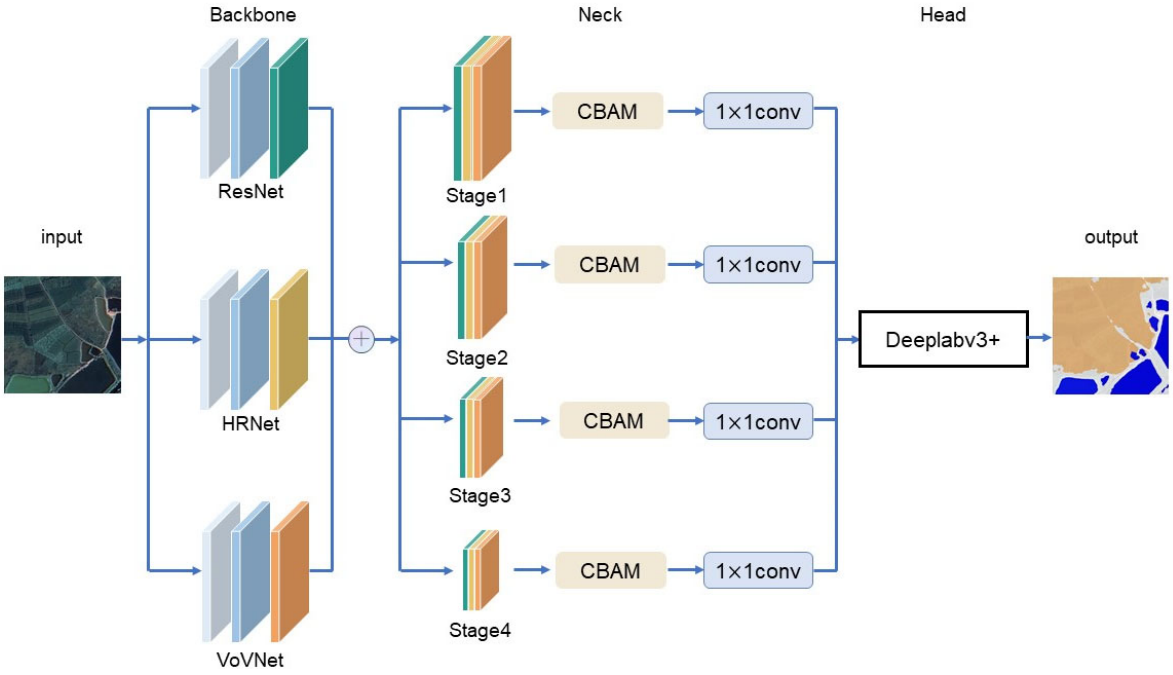

Figure 1. Illustration of the proposed aggregated backbone network (ABNet) architecture for semantic segmentation.

3.2. Basic Backbone Networks

3.2.1. Residual Network (ResNet)

In a conventional convolutional neural network, the output of the layer is used as the input to the neural network in layer $(i + 1)$ during forward propagation. The expression is defined as follows: $xi = Li(xi - 1)$. As the number of neural network layers increases, it leads to the problem of gradient vanishing or gradient explosion. Residual Network (ResNet) adds a jump connection to the traditional convolutional neural network and constructs the residual module by nonlinear transformation and the activation function, whose expression is defined as $xi = Li(xi - 1) + xi - 1$. One of the advantages of ResNet is that the gradient can flow directly from deep to shallow layers, thus avoiding the problem of gradient vanishing and gradient explosion that occurs when the model becomes deeper, and thus it has become the mainstream encoder backbone network.

3.2.2. High-Resolution Network (HRNet)

Typically, deep convolutional neural networks gradually generate features from high to low resolution through cascaded convolutional blocks and downsampling or pooling operations at different stages. However, high-resolution networks employ parallel connections of convolutional streams from high to low resolution, where the output at each resolution is the sum of the already generated resolution in the previous stage, mathematically defined as $R_s^r = f(R_{s-1}^1) + f(R_{s-1}^2) \cdots + f(R_{s-1}^{r-1})$, where s represents the different stages and r represents the different resolutions. Furthermore, as the resolution decreases,

information from different resolution features is exchanged repeatedly. HRNet gains richer semantic representation, more precise spatial information, and a stronger feature extraction capability by exchanging information across resolutions repeatedly.

3.2.3. Variety of View Network (VoVNet)

To enhance the fusion of shallow and deep features in the neural networks, a Dense Connected Network (DenseNet) has been proposed. In DenseNet, each subnetwork layer takes the outputs of the previous layers as inputs after concatenating them in the channel dimension, i.e., the input of layer i is the output of layers 1 to $(i-1)$, whose expression is defined as follows: $x_i = L_i([x_0, x_1, \ldots x_{i-1}])$. However, too many dense connections in the intermediate layers leads to the computational inefficiency of DenseNet. VoVNet proposes the strategy of one-shot aggregation, which not only adopts the advantage of DenseNet, i.e., the advantage of representing diverse features with multiple sensory fields, but also overcomes the inefficiency of dense connections. One-shot aggregation is performed only once for all features in the last feature map of each stage, and its expression is defined as $x_s = \sum_{1}^{i-1} L_i([x_0, x_1, \ldots x_{i-1}])$.

While reducing the model complexity, VoVNet still maintains the effective use of global information, and this global context summarization capability improves the interpretation and characterization of fine and complex features.

3.3. Backbone Ensemble

The majority of ensemble methods necessitate a voting strategy to compare output results from various models and enhance the final prediction [49]. This research introduces a novel approach by amalgamating multiple existing backbone networks through intricate connections between adjacent backbone networks, thereby creating a more robust composite backbone network known as the aggregated backbone network (ABNet). Through this process, ABNet includes output features from different backbone networks as part of the input features, progressively fusing them for the model and ultimately conducting semantic segmentation via the decoder's feature mapping.

Model diversity implies the significance of training models with distinct architectures or techniques, which has been widely acknowledged within the model ensemble [50]. Hence, this study adopts the same-stage fusion approach to combine the backbones of three different architectures, namely, HRNet, ResNet, and VoVNet. Specifically, HRNet employs the HRNet-48 structure with 4 stages, ResNet employs the ResNet-50 structure with 4 stages, and VoVNet employs the VoVNet-39 structure with 5 stages, as illustrated in Figure 2.

Each stage consists of several convolutional layers with the same-sized feature mappings, except the Stem layer in HRNet-48 and ResNet-50, and apart from the Stem layer and the first stage in VoVNet-39. Furthermore, the downsampling for each stage is accomplished by 3×3 max-pooling with a stride of 2. The specific structures are presented in Table 2. Upon preprocessing an input image of size C through the Stem layer, HRNet-48 and ResNet-50 engage in convolutional operations to produce the first stage output, resulting in a size of $C/4$. Similarly, VoVNet-39 requires convolutional operations in the Stem layer and the first stage to attain an output size of $C/4$, aligning it with the outputs from HRNet-48 and ResNet-50. This process of merging the outputs from various stages of HRNet-48 and ResNet-50 with the subsequent stage output of VoVNet-39 yields a multi-scale aggregated result.

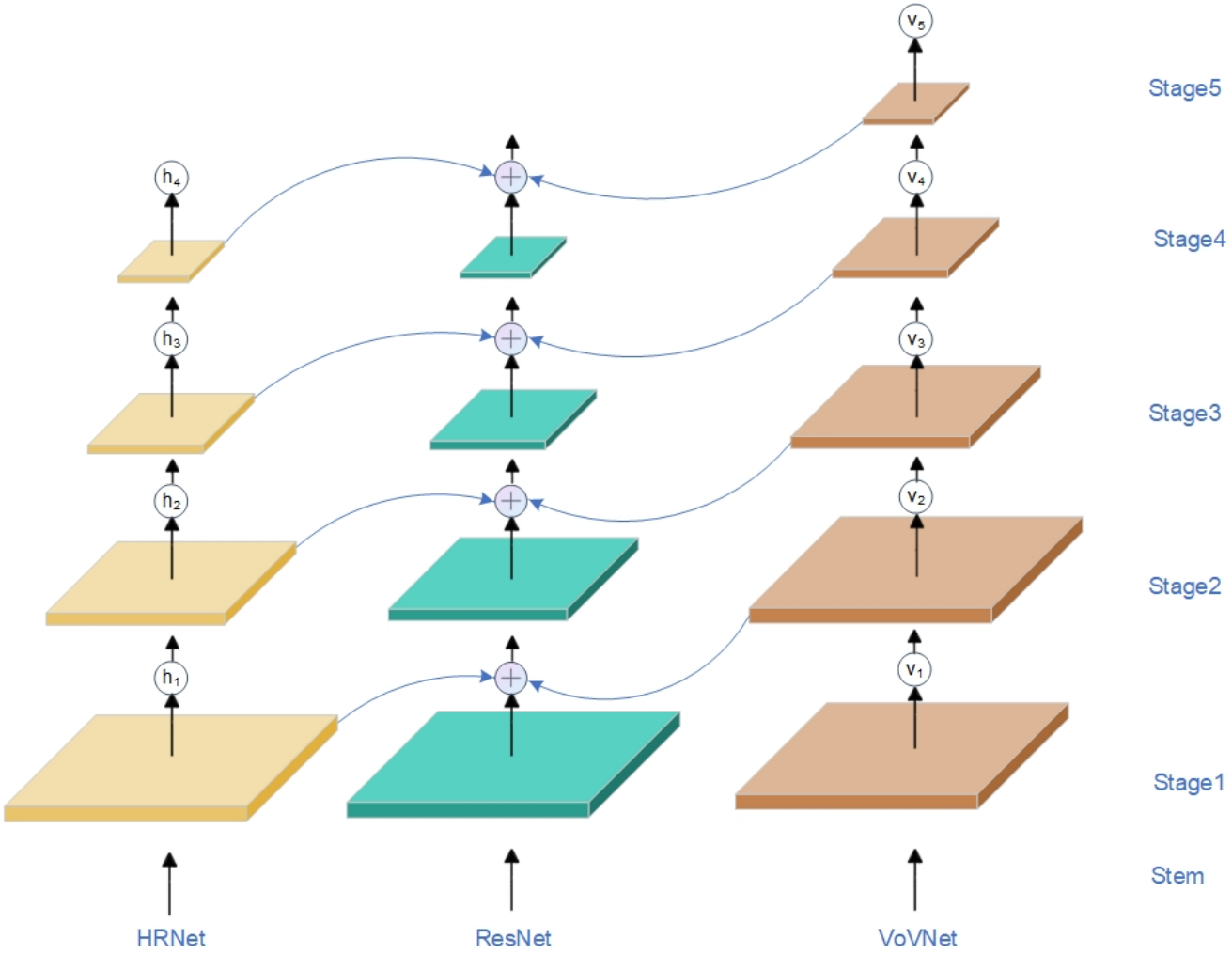

Figure 2. Same-stage fusion method for HRNet, ResNet, and VoVNet.

Table 2. Detailed module structures for each stage of HRNet-48, ResNet-50, and VoVNet-39. The downsampling for each stage is accomplished by 3 × 3 max-pooling with a stride of 2.

Stage	HRNet-48	ResNet-50	VoVNet-39
Stem	3 × 3 conv, 64, stride = 2 3 × 3 conv, 64, stride = 2	7 × 7 conv, 64, stride = 2 3 × 3 max pool, stride = 2	3 × 3 conv, 64, stride = 2
Stage1	[1 × 1conv, 48 3 × 3conv, 48 1 × 1conv, 48] ×4	[1 × 1conv, 64 3 × 3conv, 64 1 × 1conv, 256] ×3	3 × 3 conv, 64, stride = 1 3 × 3 conv, 128, stride = 1
Stage2	[3 × 3conv, 96 3 × 3conv, 96] ×4	[1 × 1conv, 128 3 × 3conv, 128 1 × 1conv, 512] ×4	[3 × 3conv, 128, × 5 concate & 1 × 1conv, 256] ×1
Stage3	[3 × 3conv, 192 3 × 3conv, 192] ×4	[1 × 1conv, 256 3 × 3conv, 256 1 × 1conv, 1024] ×6	[3 × 3conv, 160, × 5 concate & 1 × 1conv, 512] ×1
Stage4	[3 × 3conv, 384 3 × 3conv, 384] ×4	[1 × 1conv, 512 3 × 3conv, 512 1 × 1conv, 2028] ×3	[3 × 3conv, 192, × 5 concate & 1 × 1conv, 768] ×2
Stage5			[3 × 3conv, 224, × 5 concate & 1 × 1conv, 1024] ×2

3.4. Convolutional Block Attention Module (CBAM)

To enhance the feature representation after merging multiple backbone networks and mitigating redundant information within the merged feature layer, we propose an

optimization method through the utilization of the convolutional block attention model [51] (as shown in Figure 3). This approach aims to focus on essential parts of the feature map while suppressing noise and uncertainty. The convolutional attention model consists of the following two key components: the channel attention and spatial attention models. These models generate the channel attention weights and spatial attention weights, respectively. Subsequently, the attention weights are multiplied by the input feature map to achieve adaptive feature refinement. The specific formula is as follows:

$$F_c = M_c(F) \otimes F \quad (1)$$

$$F_{out} = M_s(F_c) \otimes F_c \quad (2)$$

F represents the input feature map, F_c represents the input feature map after channel attention, M_c represents the channel attention model, M_s represents the spatial attention model, $\otimes$ represents the pixel-by-pixel multiplication, and F_{out} represents the final optimized output.

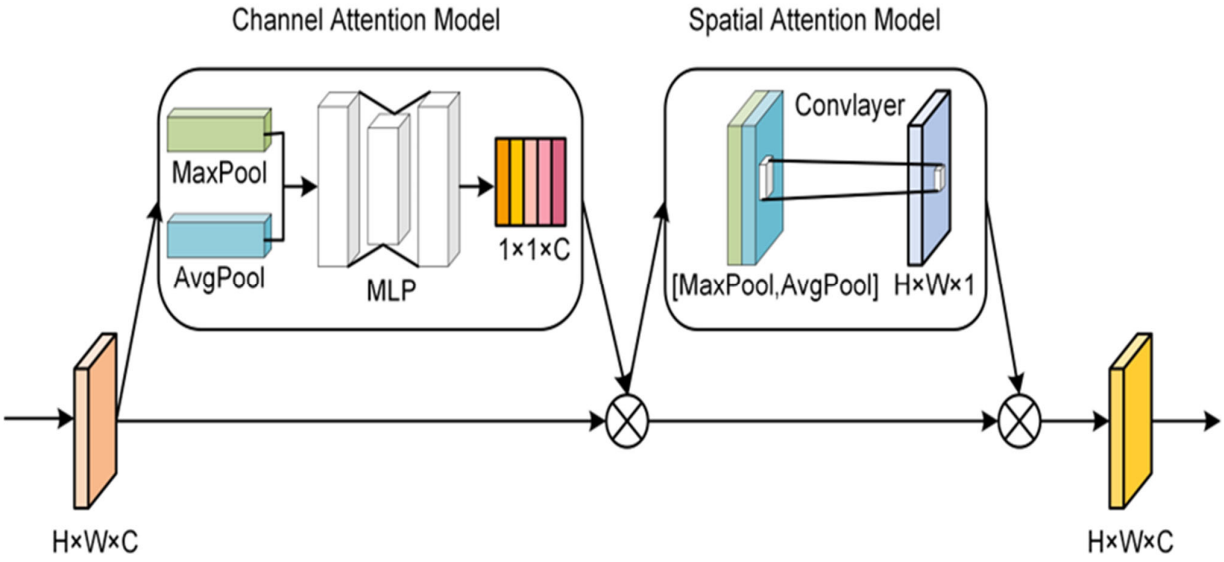

Figure 3. Diagram of the convolutional block attention module (CBAM) structure. H, W, and C represent the height, width, and number of channels of the feature maps. MLP stands for Multi-Layer Perceptron.

4. Experiments and Results

4.1. Datasets

4.1.1. LoveDA Dataset

LoveDA is a well-labeled and challenging landcover classification dataset [52]. The images were collected from Nanjing, Changzhou, and Wuhan cities, covering 18 different administrative districts. LoveDA dataset were acquired by Spaceborne sensors and collected from two main scenarios, urban and rural, containing a total of 2522 training samples and 1669 test samples. Each image contains red, green, and blue bands with a size of 1024 × 1024 pixels and a spatial resolution of 0.3 m. The labels contain the following seven categories: background, building, road, water, barren, forest, and agricultural land (Figure 4). The LoveDA dataset has become one of the most challenging landcover datasets in the field of semantic segmentation for remote sensing due to its multi-scale targets, complex background, and discontinuous category distribution.

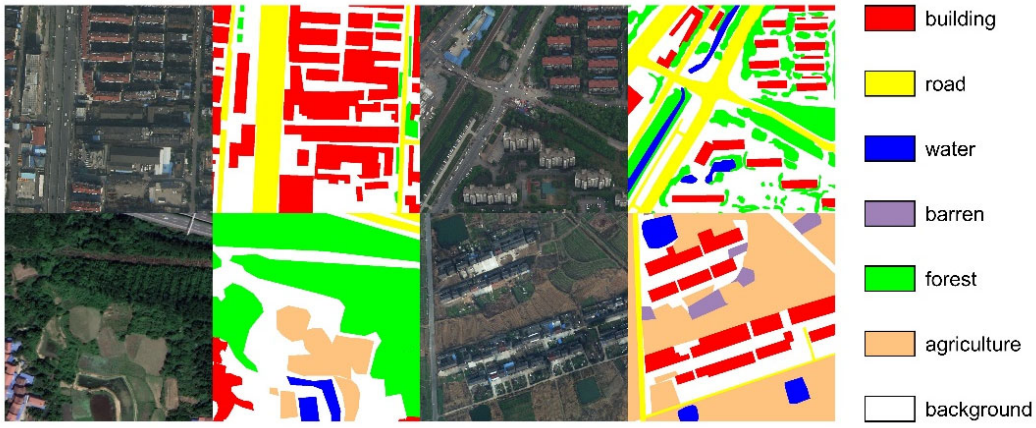

Figure 4. Some typical sample images of the LoveDA dataset and their corresponding labels.

4.1.2. Gaofen Image Dataset (GID)

The Gaofen Image Dataset (GID) is a large-scale landcover classification dataset [53] which contains widely distributed Gaofen-2 satellite images. GID contains 150 high-quality GF-2 images acquired from more than 60 different cities in China. The Gaofen image dataset contains a large-scale classification (LSC) dataset and a fine landcover classification (FLC) dataset. To verify the effectiveness of the extracted method, the fine landcover classification set providing 15 categories was selected as the experimental dataset in this study (Figure 5), which is known as the GID15 dataset. The GID15 dataset contains a total of 10 pixel-level labeled images of 7200 × 6800 size, a spatial resolution of 0.81 m, and four spectral bands of IRRGB. This dataset was first cropped into non-overlapping 2100 patches of 512 × 512 size, and then these patches were divided into 1680 training samples and 420 validation samples.

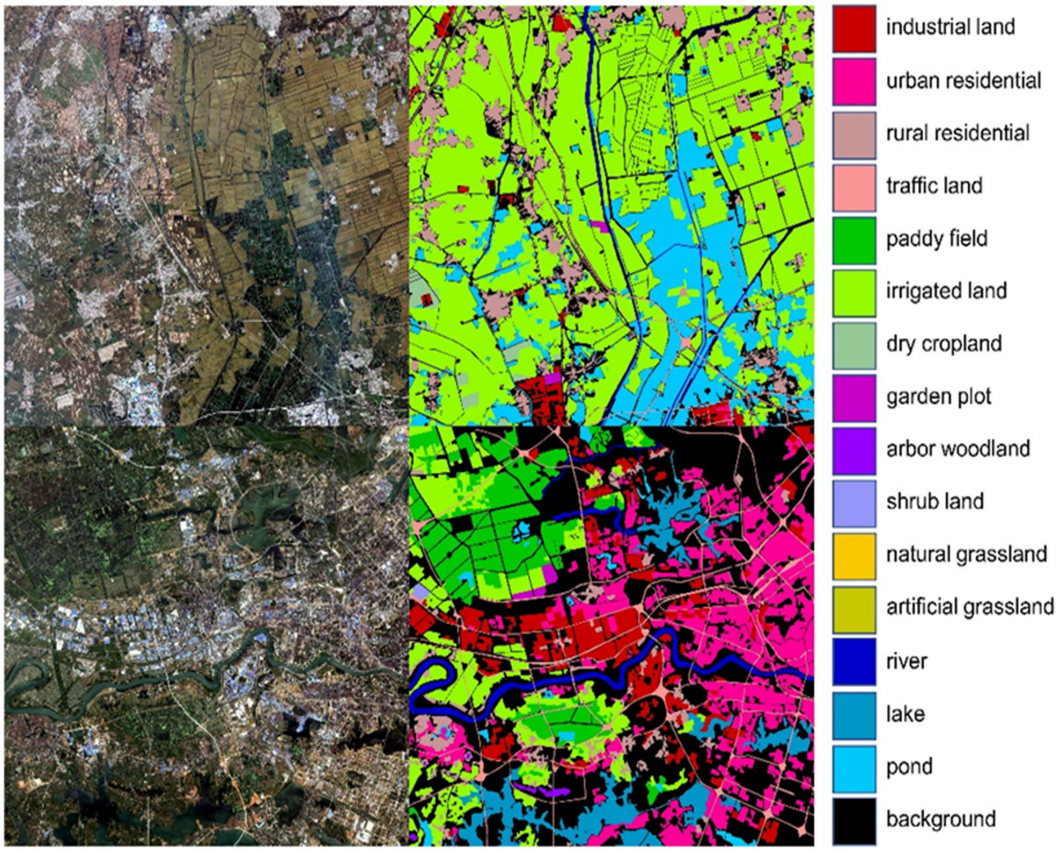

Figure 5. Some typical sample images of the GID15 dataset and their corresponding labels.

4.2. Implementation Details

All experiments were implemented by the Python 3.8 and Pytorch 1.10.0 framework and trained on a 1 NVIDIA RTX A6000 GPU with 48 G of RAM. The batch size was set to eight, the number of working threads was four, and the cross-entropy loss function was used to calculate the loss values for all networks [54]. The optimizer used an SGD with a momentum value of 0.9 and weight decay of 0.0005. The learning rate was initialized with 0.01 and scheduled by the poly strategy. In addition to expanding the dataset, all training samples performed image enhancement operations including a random horizontal flip, vertical flip, and random scaling in the range of [0.5, 0.75, 1.0, 1.25, 1.5, 1.75] (all with equal probability). For all backbone networks, we used the models pre-trained on the ImageNet dataset as the initial weight files for the network training.

4.3. Evaluation Metrics

In this study, we use the overall accuracy (OA), intersection over union (IoU), mean intersection over union (mIoU), and mean accuracy (mAcc) to measure the performance of ABNet, as shown in Equations (3)–(6). The OA denotes the overall accuracy of the classification result; the IoU denotes the similarity between the predicted results and the true values; the mIoU denotes the average IoU value across all categories; the mAcc denotes the average classification accuracy across all categories.

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{3}$$

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{4}$$

$$\mathrm{mIoU} = \frac{1}{\mathrm{N}} \sum_{\mathrm{c}=1}^{\mathrm{N}} \frac{\mathrm{TP_c}}{\mathrm{TP_c} + \mathrm{FP_c} + \mathrm{FN_c}} \tag{5}$$

$$\mathrm{mAcc} = \frac{1}{\mathrm{N}} \sum_{\mathrm{c}=1}^{\mathrm{N}} \frac{\mathrm{TP_c}}{\mathrm{TP_c} + \mathrm{FN_c}} \tag{6}$$

where TP and TN are the positive and negative examples of correct predictions; FP and FN are the positive and negative examples of incorrect predictions.

Furthermore, the following two criteria have been adopted to evaluate the model complexity: (1) floating-point operations (FLOPs), which is the number of floating-point operations accounting for the computational complexity; (2) parameters, which is the number of trainable parameters representing the model size.

4.4. Ablation Studies

To assess the efficacy of both single backbone networks and aggregated backbone networks, we employed the Deeplabv3+ decoding model as a unified segmentation head. Experiments were individually performed utilizing ResNet-50, HRNet-48, VoVNet-39, and the aggregated backbone as the encoder's backbone network. Ablation experiments were executed on both the LoveDA dataset and the GID15 dataset, with the ensuing experimental results expounded upon in Sections 4.4.1 and 4.4.2, respectively.

4.4.1. LoveDA Dataset

Illustrated in Figure 6, the classification accuracy of the aggregated backbone network demonstrates notable improvements in contrast to ResNet-50, HRNet-48, and VoVNet-39. The aggregated backbone network surpasses ResNet-50, HRNet-48, and VoVNet-39 by over 3% in terms of the overall accuracy (OA), mean accuracy (mAcc), and mean intersection over union (mIoU). Consequently, the amalgamation of diverse backbone networks proves to be an effective strategy for enhancing the performance of semantic segmentation models in the context of encoder–decoders.

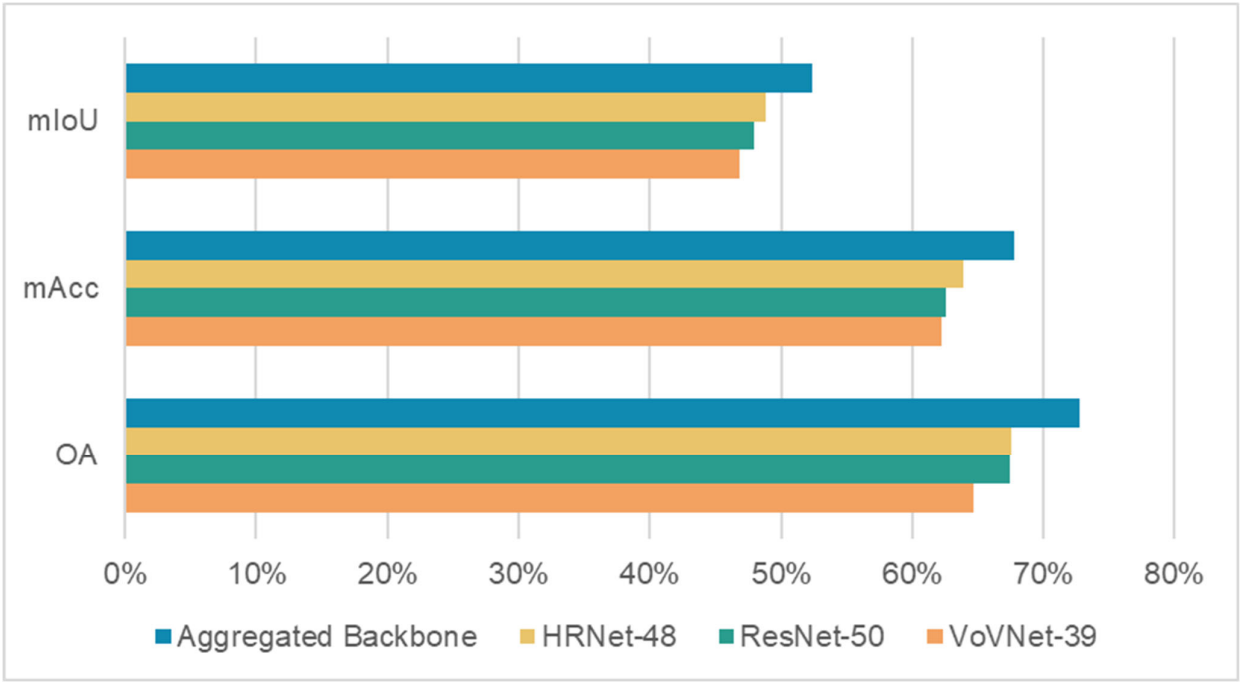

Figure 6. Performance comparison of ABNet with the unaggregated backbones.

As depicted in Table 3, the aggregated backbone network exhibits superior performance over ResNet-50, HRNet-48, and VoVNet-39 across all categories of the LoveDA dataset. Remarkably, ABNet notably enhances the segmentation performance of water bodies and bare land by a minimum of 7.14% and 4.63% in comparison to the other three backbone networks. Moreover, the aggregated backbone network demonstrates varied degrees of improvement across categories, including buildings, roads, forests, agricultural land, and background.

Table 3. IoU (%) for ablation experiments on the LoveDA dataset.

Backbone	Background	Building	Road	Water	Barren	Forest	Agricultural
ResNet-50	52.19	56.03	51.51	62.85	22.4	39.83	50.56
HRNet-48	52.87	59.63	53.01	60.85	29.23	38.58	47.66
VoVNet-39	48.98	62.51	53.51	61.32	18.06	37.48	45.65
ABNet	53.85	63.69	54.06	69.99	33.86	40.23	51.33

To more effectively demonstrate the advantages of the aggregated backbone network in comparison to other backbone networks, we present the visualization results of different backbone networks in Figure 7. In Figure 7a, the aggregated backbone network excels in recognizing complex shapes of bare ground. Moreover, in Figure 7b, the aggregated backbone network demonstrates superior recognition of small-scale water bodies. Figure 7c illustrates the aggregation backbone network's ability to minimize misclassifications of forests, buildings, and roads. While in Figure 7d, the network effectively reduces misclassifications of agricultural land, water bodies, and forests. Combining the observations from Table 3 and Figure 7, it is evident that the aggregated backbone network not only excels in recognizing intricate features that pose challenges to individual backbone networks but also enhances the recognition capabilities across multi-scale targets.

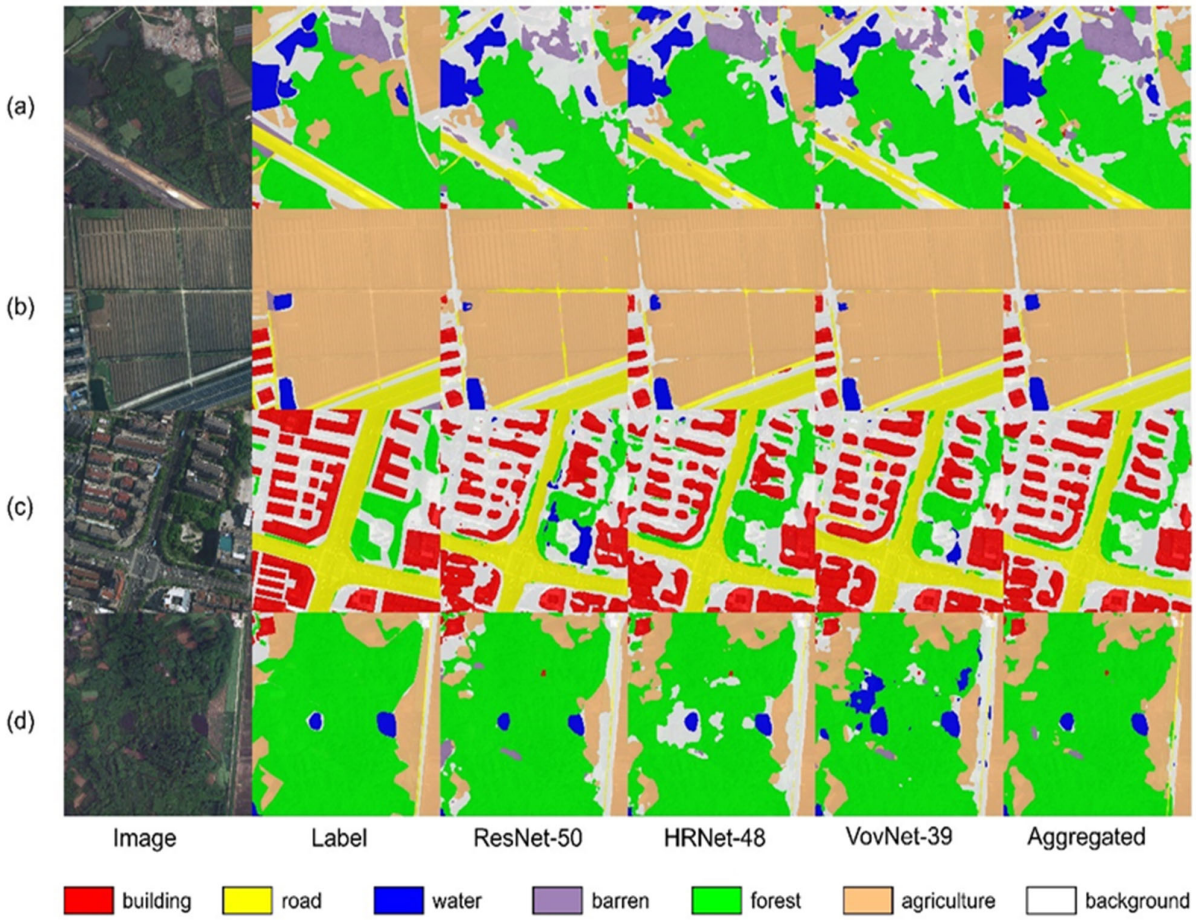

Figure 7. Prediction results of ablation experiments on the LoveDA dataset. (**a**–**d**) shows the typical case that it is difficult to identify barren, water, building, and agriculture.

4.4.2. GID15 Dataset

To further verify the generalization performance of the aggregated backbone network, we conducted ablation experiments using the GID15 dataset, which encompasses 15 categories. In this study, Table 4 and Figure 8 present the IoU and visualization results from the classification of the GID15 dataset, leveraging ResNet-50, HRNet-48, VoVNet-39, and the aggregated backbone as the backbone network, respectively.

As observed in Table 4, the segmentation accuracy of the aggregated backbone network on the GID15 dataset significantly surpasses that of ResNet-50, HRNet-48, and VoVNet-39, with the mean intersection over union improved by 4.59%, 3.45%, and 3.24%, respectively. Except for rural residential land and artificial grassland, the IoU of segmentation by ABNet exhibits differing degrees of enhancement compared to the other three backbone networks. Specifically, ABNet proves highly effective in classifying shrub land, demonstrating a minimum increase of 6.15% in the IoU. Furthermore, ABNet also enhances the classification accuracy of garden plots, natural grassland, ponds, paddy fields, and dry cropland, with minimum increases of 4.68%, 2.95%, 2.85%, 2.42%, and 2.05%, respectively. Additionally, ABNet improves the segmentation accuracy of the background, industrial land, urban residential land, traffic land, irrigated land, arbor woodland, river, and lake classes to a certain extent.

From Figure 8a–c,e,g, it is evident that our results better preserve the geometric features of the objects, aligning more closely with the manually labeled results and exhibiting more regular shapes. Figure 8d demonstrates that our proposed method reduces misclassifications of small-scale targets in ponds. Furthermore, Figure 8f visually illustrates a substantial improvement in the segmentation accuracy of shrub land, providing compelling evidence for the superiority of ABNet.

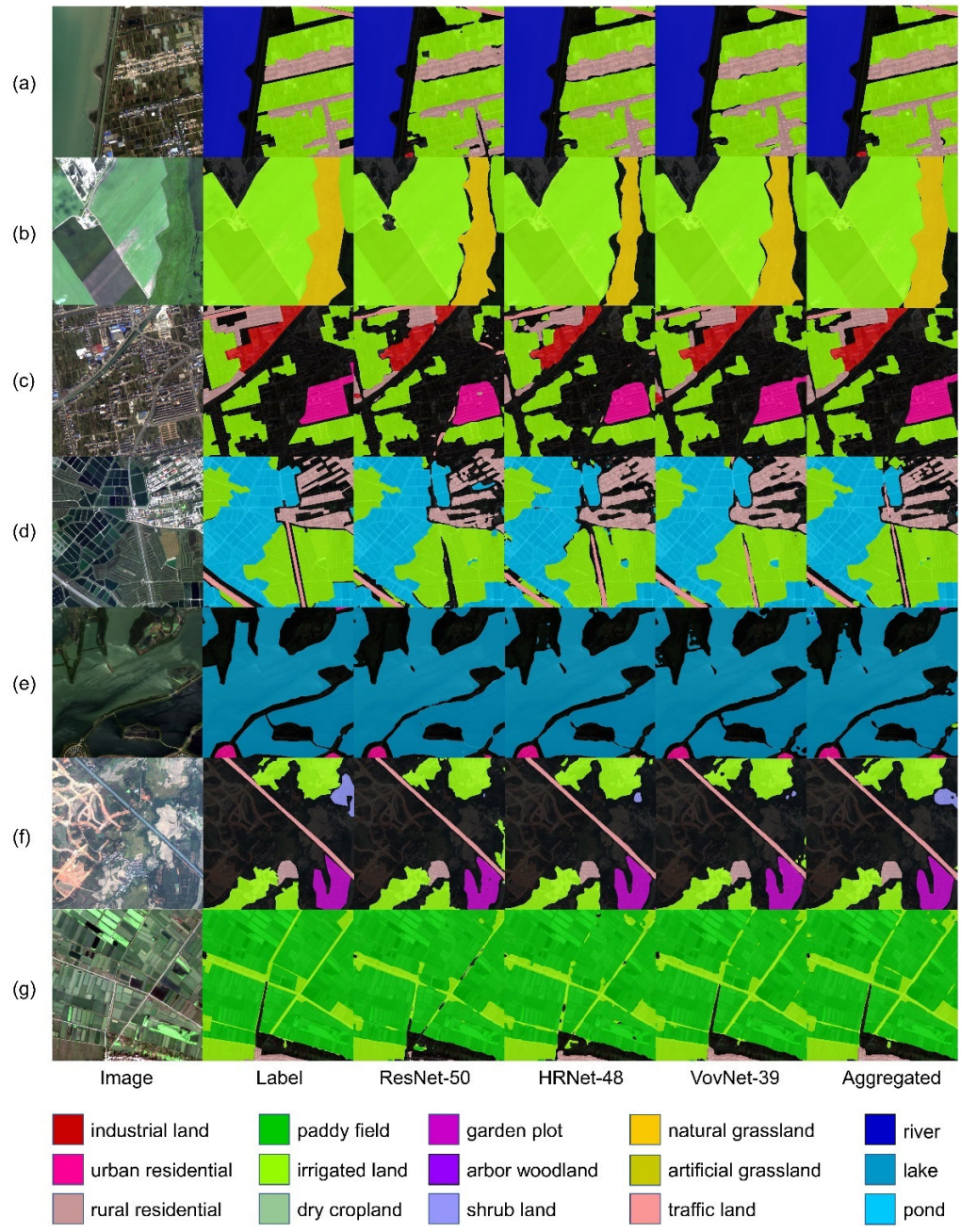

Figure 8. Prediction results of ablation experiments on GID15 dataset. (**a**–**g**) shows the typical case that it is difficult to identify rural residential, natural grassland, industrial land, pond, lake, shrub land, and paddy field.

Table 4. IoU (%) for ablation experiments on the GID15 dataset.

Class	ResNet-50	HRNet-48	VoVNet-39	ABNet
Background	70.01	69.61	70.47	73.02
Industrial land	59.51	60.03	61.93	63.43
Urban residential	64.34	63.58	66.57	67.38
Rural residential	56.03	54.18	58.09	57.81
Traffic land	59.93	63.86	63.9	64.05
Paddy field	63.07	65.8	64.52	68.22
Irrigated land	71.37	71.29	71.24	73.27
Dry cropland	59.01	56.68	58.76	61.06
Garden plot	31.38	39.54	24.02	44.22
Arbor woodland	78.35	78.64	75.52	79.39
Shrub land	13.67	23.15	21.67	29.3
Natural grassland	61.94	62.54	64.06	67.01
Artificial grassland	24.37	24.53	32.62	30.03
River	79.85	78.81	80.48	82.7
Lake	70.99	67.99	69.79	71.69
Pond	62.29	64.01	63.99	66.86
mIoU	57.88	59.02	59.23	62.47

4.5. Comparing with the State-of-the-Art

To showcase the superior performance of our method, a series of experiments were conducted to compare it with other state-of-the-art semantic segmentation models. This study selected mainstream single backbone network models and a composite backbone network model for experiments with equivalent training configurations on the LoveDA and GID15 datasets. The single backbone network models included UNet, FPN, PSPNet, DANet, CCNet, and UPerNet utilizing the widely used ResNet-50 as the backbone network for experimentation. CBNet, as a representative of composite backbone network models, involved the composite connection of ResNet and Res2Net as the backbone network for experimentation. UNet is a classic encoder–decoder semantic segmentation model; FPN, PSPNet, and UPerNet represent feature pyramid methods; DANet and CCNet represent attention mechanism methods; and CBNet represents an approach to convergence of multiple backbone networks.

4.5.1. LoveDA Dataset

Table 5 illustrates the accuracy results obtained from the different models after training on the LoveDA dataset. Whether considering the OA, mIoU, or mAcc, the segmentation accuracy of ABNet outperforms the other models. In comparison to UNet, the mIoU demonstrates a 7.33% improvement, and when compared with CCNet, there is a 1.51% increase in the mIoU. The mIoU of FPN, PSPNet, DANet, and UPerNet all surpass that of UNet, but are lower than that of CCNet.

Table 5. Accuracy results of different models on the LoveDA dataset (%).

Model	Backbone	OA	mIoU	mAcc
UNet	ResNet-50	63.50	45.10	61.76
FPN	ResNet-50	67.88	49.76	62.99
PSPNet	ResNet-50	68.82	50.53	62.08
DANet	ResNet-50	68.19	48.73	60.13
CBNet	CB-ResNet50	68.26	50.28	62.39
UPerNet	ResNet-50	68.72	50.36	62.03
CCNet	ResNet-50	68.87	50.92	63.72
ABNet	Aggregated Backbone	72.75	52.43	67.82

From Figure 9a, it is evident that ABNet significantly extracts the barren feature compared to the other models, showcasing that features that are challenging to distinguish by a single backbone network can be more effectively identified by the composite backbone network. In Figure 9b, it is noticeable that ABNet is less adept than CCNet and UPerNet in extracting small-sized water bodies. However, as depicted in Figure 9c,d, ABNet delineates agricultural land more effectively than CCNet and UPerNet. This shows that the proposed model can effectively improve the task of semantic segmentation of remote sensing images by fusing the structure of different backbone networks.

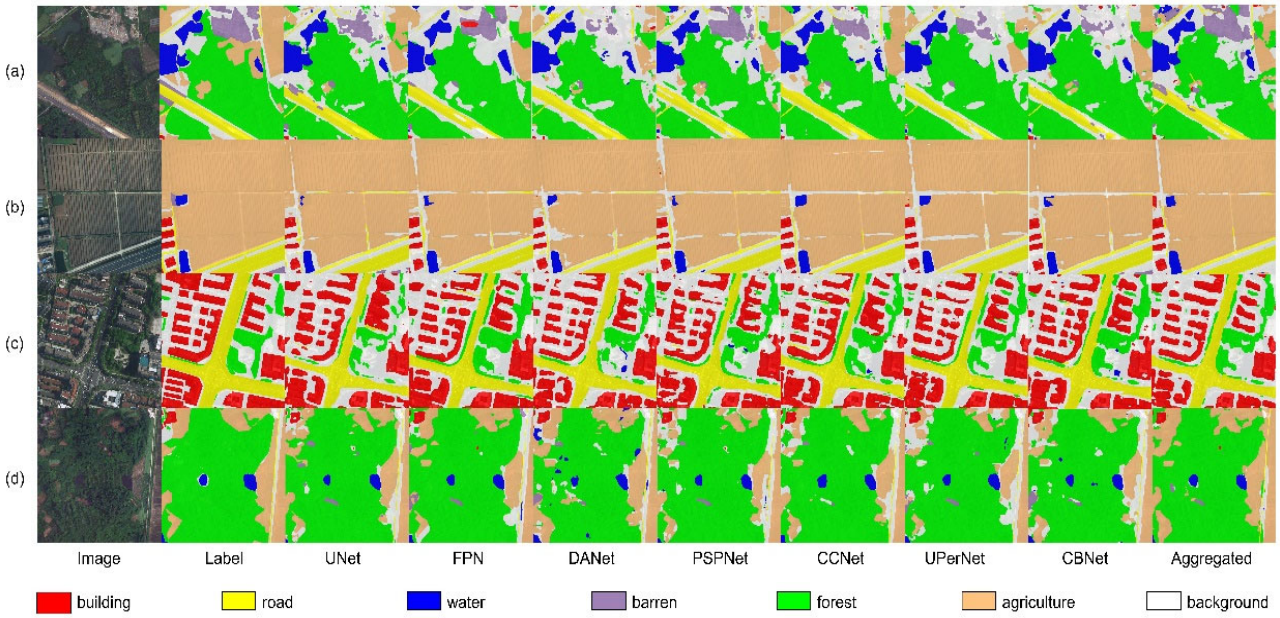

Figure 9. Comparison of prediction results of different state-of-the-art models on the LoveDA dataset. (**a**–**d**) shows the typical case that it is difficult to identify barren, water, building, and agriculture.

4.5.2. GID15 Dataset

In this study, we conducted experiments on the GID15 dataset, which contains finer categories, to evaluate the efficacy of ABNet. Table 6 demonstrates the segmentation accuracy obtained after training various models on the GID15 dataset. Across metrics such as the overall accuracy (OA), mean intersection over union (mIoU), and mean accuracy (mAcc), ABNet consistently outperforms the other models. Notably, in comparison to UNet, the mIoU is improved by 7.58%, while, compared with CBNet, there is a 1.20% enhancement in the mIoU. Additionally, the mIoU values of FPN, PSPNet, DANet, CCNet, and UPerNet surpass that of UNet, although they are lower than that of CBNet.

Table 6. Accuracy results of different models on the GID15 dataset (%).

Model	Backbone	OA	mIoU	mAcc
UNet	ResNet-50	81.33	54.89	65.39
FPN	ResNet-50	81.74	59.66	67.77
PSPNet	ResNet-50	82.05	60.78	68.48
DANet	ResNet-50	82.19	60.84	68.83
CCNet	ResNet-50	82.29	60.46	69.09
UPerNet	ResNet-50	82.27	61.09	69.54
CBNet	CB-ResNet50	82.83	61.27	68.92
ABNet	Aggregated Backbone	83.27	62.47	72.74

Moreover, in Figure 10, we provide a comprehensive comparison result. Specifically, Figure 10a emphasizes the superior capability of ABNet in correctly recognizing small-

sized features, a task in which the other models encounter challenges. Furthermore, Figure 10b demonstrates ABNet's precise classification of water body edges, surpassing the performance of the other models. Similarly, in Figure 10c, the result indicates that ABNet excels in extracting shrub land, implying its effectiveness in classifying intricate feature targets. Finally, Figure 10d illustrates the consistency of ABNet in extracting the contour boundaries of rice fields, showcasing its ability to achieve superior segmentation performance, particularly for finer feature classes.

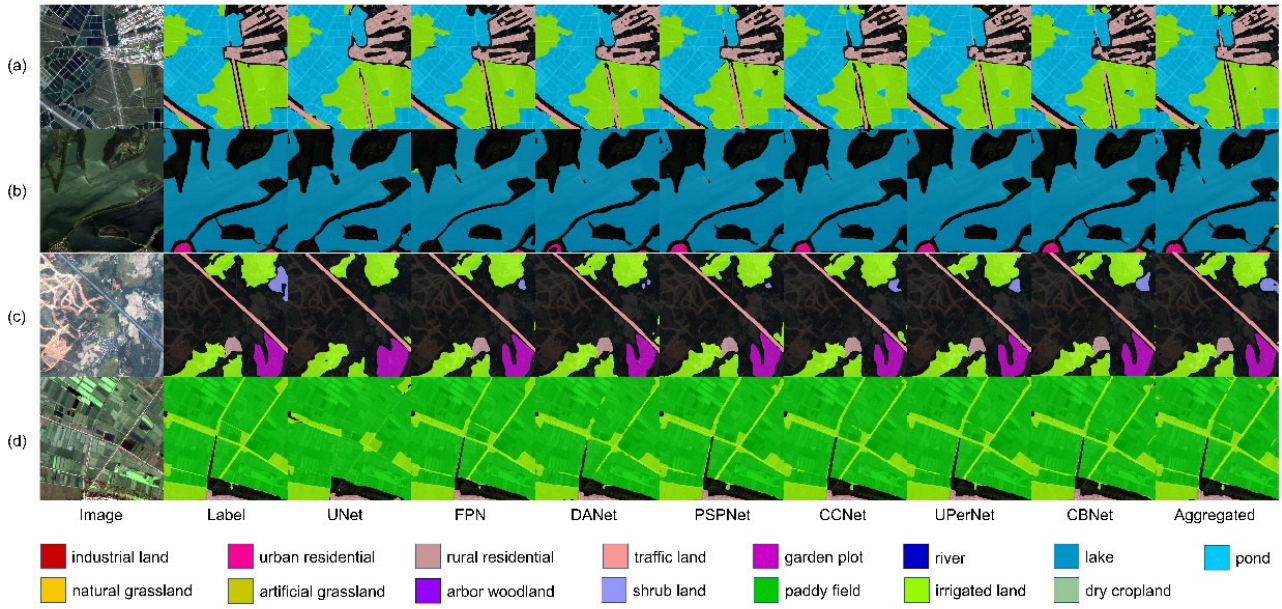

Figure 10. Comparison of prediction results of different state-of-the-art models on the GID15 dataset. (**a**) shows the typical case that it is difficult to identify pond in small-scale feature. (**b**–**d**) shows the typical case that it is difficult to identify lake, shrub land, and paddy field.

5. Discussion

5.1. Advantages and Limitations

5.1.1. Advantages

ResNet utilizes residual connections in convolutional blocks to ensure the stability of neural network convergence at each stage. However, it lacks feature fusion within and between stages of the backbone network. In contrast, VoVNet achieves feature integration within each stage through the one-shot aggregation of each convolutional block in the backbone network. Nonetheless, it does not fuse features between different stages. Meanwhile, HRNet achieves the feature fusion of outputs from different stages through parallel concatenation. However, as the number of stages increases, the output features of each stage become more complex after multi-scale feature fusion, leading to some loss of original information within each stage. The aggregation backbone network (ABNet) uniquely combines the deep feature characterization capabilities of ResNet with the structural features of HRNet's string–parallel combination and VoVNet's global one-shot aggregation. This integration significantly expands the receptive field of the convolutional blocks at each stage, thereby enhancing the model's capacity to leverage spatial context information [55]. As a result, ABNet consistently enhances the classification accuracy across all classes within the LovaDA and GID15 datasets. Furthermore, owing to the multi-stage output structure of ABNet, it efficiently generates merged features of varying scale sizes, which efficiently utilizes the semantic information from both deep and shallow features and improves the classification accuracy of the model for small-scale targets [56]. Additionally, to mitigate information redundancy within the merged features, the convolutional block attention model (CBAM) optimizes the multi-scale merged features across both spatial and channel dimensions. This optimization enhances the distinctive characterizing abilities of the fea-

tures at each stage [57]. Consequently, when applied to landcover semantic segmentation tasks, this approach yields superior segmentation results and bolsters the model's capability to discern features that pose challenges to a single backbone network's recognition ability.

5.1.2. Limitations

Despite the proven efficacy of ABNet on landcover datasets, its substantial model parameters and higher computational time complexities present notable considerations. To comprehensively assess the practicality of these models, this study quantified the parameters and floating-point operations (FLOPs) for ABNet and comparable models using a sample input image size of 1024 × 1024 [58]. As illustrated in Table 7, the parameters and FLOPs for ABNet notably surpass those of other models, with ABNet exceeding the second-largest CBNet model by 100.55 M and 0.33 T FLOPs. The abundance of model parameters and computational cost contributes to heightened challenges in model convergence, especially when computational resources are constrained. Consequently, the suitability of ABNet for complex scene segmentation tasks may diminish under limited computational resources.

Table 7. Comparison with other networks on parameters and FLOPs.

Method	Parameters (M)	FLOPs (T)
Deeplabv3plus_Res-50	41.22	0.71
Deeplabv3plus_HRNet-48	68.59	1.01
Deeplabv3plus_VoVNet-39	34.27	0.88
DANet	47.93	0.96
PSPNet	46.61	0.72
CCNet	47.46	0.84
UPerNet	64.04	0.95
CBNet	69.71	1.08
ABNet	170.26	1.41

5.2. Potential Improvements

To further optimize ABNet, this study explores three potential directions for enhancing its performance. Firstly, without compromising the model performance, reducing the number of model parameters can be achieved by decreasing the number of layers in the backbone network or minimizing the involvement of the backbone network models in fusion. Secondly, integrating advanced decoders, such as the feature pyramid network (FPN), can facilitate a more effective fusion of deep and shallow features, thereby enhancing classification performance. Lastly, enhancing the fusion of different backbone networks presents an opportunity for improvement. In addition to merging the output layers of the same stage, combining the outputs of adjacent stages of diverse backbone networks through upsampling or downsampling operations can promote the deep fusion of deep and shallow features, fostering further performance enhancements.

6. Conclusions

This paper introduced the aggregated backbone network (ABNet), a novel convolutional neural network designed for fine-grained landcover classification. ABNet leverages the Deeplabv3+ head and reconstructs the backbone network through the same-stage fusion of ResNet-50, HRNet-48, and VoVNet-39. The aggregated backbone network yields merged features in multi-scale sizes, which are further optimized using the convolutional block attention mechanism.

On the LoveDA and GID15 datasets, the aggregated backbone network demonstrates superior classification accuracy compared to a single backbone network, outperforming the overall accuracy (OA), mean intersection over union (mIoU), and mean accuracy (mAcc). Specifically, the mIoU derived from the aggregated backbone network surpasses that of the single backbone network by at least 3%. Notably, ABNet enhances the classification

accuracy of diverse land classes and improves the discernment of challenging classes often difficult for a single backbone network to recognize. Moreover, when compared with popular semantic segmentation models (UNet, FPN, PSPNet, DANet, CBNet, CCNet, and UPerNet), the aggregated backbone network exhibits leading performance, affirming the efficacy of the proposed method. Although the ABNet proposed in this paper exhibits excellent segmentation performance on the LovaDA and GID15 datasets, the model has no advantages in terms of the parameters and FLOPs of the algorithm.

Consequently, this approach presents an effective strategy to enhance the classification performance of remote sensing deep learning models. Researchers can further explore and customize the aggregated structure by integrating different types and quantities of backbone networks to tackle more intricate landcover classification challenges. Meanwhile, scholars can investigate how to reduce the parameters of the integrated model while maintaining efficient classification performance.

Author Contributions: Conceptualization, B.S.; methodology, B.S.; software, B.S.; validation, B.S., formal analysis, B.S.; investigation, B.S.; writing—original draft preparation, B.S.; writing—review and editing, B.S. and Z.Y.; visualization, B.S. and Z.W.; supervision, K.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: We analyzed publicly available datasets in this study. Data are freely available here: (1) LoveDA dataset: https://github.com/Junjue-Wang/LoveDA (accessed on 14 March 2023) (2) GID Dataset: http://www.captain-whu.com/repository.html (accessed on 14 March 2023).

Acknowledgments: Thanks to Ziran Ye for suggesting revisions to this paper.

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

References

1. Wang, Y.Z.; Sun, Y.H.; Cao, X.Y.; Wang, Y.H.; Zhang, W.K.; Cheng, X.L. A review of regional and Global scale Land Use/Land Cover (LULC) mapping products generated from satellite remote sensing. *ISPRS J. Photogramm. Remote Sens.* **2023**, *206*, 311–334. [CrossRef]
2. Su, Y.; Qian, K.; Lin, L.; Wang, K.; Guan, T.; Gan, M.Y. Identifying the driving forces of non-grain production expansion in rural China and its implications for policies on cultivated land protection. *Land Use Policy* **2020**, *92*, 104435. [CrossRef]
3. Boguszewski, A.; Batorski, D.; Ziemba-Jankowska, N.; Dziedzic, T.; Zambrzycka, A. LandCover.ai: Dataset for Automatic Mapping of Buildings, Woodlands, Water and Roads from Aerial Imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 1102–1110.
4. Tong, X.Y.; Xia, G.S.; Zhu, X.X. Enabling country-scale land cover mapping with meter-resolution satellite imagery. *ISPRS J. Photogramm. Remote Sens.* **2023**, *196*, 178–196. [CrossRef] [PubMed]
5. Sertel, E.; Ekim, B.; Osgouei, P.E.; Kabadayi, M.E. Land Use and Land Cover Mapping Using Deep Learning Based Segmentation Approaches and VHR Worldview-3 Images. *Remote Sens.* **2022**, *14*, 4558. [CrossRef]
6. Ma, L.; Liu, Y.; Zhang, X.L.; Ye, Y.X.; Yin, G.F.; Johnson, B.A. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 166–177. [CrossRef]
7. Kampffmeyer, M.; Salberg, A.B.; Jenssen, R. Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks. In Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 680–688.
8. Ye, Z.R.; Fu, Y.Y.; Gan, M.Y.; Deng, J.S.; Comber, A.; Wang, K. Building Extraction from Very High Resolution Aerial Imagery Using Joint Attention Deep Neural Network. *Remote Sens.* **2019**, *11*, 2970. [CrossRef]
9. Fu, Y.Y.; Liu, K.K.; Shen, Z.Q.; Deng, J.S.; Gan, M.Y.; Liu, X.G.; Lu, D.M.; Wang, K. Mapping Impervious Surfaces in Town-Rural Transition Belts Using China's GF-2 Imagery and Object-Based Deep CNNs. *Remote Sens.* **2019**, *11*, 280. [CrossRef]
10. Zhang, D.J.; Pan, Y.Z.; Zhang, J.S.; Hu, T.G.; Zhao, J.H.; Li, N.; Chen, Q. A generalized approach based on convolutional neural networks for large area cropland mapping at very high resolution. *Remote Sens. Environ.* **2020**, *247*, 111912. [CrossRef]
11. Kumar, D.G.; Chaudhari, S. Comparison of Deep Learning Backbone Frameworks for Remote Sensing Image Classification. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 7763–7766.
12. Wieland, M.; Martinis, S.; Kiefl, R.; Gstaiger, V. Semantic segmentation of water bodies in very high-resolution satellite and aerial images. *Remote Sens. Environ.* **2023**, *287*, 113452. [CrossRef]

13. Liu, Y.D.; Wang, Y.T.; Wang, S.W.; Liang, T.T.; Zhao, Q.J.; Tang, Z.; Ling, H.B. CBNet: A Novel Composite Backbone Network Architecture for Object Detection. In Proceedings of the 34th AAAI Conference on Artificial Intelligence/32nd Innovative Applications of Artificial Intelligence Conference/10th AAAI Symposium on Educational Advances in Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 11653–11660.
14. Liang, T.T.; Chu, X.J.; Liu, Y.D.; Wang, Y.T.; Tang, Z.; Chu, W.; Chen, J.D.; Ling, H.B. CBNet: A Composite Backbone Network Architecture for Object Detection. *IEEE Trans. Image Process.* **2022**, *31*, 6893–6906. [CrossRef]
15. Elharrouss, O.; Akbari, Y.; Almaadeed, N.; Al-Maadeed, S. Backbones-Review: Feature Extraction Networks for Deep Learning and Deep Reinforcement Learning Approaches. *arXiv* **2022**, arXiv:2206.08016.
16. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
17. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
18. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 27–30 June 2016; pp. 2818–2826.
19. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
20. Lee, Y.W.; Hwang, J.W.; Lee, S.; Bae, Y.; Park, J. An Energy and GPU-Computation Efficient Backbone Network for Real-Time Object Detection. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 752–760.
21. Sun, K.; Xiao, B.; Liu, D.; Wang, J.D. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 5686–5696.
22. Tan, M.X.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning (ICML), Long Beach, CA, USA, 9–15 June 2019.
23. Gao, S.H.; Cheng, M.M.; Zhao, K.; Zhang, X.Y.; Yang, M.H.; Torr, P. Res2Net: A New Multi-Scale Backbone Architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 652–662. [CrossRef] [PubMed]
24. Ye, M.; Ruiwen, N.; Chang, Z.; He, G.; Tianli, H.; Shijun, L.; Yu, S.; Tong, Z.; Ying, G. A Lightweight Model of VGG-16 for Remote Sensing Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 6916–6922. [CrossRef]
25. Tao, C.X.; Meng, Y.Z.; Li, J.J.; Yang, B.B.; Hu, F.M.; Li, Y.X.; Cui, C.L.; Zhang, W. MSNet: Multispectral semantic segmentation network for remote sensing images. *GIScience Remote Sens.* **2022**, *59*, 1177–1198. [CrossRef]
26. Liu, K.; Yu, S.T.; Liu, S.D. An Improved InceptionV3 Network for Obscured Ship Classification in Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 4738–4747. [CrossRef]
27. Xu, Y.Y.; Xie, Z.; Feng, Y.X.; Chen, Z.L. Road Extraction from High-Resolution Remote Sensing Imagery Using Deep Learning. *Remote Sens.* **2018**, *10*, 1461. [CrossRef]
28. Zhao, L.R.; Niu, R.Q.; Li, B.Q.; Chen, T.; Wang, Y.Y. Application of Improved Instance Segmentation Algorithm Based on VoVNet-v2 in Open-Pit Mines Remote Sensing Pre-Survey. *Remote Sens.* **2022**, *14*, 2626. [CrossRef]
29. Guo, S.C.; Yang, Q.; Xiang, S.M.; Wang, P.F.; Wang, X.Z. Dynamic High-Resolution Network for Semantic Segmentation in Remote-Sensing Images. *Remote Sens.* **2023**, *15*, 2293. [CrossRef]
30. Hu, S.; Liu, J.; Kang, Z.W. DeepLabV3+/Efficientnet Hybrid Network-Based Scene Area Judgment for the Mars Unmanned Vehicle System. *Sensors* **2021**, *21*, 8136. [CrossRef]
31. Das, A.; Chandran, S. Transfer Learning with Res2Net for Remote Sensing Scene Classification. In Proceedings of the 11th International Conference on Cloud Computing, Data Science and Engineering (Confluence), Uttar Pradesh, India, 28–29 January 2021; pp. 796–801.
32. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
33. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; pp. 234–241.
34. Chen, L.C.E.; Zhu, Y.K.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 833–851.
35. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.M.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
36. Zhao, H.S.; Shi, J.P.; Qi, X.J.; Wang, X.G.; Jia, J.Y. Pyramid Scene Parsing Network. In Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6230–6239.
37. Fu, J.; Liu, J.; Tian, H.J.; Li, Y.; Bao, Y.J.; Fang, Z.W.; Lu, H.Q. Dual Attention Network for Scene Segmentation. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 3141–3149.

38. Xiao, T.T.; Liu, Y.C.; Zhou, B.L.; Jiang, Y.N.; Sun, J. Unified Perceptual Parsing for Scene Understanding. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 432–448.
39. Huang, Z.L.; Wang, X.G.; Wei, Y.C.; Huang, L.C.; Shi, H.; Liu, W.Y.; Huang, T.S. CCNet: Criss-Cross Attention for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 6896–6908. [CrossRef]
40. Liang, C.B.; Xiao, B.H.; Cheng, B.; Dong, Y.Y. XANet: An Efficient Remote Sensing Image Segmentation Model Using Element-Wise Attention Enhancement and Multi-Scale Attention Fusion. *Remote Sens.* **2023**, *15*, 236. [CrossRef]
41. Wang, D.; Yang, R.H.; Liu, H.H.; He, H.Q.; Tan, J.X.; Li, S.D.; Qiao, Y.C.; Tang, K.Q.; Wang, X. HFENet: Hierarchical Feature Extraction Network for Accurate Landcover Classification. *Remote Sens.* **2022**, *14*, 4244. [CrossRef]
42. Chen, C.; Zhao, H.L.; Cui, W.; He, X. Dual Crisscross Attention Module for Road Extraction from Remote Sensing Images. *Sensors* **2021**, *21*, 6873. [CrossRef] [PubMed]
43. Ye, Z.R.; Si, B.; Lin, Y.; Zheng, Q.M.; Zhou, R.; Huang, L.; Wang, K. Mapping and Discriminating Rural Settlements Using Gaofen-2 Images and a Fully Convolutional Network. *Sensors* **2020**, *20*, 6062. [CrossRef]
44. Kotaridis, I.; Lazaridou, M. Cnns in land cover mapping with remote sensing imagery: A review and meta-analysis. *Int. J. Remote Sens.* **2023**, *44*, 5896–5935. [CrossRef]
45. Bigdeli, B.; Pahlavani, P.; Amirkolaee, H.A. An ensemble deep learning method as data fusion system for remote sensing multisensor classification. *Appl. Soft Comput.* **2021**, *110*, 107563. [CrossRef]
46. Fan, R.Y.; Feng, R.Y.; Wang, L.Z.; Yan, J.N.; Zhang, X.H. Semi-MCNN: A Semisupervised Multi-CNN Ensemble Learning Method for Urban Land Cover Classification Using Submeter HRRS Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 4973–4987. [CrossRef]
47. Cao, Y.; Huo, C.L.; Xu, N.; Zhang, X.; Xiang, S.M.; Pan, C.H. HENet: Head-Level Ensemble Network for Very High Resolution Remote Sensing Images Semantic Segmentation. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6506005. [CrossRef]
48. Ekim, B.; Sertel, E. Deep neural network ensembles for remote sensing land cover and land use classification. *Int. J. Digit. Earth* **2021**, *14*, 1868–1881. [CrossRef]
49. Mao, M.; Zhang, B.; Doermann, D.; Guo, J.; Han, S.; Feng, Y.; Wang, X.; Ding, E. Probabilistic Ranking-Aware Ensembles for Enhanced Object Detections. *arXiv* **2021**, arXiv:2105.03139.
50. Chen, M.H.; Fu, J.L.; Ling, H.B. One-Shot Neural Ensemble Architecture Search by Diversity-Guided Search Space Shrinking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 16525–16534.
51. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. *arXiv* **2018**, arXiv:1807.06521.
52. Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; Zhong, Y. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv* **2021**, arXiv:2110.08733.
53. Tong, X.Y.; Xia, G.S.; Lu, Q.K.; Shen, H.F.; Li, S.Y.; You, S.C.; Zhang, L.P. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sens. Environ.* **2020**, *237*, 111322. [CrossRef]
54. Xia, L.; Zhao, F.; Chen, J.; Yu, L.; Lu, M.; Yu, Q.Y.; Liang, S.F.; Fan, L.L.; Sun, X.; Wu, S.R.; et al. A full resolution deep learning network for paddy rice mapping using Landsat data. *ISPRS J. Photogramm. Remote Sens.* **2022**, *194*, 91–107. [CrossRef]
55. Qiang, J.; Liu, W.J.; Li, X.X.; Guan, P.; Du, Y.L.; Liu, B.; Xiao, G.L. Detection of citrus pests in double backbone network based on single shot multibox detector. *Comput. Electron. Agric.* **2023**, *212*, 108158. [CrossRef]
56. Neupane, B.; Horanont, T.; Aryal, J. Deep Learning-Based Semantic Segmentation of Urban Features in Satellite Images: A Review and Meta-Analysis. *Remote Sens.* **2021**, *13*, 808. [CrossRef]
57. Cui, Z.Y.; Li, Q.; Cao, Z.J.; Liu, N.Y. Dense Attention Pyramid Networks for Multi-Scale Ship Detection in SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 8983–8997. [CrossRef]
58. Zhang, Y.H.; Lu, H.Y.; Ma, G.Y.; Zhao, H.J.; Xie, D.L.; Geng, S.T.; Tian, W.; Sian, K. MU-Net: Embedding MixFormer into Unet to Extract Water Bodies from Remote Sensing Images. *Remote Sens.* **2023**, *15*, 3559. [CrossRef]