remote sensing

MDPI

Article

C3Net: Cross-Modal Feature Recalibrated, Cross-Scale Semantic Aggregated and Compact Network for Semantic Segmentation of Multi-Modal High-Resolution Aerial Images

Zhiying Cao [1,2,3,4], Wenhui Diao [1,2,*], Xian Sun [1,2,3,4], Xiaode Lyu [1,5], Menglong Yan [1,2] and Kun Fu [1,2,3,4]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China; caozhiying16@mails.ucas.ac.cn (Z.C.); sunxian@mail.ie.ac.cn (X.S.); Louee@mail.ie.ac.cn (X.L.); yanmenglong@foxmail.com (M.Y.); fukun@mail.ie.ac.cn (K.F.)
[2] Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China
[3] University of Chinese Academy of Sciences, Beijing 100190, China
[4] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China
[5] Key Laboratory on Microwave Imaging Technology, Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China
* Correspondence: diaowh@aircas.ac.cn

check for updates

Citation: Cao, Z.; Diao, W.; Sun, X.; Lyu, X.; Yan, M.; Fu, K. C3Net: Cross-Modal Feature Recalibrated, Cross-Scale Semantic Aggregated and Compact Network for Semantic Segmentation of Multi-Modal High-Resolution Aerial Images. *Remote Sens.* **2021**, *13*, 528. https://doi.org/10.3390/rs13030528

Academic Editor: Jon Atli Benediktsson
Received: 30 December 2020
Accepted: 29 January 2021
Published: 2 February 2021

Abstract: Semantic segmentation of multi-modal remote sensing images is an important branch of remote sensing image interpretation. Multi-modal data has been proven to provide rich complementary information to deal with complex scenes. In recent years, semantic segmentation based on deep learning methods has made remarkable achievements. It is common to simply concatenate multi-modal data or use parallel branches to extract multi-modal features separately. However, most existing works ignore the effects of noise and redundant features from different modalities, which may not lead to satisfactory results. On the one hand, existing networks do not learn the complementary information of different modalities and suppress the mutual interference between different modalities, which may lead to a decrease in segmentation accuracy. On the other hand, the introduction of multi-modal data greatly increases the running time of the pixel-level dense prediction. In this work, we propose an efficient C3Net that strikes a balance between speed and accuracy. More specifically, C3Net contains several backbones for extracting features of different modalities. Then, a plug-and-play module is designed to effectively recalibrate and aggregate multi-modal features. In order to reduce the number of model parameters while remaining the model performance, we redesign the semantic contextual extraction module based on the lightweight convolutional groups. Besides, a multi-level knowledge distillation strategy is proposed to improve the performance of the compact model. Experiments on ISPRS Vaihingen dataset demonstrate the superior performance of C3Net with 15× fewer FLOPs than the state-of-the-art baseline network while providing comparable overall accuracy.

Keywords: semantic segmentation; multi-modal learning; deep neural network design

1. Introduction

Remote sensing images are widely used in a variety of applications, such as military reconnaissance [1,2], urban planning [3–9], disaster monitoring [10,11], and meteorological monitoring [12,13]. As one of the significant methods for automatic analysis and interpretation of remote sensing images, semantic segmentation, namely pixel-wise classification, aims to assign each pixel with a semantic label. In the past few years, semantic segmentation has benefited a lot from deep learning methods in the computer vision fields of natural RGB images. Indeed, methods based on deep learning, especially Convolutional Neural Networks (CNNs), have greatly improved the accuracy of semantic segmentation in remote sensing images. However, large-scope remote sensing scene usually contains complex backgrounds and multi-scale objects, which brings great challenges to semantic segmentation. The third column in Figure 1 shows the segmentation results of the remote sensing images with the model trained on the natural RGB images. As shown in Figure 1, misclassification results appear in objects with similar color and texture.

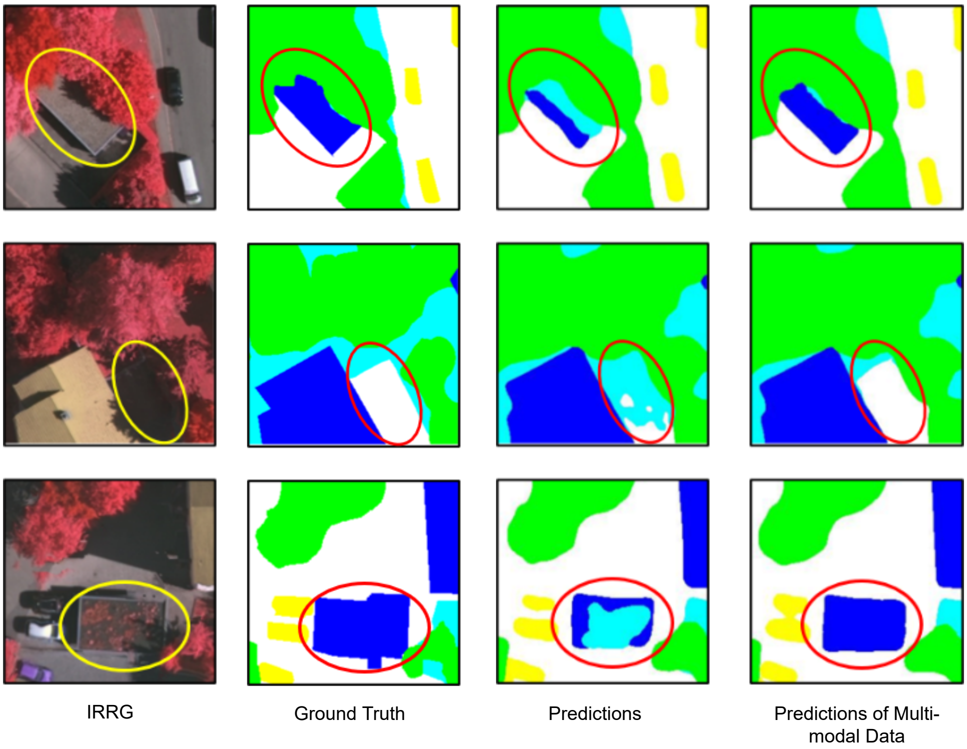


Figure 1. Predictions of objects with similar color and texture.

In addition to common RGB data, additional remote sensing data is also widely used for semantic segmentation, such as Synthetic Aperture Radar (SAR) images [14–17]. and Digital Surface Models (DSM) [4,18–20]. These multi-modal data usually refer to a collection of multi-channel data collected by different sensors and can reflect different characteristics of the objects. In order to improve the performance of semantic segmentation based on single modal data, we utilize the rich complementary features of multi-modal remote sensing images including IRRG (Near Infrared-Red-Green) and DSM. Semantic segmentation based on multi-modal images aims to use the complementary characteristics of different modalities to improve classification accuracy while reducing the influence of inherent noise in single-modal data and improving the generalization performance in complex application scenes.

There are two main challenges in the utilization of multi-modal images. One is how to effectively extract multi-modal features. Typical methods directly concatenate multiple channels as the input or integrate the prediction results of different models corresponding to multi-modal images. However, indiscriminately fuse the features at data-level or prediction-level does not fully extract the complementary characteristics of multi-modality, and even introduces redundant features and aggravates the impact of image noise. In this work, we extract multi-modal information by feature-level fusion method. Furthermore, we introduce a cross-modal feature recalibration module (CFR) to improve the quality of multi-modal representation by transforming the features of corresponding modalities first and then recalibrate and aggregate the informative features as the fusion feature.

The other challenge of semantic segmentation of multi modal images is the efficiency of the algorithm. The semantic segmentation task is originally a pixel-level dense classification task. In addition, the segmentation model designed for multi-modal data introduces more modalities and corresponding feature extraction networks. The huge model scale and extensive computational burden limit the application in edge devices and scenarios with real-time requirements. Considering this, in quest of an efficient model that strikes a balance between speed and accuracy, we redesign a lightweight cross-scale semantic aggregation module proposed in our previous work. Besides, we introduce a multi-level knowledge distillation (KD) strategy to obtain an accurate and compact network for dense prediction.

Our contributions are the following:

- We propose a multi-modal semantic segmentation network, C3Net, which takes into account both efficiency and accuracy. Compared with previous works, our network can be up to 15× fewer FLOPs with comparable accuracy.
- In order to extract complementary features of different modalities, a cross-modal feature recalibration module (CFR) is designed to aggregate information of multi-modality and form discriminative representations for classification.
- A lightweight cross-scale semantic aggregation module (CSA) is introduced for adaptive multi-scale semantic context information propagation. Compared with the previous version, it can greatly reduce parameters and running time without loss of accuracy.
- A multi-level knowledge distillation strategy is utilized on a variety of compact backbones, which further reduces the model size and improves the segmentation performance of lightweight architectures.

The remainder of this paper is organized as follows—Section 2 introduces the background of the semantic segmentation methods based on RGB images and multi-modal images, lightweight network design, and knowledge distillation. Section 3 describes the proposed C3Net in details. Section 4 illustrates the experimental settings and analyses the experimental results, followed the conclusions in Section 5.

2. Related Work

In this section, we first briefly review the development of deep learning methods in the field of semantic segmentation, including both the segmentation tasks of common RGB images and remote sensing multi-modal images. We then review several mainstream handcrafted networks and corresponding lightweight convolutional modules. Finally, we review the related work in the field of knowledge distillation, which is also an active direction in the field of model compression and acceleration.

2.1. Semantic Segmentation

Research on semantic segmentation has received a tremendous performance boost since the design of the fully convolutional network [21]. The two major factors restricting the improvement of semantic segmentation accuracy are the extraction of spatial information and semantic information. In order to recover the spatial detail information, the low-level features are usually introduced in the form of skip-connection in the network based on the encoder-decoder architecture [22–25]. The methods based on skip-connection can directly introduce spatial information without many additional parameters. However, simply merging low-level features runs the risk of introducing redundant features. Especially the spatial information of multi-modal features needs to be extracted selectively to prevent the introduction of redundant features and noise. Previous work [4] utilize artificial prior knowledge to concatenate low-level features, which still ignore the complementary features of multi-modal data. In order to make full use of the complementary features of multi-modal data, we proposed a cross-modal feature recalibration module to extract informative spatial information.

The key to learning semantic information is to extract appropriate receptive fields. The method based on pooling operation [26,27] can learn the scene-level global context, but result in the loss of spatial information. The method based on attention mechanism use "attention" descriptors to focus feature learning on global semantic information [28–30]. It can extract the global receptive field but increase the computational burden. Dilated convolution [31] is another method that can enlarge the receptive fields without significantly increasing the calculations [27,32]. However, the cascade of dilated convolution may lead to grid effect, which result in the reduction of the effective features. Our previous work [33] densely connect dilated convolutions to eliminate the grid effect and obtain multi-scale context information. However, it introduces a large number of parameters resulting in low computational efficiency. In order to solve the problems above-mentioned in the process of semantic information extraction, we design a lightweight cross-scale semantic aggregation module. The proposed module has a better performance for multi-scale objects with fewer parameters.

With the development of remote sensing imaging technology and deep learning, recent work on the semantic segmentation of multi-modal remote sensing images has made considerable progress. Rifcn [34] simply combine near-infrared, red, green (IRRG) spectrum and DSM as the fusion input to the network. However, it does not fully exploit the relationship between heterogeneous features and may introduce redundant features in training. Some works [18,35] based on parallel branch architecture use multi-backbones to process the multi-modal data separately. Such parallel branch fusion architecture can effectively extract informative features; however, the huge model scale brings a large number of parameters and reduce model inference speed. In [19], backend processing is utilized to fuse the DSM feature to extract complementary information and refine the segmentation details. However, a multi-stage training method is difficult to guarantee the global optimal solution. In this work, we aim to design a lightweight and end-to-end architecture which can effectively extract cross-modal complementary information and achieve more accurate segmentation results.

2.2. *Lightweight Networks*

In recent years, designing handcrafted convolutional neural network architecture for the optimal trade-off between accuracy and efficiency has been developed into an active research field. SqueezeNet [36] utilizes 1 × 1 convolutions in squeeze and expand module to reduce the number of parameters. Group convolution is proposed in [37] for parallelization to accelerate the training process. Many state-of-the-art networks [38–40] integrate group convolution into model architecture to reduce model scale and parameters. Shufflenet introduces channel shuffle operation to enhance the flow of information between channels and improve the performance of group convolution. Furthermore, depthwise separable convolution [41] is introduced as an efficient replacement for the traditional convolution layers and widely used in [42–44]. An inverted residual bottlenecks module is proposed [45] as an extension of depthwise separable convolution and used in state-of-the-art efficient architectures [46–48]. In this work, we choose a variety of lightweight networks as the backbone network to explore the optimal knowledge distillation strategy and the optimal architecture of the student network. More quantitatively, comparisons can be found in Section 4.3.3. Besides, we also design a plug-and-play module, namely lightweight convolution group (LCG). As the basic architecture of the cross-scale semantic aggregation module, LCG can drastically reduce model parameters while retaining multi-scale semantic information.

2.3. *Knowledge Distillation*

Knowledge Distillation aims to compress the model and improve the accuracy by transferring knowledge from a teacher network (cumbersome model) to a student network (compact model). It has been widely used in image classification and has been proven to achieve good performance improvements [49–51]. The research of [49] is the pioneering work that exploits intra-class similarity and inner-class diversity from teacher network as the useful knowledge to guide the learning processing of the student network. Following [49], many other methods are proposed for knowledge distillation. Fitnet [50] considers the intermediate features as the informative knowledge for the first time, which directly aligns feature maps to learn intermediate representation. Subsequently, the attention map aggregated from the response maps is also regarded as a kind of knowledge. Attention transfer [51], which mimics the attention map between teacher network and student network, is related to our approach. However, the previous methods usually focus on the tasks of image classification and are only applied for single-modal data. In this work, we also explore knowledge distillation approaches in multi-modal semantic segmentation network. A multi-level knowledge distillation strategy based on class knowledge transfer and attention knowledge transfer is proposed to design an efficient semantic segmentation framework for multi-modal image processing.

3. Method

Following the previous work [4], proposed network architecture adopts an encoder-decoder structure, as shown in the Figure 2. In this work, ResNet101 and ResNet50 [52] are selected as the feature extraction networks of multispectral data and DSM data, respectively. In the encoder, the parallel multi-modal feature extraction network can be replaced with any CNN network. The CSA is an abbreviation for a lightweight cross-scale semantic aggregation module, which can efficiently obtain multi-scale contextual information. The CFR represents a cross-modal feature recalibration module, which is designed for multi-modal feature fusion. In order to simplify the model structure, the architecture of decoder follows the design in [25]. In this section, we will introduce CFR module, CSA module, and a multi-modal knowledge distillation strategy respectively.

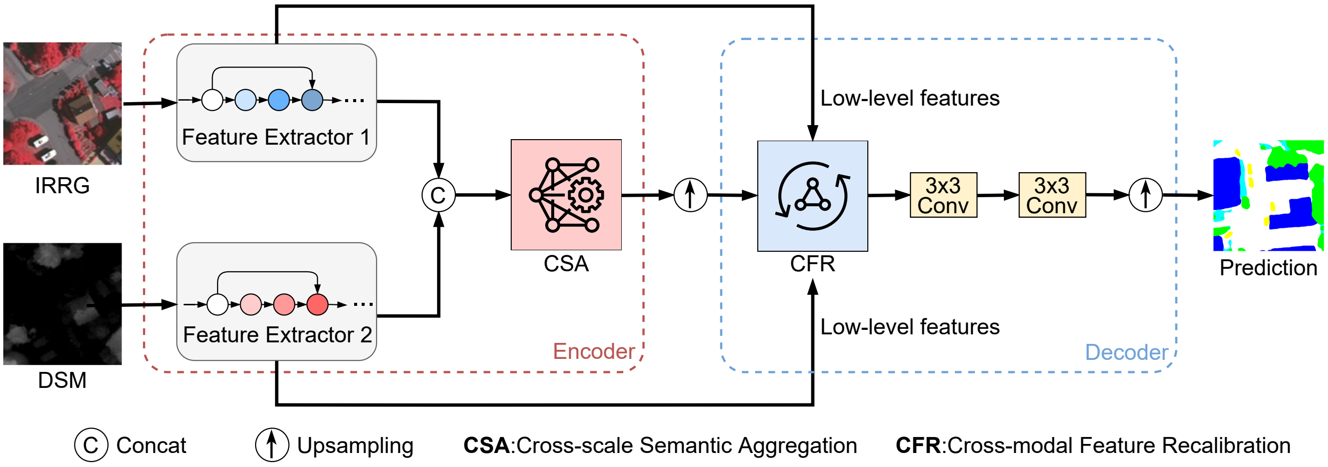


Figure 2. Framework overview of the proposed method.

3.1. Cross-Modal Feature Recalibration Module (CFR)

The CFR module is designed to jointly learn the multi-modal features while reducing the influence of inherent noise of different modalities. The simple concatenation of multi-modal feature [34] cannot learn cross-modal complementary characteristics, and even the inherent noise and redundant feature will reduce the ability of feature expression [4]. In order to extract informative features of different modalities, the CFR module is designed with feature aggregation, feature recalibration, and feature reconstruction operation. The architecture of CFR is shown as Figure 3. We denote $\boldsymbol{X}_1 \in \mathbb{R}^{\mathrm{C} \times \mathrm{H} \times \mathrm{W}}$ and $\boldsymbol{X}_2 \in \mathbb{R}^{\mathrm{C} \times \mathrm{H} \times \mathrm{W}}$ as the input feature maps of two branches respectively.

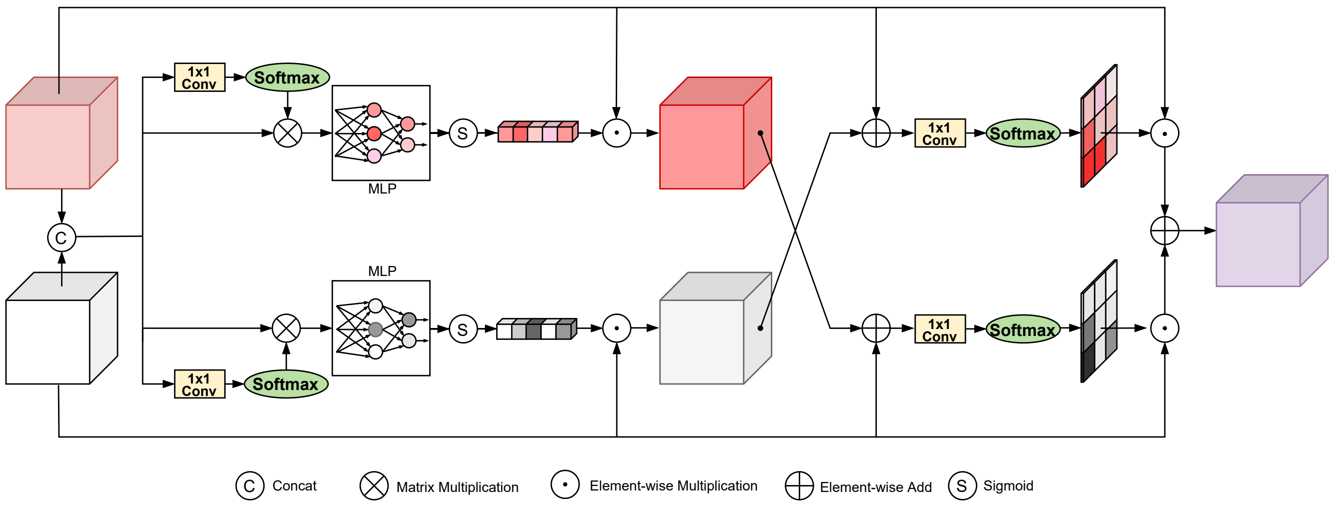


Figure 3. The architecture of cross-modal feature recalibration module.

First, fusion feature map $\boldsymbol{X}_c$ is calculated as

$$\boldsymbol{X}_c = [\boldsymbol{X}_1, \boldsymbol{X}_2], \tag{1}$$

where [•] denotes concatenate operation along the channel dimension. To ensure the validity of the features, we utilize the attention mechanism [28,53] to transform the feature maps into the embedding space and regard it as an attention vector. Attention vector represents the high confident activations in an original feature map, which can effectively focus the feature learning process on most informative features and suppress the importance of noisy features. We take one branch as an example, the attention vector is calculated as

$$\boldsymbol{W}_1 = \sigma(F_M(F_G(\boldsymbol{X}_c))), \tag{2}$$

where F_M denotes a common MLP network which includes two 1×1 convolutional layers and σ denotes sigmoid function. F_G denotes long-range context modeling operation. By doing this, the global semantic information is obtained for feature aggregation.

$$F_G(\boldsymbol{X}_c) = \sum_{i=1}^{\mathrm{H}\times\mathrm{W}} \frac{exp\left(W_C \boldsymbol{X}_C^i\right)}{\sum_{j=1}^{\mathrm{H}\times\mathrm{W}} exp\left(W_C \boldsymbol{X}_C^j\right)} \cdot W_C \boldsymbol{X}_C^i, \tag{3}$$

where $\boldsymbol{X}_C^i$ denotes the feature vector in position i and W_C denotes linear transformation matrix. After calculating the attention vector, the aggregate feature can be formulated as

$$\boldsymbol{X}_1' = \boldsymbol{X}_1 \odot \boldsymbol{W}_1, \tag{4}$$

where $\odot$ denotes element-wise multiplication. The aggregate features can be regarded as the initial calibration of the features from the same modality. Compared with original feature maps, aggregate feature maps contain more informative features and less noisy features. The feature recalibration is designed to exchange and recalibrate the features of different modalities.

$$\boldsymbol{X}_2'' = \boldsymbol{X}_1' + \boldsymbol{X}_2. \tag{5}$$

The process of recalibration completes the cross-modal information interaction. More related discussions can be found in Section 4.3.1.

The previous works usually directly introduce low-level features from different modalities to the decoder for rich spatial information. However, the spatial information of different modalities is usually not aligned well, which seriously affects the feature reconstruction. Inspired by [54], a spatial-wise gate mechanism is designed to obtain the multi-modal fusion feature. A 1×1 convolutional layer is utilized for embedding space mapping and a SoftMax function is applied to obtain gate weights:

$$\boldsymbol{G}_1 = \sum_{i=1}^{C} \frac{exp(W_S \boldsymbol{X}''_{1i})}{\sum_{j=1}^{C} exp\left(W_S \boldsymbol{X}''_{1j}\right)} \cdot W_S \boldsymbol{X}''_{1i}, \tag{6}$$

where $\boldsymbol{X}''_{1i}$ denotes the feature vector in channel i from modality 1 and W_S denotes linear transformation matrix. The final fusion feature is the summation of the weighted features.

$$\boldsymbol{X}_o = \sum_{i=1}^{m} \boldsymbol{X}_i \odot \boldsymbol{G}_i, \tag{7}$$

where m denotes the total number of modalities, in our case, $m = 2$. $\boldsymbol{X}_o$ is regarded as the low-level feature in the decoder for up-sampling and spatial information recoverying.

3.2. Cross-Scale Semantic Aggregation Module (CSA)

Semantic contextual information is critical to the pixel-level classification. In recent years, parallel spatial pyramid pooling structures [25,27] are widely used to obtain multi-scale receptive fields. However, it is still difficult to obtain a large and dense receptive field in remote sensing scenes. To remedy this problem, the CSA module is designed to capture dense contextual information in a cross-scale form through multiple sets of dilated convolutions. Besides, the CSA module is composed of a lightweight convolution group (LGC), which can reduce the computational burden of multi-modal high-resolution images. The architecture of CSA is shown as Figure 4. CSA contains six contextual information extraction branches. We utilize five LCGs with different dilated rates to capture multiple receptive fields and a multi-shape pooling (MSP) to obtain global semantic information. We denote $\boldsymbol{X}_s$ and $\boldsymbol{Y}_s$ as the input feature map and output feature map of CSA module respectively. Above operations can be formulated as

$$\boldsymbol{Y}_s = [\boldsymbol{Y}_0, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_l], \tag{8}$$

where $[\bullet]$ denotes concatenate operation along the channel dimension. $\boldsymbol{Y}_i$ denotes different branches of CSA module.

$$\boldsymbol{Y}_i = \begin{cases} \boldsymbol{X}_s, & i = 0 \\ F_M([\boldsymbol{Y}_0, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{l-1}]), & i = l \\ F_L^{k,d}([\boldsymbol{Y}_0, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{i-1}]), & others, \end{cases} \tag{9}$$

where F_M denotes the multi-shape pooling operation and $F_L^{k,d}$ denotes the lightweight convolution group. The architecture of the lightweight convolution group is shown as Figure 4. LCG consists of two 1×1 convolution layers for channel reduction and expansion respectively.Besides, a 3×3 convolution layer with different dilated rate r is utilized for semantic information encoding and an another 3×3 convolution layer is added to encode more spatial contextual information. The architecture of multi-shape pooling is shown as Figure 4. MSP utilizes 1×1, 3×1, and 1×3 convolutional layers to model global semantic. In particular, it encodes the globally horizontal and vertical information to capture complex objects with multiple shapes in remote sensing scenes. Assuming that the input of the module is $\boldsymbol{X}_{in}$, the above operations can be formulated as

$$F_M(\boldsymbol{X}_{in}) = \sigma(W_{1\times1}(\sum_{i=1}^{3} W_i(P_i(\boldsymbol{X}_{in})))) \odot \boldsymbol{X}_{in}, \tag{10}$$

where σ denotes sigmoid function, $W_{1\times1}$ denotes convolution layer with 1×1 kernel, W_i, P_i denote the convolution and pooling operations corresponding to the three branches.

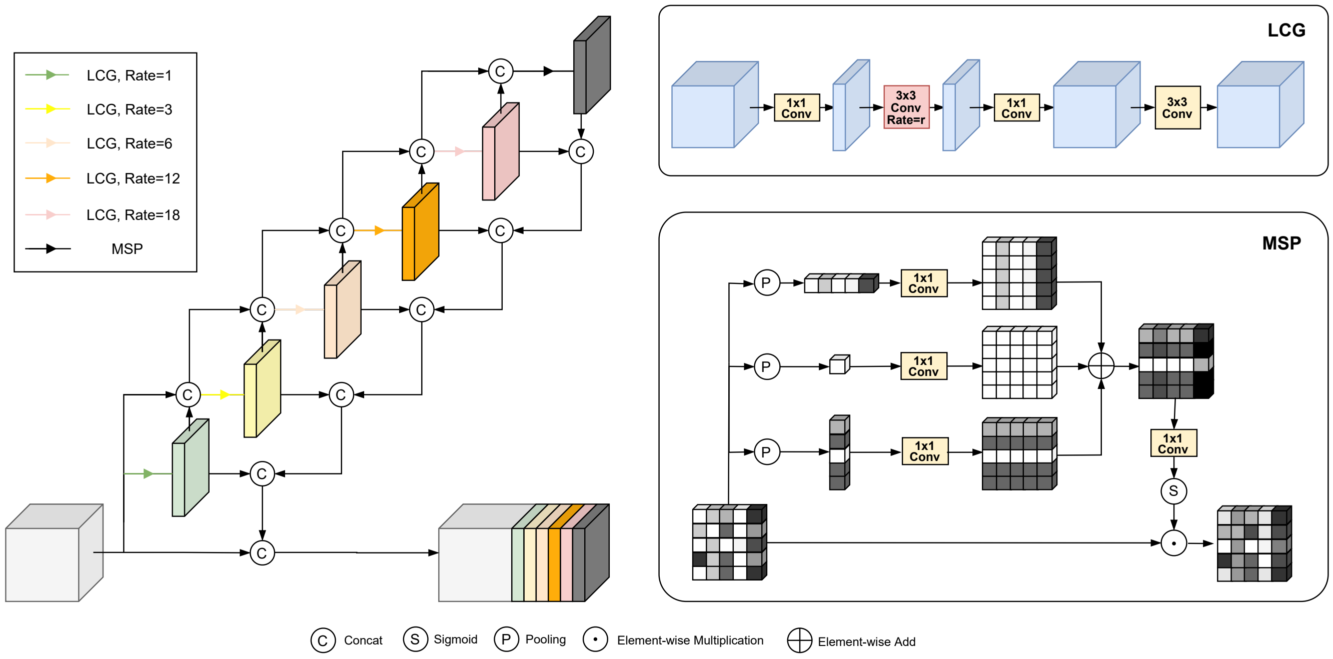


Figure 4. The architecture of cross-scale semantic aggregation module.

3.3. *Multi-Level Knowledge Distillation Strategy*

In order to obtain a compact model, knowledge distillation has been widely used in CNN designed for natural RGB images, however, there are few related research on the knowledge distillation strategy of multi-modal images. In this section, we explore knowledge distillation strategies for networks that process multi-modal images and introduce a multi-level knowledge distillation strategy. As shown in Figure 5, the proposed knowledge distillation strategy includes class knowledge transfer and attention knowledge transfer. Note that only a single branch of the multi-modal network is shown as an example.

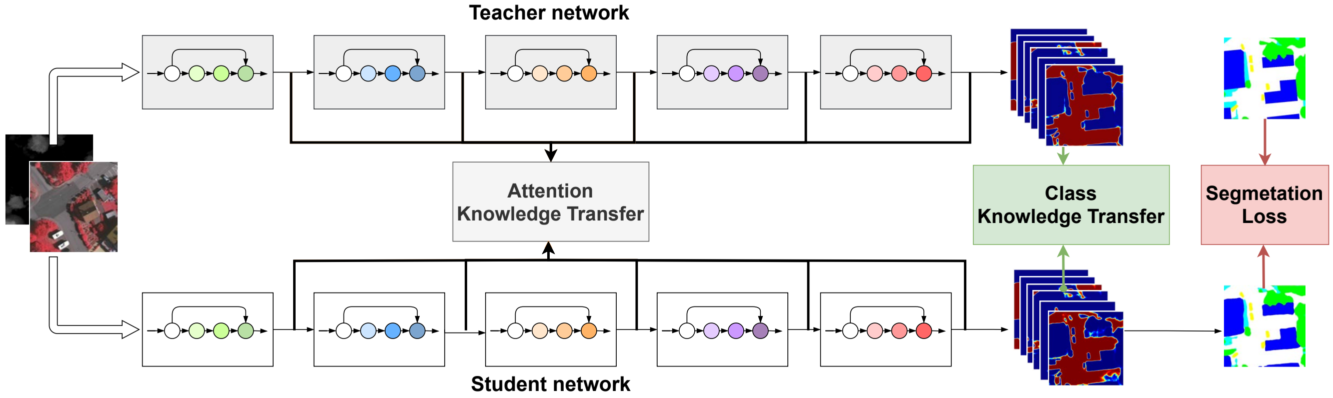


Figure 5. Illustration of our knowledge distillation strategy.

3.3.1. Class Knowledge Transfer

Semantic segmentation can be regarded as a pixel-wise classification task. Therefore, class probabilities produced from the teacher network can be utilized to transfer the informative knowledge to the student network. We follow [49] and regard the class probabilities as a soft target which provides much more information than hard targets. We denote $\mathcal{S}$ and $\mathcal{T}$ as the compact student network and pre-trained teacher network. The loss function of the class knowledge transferring process is as follows

$$\ell_c = \frac{1}{H \times W} \sum_{i \in \mathbb{R}} KL(\boldsymbol{p}_{\mathcal{S},i}, \boldsymbol{p}_{\mathcal{T},i}), \mathbb{R} \in \{1, 2, \ldots, H \times W\}, \tag{11}$$

where $H \times W$ denotes the spatial size of softmax layer output. $KL(\cdot, \cdot)$ is the Kullback-Leibler divergence. $\boldsymbol{p}_{\mathcal{S},i}$ and $\boldsymbol{p}_{\mathcal{T},i}$ denote the soft target of the i-th pixel produced from student and teacher network, which can be formulate as

$$\boldsymbol{p}_j = \frac{exp(z_j/T)}{\sum_{m=1}^{C} exp(z_m/T)}, \tag{12}$$

where j denotes the j-th class of C classes, z denotes the output of the softmax layer. T denotes the temperature factor that is utilized to control the probability distribution over classes.

3.3.2. Attention Knowledge Transfer

Class knowledge transferring performs knowledge distillation for each pixel separately, which may lead to sub-optimal result for the dense prediction task. The knowledge extracted from deep features only focuses on the high-level information and neglect the detailed spatial knowledge from low-level features. Besides, the activation of different modal features varies greatly. It is necessary to transfer the knowledge for multi-modal features. Due to the above fact, we explore a multi-level knowledge distillation strategy based on attention knowledge, as shown in Figure 5.

$$\ell_A = \sum_{m=1}^{M} \sum_{l \epsilon \mathcal{I}} \left|\left| \frac{\sum_{k=1}^{C} |A^l_{\mathcal{S},k}|^2}{||\sum_{k=1}^{C} |A^l_{\mathcal{S},k}|^2||_2} - \frac{\sum_{k=1}^{C} |A^l_{\mathcal{T},k}|^2}{||\sum_{k=1}^{C} |A^l_{\mathcal{T},k}|^2||_2} \right|\right|_2, \tag{13}$$

where m denotes the m-th modality in the M modalities, l denotes the j-th pair of teacher attention maps $A^l_{\mathcal{T},k}$ and student attention maps $A^l_{\mathcal{S},k}$, k denotes the k-th class in the C classes. $||\cdot||_2$ denotes l_2-normalization which is utilized during the knowledge transferring.

The training process of our method is presented in Algorithm 1. Given the pre-trained teacher network, the parameters of the teacher network are kept frozen during the training process. The student network is supervised by three losses: common cross entropy loss ℓ_s with ground truth, class knowledge transferring loss ℓ_c in Equation (11) and attention transferring loss ℓ_A in Equation (13). Scaling factor α (10) and β (1000) are utilized to make these loss value range comparable.

Algorithm 1 Training Process of Proposed Method.

```
Stage 1: Training teacher network $\mathcal{T}$
Input: Images $\mathcal{D}$,Ground truths $\mathcal{G}$
  $W_{\mathcal{T}} = argmin_{W_{\mathcal{T}}} \ell_s(\mathcal{D}, \mathcal{G})$
Stage 2: Training student network $\mathcal{S}$
Input: Images $\mathcal{D}$,Ground truths $\mathcal{G}$, Class-Knowledge $\boldsymbol{p}_{\mathcal{T}}$, Attetion-knowledge $A_{\mathcal{T}}$
  $W_{\mathcal{S}} = argmin_{W_{\mathcal{S}}} \ell_s(\mathcal{D}, \mathcal{G}) + \alpha \ell_c(\boldsymbol{p}_{\mathcal{T}}) + \beta \ell_A(A_{\mathcal{T}})$.
```

4. Experiments

4.1. Dataset

4.1.1. ISPRS Vaihingen

The Vaihingen Dataset contains 33 very high-resolution true orthophoto (TOP) tiles. Each tile contains approximately 2100 × 2100 pixels with a resolution of 9cm/pixel. Tiles consist of Infrared-Red-Green (IRRG) images. Besides, corresponding Digital Surface Models (DSM) and normalized Digital Surface models (nDSM) are also provided. The

Vaihingen Dataset provides 16 labeled tiles for training. Following other works, the training set consists of 11 tiles (region numbers 1, 3, 5, 7, 13, 17, 21, 23, 26, 32, 37) and the validation set consists of five tiles (region numbers 11,15, 28, 30, 34). Limited by computing resources, we crop the large tiles into patches of 513 × 513 pixels using an overlapped (50%) sliding window. Finally, we get 696 patches for training and 297 patches for validation. Note that in the testing phase, we use all patches for training.

4.1.2. ISPRS Potsdam

The Potsdam Dataset contains 38 very high true orthophoto (TOP) tiles. Each tile contains 6000 × 6000 pixels with a resolution of 5cm/pixel. Tiles consist of Infrared-Red-Green-Blue (IRRGB) data, DSM data and nDSM data. The Potsdam Dataset consists of 24 labeled tiles for training. In the following experiments, 6 tiles (region numbers 2_12, 3_12, 4_12, 5_12, 6_12, 7_12) are removed from the training set as the validation set. During the training process, the large tiles are cropped into 9522 patches for training and 3174 patches for validation according to the method mentioned above. Note that in the testing phase, we use all patches for training.

4.1.3. Dataset Augmentation

Standard data augmentation is applied during the training process, including random flipping, random scaling (from 0.5 to 2), and random rotating (between −10 and 10 degrees).

4.1.4. Evaluation

According to the benchmark rules, overall accuracy (OA) and F1 score are used to quantitatively evaluate the model performance. OA is the normalization of the trace of pixel-based confusion matrices. $F1$ score is calculated as follows:

$$F1 = 2 * \frac{PR}{P + R} \tag{14}$$

where P, R denotes precision and recall respectively.

$$P = \frac{TP}{TP + FP} \tag{15}$$

$$R = \frac{TP}{TP + FN}, \tag{16}$$

where TP denotes the number of true positive, FP denotes the number of false positive and FN denotes the number of false negative. All these metrics can be calculated by a pixel-based accumulated confusion matrix. The number of float-point operations (FLOPs) is also applied to investigate the computation complexity.

4.2. Implementation DETAILS

During the training stage, proposed network utilizes stochastic gradient descent (SGD) with the momentum (0.9) and the weight decay (0.0005) as the optimization strategy and be trained for 100 epochs. The basic learning rate is set to 0.01 with the "poly" learning rate policy and the power is set to 0.9. Our experiments are implemented on a Tesla P100 GPU while the batch-size is set to 4. For a fair comparison, we make the ablation study with 513 × 513 patches.

4.3. Results and Discussion

In this section, we report the effects of cross-modal feature recalibration module, cross-scale semantic aggregation module and multi-level knowledge distillation strategy and analyze these proposed methods from both qualitative and quantitative perspective.

4.3.1. Effects of CFR

In order to further reveal the role of CFR, we visualize the response of DSM features in Figure 6. We average the channel dimension and convert it to one dimension for visualization. As shown in the second column of Figure 6, due to the imaging quality and the inherent characteristics of the data, DSM images lack spatial detail information. Comparing the fifth and sixth columns in Figure 6, the DSM feature contains more significant detailed information after CFR feature calibration. In particular, there is a higher activation response at the edges between different classes and the outline of the object is more precise. In addition, the third row of Figure 6 shows that the CFR module makes the feature map shows a good response to the car.

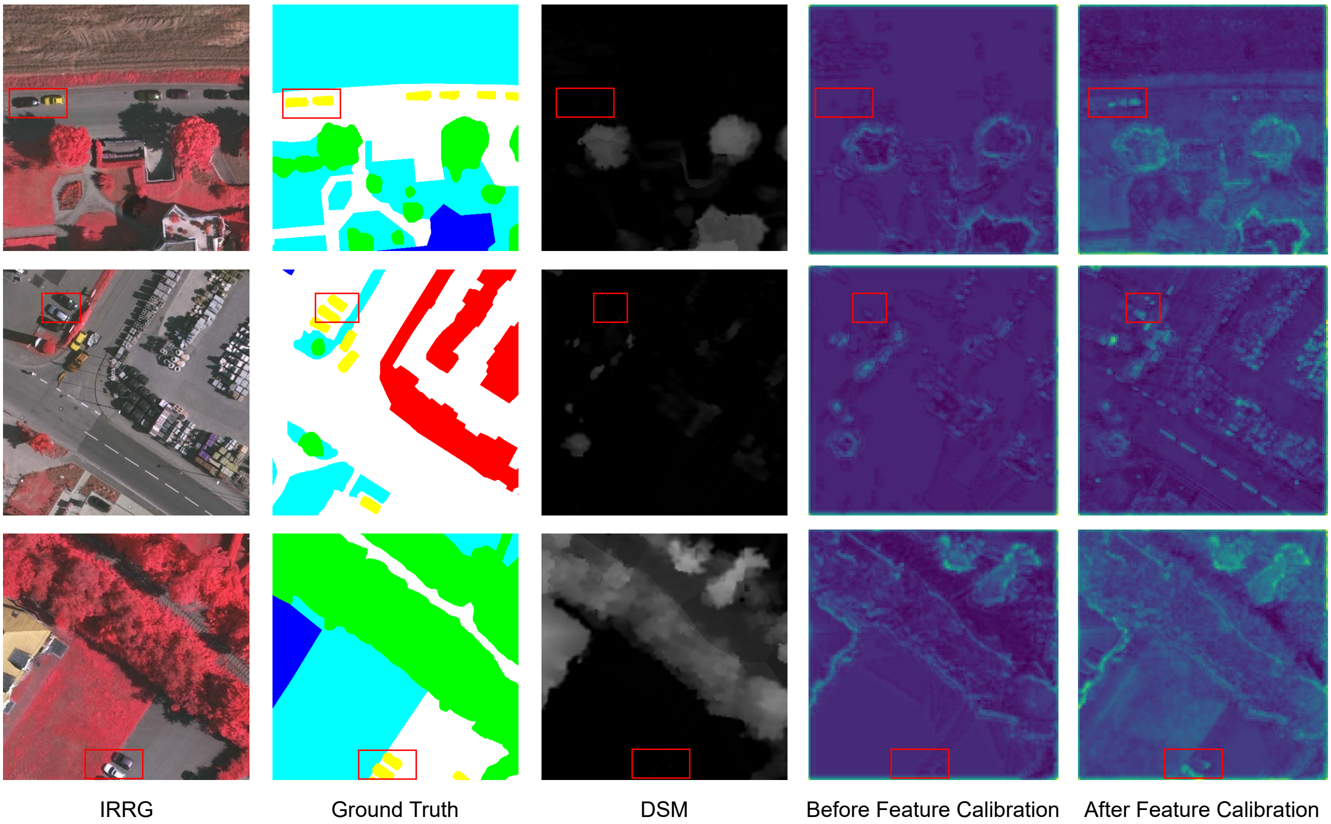


Figure 6. Visualization of Digital Surface Model (DSM) features before and after Feature Calibration operation.

Table 1 shows the results on ISPRS Vaihingen testing dataset with different multi-modal fusion methods based on the baseline architecture. The results of IRRG based methods are also listed for reference in the first row. As shown in the second row of Table 1, due to the noisy and redundant features of DSM modality, directly concatenate the multi-modal images as the input leads to worse performance in car. This simple multi-modal fusion mechanism cannot explore the complementarity of different modalities to boost performance. As shown in the third row of Table 1, the design of the baseline is inspired by previous work [4] which has achieved good segmentation performance. Based on the above observation, the baseline only introduces the IRRG features to the decoder. In order to utilize the strength of multi-modal information, CFR is designed to filter the useless information and fusion the cross-modal features in an effective way. Thus, CFR still gains 0.47% OA improvement compared with baseline. In particular, CFR can obtain a higher F1 value of car class. CFR uses the rich semantic information of DSM modality in the decoder to reconstruct the low-level features and achieve better segmentation results on small targets that are likely to lose spatial detail information while encoding.

Table 1. Comparison of F1 score and overall accuracy on ISPRS Vaihingen dataset.

Methods	Impervious Surface	Building	Low Vegetation	Tree	Car	OA
Baseline (IRRG)	91.84	95.27	82.47	88.99	76.39	89.83
Baseline (Directly Concatenate)	91.15	95.02	84.12	89.60	70.13	89.85
Baseline	92.36	95.83	84.06	89.69	82.59	90.61
Baseline + CFR	92.79	96.01	84.99	90.06	83.29	91.08

In summary, the cross-modal feature recalibration module could aggregate information of multiple modalities and form discriminative representations for segmentation. Furthermore, the CFR module boosts the segmentation performances while only adds a small parameter and computation overhead. It can also be regarded as a plug-and-play module to effectively fuse multi-modal images.

4.3.2. Effects of CSA

Contextual semantic information plays an important role in classification, especially for dense prediction tasks. Existing methods usually only contain limited scales of receptive field. It is difficult to deal with remote sensing application scenarios that include multiple complex objects and large scope. The CSA module uses multiple dilated convolutions to obtain multi-scale receptive fields in the form of dense connections. In the experiment, we empirically choose 1, 3, 6, 12, 18 as the dilation rates. As shown in Figure 7, SPP [27] contains four scales of receptive fields, ASPP [25] contains five scales of receptive fields, while the proposed CAS and MCA contains 25 types of receptive fields. The CAS module can not only extract multi-scale contextual information but also obtain global semantic with multi-shape pooling.

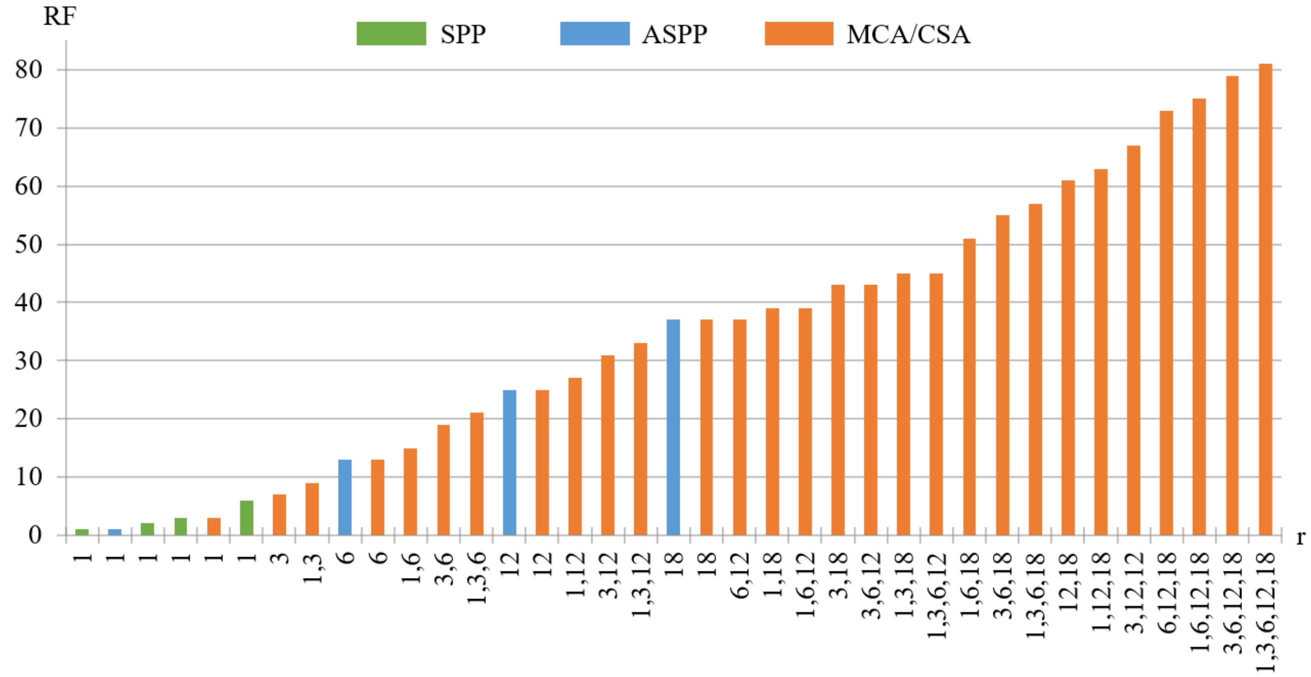


Figure 7. Comparison of the receptive fields corresponding to SPP, ASPP, MCA and proposed cross-scale semantic aggregation module (CSA) with different combinations of dilated convolutions. Note that only the receptive fields with different combinations of dilated convolutions are listed in the bar graph.

We examine the role of the cross-scale semantic aggregation module on ISPRS Vaihingen testing dataset. In order to verify the effectiveness of CSA, we re-implement several mainstream semantic context extraction modules based on the baseline network. Table 2 shows the results of the baseline containing different semantic context extraction modules. The performance of the network without the semantic aggregation module is also shown in the first row of Table 2 for comparation. Compared with ASPP, SPP, and MCA, the proposed CSA gain 91.15% OA improvement based on the same baseline network. CSA and MCA have the same receptive field scale, but CSA utilizes a lightweight convolution group instead of traditional convolution, which greatly reduces the number of parameters of the semantic extraction module. As shown in Table 2, compared to MCA, CSA can reduce the amount of parameters by 55.72% while slightly improving performance by 0.12%.

Table 2. Comparison of several semantic extraction modules' efficiency on ISPRS Vaihingen dataset.

Module	Parameters	GFLOPs	Overall Accuracy
-	0	0	90.36
ASPP	30,737,920	8.582	90.61
SPP	68,157,696	1.201	90.63
MCA	49,613,312	28.08	91.03
CSA	21,969,888	2.327	91.15

4.3.3. Effects of Multi-Level Knowledge Distillation Strategy

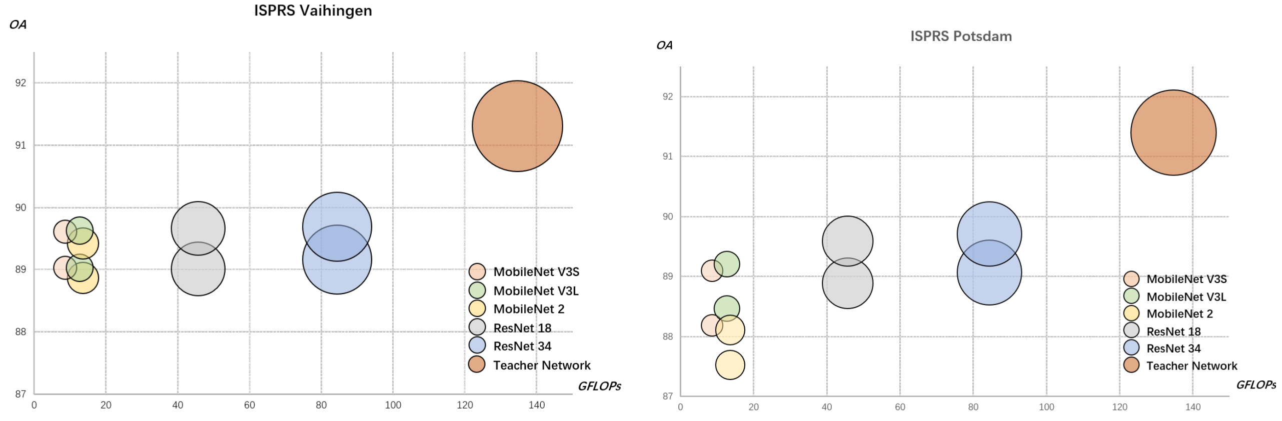
The multi-level knowledge distillation strategy contains two types of distillation methods: class knowledge transfer and attention knowledge transfer. We conduct ablation experiments on these two distillation methods respectively. In the following experiments, we choose ResNet101 as the IRRG feature extraction network and ResNet50 as the DSM feature extraction network for the teacher network. As for the student network, in order to compress the model parameters as much as possible and obtain the relevant knowledge of the corresponding feature pairs, we choose ResNet18 as the feature extraction network for two modalities.

In order to verify the effect of feature maps at different levels in attention knowledge transfer strategy, we distillation the attention knowledge at multiple levels of feature map pairs. As shown in Table 3, the feature map pair from level 5 is distilled as the high-level attention knowledge. Then the feature pair attention knowledge distilled from level 4, level 3, level 2, and layer 1 are introduced in sequence. Here, feature map pairs from different levels represent the output feature map pairs from corresponding different stages in Resnet. As expected, the effect of distillation gradually increased with the introduction of low-level attention maps. This also indirectly proves that the detailed knowledge contained in the low-level attention maps could help to improve the performance of dense prediction tasks.

Table 3. Ablations for the different level feature maps on ISPRS Vaihingen dataset.

Level 5	Level 4	Level 3	Level 2	Level 1	Overall Accuracy
					89.01
✓					89.42
✓	✓				89.51
✓	✓	✓			89.56
✓	✓	✓	✓		89.57
✓	✓	✓	✓	✓	89.66

During the knowledge distillation training process, the model scale and data set scale have a great influence on information transferring. In order to demonstrate the effectiveness of the proposed multi-level knowledge distillation strategy, we ablate a variety of backbone models [45,47,52] on Vaihingen and Potsdam data sets. As shown in Figure 8, we compare the accuracy and efficiency of different models and quantify them with OA and GFLOPs in the bubble chart. Note that the area of the bubble represents the number of parameters. Experimental results on various backbones and data sets show that the proposed distillation strategy can effectively improve the segmentation accuracy of compact models. The results also show that the proposed distillation strategy can obtain greater performance gains on large-scale data sets, which implies that the proposed method has better generalization performance in big-data application scenarios.



(**a**) ISPRS Vaihingen (**b**) ISPRS Potsdam

Figure 8. Ablation study on the backbone.

4.3.4. Comparing with State-of-the-Arts

We report the results on the Vaihingen dataset in Figure 9 and Table 4. In order to demonstrate the effectiveness of the cross-scale semantic aggregation module (CSA), we also train C1Net (baseline with CSA) with IRRG data (CH3) to compare with other state-of-art methods. As shown in Table 4, C1Net shows superior segmentation results than the other methods. In multi-modal data experiments (CH4), the proposed C2Net (Baseline with CSA and CFR) ranks 1st both in mean F1 score and overall accuracy compared with all the other published works. Note that neither the test time augmentation nor the CRF for post-processing is used in the proposed methods. This is based on the consideration of fast inferencing and designing compact models. As shown in Figures 8 and 9, it is worth noting that the C3Net (backbone with MobileNet V3S) obtained by knowledge distillation can still achieve competitive segmentation accuracy when the amount of parameters is reduced by 93.6% and the inference speed is increased by 1.6 times. It turns out that C3Net could achieve a good balance between accuracy and efficiency.

Table 4. State-of-the-art comparison experiments on ISPRS Vaihingen test set.

	Methods	Impervious Surface	Building	Low Vegetation	Tree	Car	OA
CH3	FCN [55]	89.4	93.8	76.5	86.6	71.3	86.8
	SegNet [22]	90.2	94.1	77.4	87.4	77.3	87.6
	DANet [56]	94.1	90.8	81.4	87.4	75.9	88.6
	DeepLab V3p [25]	94.3	91.4	81.3	87.8	78.1	88.9
	PSPNet [27]	94.4	91.4	81.5	87.9	78.0	89.0
	Baseline(ResNet101)	91.8	95.3	82.5	89.0	76.4	89.8
	HRNet+ASP+SR [57]	94.7	92.9	83.2	88.9	84.3	90.1
	Baseline+CSA	92.3	95.6	83.6	89.3	80.7	90.4
CH4	FPL [58]	90.4	94.6	78.1	86.8	66.8	87.7
	HSN+OI+WBP [59]	91.3	94.9	79.8	88.3	83.6	88.8
	FCN [60]	90.5	93.7	83.4	89.2	72.6	89.1
	SegNet-RC [18]	91.0	94.5	84.4	89.9	77.8	89.8
	FCN+fusion+boundaries [20]	92.3	95.2	84.1	90.0	79.3	90.3
	FCN_MFS_DSMBackend [19]	92.3	95.8	83.8	89.6	86.4	90.6
	Baseline	92.4	95.8	84.1	89.7	82.6	90.6
	C2Net	93.0	96.1	85.4	90.3	85.4	91.3

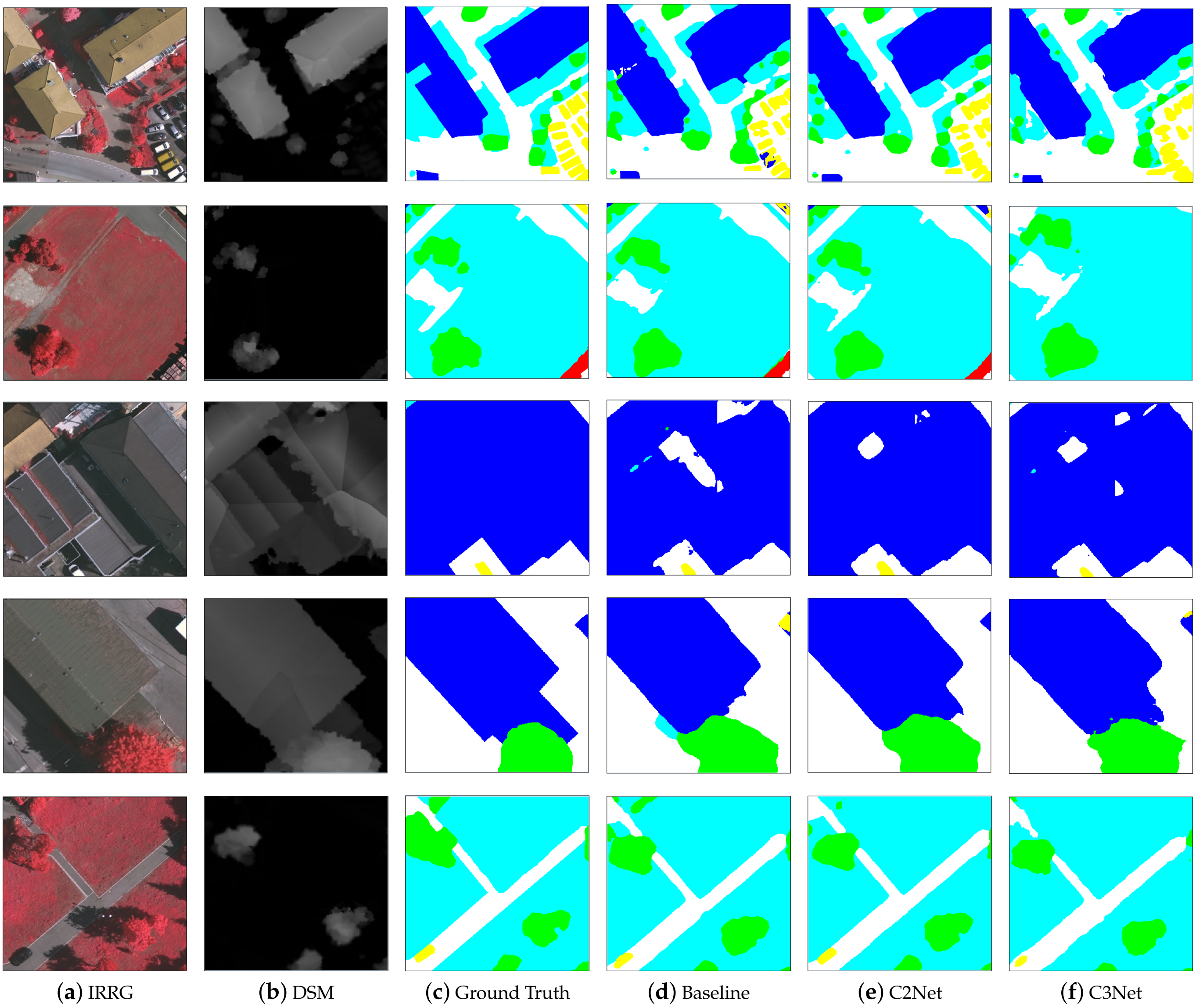


Figure 9. Semantic segmentation results on the ISPRS Vaihingen dataset.

5. Conclusions

In this paper, a novel semantic segmentation framework is presented for multi-modal remote sensing images. The major contribution of this work is to address two key challenges in the application of multi-modal images, *i.e.*, the effective joint representation of different modalities and the compact model for efficiently inferencing. A cross-modal feature recalibration module (CFR) is designed to recalibrate the noisy modalities and extract cross-modal complementary features. Besides, lightweight cross-scale semantic aggregation module (CSA) and multi-level knowledge distillation strategy are utilized to obtain a compact model while still remaining segmentation accuracy. The experimental results indicate that the proposed C2Net achieves state-of-the-art segmentation performance and the compact C3Net achieves 8.66 GFLOPs while keeping the performance levels almost intact. Note that the proposed networks are only applicable to IRRG and DSM data. In future work, we will introduce more informative data such as SAR and Light Detection and Ranging (LiDAR) data in semantic segmentation task to improve the generalization performance in complex remote sensing scenarios. In addition, the proposed lightweight model C3Net still takes up a lot of memory usage on an GPU platform during training process. The future works include lightweight feature extractor designing and knowledge distillation based on structured information of multi-scale objects, which aim to further compress the model for less running time while remaining segmentation accuracy.

Author Contributions: Z.C. conceived and designed the experiments; Z.C. and W.D. performed the experiments and analyzed the data; Z.C. wrote the paper; X.S.; X.L. and K.F. contributed materials; W.D. and K.F. and M.Y. supervised the study and reviewed this paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grant 41701508 and 61725105.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The Vaihingen and Potsdam dataset were provided by the International Society for Photogrammetry and Remote Sensing (ISPRS): https://www2.isprs.org/commissions/comm2/wg4/benchmark/semantic-labeling. The authors would like to thank all the colleagues for the fruitful discussions on semantic segmentation and knowledge distillation.

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

Abbreviations

The following abbreviations are used in this manuscript:

ISPRS	International Society for Photogrammetry and Remote Sensing
CNNs	Convolutional Neural Networks
SAR	Synthetic Aperture Radar
LiDAR	Light Detection and Ranging
CFR	Cross-modal feature recalibration module
KD	Knowledge distillation
OA	Overall accuracy
CSA	Cross-scale semantic aggregation module
FCN	Fully convolutional network
IRRG	Near Infrared-Red-Green
IRRGB	Near Infrared-Red-Green-Blue
LCG	lightweight convolution group
MSP	Multi-shape pooling
TOP	True orthophoto
DSM	Digital Surface models
nDSM	Normalized Digital Surface models
FLOPs	Float-point operations
GFLOPs	Giga Float-point operations

References

1. Fu, K.; Li, Y.; Sun, H.; Yang, X.; Xu, G.; Li, Y.; Sun, X. A Ship Rotation Detection Model in Remote Sensing Images Based on Feature Fusion Pyramid Network and Deep Reinforcement Learning. *Remote Sens.* **2018**, *10*, 1922. [CrossRef]
2. Yang, L.; Kun, F.; Hao, S.; Xian, S. An Aircraft Detection Framework Based on Reinforcement Learning and Convolutional Neural Networks in Remote Sensing Images. *Remote Sens.* **2018**, *10*, 243.
3. Chen, K.; Fu, K.; Yan, M.; Gao, X.; Sun, X. Semantic Segmentation of Aerial Images With Shuffling Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 173–177. [CrossRef]
4. Cao, Z.; Fu, K.; Lu, X.; Diao, W.; Sun, H.; Yan, M.; Yu, H.; Sun, X. End-to-End DSM Fusion Networks for Semantic Segmentation in High-Resolution Aerial Images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1766–1770. [CrossRef]
5. Sun, X.; Liu, Y.; Yan, Z.; Wang, P.; Diao, W.; Fu, K. SRAF-Net: Shape Robust Anchor-Free Network for Garbage Dumps in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2020**, 1–15. [CrossRef]
6. Sun, X.; Shi, A.; Huang, H.; Mayer, H. BAS^{4}Net: Boundary-Aware Semi-Supervised Semantic Segmentation Network for Very High Resolution Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 5398–5413. [CrossRef]

7. Wang, P.; Sun, X.; Diao, W.; Fu, K. FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 3377–3390. [CrossRef]
8. Fu, K.; Chang, Z.; Zhang, Y.; Xu, G.; Zhang, K.; Sun, X. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *161*, 294–308. [CrossRef]
9. Fu, K.; Chang, Z.; Zhang, Y.; Sun, X. Point-Based Estimator for Arbitrary-Oriented Object Detection in Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2020**, 1–18. [CrossRef]
10. Liu, J.; Chen, K.; Xu, G.; Li, H.; Yan, M.; Diao, W.; Sun, X. Semi-Supervised Change Detection Based on Graphs with Generative Adversarial Networks. In Proceedings of the 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 74–77.
11. Ma, H.; Liu, Y.; Ren, Y.; Yu, J. Detection of Collapsed Buildings in Post-Earthquake Remote Sensing Images Based on the Improved YOLOv3. *Remote Sens.* **2019**, *12*, 44. [CrossRef]
12. Chai, Y.; Fu, K.; Sun, X.; Diao, W.; Wang, L. Compact Cloud Detection with Bidirectional Self-Attention Knowledge Distillation. *Remote Sens.* **2020**, *12*, 2770. [CrossRef]
13. Yan, Z.; Yan, M.; Sun, H.; Fu, K.; Sun, X. Cloud and Cloud Shadow Detection Using Multilevel Feature Fused Segmentation Network. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1600–1604. [CrossRef]
14. Lang, F.; Yang, J.; Yan, S.; Qin, F. Superpixel Segmentation of Polarimetric Synthetic Aperture Radar (SAR) Images Based on Generalized Mean Shift. *Remote Sens.* **2018**, *10*, 1592. [CrossRef]
15. Ciecholewski, M. River channel segmentation in polarimetric SAR images: Watershed transform combined with average contrast maximisation. *Expert Syst. Appl. Int. J.* **2017**, *82*, 196–215. [CrossRef]
16. Braga, A.M.; Marques, R.C.P.; Rodrigues, F.A.A.; Medeiros, F.N.S. A Median Regularized Level Set for Hierarchical Segmentation of SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1171–1175. [CrossRef]
17. Jin, R.; Yin, J.; Zhou, W.; Yang, J. Level Set Segmentation Algorithm for High-Resolution Polarimetric SAR Images Based on a Heterogeneous Clutter Model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 4565–4579. [CrossRef]
18. Audebert, N.; Le Saux, B.; Lefèvre, S. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *140*, 20–32. [CrossRef]
19. Sun, W.; Wang, R. Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 474–478. [CrossRef]
20. Marmanis, D.; Schindler, K.; Wegner, J.D.; Galliani, S.; Datcu, M.; Stilla, U. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS J. Photogramm. Remote Sens.* **2018**, *135*, 158–172. [CrossRef]
21. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
22. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
23. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
24. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1925–1934.
25. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
26. Liu, W.; Rabinovich, A.; Berg, A.C. Parsenet: Looking wider to see better. *arXiv* **2015**, arXiv:1506.04579.
27. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
28. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
29. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7151–7160.
30. Zhao, H.; Zhang, Y.; Liu, S.; Shi, J.; Change Loy, C.; Lin, D.; Jia, J. Psanet: Point-wise spatial attention network for scene parsing. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 267–283.
31. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv* **2016**, arXiv:1511.07122.
32. Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.P.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef] [PubMed]
33. Cao, Z.; Diao, W.; Zhang, Y.; Yan, M.; Yu, H.; Sun, X.; Fu, K. Semantic Labeling for High-Resolution Aerial Images Based on the DMFFNet. In Proceedings of the 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 1021–1024.

34. Mou, L.; Zhu, X.X. RiFCN: Recurrent network in fully convolutional network for semantic segmentation of high resolution remote sensing images. *arXiv* **2018**, arXiv:1805.02091.
35. Marmanis, D.; Wegner, J.D.; Galliani, S.; Schindler, K.; Datcu, M.; Stilla, U. Semantic segmentation of aerial images with an ensemble of CNNS. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *3*, 473–480. [CrossRef]
36. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50× fewer parameters and <0.5 MB model size. *arXiv* **2016**, arXiv:1602.07360.
37. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90.
38. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
39. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv* **2016**, arXiv:1602.07261.
40. Huang, G.; Liu, S.; Van der Maaten, L.; Weinberger, K.Q. Condensenet: An efficient densenet using learned group convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2752–2761.
41. Sifre, L.; Mallat, S. Rigid-Motion Scattering for Image Classification. Ph.D. Thesis, Ecole Polytechnique, Paris, France, 2014.
42. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
43. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
44. Wang, M.; Liu, B.; Foroosh, H. Factorized convolutional neural networks. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 545–553.
45. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
46. Tan, M.; Chen, B.; Pang, R.; Vasudevan, V.; Sandler, M.; Howard, A.; Le, Q.V. Mnasnet: Platform-aware neural architecture search for mobile. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 2820–2828.
47. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 1314–1324.
48. Tan, M.; Le, Q.V. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv* **2019**, arXiv:1905.11946.
49. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
50. Romero, A.; Ballas, N.; Kahou, S.E.; Chassang, A.; Gatta, C.; Bengio, Y. Fitnets: Hints for thin deep nets. *arXiv* **2014**, arXiv:1412.6550.
51. Zagoruyko, S.; Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv* **2016**, arXiv:1612.03928.
52. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
53. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. GCNet: Non-local Networks Meet Squeeze-Excitation Networks and Beyond. *arXiv* **2019**, arXiv:1904.11492.
54. Qian, C.; Li, H.; Zeng, G. Bi-directional Cross-Modality Feature Propagation with Separation-and-Aggregation Gate for RGB-D Semantic Segmentation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020.
55. 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 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1–9.
56. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3146–3154.
57. Zhang, J.; Lin, S.; Ding, L.; Bruzzone, L. Multi-scale context aggregation for semantic segmentation of remote sensing images. *Remote Sens.* **2020**, *12*, 701. [CrossRef]
58. Volpi, M.; Tuia, D. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2016**, *55*, 881–893. [CrossRef]
59. Yu, L.; Duc, M.N.; Nikos, D.; Wenrui, D.; Adrian, M. Hourglass-ShapeNetwork Based Semantic Segmentation for High Resolution Aerial Imagery. *Remote Sens.* **2017**, *9*, 522.
60. Sherrah, J. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv* **2016**, arXiv:1606.02585.