remote sensing

MDPI

Article

Scale-Invariant Multi-Level Context Aggregation Network for Weakly Supervised Building Extraction

Jicheng Wang [1,2,†], Xin Yan [3,†], Li Shen [3,*], Tian Lan [3], Xunqiang Gong [2] and Zhilin Li [3]

[1] Key Laboratory of Land Resources Evaluation and Monitoring in Southwest China of Ministry of Education, Sichuan Normal University, Chengdu 610068, China
[2] Key Laboratory of Mine Environmental Monitoring and Improving around Poyang Lake of Ministry of Natural Resources, East China University of Technology, Nanchang 330013, China
[3] Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu 611756, China
* Correspondence: lishen@swjtu.edu.cn
† These authors contributed equally to this work.

Abstract: Weakly supervised semantic segmentation (WSSS) methods, utilizing only image-level annotations, are gaining popularity for automated building extraction due to their advantages in eliminating the need for costly and time-consuming pixel-level labeling. Class activation maps (CAMs) are crucial for weakly supervised methods to generate pseudo-pixel-level labels for training networks in semantic segmentation. However, CAMs only activate the most discriminative regions, leading to inaccurate and incomplete results. To alleviate this, we propose a scale-invariant multi-level context aggregation network to improve the quality of CAMs in terms of fineness and completeness. The proposed method has integrated two novel modules into a Siamese network: (a) a self-attentive multi-level context aggregation module that generates and attentively aggregates multi-level CAMs to create fine-structured CAMs and (b) a scale-invariant optimization module that cooperates with mutual learning and coarse-to-fine optimization to improve the completeness of CAMs. The results of the experiments on two open building datasets demonstrate that our method achieves new state-of-the-art building extraction results using only image-level labels, producing more complete and accurate CAMs with an IoU of 0.6339 on the WHU dataset and 0.5887 on the Chicago dataset, respectively.

Keywords: building extraction; high-resolution remote sensing image; weakly supervised semantic segmentation; self-attentive aggregation; class activation map

Citation: Wang, J.; Yan, X.; Shen, L.; Lan, T.; Gong, X.; Li, Z. Scale-Invariant Multi-Level Context Aggregation Network for Weakly Supervised Building Extraction. *Remote Sens.* **2023**, *15*, 1432. https://doi.org/10.3390/rs15051432

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

Received: 7 February 2023
Revised: 1 March 2023
Accepted: 2 March 2023
Published: 3 March 2023

1. Introduction

Automatic building extraction from high-resolution images has become an active topic in the field of remote sensing in recent decades. It plays a vital role in a variety of applications, such as urban monitoring [1], population and economic estimation [2], and geospatial database making and updating [3]. Building extraction aims to classify each pixel as building or non-building, which can be regarded as binary semantic segmentation. However, this task is very challenging due to the difficulty in distinguishing between buildings with complex appearances and varying scales in high-resolution images with rich details and intra-class variance characteristics [4].

Convolutional neural networks (CNNs) have gained widespread popularity in various domains in recent years, including computer vision [5], climate change [6], and industrial detection [7,8]. With remarkable success in image segmentation, CNNs have been applied to building extraction from high-resolution remote sensing imagery using networks such as UNet [9], DeeplabV3+ [10], and PSPNet [11]. Researchers have also proposed building-specific approaches based on analyzing the characteristics of buildings, which have shown promising results [12–15]. However, these approaches have limited applications due to the

need for a large number of pixel-level annotations, which are time-consuming and labor-intensive to collect. Instead, weakly supervised semantic segmentation (WSSS) tries to alleviate this issue by utilizing weak supervision, such as image-level labels [16], bounding boxes [17], and scribbles [18]. As image-level labels are more readily available than other forms of supervision, this paper focuses on weakly supervised building extraction using image-level labels.

Image-level labels only indicate the presence or absence of buildings in the image without any localization cue, making it challenging for WSSS to achieve compelling results with fully supervised semantic segmentation. Fortunately, a class activation map (CAM) [16] is proposed to perform object localization only using image-level labels. Most advanced WSSS approaches are based on CAM and follow a three-stage learning paradigm: (1) using image-level labels to train a classification network to obtain the initial CAMs; (2) refining the initial CAMs to generate pseudo-pixel-level labels by semantic affinity-based methods [19], dense conditional random fields (CRF) [20] or saliency detection methods [21]; and (3) training a semantic segmentation network with these pseudo-labels. As supervised information is only utilized in the first stage, the key to the WSSS method is generating a promising CAM that is accurately activated on entire objects and not the background. To this point, many methods have been extensively investigated in the field of computer vision [22–26] and have shown effectiveness in processing natural scene images. However, these methods may not be suitable for high-resolution remote sensing images, which often contain vast amounts of visual information, significant spatio–spectral variability, and a much wider field of view [27].

In recent years, the remote sensing community has seen a growing interest in utilizing WSSS techniques for building extraction in high-resolution images. Advances in this area have resulted in several notable contributions, such as the SPMF-Net [28], which incorporates a superpixel-pooling mechanism to enhance the CAM and better preserve the shape and boundary information of buildings. The MSG-SR-Net [29] further advances this line of research by integrating a multi-scale generation strategy to improve the fineness of the CAM. Other methods, such as that of Li et al. [30], have utilized conditional random fields (CRF) to optimize both the CAM and segmentation results. In an effort to learn more building-specific information and encourage the network to perform more accurately, some studies have explored the exploitation of inherent relationships, such as pixel affinity [31] and adversarial information [32], to benefit building representation. Despite the demonstrated effectiveness of these techniques, the quality of the pseudo-labels generated by the CAM remains a critical factor affecting the performance of WSSS for building extraction. As illustrated in Figure 1, the CAMs often only activate the most discriminative regions, making it challenging to generate complete buildings. Furthermore, over-activation and under-activation of the CAMs can result in fuzzy boundaries, presenting opportunities for further improvement in the performance of weakly supervised building extraction.

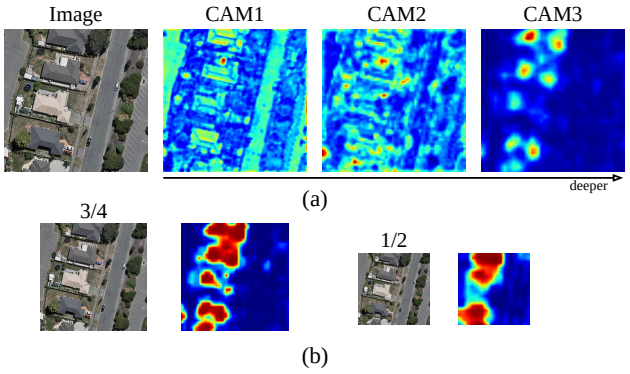


Figure 1. Visualization of CAMs generated by GradCAM++: (**a**) CAMs generated at different levels; (**b**) CAMs generated from different scales.

The limitations in building extraction using weakly supervised methods stem from the absence of localization information in image-level labels, creating a supervision gap between fully and weakly supervised methods. To overcome this challenge, it is crucial to incorporate more spatial information into the weakly supervised method, such as by utilizing the outputs of various layers in a neural network. As depicted in Figure 1, our research findings indicate that the CAM generated by lower-level layers possesses more details, and also more noise, than the CAM produced by high-level layers. By effectively combining CAMs from various levels, the accuracy of the CAMs can be significantly improved. Furthermore, our observations suggest that CAMs generated at different image scales do not always align with the scale variations of the buildings. The CAM generated at a coarser scale tends to highlight the complete area of the buildings but lacks detail at the boundaries, whereas the CAM generated at a finer scale exhibits the opposite trend. This discrepancy can serve as supervisory information to enhance the integrity of the building representations in a CAM.

Based on the above observation and analysis, we present a unified network that aims to enhance the quality of a CAM with regard to building representation. This is achieved through two key improvements to CAMs: (1) multi-level context aggregation for fine-structured refinement and (2) utilization of multi-scale supervision for integrity improvement. To achieve the first improvement, we introduce the self-attentive multi-level context aggregation module (SMCAM), which is based on GradCAM++ [23]. This module generates CAMs from multiple levels while suppressing noise in the network and uses a self-attention mechanism to effectively combine these CAMs, resulting in a more refined representation of building structures. For the second improvement, we propose the scale-invariant optimization module (SIOM) to further improve the integrity of CAMs. This module uses CAMs generated on a coarse scale as pixel-level supervision, guiding the network in learning to improve the integrity of the CAMs, thus ensuring that the CAMs are more aligned with the scale variation of buildings. Therefore, by incorporating the two enhancements, the proposed unified network is capable of generating high-quality CAMs that not only preserve fine structures but also ensure the integrity of the building representation. This leads to more reliable pixel-level training samples that are crucial for the performance of subsequent semantic segmentation steps. The main contributions of this study are summarized as follows:

- A self-attentive method that effectively generates and aggregates multi-level CAMs is proposed to produce fine-structured CAMs;
- A scale-invariant optimization method that incorporates multi-scale supervision is proposed to improve the completeness of CAMs;
- The Siamese network that integrates the above two improvements with designed losses is introduced with the aim of narrowing the supervision gap between fully and weakly supervised building extraction.

The organization of this paper is as follows: In Section 2, we give a comprehensive overview of previous studies on building extraction and weakly supervised semantic segmentation methods. The proposed network and its crucial components, SMCAM and SIOM, are thoroughly explained in Section 3. In Section 4, the performance of the proposed network is evaluated and compared with existing methods on two commonly used building datasets. A thorough analysis and discussion of the results are presented in Section 5, followed by the conclusions and future work in Section 6.

2. Related Works

This section presents a comprehensive overview of the most recent deep learning methods for building extraction from high-resolution remote sensing images. With the rapid advancements in the field, many researchers have proposed new techniques for weakly supervised semantic segmentation and building extraction specifically. Instead of giving a complete assessment of all current approaches in the field, this review only focuses on the related studies.

2.1. Building Extraction

Recently, fully convolutional networks (FCNs) have gained significant attention in the remote sensing community for their ability to perform efficient and accurate building extraction from high-resolution images. As a type of convolutional neural network (CNN), FCNs use a combination of a convolutional encoder and decoder to make dense predictions for every pixel in the input image, which makes them a promising solution for building extraction tasks. Studies have shown that FCNs for building extraction outperform traditional methods that rely on hand-crafted features in terms of accuracy and computational efficiency [12,15,33–42]. Some of these studies modify existing semantic segmentation networks to adapt them to building extraction, such as Schuegraf and Bittner [12], who combine two parallel U-Net-like [9] FCNs to extract the spatial and spectral features, respectively, and then fuse them to extract buildings and Yuan et al. [40], who improve PSPNet [11] by adding a designed feature pooling layer to capture both local and global relationships in building extraction. Others have proposed FCN models for building extraction that are specific to building characteristics. Guo et al. [42] propose a novel coarse-to-fine boundary refinement network (CBR-Net) that accurately extracts buildings. Li et al. [39] develop the CrossGeoNet with a Siamese network and a cross-geolocation attention module to provide the general building representation in different cities. There are also FCN models that utilize auxiliary data, such as digital surface models [36,38], to improve building extraction results.

The aforementioned methods can produce desirable outcomes; however, they heavily rely on large quantities of pixel-level annotations during training. To mitigate this issue, semi-supervised, unsupervised, and weakly supervised methods have emerged as alternatives for building extraction. Existing semi-supervised building extraction methods can be divided into three categories: self-training, generative adversarial network (GAN)-based, and consistency regularization methods [43]. Self-training models are trained with labeled samples, and then predictions for unlabeled samples are utilized as pseudo-labels for supervised training [4,44]. GAN-based methods mainly use limited annotated data to train the generators, which, in turn, generate synthetic annotations for the unannotated training images [45]. Consistency regularization methods can learn the distribution of unlabeled data by detecting the consistency of the output before and after perturbation [46]. However, these methods still rely on pixel-level labels in essence. Moreover, some researchers have attempted unsupervised building extraction [47]. As expected, the task presents significant challenges, and much progress must still be made before it can be accomplished successfully. More studies are focused on weakly supervised building extraction and are based on annotations that are less supervised than pixel-level labels, such as scribbles [48], image-level labels, and bounding boxes. In this paper, the approach taken is a weakly supervised segmentation method that exploits image-level labels, as described in the next subsection.

2.2. Weakly Supervised Semantic Segmentation

Weakly supervised semantic segmentation (WSSS) methods with image-level labels have gained widespread attention in the field of computer vision due to their cost-effectiveness as compared to pixel-level labeling. WSSS methods typically involve localizing objects with class activation maps (CAMs), generating pseudo-labels from CAMs, and training a semantic segmentation network. However, native CAMs [16,23] tend to produce inaccurate pseudo-labels as they only highlight the most discriminative regions, which are often incomplete and rough for objects. Hence, as the key to WSSS, many efforts have been made to improve CAMs for more complete object localization, such as iterative erasing [49], random dropping [50], super-pixel pooling [25], and local-to-global transferring [51]. Other studies have focused on refining CAMs using techniques such as multitask learning [52], region growing [53], pixel affinity [19], and inter-pixel relations [54]. Researchers also believe that the limitations of WSSS methods can be attributed to the supervision gap between the classification and segmentation tasks and have thus proposed methods to introduce auxiliary supervisory information to narrow this gap.

Wang et al. [24] propose SEAM to exploit pixel-level supervision through constraints between various affine changes. Du et al. [22] explore pixel-level supervisory signals with a combination of contrastive learning and consistency regularization. Sub-pixel supervision information is also introduced to compensate for the lack of supervision information [55]. Although these WSSS methods have achieved promising results for handling natural images in the field of computer vision, they may struggle to perform efficiently with remote sensing images as they are not specifically designed for this domain. Generalizing weakly supervised methods to different domains remains a challenge, as has been established by research [27].

Researchers have recently started exploring the potential of weak supervision in building extraction from remote sensing imagery, recognizing the need for methods specifically tailored to building characteristics to improve the accuracy and consistency of building representations in CAMs. Several studies have been conducted to address this challenge and enhance the quality of CAMs. For example, Fu et al. [56] develop WSF-Net for binary segmentation in remote sensing images and address class imbalance through a balanced binary training approach. MSG-SR-Net [29] further improves CAM fineness by integrating a multi-scale generation strategy. Meanwhile, methods such as conditional random fields by Li et al. [30] and superpixel pooling by Chen et al. [28] have been utilized to explore the spatial context and enhance building representation, respectively. The use of semantic affinity [32] and pixel affinity [31] has also been shown to improve the quality of CAMs. In contrast to these methods, we propose a novel WSSS method, based on the observations of CAM (Figure 1), which addresses two crucial aspects: multi-scale and multi-level. By generating and aggregating CAMs at multiple levels and training the images with multi-scale inputs, our method generates auxiliary pixel-level supervised information. This interplay of multi-scale and multi-level results in higher-quality CAMs compared to existing methods. It is noteworthy that MSG-SR-Net also leverages multi-level CAM fusion; however, our approach incorporates a self-attentive mechanism that enables CAMs to automatically calculate their importance during fusion, thus preserving valuable information and suppressing noise.

3. Scale-Invariant Multi-Level Context Aggregation Network

This section provides a description of the methodology and design of the network. Firstly, we define the problem context and describe the process of generating a CAM. Secondly, the overall architecture of the proposed network is presented. Thirdly, we delve into the two modules that are utilized to improve the quality of the CAMs with respect to building characteristics. Finally, we discuss the loss functions that are used in the proposed network.

3.1. Prerequisites

The problem of weakly supervised building extraction can be defined as follows: Given a set of images and corresponding image-level labels, the goal is to learn a model that can predict a pixel-level segmentation mask for the buildings in the image. Specifically, each training image, represented as $\mathbf{I}_{\mathcal{D}} \in \mathbb{R}^{W \times H \times 3}$ in dataset $\mathcal{D}$, is associated with a binary image-level label $y \in \{0, 1\}$, where 1 indicates the presence of buildings in image $\mathbf{I}$ and 0 indicates their absence. Since the image-level labels do not provide any localization information, current methods typically follow a two-stage approach to tackle this task, i.e., first training a classification network to identify regions in the image that correspond to buildings, and then using these regions to generate pseudo-labels for training a semantic segmentation network for building extraction. The majority of existing WSSS approaches rely on CAMs to produce the localization map. The training phase of the task can be formulated as follows:

$$\boldsymbol{M} = \mathcal{F}_{\mathrm{CAM}}(\mathbf{I}|\mathbf{I}_{\mathcal{D}}, \mathbf{y}_{\mathcal{D}}) \in \mathbb{R}^{W \times H \times 2}, \quad \boldsymbol{Y}_{pred} = \mathcal{F}_{\mathrm{SEG}}(\mathbf{I}|\mathbf{I}_{\mathcal{D}}, \mathcal{F}_{pseudo}(\boldsymbol{M})) \in \mathbb{R}^{W \times H \times 2} \tag{1}$$

Here, $\boldsymbol{M}$ represents the CAMs for the image $\mathbf{I}$ generated by the classification network $\mathcal{F}_{\mathrm{CAM}}$, which is trained on the image $\mathbf{I}_{\mathcal{D}}$ and corresponding image-level labels $\mathbf{y}_{\mathcal{D}}$ in dataset $\mathcal{D}$. $\mathcal{F}_{\mathrm{SEG}}$ is an FCN network for semantic segmentation, and $\mathcal{F}_{\mathrm{pseudo}}$ is used to generate the required pseudo-pixel-level labels. The final score map of the buildings is denoted by $\boldsymbol{Y}_{\mathrm{pred}}$. Accurate and complete CAMs are crucial for the success of semantic segmentation, as they significantly influence the performance of the pseudo-labels. Therefore, this paper focuses on improving the generation of CAMs for building extraction.

The proposed network is based on GradCAM++ [23] to improve the quality of CAMs. We chose GradCAM++ due to its two key features. Firstly, it can be loosely considered as a generalized version of the original CAM with an improved ability to localize features. Secondly, it does not require any modifications to the network, preserving the original classification network's learning capabilities. GradCAM++ generates a visual explanation for the specified class label c by using a weighted combination of the positive partial derivatives of the last convolutional layer's feature maps with respect to the class score, which is calculated as:

$$\boldsymbol{M} = relu(\sum_{k} \alpha_k^c \boldsymbol{A}_k) \tag{2}$$

where $\boldsymbol{A}_k$ denotes the k -th channel of the output feature map generated by the last convolutional layer. The rectified linear unit, $relu()$, filters out the features with negative values, where $relu(x) = max(0, x)$. α_k^c denotes the weight of class c corresponding to the k -th channel, which is calculated as follows:

$$\alpha_k^c = \frac{\frac{\partial^2 Y^c}{\partial \boldsymbol{A}_k^2}}{2\frac{\partial^2 Y^c}{\partial \boldsymbol{A}_k^2} + \sum \boldsymbol{A}_k \cdot \frac{\partial^3 Y^c}{\partial {\boldsymbol{A}_k}^3}} \cdot relu(\frac{\partial Y^c}{\partial \boldsymbol{A}_k}) \tag{3}$$

where $\frac{\partial Y^c}{\partial \boldsymbol{A}_k}$, $\frac{\partial^2 Y^c}{\partial \boldsymbol{A}_k^2}$, and $\frac{\partial^3 Y^c}{\partial \boldsymbol{A}_k^3}$ represent the first-order, second-order, and third-order gradients of the prediction score Y^c, respectively. For computational convenience, it is common to make $Y^c = exp(f_c)$; here, f_c is the output score of the classification network.

It should be noted that the CAMs generated by GradCAM++ have the same size as feature map $\boldsymbol{A}$ and are not normalized. To make them more suitable for subsequent processing, we resample them to the size of the input image and normalize them using the following formula:

$$\boldsymbol{M}^* = \frac{\boldsymbol{M} - MIN(\boldsymbol{M})}{MAX(\boldsymbol{M}) - MIN(\boldsymbol{M})} \tag{4}$$

where $MAX()$ represents the maximum value, and $MIN()$ represents the minimum value.

3.2. Overall Network Architecture

We proposed a network for the task of weakly supervised building extraction, as depicted in Figure 2, which leverages multi-level features and multi-scale information to improve the quality of CAMs. It consists of three main components: a Siamese classification network, the self-attentive multi-level context aggregation module (SMCAM), and the scale-invariant optimization module (SIOM). The Siamese network utilizes shared weights to classify two input images with different scales, utilizing the ground-truth image-level labels as the target. Both images are processed simultaneously using the same network architecture, which can be designed based on well-known networks, such as ResNet [57] and VGG [58]. The main purpose of the Siamese network is to generate multi-level features at different scales through training with an image-level labeled dataset. The other two modules are designed to utilize these features to guide the improvement of the CAM. Specifically, SMCAM generates multi-level CAMs based on GradCAM++ and fuses these CAMs through a self-attention mechanism. SIOM further exploits these features to improve the integrity of CAMs by mutual learning between different scale supervisions. Both modules are described in more detail in the subsequent subsections.

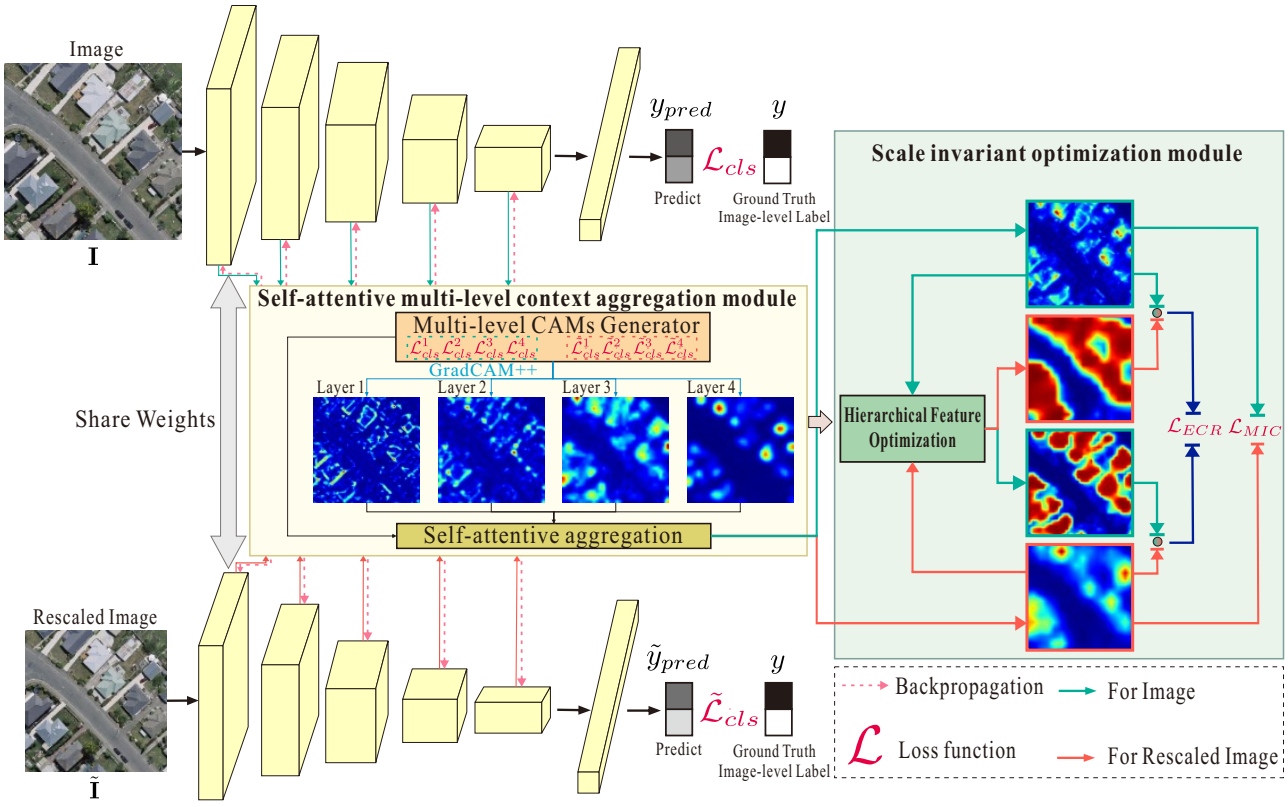


Figure 2. The framework of the proposed network.

It is noteworthy that the proposed network takes two images of different scales as input during the training phase in order to provide multi-scale supervision. During the inference phase, the network only requires a single input image, and the CAM is generated through sequential optimization of SMCAM and SIOM.

3.3. Self-Attentive Multi-Level Context Aggregation Module

The proposed module, SMCAM, is designed to address two major problems in building extraction tasks: the generation of multi-level CAMs and the aggregation of these CAMs. The main aim of SMCAM is to utilize GradCAM++ to generate CAMs at different levels of a deep neural network and to aggregate these CAMs to produce more detailed CAMs. To achieve the first goal, we utilize GradCAM++ to backpropagate gradients from score maps to any nodes of the network (such as the red dot on each ResNet Block in Figure 3). However, it is observed that CAMs generated from low-level features may contain a significant amount of noise, making them challenging to use directly. This noise is mainly due to two factors: disturbance introduced during gradient backpropagation due to the long path and presence of noise in the low-level features themselves. To tackle these issues, SMCAM introduces auxiliary classification branches at each node to improve the semantic depth of low-level features and shorten the gradient backpropagation path. These branches consist of a 1x1 convolutional layer with 1024 output channels, followed by an average pooling layer, which ultimately outputs an image-level score. The auxiliary classification branches enhance the semantic information in low-level features, making the generated CAMs more useful for the building extraction task.

To achieve the second goal of aggregating multi-level CAMs, our proposed module incorporates an attention-based mechanism. Unlike traditional fusion methods, such as averaging and concatenation, which ignore the relative importance of features at different levels, our model aims to learn the varying contributions of each level for every pixel in the CAMs. Specifically, the process starts with passing the feature maps generated by the average pooling layer of each auxiliary branch through the channel average pooling

layer, generating a score map for each level, $l \in \{1, ..., L\}$, where each score map has a single channel. The score maps are then upsampled to match the size of the CAMs. Mathematically, let $\boldsymbol{H}_l$ denote the weight score generated at level l. A softmax function is applied across the levels to compute the specific weights, $\mathbf{w}_l$, of the CAMs for each level:

$$\mathbf{w}_l = \frac{exp(\boldsymbol{H}_l)}{\sum_{l=1}^{L} exp(\boldsymbol{H}_l)}. \tag{5}$$

Finally, the CAMs are aggregated into a single map, $\boldsymbol{M}$, through a weighted sum of the score maps of all levels, as calculated by:

$$\boldsymbol{M} = \sum_{l=1}^{L} \mathbf{w}_l \boldsymbol{M}^l, \tag{6}$$

where $\boldsymbol{M}^l$ represents the CAM generated by GradCAM++ at the auxiliary branch of level l, and the weights are determined by the softmax function applied to the weight scores generated at each level.

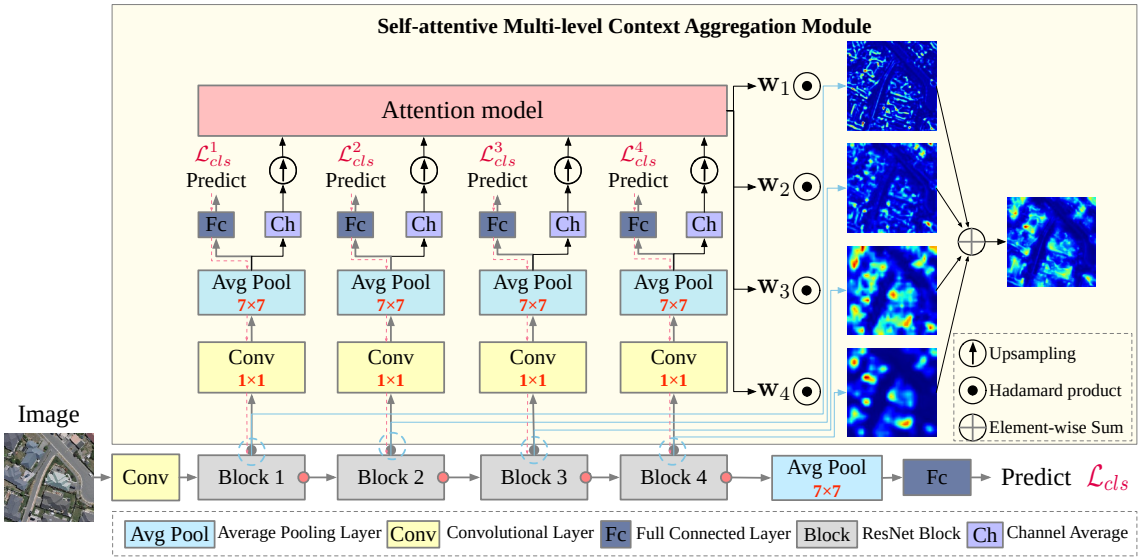


Figure 3. The structure of self-attentive multi-level context aggregation module.

3.4. Scale-Invariant Optimization Module for Improving Integrity of CAMs

Although the CAMs generated by SMCAM utilize multi-level features, there still exists a drawback in that the generated CAMs may not cover the entire building object, as they often struggle to activate the most discriminative regions without adequate pixel-level supervision. One of our key observations is that CAMs generated by the classification network for different scales of input images activate different regions for building objects, providing valuable additional supervisory information. Therefore, we proposed the scale-invariant optimization module (SIOM) to leverage this multi-scale information. As illustrated in Figure 2, SIOM consists of two crucial parts: a mutual learning mechanism of CAMs and hierarchical feature optimization. For the former, a mutual learning mechanism benefits from the architecture of the Siamese network, which consists of two subnetworks with shared parameters that process two different scale images simultaneously and generate a series of CAMs at different scales under two scales with SMCAM. By motivating CAMs of different scales to be similar at multiple levels, the network can learn scale-invariant representations that better capture complete building objects. Therefore, we proposed the multi-level invariant constraint loss $\mathcal{L}_{MIC}$ by extending the equivariant constraint to multiple levels, as presented in the next subsection.

For the latter, although the multi-level invariant constraint loss can provide additional pixel-level supervision at multiple levels, the CAM generation is still limited to the framework of the classification networks. To alleviate this, we further propose a separate learnable branch, or hierarchical feature optimization, for enhancing the completeness of a CAM. In particular, it optimizes a CAM in a progressive manner, leveraging progressively the image and multi-level features generated by the classification network through a three-stage process, as shown in Figure 4. The first two stages utilize self-attention mechanisms to uncover non-local relationships [59] within the multi-level features of the classification network, resulting in improved object integrity. Each stage consists of two convolutional layers with learnable parameters. The optimization process can be mathematically expressed as follows:

$$\boldsymbol{M}' = softmax(\mathbf{x}^T W_1^T W_2 \mathbf{x})\boldsymbol{M}, \tag{7}$$

where $\boldsymbol{M}'$ denotes the optimized CAM. W_1 and W_2 denote the parameters of the first and second convolutional layers, respectively. The utilized features are indicated by $\mathbf{x}$, for stage 1, $\mathbf{x} = [\mathbf{x}_4, Down(\mathbf{x}_2)]$ and for stage 2, $\mathbf{x} = [\mathbf{x}_3, Down(\mathbf{x}_1)]$, where $[]$ and $Down()$ denote the concatenating and downsampling, respectively. This self-attention mechanism essentially uses the relationships between all of the pixels in feature maps to refine the CAM and has been shown to be particularly effective at mining global relationships. However, as the size of the CAM increases, the size of the attention map $W_1^T W_2$ increases exponentially, leading to memory explosions on the graphics card during training; hence, we only use this mechanism in the first two stages. In the third stage, we fully exploit the local information in the image, which is used to enhance the detail stages of the CAM. We use a fixed-parameter pixel adaption convolution [60], and the optimization process can be represented by the following equation:

$$M_i'' = \sum_{j \in \mathcal{N}(i)} D_{i,j} M_{i,j}', \tag{8}$$

where M_i denotes the value at pixel i in the CAM; $\mathcal{N}(i)$ denotes the set of neighboring pixels of pixel i; and $\mathbf{D}$ denotes the affinity map, where $D_{i,j}$ denotes the similarity relationship between pixel i and pixel j, as calculated by the following equation:

$$D_{i,j} = \frac{exp(k(\mathbf{I}_i, \mathbf{I}_j))}{\sum_{j \in \mathcal{N}(i)} exp(k(\mathbf{I}_i, \mathbf{I}_j))}, \tag{9}$$

where $\mathbf{I}_i$ denotes the spectral vector at image pixel i; k is the kernel function, where the Gaussian function is used, i.e., $k(\mathbf{I}_i, \mathbf{I}_j) = exp(-\frac{1}{2}(\mathbf{I}_i - \mathbf{I}_j)^T(\mathbf{I}_i - \mathbf{I}_j))$; and the softmax function is used to normalize. It is important to note that this operation is distinct from the Hadamard product, as indicated by the notation $\odot$ in Figure 4. The parameters of SIOM are learned by comparing optimized and unoptimized CAMs on two scales. The gradient propagation path of this module is detached from the classification network to prevent interference with learning classification representation. By optimizing the hierarchy of non-local and local information, the problem of building edges and integrity can be solved to a large extent, effectively improving the quality of the CAM for building extraction.

It should be noted that the basic architecture of SIOM is inspired by SEAM [24] but has three differences. Firstly, we are based on GradCAM++, which is more generalized than the original CAM applied in SEAM. Secondly, we propose hierarchical feature optimization, which can utilize multi-level features and images to optimize CAM with regard to building extraction. Thirdly, our loss function is specifically designed to address the objective of building extraction. Therefore, the proposed SIOM is more appropriate for weakly supervised building extraction than SEAM.

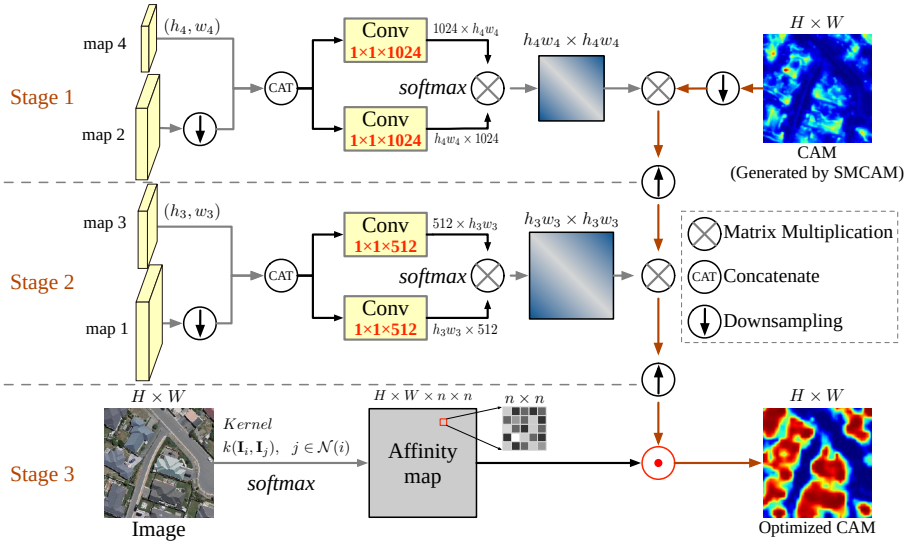


Figure 4. The structure of scale-invariant optimization module.

3.5. The Design of Loss Function

Since only image-level labels are available, the design of the loss function is exceptionally important in order to introduce more supervised information, especially some pixel-level supervisions. In summary, the total loss of the proposed network is defined as follows:

$$\mathcal{L}_{ALL} = \mathcal{L}_{CLS} + \mathcal{L}_{MIC} + \mathcal{L}_{ECR}, \tag{10}$$

where the classification loss is denoted as $\mathcal{L}_{CLS}$, which exploits image-level supervision, and $\mathcal{L}_{MIC}$ is the multi-level invariant constraint loss in SIOM. The equivariant cross regularization loss $\mathcal{L}_{ECR}$ is adopted to train the branch of hierarchical feature optimization.

The binary cross-entropy loss is utilized as the classification loss in the proposed network and is defined as follows:

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{n=1}^{N}(y_n log(y_{pred,n}) + (1 - y_n) log(1 - y_{pred,n})), \tag{11}$$

where N is the batch size; y_{pred} denotes the output of the classification network and y_n is the corresponding image-level label. The network is a Siamese network and can generate two scores for the original image $\mathbf{I}$ and the rescaled image $\tilde{\mathbf{I}}$, respectively. In addition, the proposed SMCAM module has multiple output branches for various levels. The classification loss of the network is made up of many sub-losses, which is described below:

$$\mathcal{L}_{CLS} = \mathcal{L}_{cls} + \tilde{\mathcal{L}}_{cls} + \frac{1}{L}\sum_{l=1}^{L}(\mathcal{L}_{cls}^{l} + \tilde{\mathcal{L}}_{cls}^{l}), \tag{12}$$

where $\mathcal{L}_{cls}$ and $\tilde{\mathcal{L}}_{cls}$ are the losses calculated for the Siamese network, while SMCAM at level l has its own losses, represented by $\mathcal{L}_{cls}^{l}$ and $\tilde{\mathcal{L}}_{cls}^{l}$.

The multi-level invariant constraints (MIC) loss is built with a multi-level equivariant metric based on the L1 loss. The aim is to enhance the CAMs generated by different branches of the Siamese network through multiple levels of features, leading to scale-invariant features that result in more complete building representations in the CAMs. The calculation of this loss is described below.

$$\mathcal{L}_{MIC} = \frac{1}{L}\sum_{l=1}^{L}\|\boldsymbol{M}^{l} - up(\tilde{\boldsymbol{M}}^{l})\|_1, \tag{13}$$

where $\|\cdot\|$ denotes the L1-norm, and $up(\cdot)$ denotes the upsampling.

For the equivariant cross regularization loss, we refer to the SEAM [24], which can be represented as:

$$\mathcal{L}_{ECR} = \frac{1}{2}(\|\boldsymbol{M}'' - up(\tilde{\boldsymbol{M}})\|_1 + \|up(\tilde{\boldsymbol{M}}'') - \boldsymbol{M}\|_1), \tag{14}$$

where $\boldsymbol{M}''$ is the CAM optimized by hierarchical feature optimization; $\tilde{\boldsymbol{M}}$ denotes the CAM generated by SMCAM the from rescaled image. This loss can further improve the quality of the CAMs and prevent degeneration.

4. Experiments

4.1. Experimental Datasets

The proposed method is evaluated using two publicly available high-resolution remote sensing datasets, namely the WHU Building Dataset [34] and the Inria Aerial Image Labeling Dataset [61]. The WHU Building Dataset (abbreviated as "WHU dataset") consists of 8189 tiles with 512 × 512 pixels, covering approximately 450 km^2 in the Christchurch area. Each image has three channels (RGB) and a spatial resolution of 0.3 m. These images are divided into 4736, 1036, and 2416 patches with corresponding ground-truth labels, which are then split into training, validation, and test sets. The Inria Aerial Image Labeling Dataset includes images and building labels for 10 cities, with a subset dataset of Chicago (abbreviated as "Chicago dataset") selected for the experiment. It contains 36 color image tiles of 5000 × 5000 pixels with a 0.3 m spatial resolution. All of these images have corresponding building annotations and are divided into a training set of 24 images, a testing set of 8 images, and a validation set of 4 images in this experiment.

Both datasets provide pixel-level labels and are primarily used to test and evaluate fully supervised building extraction methods. For the proposed weakly supervised method, only image-level labels are utilized to train the model. Therefore, pixel-level labels are used to generate image-level annotations for evaluating the method. Following the preprocessing in previous works [29] and [32], the images are cropped into patches with 256 × 256 pixels and a stride of 128 pixels. The image-level labels of these patches are determined by the percentage of building pixels in the pixel-level annotations. Specifically, patches with more than 22 percent of building pixels (about 2/9) are labeled as building, while patches with a percentage of 0, i.e., without any building pixels, are labeled as non-building. The remaining patches with a percentage between 0 and 22 are simply ignored. After processing, the WHU dataset contains 27,879 image patches for training, with 14,316 building patches, 13,563 non-building patches, and 18,364 patches for testing. The Chicago dataset contains 24,736 patches for training, including 12,793 building patches, 12,943 non-building patches, and 11,020 patches for testing. Some processed examples from both datasets are shown in Figure 5. To gain a better understanding of these two datasets, we utilize t-SNE [62] to visualize the feature distributions of both datasets. The features are extracted from the last layer output of a residual neural network [57]. Despite having the same resolution and purpose, the t-SNE visualization (Figure 5b) reveals that the feature distributions of the two datasets are notably distinct.

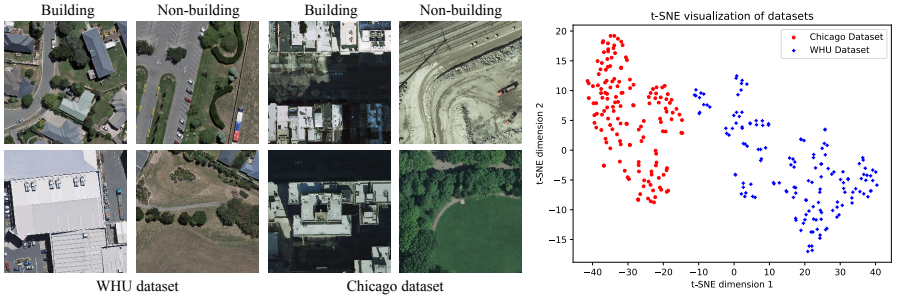


Figure 5. Sample images (**a**) and t-SNE visualization (**b**) for two datasets.

4.2. Experimental Setup

4.2.1. Methods for Comparison

In the experiments, seven WSSS methods are compared with the proposed method for the selected datasets. The main information about these methods, including ours, is summarized as follows:

- CAM-GAP: Zhou et al. [16] propose CAM-GAP for discriminative localization by adding a global average pooling modification to the network. It actually builds a generic localizable deep representation that can be applied to weakly supervised semantic segmentation;
- GradCAM++: Chattopadhay et al. [23] propose GradCAM++ based on gradients without changing the network structure. Its goal is to provide a visual interpretation for CNN-based models and can also be used for WSSS. It can be regarded as the generalization of a CAM;
- WILDCAT: Durand et al. [26] introduce WILDCAT to simultaneously align image regions for spatial invariance and learn strongly localized features for WSSS. It uses a single generic training scheme for classification, object localization, and semantic segmentation;
- SPN: The superpixel pooling network (SPN), proposed by Kwak et al. [25], utilizes superpixel segmentation of the input image as a pooling layout to cooperate with low-level features for semantic segmentation learning and inferring. The network architecture decouples the semantic segmentation task into classification and segmentation, allowing the network to learn class-agnostic shapes prior to the noisy annotations. It achieves outstanding performance on the challenging PASCAL VOC 2012 segmentation benchmark;
- WSF-Net: WSF-Net [56] is proposed for binary segmentation in remote sensing images with the potential to handle class imbalance through a balanced binary training strategy. It introduces a feature fusion strategy to adapt to the characteristics of objects in remote sensing images. The experiments achieve a promising performance for water and cloud extraction;
- MSG-SR-Net: MSG-SR-Net is proposed by Yan et al. [29] for weakly supervised building extraction from high-resolution images. It integrates two modules, i.e., multiscale generation and superpixel refinement, to generate high-quality CAMs so as to provide reliable pixel-level training samples for subsequent semantic segmentation. It achieves excellent performance for building extraction;
- SEAM: The self-supervised equivariant attention mechanism (SEAM) is proposed by Wang et al. [24] for WSSS. It embeds self-supervision into the weakly supervised learning framework through equivariant regularization, which forces CAMs generated from various transformed images to be consistent. It achieved state-of-the-art performance on the PASCAL VOC 2012 dataset;
- Ours–SIOM: The proposed method that only utilized SIOM. The CAM is generated by GradCAM++ at the last convolutional layer;
- Ours–SMCAM: The proposed method that only utilized SMCAM, including mutual learning between different scales ($\mathcal{L}_{MIC}$);
- Ours: The proposed network with SMCAM and SIOM.

As most methods focus on improving the quality of a CAM for target extraction as well as ours, we mainly compare the completeness and fineness of buildings in a CAM.

4.2.2. Evaluation Criteria

In the evaluation of our proposed methods, we employ three commonly used quantitative criteria: the F1 score, overall accuracy (OA), and intersection over union (IoU). The F1 score is computed as the harmonic mean of precision and recall, which are defined as follows:

$$\mathrm{F1} = \frac{2}{recall^{-1} + precision^{-1}}, \tag{15}$$

where *precision* and *recall* are defined as:

$$precision = \frac{TP}{TP + FP}, \quad recall = \frac{TP}{TP + FN}, \tag{16}$$

where TP and TN, respectively, represent a true positive and true negative, while FP is a false positive and FN is a false negative.

The OA is calculated as the ratio of correctly classified instances to the total instances, as represented by the following equation:

$$\text{OA} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{17}$$

The IoU is calculated as the ratio of the area of intersection between the predicted and ground-truth segmentation divided by the area of union between the predicted and ground-truth segmentation and can be calculated as follows:

$$\text{IoU} = \frac{TP}{TP + FP + FN}. \tag{18}$$

The above three metrics are used to quantitatively evaluate the results of the proposed methods, both for CAMs and building extractions. Additionally, comparisons of visualizations are also utilized to evaluate the results.

4.2.3. Implementation Details

The proposed method employs ResNet50 [57] as the backbone of the Siamese network, and the pre-trained weights from ImageNet (provided by PyTorch) are utilized for the backbone initialization. The generation of multi-scale CAMs is depicted in Figure 3 and derived from the blocks of ResNet50. The training process is carried out by using stochastic gradient descent (SGD) with momentum, where a momentum coefficient of 0.9 and weight decay of 0.0005 are set [11,29,31]. In accordance with the approach presented in [11], a poly-like learning rate rule is employed, where the learning rate is defined as base one multiplying $(1 - t/T)^{power}$ with the base set to 0.01, power to 0.9, t denoting the current iteration, and T denoting the maximum iteration. The network is trained with a batch size of 8, over a total of 30 epochs, using 2 GPUs. In the initial 5 epochs, only cross-entropy loss ($\mathcal{L}_{CLS}$) is employed, while the total loss is utilized in the subsequent epochs. Adhering to SEAM ([24]), the proposed method employs online hard example mining (OHEM) on the ECR loss ($\mathcal{L}_{ECR}$), retaining the top 20% of pixel losses. Random data augmentation, including mirroring, rotation, Gaussian blurring, and color jittering, is applied during the training process, and the Siamese structure expands the matched augmentation to the rescaled image. During inference, the shared-weight Siamese network only requires one branch to be restored.

To ensure a fair comparison, all comparison methods are reimplemented based on the ResNet50 [57] backbone, with the exception of SPN, which presents some difficulties in separating its main modules from the classification network. Both SPN and MSG-SR-Net employ the simple linear iterative clustering algorithm [63] to pre-segment the images into approximately 256 superpixels per image. Furthermore, DeepLabV3+ (based on Resnet50) is utilized as the semantic segmentation network $\mathcal{F}_{\text{SEG}}$ for all methods. To train DeepLabV3+, we employ cross-entropy as the loss function along with the same SGD optimizer as mentioned previously and with the same parameters. However, the initial learning rate is set to 0.007 for this process [10]. The training batch size is set to 8, and a total of 20 epochs are trained. For data augmentation, we only apply random mirroring and rotation. All experimental settings are made consistent across methods as much as possible, though some hyperparameters (e.g., learning rate and momentum) specific to each model may have been adjusted to enable efficient convergence during training. It should be noted that the background thresholds of CAMs for each approach differ as well, and one should

be chosen with the best F1 score of pseudo-labels after traversing all background threshold options on the validation set.

The proposed method and all comparison methods are implemented on a Linux 5.15 platform, using Python 3.9 and PyTorch for deep learning. The CUDA 11.6 version is utilized for GPU acceleration. The experiments are run on a Linux platform equipped with two Intel Xeon 8-core CPUs @ 2.2 GHz and two NVIDIA RTX 4000 GPUs with 8 GB memory and 2304 shading units.

4.3. Comparison of CAM Results

4.3.1. Results of Chicago Dataset

For the quantitative aspect, Table 1 reports three evaluation criteria of the CAM results in terms of building extraction accuracy. The proposed method yields the highest values of F1, IoU, and OA when compared to other methods, achieving the best pseudo-label accuracy. In particular, CAM-GAP and GradCAM++ obtain pool accuracy due to the fact that they only activate the most discriminative regions. However, WILDCAT achieves the worst results, which demonstrates that this method is not suitable for building extraction in the remote sensing field. Ours–SMCAM still achieves remarkable results with 0.6774 of F1 and 0.8081 of OA, which is slightly behind MSG-SR-NET, which benefits from the superpixel's ability to utilize the low-level features. It also illustrates the effectiveness of the attentive aggregation of multi-scale CAMs since we do not employ the superpixel-like technique. As for ours–SIOM, it achieves slightly better results than SEAM. Both of these two methods utilize the Siamese networks. Overall, the results indicate the effectiveness of our method, which incorporates SMCAM and SIOM for optimal performance.

Table 1. Evaluation of CAMs generated by various methods with IoU (%), OA (%), and F1 score (%) on WHU dataset and Chicago dataset.

Method	WHU Dataset			Chicago Dataset		
	IoU	F1	OA	IoU	F1	OA
CAM-GAP	41.85	59.01	84.56	36.17	53.13	73.72
GradCAM++	42.73	59.88	84.13	35.74	52.66	73.29
WILDCAT	37.11	54.13	85.51	33.61	50.31	71.22
SEAM	51.83	68.27	90.97	46.82	63.78	79.56
SPN	40.79	57.94	85.31	44.76	61.84	75.93
MSG-SR-NET	53.76	69.93	91.90	51.25	67.77	80.98
WSF-Net	46.58	63.56	87.77	45.27	62.33	79.04
Ours-SIOM	51.91	68.34	91.78	47.48	64.39	80.38
Ours-SMCAM	54.38	70.45	92.21	51.22	67.74	80.01
Ours	58.04	73.45	93.29	52.68	69.01	81.53

Five image patches containing buildings with different sizes and shapes are selected for visual comparison, as shown in Figure 6. It can be observed that WILDCAT performed the worst in identifying various sizes of buildings, making it difficult to separate the buildings from the background. As inferred above, CAM-GAP and GradCAM++ only activate the most discriminative parts of buildings (e.g., edges and texture); the difference is that the former is relatively complete and the latter relatively fine. SEAM achieves promising results in terms of building completeness but still has some over-activation and under-activation issues. The CAMs generated by WSF-Net are relatively coarse and not suitable for building extraction, as this method is designed for water and cloud extraction in remote sensing images. Both SPN and MSG-SR-Net benefit from using superpixel for more detailed information but are also limited by the inaccurate activation of non-building regions. In addition, the results of MSG-SR-Net are notably better than SPN for building extraction. The CAMs obtained by the proposed methods are better than those obtained by the other methods. Specifically, ours–SIOM obtained a relatively integral building in

CAMs, while ours–SMCAM contained more detailed information. The proposed method combines the advantages of both modules to obtain the best overall results.

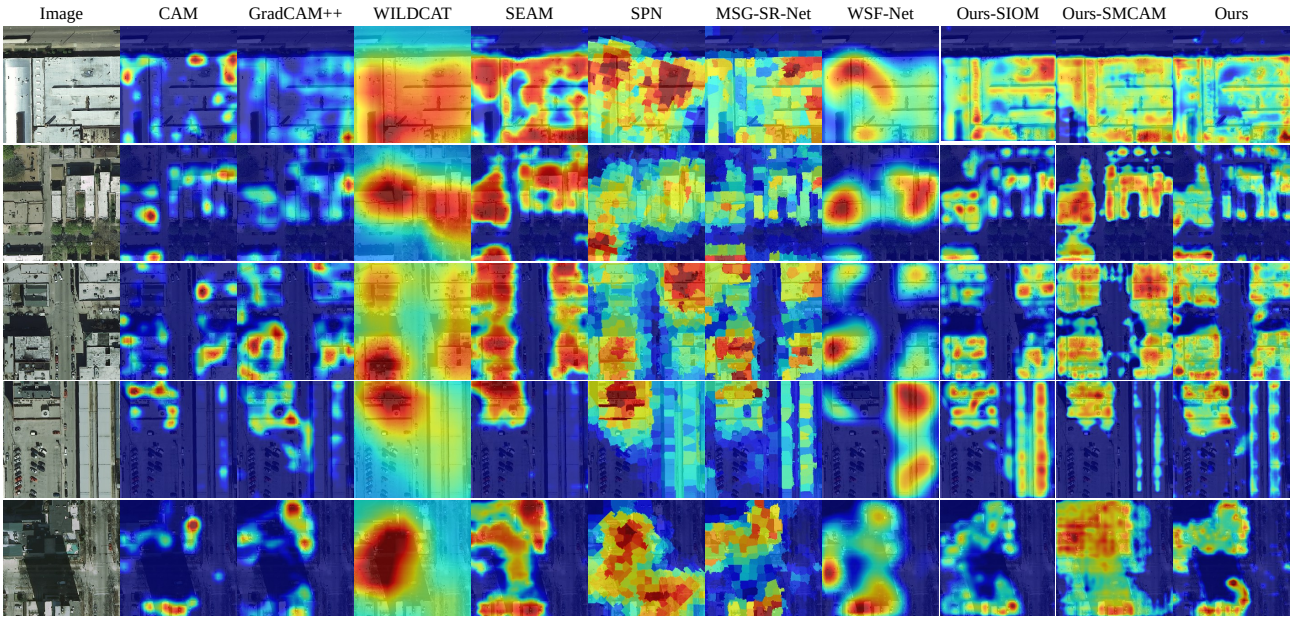


Figure 6. Qualitative comparison of the CAM results obtained by various methods on the Chicago dataset.

4.3.2. Results of WHU Dataset

Table 1 also reports the F1 score, IoU, and OA of different methods on the WHU dataset. It can be observed that: (1) WILDCAT cannot successfully generate a CAM for buildings. Other than that, CAM-GAP, GradCAM++, and WILDCAT have the worst results, and MSG-SR-Net, designed for remote sensing images, achieves remarkable results, as well as the Chicago dataset. (2) The proposed method outperforms all other methods, but with a larger lead than the results of the Chicago dataset. (3) Ours-SMCAM also shows a comparative performance with other methods. This reveals the advantages of the proposed module. (4) Ours-SIOM shows a better performance than SEAM, with a higher F1 score and IoU and an improvement of over one percent in OA, demonstrating its effectiveness in distinguishing buildings from the background.

In order to visually compare the performances of the different methods, five image patches featuring buildings of various sizes and shapes have been selected, as shown in Figure 7. Notably, the quality of images and labels in the WHU dataset is better than that in the Chicago dataset. The images in the WHU dataset are well-calibrated and contain fewer shadows from the buildings. Since buildings and shadows often accompany each other, almost all methods are difficult to distinguish them from one another due to the lack of relevant supervision information. Therefore, both in terms of quality and quantity, the results for the WHU dataset are superior to those of the Chicago dataset. Additionally, the corresponding pixel-level labels are more accurate for producing better image-level labels. Similar to the results for the Chicago dataset, WILDCAT, CAM-GAP, and GradCAM++ still struggle to accurately identify buildings. Specifically, WILDCAT has over-activated too many regions, which overwhelms the buildings. Additionally, CAM-GAP and GradCAM++ still highlight the most distinctive areas. SEAM performs better than the others but still shows over-activation in some areas. The CAMs generated by WSF-Net are coarse for buildings and more suitable for large targets. Although MSG-SR-Net outperforms SPN when utilizing superpixel segmentation, it still results in false negatives along building boundaries. Our proposed methods outperform the other methods, as seen in the results for the Chicago dataset. In particular, ours–SIOM generates more complete building activations and distinguishes buildings from backgrounds more effectively. However, the edges of the buildings remain somewhat blurred. Conversely, ours–SMCAM preserves the structural

details and boundaries of the buildings well but may result in some false activations. By combining both modules, the proposed method is able to achieve the best results, as it fuses multi-level and multi-scale features to enhance the quality of CAMs for building extraction. The results confirm the effectiveness of the proposed method for building extraction in remote sensing images.

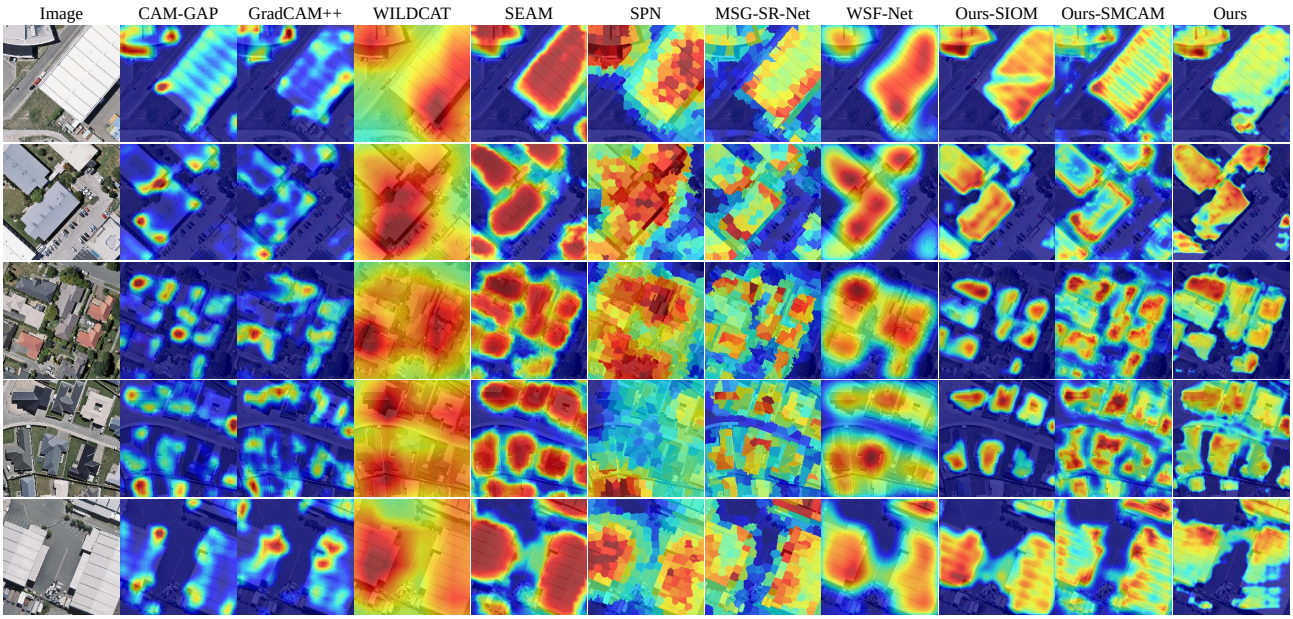


Figure 7. Qualitative comparison of the CAM results obtained by various methods on the WHU dataset.

4.4. Comparison of Building Extraction Results

A CAM can generate building extraction results directly with the use of a suitable threshold, but such results often have significant limitations and are therefore are used commonly as pseudo-labels to train a segmentation network. For a more comprehensive comparison, we trained fully convolutional networks for each method to perform building extraction. Table 2 presents the building extraction results on the two datasets. Firstly, when comparing the CAM results listed in Table 1, almost all methods show improvement in their building extraction results after training the fully convolutional network, with the exception of WILDCAT, which produces a poor CAM that results in ineffective training of the segmentation network. Secondly, the building extraction accuracies of SEAM and MSG-SR-Net are relatively good. Both methods achieve high overall accuracy (OA) scores, demonstrating their capabilities in effectively distinguishing between buildings and non-building objects. Thirdly, when applied to the WHU dataset, our proposed method achieved an F1 score of 0.7759, an OA score of 0.9457, and an IoU score of 0.6339. Similarly, on the Chicago dataset, our method achieved an F1 score of 0.7411, an OA score of 0.8436, and an IoU score of 0.5887. Both outperform the other methods significantly. This highlights the superior effectiveness and robustness of our proposed method for building extraction from remote sensing images, which can be attributed to its ability to produce high-quality CAMs.

The visual comparison of building extraction results from three image patches with varying building densities from each dataset is presented in Figure 8. The results reveal a strong correlation between the quality of the building extraction and the accuracy of the CAM. The higher quality of the annotations and images in the WHU dataset may explain the improved performance compared to the Chicago dataset. Unfortunately, the results of WILDCAT and SPN are below expectations and fail to effectively extract the buildings. On the other hand, CAM, GradCAM++, WSF-Net, and SEAM tend to over-highlight building regions, resulting in coarse outcomes. Although MSG-SR-Net performs better than the aforementioned methods, it still has some inaccuracies and incomplete building extractions. Our proposed method, however, demonstrates the best results, especially on the WHU

dataset, where individual buildings are clearly distinguishable and have well-defined edges. This highlights the superior effectiveness of our approach for WSSS building extraction.

Table 2. Evaluation of building extraction results generated by various methods with IoU (%), OA (%), and F1 score (%) on WHU dataset and Chicago dataset.

Method	WHU Dataset			Chicago Dataset		
	IoU	**F1**	**OA**	**IoU**	**F1**	**OA**
CAM	45.11	62.17	88.99	40.12	57.27	75.33
GradCAM++	42.19	59.34	87.62	38.30	55.39	74.49
WILDCAT	30.28	46.49	64.33	28.38	44.21	67.11
SEAM	55.44	71.33	91.94	51.27	67.79	81.56
SPN	43.25	60.38	86.6	50.40	67.02	82.35
MSG-SR-NET	57.07	72.67	92.34	57.68	73.16	83.28
WSF-Net	49.60	66.31	91.03	50.15	66.80	80.24
Ours	63.39	77.59	94.57	58.87	74.11	84.36

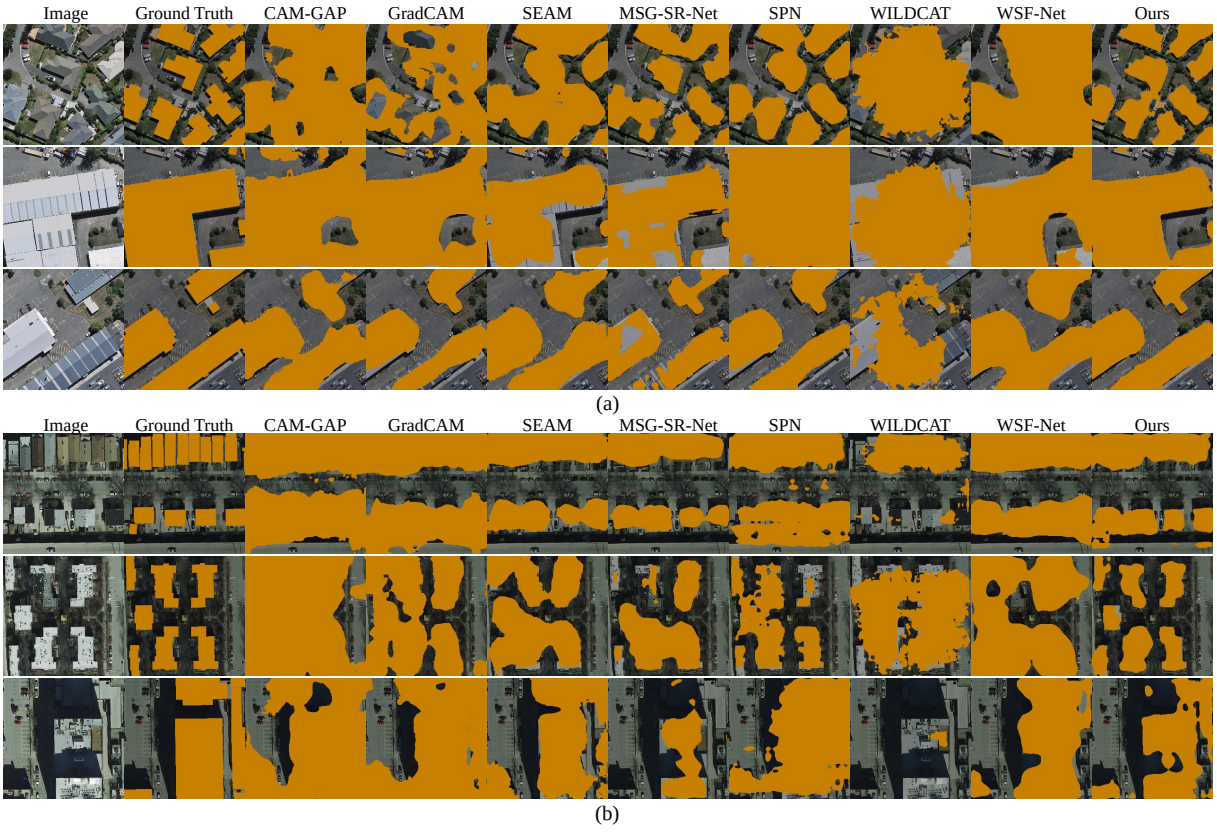


Figure 8. Examples of building extraction results of the proposed method and other comparison methods on two datasets: (**a**) WHU dataset; (**b**) Chicago dataset.

5. Discussion

5.1. The Influence of Auxiliary Branches for Classification Network

The proposed SMCAM introduces auxiliary classification branches that enhance the performance of GradCAM++ by shortening the gradient propagation path and improving the semantic representation of low-level features. This allows for the effective fusion of CAMs at multiple levels. However, the alteration of the original classification network structure and the reliance on image-level annotations for supervised learning raise concerns about the impact on the network's feature learning capabilities. In order to assess this, the classification accuracies of the trunks and branches of the Siamese network are calculated,

and the results are presented in Table 3. The results indicate that the accuracies of all branches and trunks are higher than 0.97, demonstrating that the auxiliary branches do not compromise the training accuracy of the classification network. Specifically, branch 1 has slightly lower training and testing accuracies compared to the other branches and trunks, due to its weaker feature map generalization capabilities, yet it still meets the accuracy requirement.

Table 3. The training and testing accuracies of branches and trunks in classification network.

Position	Subnetwork 1 (**I**)		Subnetwork 2 (**Ĩ**)	
	Training Acc.	**Testing Acc.**	**Training Acc.**	**Testing Acc.**
Branch 1	0.971	0.980	0.975	0.972
Branch 2	0.990	0.993	0.991	0.989
Branch 3	0.992	0.995	0.995	0.995
Branch 4	0.993	0.994	0.992	0.995
Trunk	0.990	0.992	0.992	0.989

As shown in Figure 9, the different levels of CAMs obtained from the branching network are also satisfactory, effectively suppressing the noise of low-level CAMs while highlighting the regions related to the buildings. In summary, the auxiliary branches introduced by SMCAM do not negatively impact the performance of the classification network.

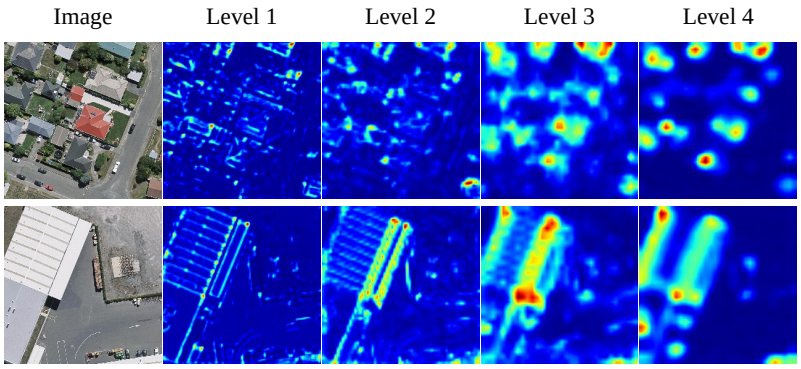


Figure 9. Visulizations of CAMs generated at various levels.

5.2. Comparison of Different Fusion Strategies in SMCAM

The main objective of SMCAM is to attain a fine-structure CAM by aggregating multi-level CAMs generated by GradCAM++. A critical aspect of this challenge is to effectively fuse these CAMs. In this study, we propose a self-attention mechanism for fusing multi-level CAMs by exploiting multi-level features. To evaluate the effectiveness of the proposed self-attention mechanism (denoted as "Attention"), we conduct experiments and compare them with direct addition fusion (denoted as "Addition") in the framework of the proposed method. Additionally, we include the CAM generated from the last convolutional layer (denoted as "Last Conv.") for comparison. The results shown in Table 4 reveal that the self-attention mechanism outperforms Addition, demonstrating its suitability for fusing multi-level CAMs. Additionally, the accuracies of both Attention and Addition are significantly improved compared to Last Conv., implying that utilizing multi-level CAMs enhances the quality of CAMs. The comparison with the results of GradCAM++ listed in Table 1 further highlights the effectiveness of the mutual learning mechanism ($\mathcal{L}_{MIC}$) in the Siamese network in enhancing the quality of the CAM. Furthermore, the improvement in the accuracies of Addition, Attention, and Last Conv. after incorporating SIOM further verifies the effectiveness of SIOM.

Visualizations of the CAMs in this experiment can be seen in Figure 10, with the self-attention method producing more complete and fine-grained CAMs than the others.

To conclude, the proposed self-attention mechanism effectively fuses multi-level CAMs and can be used in conjunction with SIOM to produce better results.

Table 4. The IoU (%), OA (%), and F1 score (%) of different fusion strategies on WHU Dataset.

	IoU	**F1**	**OA**
Last Conv.	48.63	65.44	90.01
Addition	51.73	68.19	92.07
Attention	54.38	70.45	92.21
Last Conv. + SIOM (Ours-SIOM)	51.91	68.34	91.78
Addition + SIOM	56.48	72.19	92.98
Attention + SIOM (Ours)	58.04	73.45	93.29

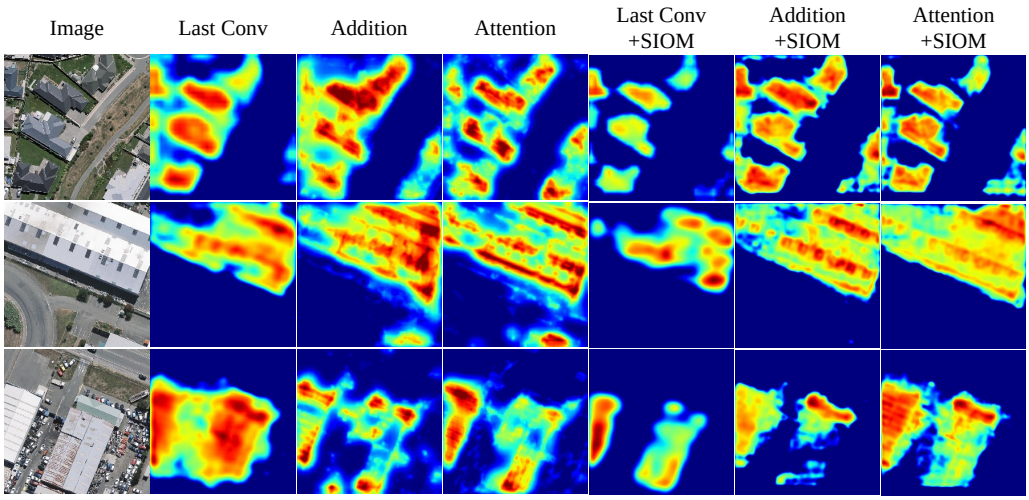


Figure 10. Visulizations of CAMs of different fusion strategies.

5.3. Performance of Hierarchical Feature Optimization in SIOM

The role of SIOM is to utilize the CAM with multi-scale information, which is effective in improving the integrality of the CAM. The core component of SIOM is hierarchical feature optimization, which is a learnable module that leverages the multi-level features to further improve the quality of CAM. It consists of three stages of optimization: the first two stages focus on mining the non-local relationships between the multi-level features to improve the integrity of the CAM, while the third stage leverages the local image information to further improve the fineness of the CAM. To evaluate the efficacy of hierarchical feature optimization, each stage of optimization is analyzed. The pixel correlation module (PCM) in SEAM is also included in the experiments for comparison, which also leverages feature relationships but only considers features from the last two convolutional layers. The experimental results are listed in Table 5, showing that the CAMs generated with hierarchical feature optimization in SIOM are better than those generated in PCM and GradCAM++. Specifically, the results of SIOM are better as the number of stages increases, especially when the third stage of local image information is included, which demonstrates the effectiveness of hierarchical optimization.

Table 5. The IoU (%), OA (%), and F1 score (%) of hierarchical feature optimization settings on WHU dataset.

	IoU	F1	OA
GradCAM++ (4 levels)	49.75	66.44	90.01
SEAM (SMCAM + PCM)	55.17	71.11	92.07
SIOM (Stage 1)	55.69	71.54	92.61
SIOM (Stage 1 + 2)	56.12	71.90	93.11
SIOM (Stage 1 + 2 + 3)	58.04	73.45	93.29

The CAMs produced by each stage of SIOM are displayed in Figure 11, and it is evident that the CAMs generated by SIOM are more comprehensive and clearer than those generated by GradCAM++ and PCM. As the number of stages increases, the clarity and fineness of the CAMs also improve. It is noteworthy that after including the image features in the third stage, the fineness is further enhanced, but the details are magnified to an excessive extent, which may not align with expectations (e.g., the texture of a roof). Despite this, the difference in semantic representation between the layers is still substantial, and, in practical applications, the direct use of image features may require additional processing.

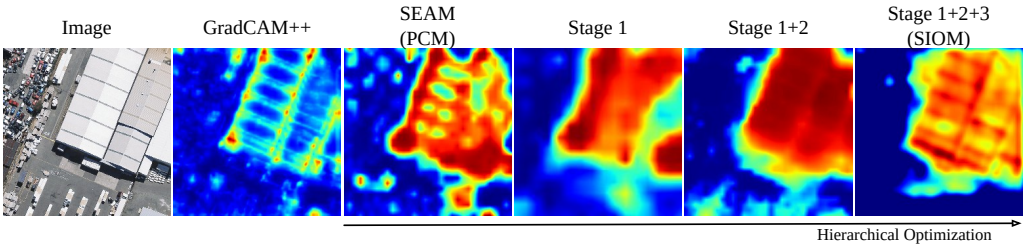
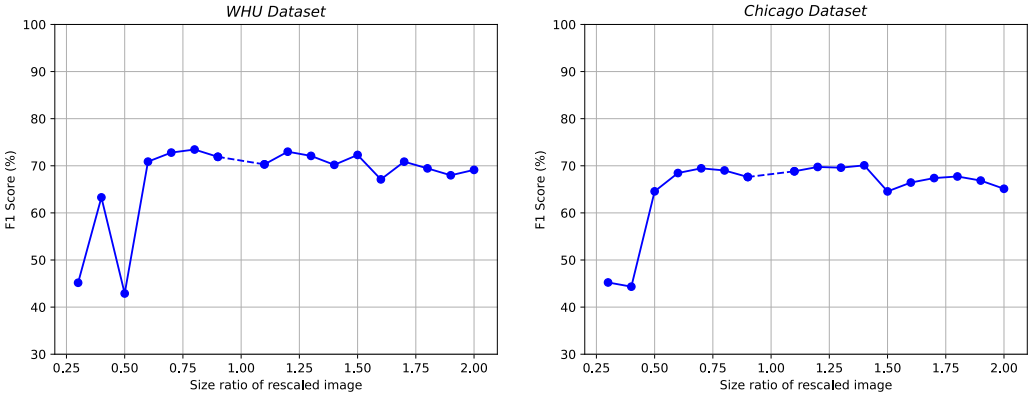


Figure 11. Visulizations of CAMs of different stages settings in hierarchical feature optimization.

5.4. Effect of Scale Setting

The proposed method employs a Siamese network that is trained using both the original image and a rescaled image from a multi-scale representation. It is crucial to discuss the potential impact of different scale settings on the method's performance. To achieve this, experiments are conducted on two datasets by varying the scale of the rescaled image while keeping the original image as the input. The scale of the rescaled image represents the ratio size of the original image. Since the task only involves building extraction, the F1 score is selected as the evaluation metric for these experiments.

The results are depicted in Figure 12. The Siamese network struggles to train effectively when the rescaled image and the original image are of the same size, as indicated by the dashed lines in the figure. From the results of the WHU dataset, the best accuracy is achieved when the scale of the rescaled image is within the range from [0.7, 0.8] to [1.2, 1.3]; secondly, the F1 score decreases and even fails when the scale is less than 0.6, while the F1 score also decreases slowly when the scale is larger than 1.5; thirdly, the F1 score decreases as the scale approaches one, as previously discussed. These observations suggest that the scale difference between the rescaled image and the original image must not be too large or too small. The results from the Chicago dataset are similar, except that the points with the highest scores are 0.8 for the WHU dataset and 0.7 and 1.4 for the Chicago dataset, which highlights the importance of considering dataset bias in the scale setting. In our initial experiments with the same scale setting as SEAM (0.3), the network fails to complete the training and even experiences a gradient explosion, further emphasizing the importance of finding an appropriate scale setting. For this reason, we set the scale ratio of the rescaled image to 0.8 in our experiments.



(a) WHU Dataset (b) Chicago Dataset

Figure 12. F1 score of CAM results under different scale settings on two datasets.

5.5. Limitations and Future Work

From the experimental results and the above analysis, the proposed method stands out among other WSSS methods by not only achieving highly accurate buildings but also ensuring completeness. However, the accuracy of building extraction using the proposed method is still considerably distant from that of the fully supervised segmentation approach. As shown in Figures 6 and 7, the CAM generated by the proposed method can sometimes blur the distinction between buildings and shadows, leading to inaccurate building edges. We hypothesize that this could be attributed to insufficient availability of supervisory information for differentiating buildings from shadows. Additionally, we have observed that our method performs better on the WHU dataset than on the Chicago dataset (Figure 5b), indicating that our approach is sensitive to dataset bias, which can hinder the generalization capability.

To further improve the performance of building extraction, incorporating more supervised information derived from underlying image features or semantic supervised information, such as additional category labels and image restoration information, may be a potential solution. Utilizing these extra supervisions is crucial for effectively and accurately distinguishing between buildings and non-buildings. Additionally, to overcome dataset bias, an unsupervised domain adaption approach using generative adversarial networks could be incorporated, thereby enabling the weakly supervised approach to be applied to a broader range of domains. However, developing specific models and conducting corresponding experiments is left for future work.

6. Conclusions

In this work, we introduce a novel scale-invariant multi-level context aggregation network for the task of weakly supervised building extraction from high-resolution imagery. Our approach focuses on two main aspects: (1) The self-attentive multi-level context aggregation module (SMCAM) is proposed to generate noise-free, fine-structured CAMs by shortening the back-gradient path and utilizing a self-attention mechanism to aggregate multi-level CAMs, allowing the model to learn the contributions of different CAMs for each pixel. (2) The scale-invariant optimization module (SIOM) is designed to narrow the supervision gap between segmentation and classification by leveraging multi-scale representations. This module includes a mutual learning mechanism for pixel-level auxiliary supervision through optimization with a multi-level constraint loss and a hierarchical optimization module to improve the CAM's completeness in a coarse-to-fine manner. The Siamese network integrates these two modules, leading to improved CAM quality in terms of both completeness and fineness, allowing for accurate building extraction results with only image-level labels.

The proposed method is evaluated through two experiments: CAM generation and building extraction. The results of the CAM generation experiment show that the proposed method outperforms state-of-the-art methods in the field of remote sensing and computer vision, as evidenced by its improved performance on both the WHU and Chicago datasets. The building extraction experiment results are similarly impressive, with the proposed method achieving an F1 score of 0.7759, an OA score of 0.9457, and an IoU score of 0.6339 for the WHU dataset, and an F1 score of 0.7411, an OA score of 0.8436, and an IoU score of 0.5887 for the Chicago dataset. These results demonstrate the effectiveness of the proposed method in accurately extracting buildings using only image-level labels.

Additionally, we present an analysis of the proposed modules. Our results indicate that the self-attentive aggregation component in SMCAM and the hierarchical optimization module in SIOM are both reasonable and effective. Furthermore, we examine the influence of the auxiliary classification branches on the network and the scale setting.

In our study, we have observed limitations in the ability of the proposed method to distinguish between shadows and buildings, as well as its sensitivity to dataset bias. In future work, we plan to incorporate additional supervision information to improve shadow differentiation and investigate domain adaptation techniques to mitigate the impact of dataset bias with the aim of achieving enhanced performance.

Author Contributions: Conceptualization, J.W. and X.Y.; methodology, J.W.; validation, X.Y., T.L. and X.G.; formal analysis, J.W. and L.S.; writing—original draft preparation, J.W. and X.Y.; writing—review and editing, L.S., T.L. and X.Y.; visualization, J.W. and X.Y.; funding acquisition, Z.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Open Fund of Key Laboratory of Mine Environmental Monitoring and Improving around Poyang Lake, Ministry of Natural Resources (Grant No. MEMI-2021-2022-09) and partially by the Interdisciplinary Cultivation Fund under Project SWJTU (Southwest Jiaotong University).

Data Availability Statement: The data presented in this study are openly available in the WHU Building Dataset at gpcv.whu.edu.cn/data/building_dataset, accessed on 3 July 2018, and the Inria Aerial Image Labeling Dataset at project.inria.fr/aerialimagelabeling, accessed on 1 January 2017.

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

References

1. Maktav, D.; Erbek, F.; Jürgens, C. Remote Sensing of Urban Areas. *Int. J. Remote Sens.* **2005**, *26*, 655–659. [CrossRef]
2. Tomás, L.; Fonseca, L.; Almeida, C.; Leonardi, F.; Pereira, M. Urban Population Estimation based on Residential Buildings Volume Using IKONOS-2 Images and Lidar Data. *Int. J. Remote Sens.* **2016**, *37*, 1–28. [CrossRef]
3. Li, J.; Huang, X.; Tu, L.; Zhang, T.; Wang, L. A Review of Building Detection from Very High Resolution Optical Remote Sensing Images. *GISci. Remote Sens.* **2022**, *59*, 1199–1225. [CrossRef]
4. Guo, H.; Shi, Q.; Marinoni, A.; Du, B.; Zhang, L. Deep Building Footprint Update Network: A Semi-supervised Method for Updating Existing Building Footprint from Bi-temporal Remote Sensing Images. *Remote Sens. Environ.* **2021**, *264*, 112589. [CrossRef]
5. Haq, M.A. CNN Based Automated Weed Detection System Using UAV Imagery. *Comput. Syst. Sci. Eng.* **2022**, *42*, 837–849.
6. Kim, J.; Lee, M.; Han, H.; Kim, D.; Bae, Y.; Kim, H.S. Case Study: Development of the CNN Model Considering Teleconnection for Spatial Downscaling of Precipitation in a Climate Change Scenario. *Sustainability* **2022**, *14*, 4719. [CrossRef]
7. Haq, M.A.; Khan, M.A.R.; Talal, A. Development of PCCNN-based network intrusion detection system for EDGE computing. *Comput. Mater. Contin.* **2022**, *71*, 1769–1788.
8. Haq, M.A.; Khan, M.A.R. DNNBoT: Deep neural network-based botnet detection and classification. *Cmc-Comput. Mater. Continua* **2022**, *71*, 1729–1750.
9. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention, Munich, German, 21–26 July 2017; pp. 234–241.
10. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
11. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
12. Schuegraf, P.; Bittner, K. Automatic Building Footprint Extraction from Multi-resolution Remote Sensing Images Using a Hybrid FCN. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 191. [CrossRef]

13. Feng, W.; Sui, H.; Hua, L.; Xu, C.; Ma, G.; Huang, W. Building Extraction from VHR Remote Sensing Imagery by Combining an Improved Deep Convolutional Encoder-decoder Architecture and Historical Land Use Vector Map. *Int. J. Remote Sens.* **2020**, *41*, 6595–6617. [CrossRef]
14. Shao, Z.; Tang, P.; Wang, Z.; Saleem, N.; Yam, S.; Sommai, C. BRRNet: A Fully Convolutional Neural Network for Automatic Building Extraction from High-resolution Remote Sensing Images. *Remote Sens.* **2020**, *12*, 1050. [CrossRef]
15. Ma, J.; Wu, L.; Tang, X.; Liu, F.; Zhang, X.; Jiao, L. Building Extraction of Aerial Images by a Global and Multi-scale Encoder-decoder Network. *Remote Sens.* **2020**, *12*, 2350. [CrossRef]
16. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
17. Saleh, F.S.; Aliakbarian, M.S.; Salzmann, M.; Petersson, L.; Álvarez, J.M.; Gould, S. Incorporating Network Built-in Priors in Weakly-Supervised Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1382–1396. [CrossRef] [PubMed]
18. Wei, Y.; Ji, S. Scribble-based Weakly Supervised Deep Learning for Road Surface Extraction from Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–12. [CrossRef]
19. Ahn, J.; Kwak, S. Learning Pixel-Level Semantic Affinity with Image-Level Supervision for Weakly Supervised Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
20. Cao, Y.; Huang, X. A Coarse-to-fine Weakly Supervised Learning Method for Green Plastic Cover Segmentation Using High-resolution Remote Sensing Images. *ISPRS J. Photogramm. Remote Sens.* **2022**, *188*, 157–176. [CrossRef]
21. Zeng, Y.; Zhuge, Y.; Lu, H.; Zhang, L. Joint Learning of Saliency Detection and Weakly Supervised Semantic Segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 Octorber–2 November 2019.
22. Du, Y.; Fu, Z.; Liu, Q.; Wang, Y. Weakly Supervised Semantic Segmentation by Pixel-to-Prototype Contrast. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 4320–4329.
23. Chattopadhay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Generalized Gradient-based Visual Explanations for Deep Convolutional Networks. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Lake Tahoe, NV, USA, 12–15 March 2018; pp. 839–847.
24. Wang, Y.; Zhang, J.; Kan, M.; Shan, S.; Chen, X. Self-supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020.
25. Kwak, S.; Hong, S.; Han, B. Weakly Supervised Semantic Segmentation Using Superpixel Pooling Network. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
26. Durand, T.; Mordan, T.; Thome, N.; Cord, M. WILDCAT: Weakly Supervised Learning of Deep Convnets for Image Classification, Pointwise Localization and Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 642–651.
27. Chan, L.; Hosseini, M.S.; Plataniotis, K.N. A Comprehensive Analysis of Weakly-supervised Semantic Segmentation in Different Image Domains. *Int. J. Comput. Vis.* **2021**, *129*, 361–384. [CrossRef]
28. Chen, J.; He, F.; Zhang, Y.; Sun, G.; Deng, M. SPMF-Net: Weakly Supervised Building Segmentation by Combining Superpixel Pooling and Multi-scale Feature Fusion. *Remote Sens.* **2020**, *12*, 1049. [CrossRef]
29. Yan, X.; Shen, L.; Wang, J.; Deng, X.; Li, Z. MSG-SR-Net: A Weakly Supervised Network Integrating Multiscale Generation and Superpixel Refinement for Building Extraction from High-Resolution Remotely Sensed Imageries. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *15*, 1012–1023. [CrossRef]
30. Li, Z.; Zhang, X.; Xiao, P.; Zheng, Z. On the Effectiveness of Weakly Supervised Semantic Segmentation for Building Extraction from High-resolution Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 3266–3281. [CrossRef]
31. Yan, X.; Shen, L.; Wang, J.; Wang, Y.; Li, Z.; Xu, Z. PANet: Pixelwise Affinity Network for Weakly Supervised Building Extraction From High-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
32. Fang, F.; Zheng, D.; Li, S.; Liu, Y.; Zeng, L.; Zhang, J.; Wan, B. Improved Pseudomasks Generation for Weakly Supervised Building Extraction from High-Resolution Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 1629–1642. [CrossRef]
33. Wang, J.; Shen, L.; Qiao, W.; Dai, Y.; Li, Z. Deep Feature Fusion with Integration of Residual Connection and Attention Model for Classification of VHR Remote Sensing Images. *Remote Sens.* **2019**, *11*, 1617. [CrossRef]
34. Ji, S.; Wei, S.; Lu, M. Fully Convolutional Networks for Multisource Building Extraction from an Open Aerial and Satellite Imagery Data Set. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 574–586. [CrossRef]
35. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Convolutional Neural Networks for Large-scale Remote-sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2016**, *55*, 645–657. [CrossRef]
36. Maltezos, E.; Doulamis, A.; Doulamis, N.; Ioannidis, C. Building Extraction from LiDAR Data Applying Deep Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *16*, 155–159. [CrossRef]

37. Zhang, Z.; Wang, Y. JointNet: A Common Neural Network for Road and Building Extraction. *Remote Sens.* **2019**, *11*, 696. [CrossRef]
38. Hosseinpour, H.; Samadzadegan, F.; Javan, F.D. CMGFNet: A Deep Cross-modal Gated Fusion Network for Building Extraction from Very High-resolution Remote Sensing Images. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 96–115. [CrossRef]
39. Li, Q.; Mou, L.; Hua, Y.; Shi, Y.; Zhu, X.X. CrossGeoNet: A Framework for Building Footprint Generation of Label-Scarce Geographical Regions. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *111*, 102824. [CrossRef]
40. Yuan, W.; Wang, J.; Xu, W. Shift Pooling PSPNet: Rethinking PSPNet for Building Extraction in Remote Sensing Images from Entire Local Feature Pooling. *Remote Sens.* **2022**, *14*, 4889. [CrossRef]
41. Chen, M.; Wu, J.; Liu, L.; Zhao, W.; Tian, F.; Shen, Q.; Zhao, B.; Du, R. DR-Net: An Improved Network for Building Extraction from High Resolution Remote Sensing Image. *Remote Sens.* **2021**, *13*, 294. [CrossRef]
42. Guo, H.; Du, B.; Zhang, L.; Su, X. A Coarse-to-fine Boundary Refinement Network for Building Footprint Extraction from Remote Sensing Imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *183*, 240–252. [CrossRef]
43. Shu, Q.; Pan, J.; Zhang, Z.; Wang, M. MTCNet: Multitask Consistency Network with Single Temporal Supervision for Semi-supervised Building Change Detection. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *115*, 103110. [CrossRef]
44. Xia, L.; Zhang, X.; Zhang, J.; Yang, H.; Chen, T. Building Extraction from Very-high-resolution Remote Sensing Images Using Semi-supervised Semantic Edge Detection. *Remote Sens.* **2021**, *13*, 2187. [CrossRef]
45. Zheng, Y.; Yang, M.; Wang, M.; Qian, X.; Yang, R.; Zhang, X.; Dong, W. Semi-supervised Adversarial Semantic Segmentation Network Using Transformer and Multiscale Convolution for High-resolution Remote Sensing Imagery. *Remote Sens.* **2022**, *14*, 1786. [CrossRef]
46. Li, Q.; Shi, Y.; Zhu, X.X. Semi-supervised Building Footprint Generation with Feature and Output Consistency Training. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–17.
47. Meng, Y.; Chen, S.; Liu, Y.; Li, L.; Zhang, Z.; Ke, T.; Hu, X. Unsupervised Building Extraction from Multimodal Aerial Data Based on Accurate Vegetation Removal and Image Feature Consistency Constraint. *Remote Sens.* **2022**, *14*, 1912. [CrossRef]
48. Chen, H.; Cheng, L.; Zhuang, Q.; Zhang, K.; Li, N.; Liu, L.; Duan, Z. Structure-aware Weakly Supervised Network for Building Extraction from Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–12. [CrossRef]
49. Wei, Y.; Feng, J.; Liang, X.; Cheng, M.M.; Zhao, Y.; Yan, S. Object Region Mining with Adversarial Erasing: A Simple Classification to Semantic Segmentation Approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1568–1576.
50. Lee, J.; Kim, E.; Lee, S.; Lee, J.; Yoon, S. FickleNet: Weakly and Semi-supervised Semantic Image Segmentation Using Stochastic Inference. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5267–5276.
51. Jiang, P.T.; Yang, Y.; Hou, Q.; Wei, Y. L2G: A Simple Local-to-global Knowledge Transfer Framework for Weakly Supervised Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 16886–16896.
52. Kolesnikov, A.; Lampert, C.H. Seed, Expand and Constrain: Three Principles for Weakly-supervised Image Segmentation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 695–711.
53. Huang, Z.; Wang, X.; Wang, J.; Liu, W.; Wang, J. Weakly-supervised Semantic Segmentation Network with Deep Seeded Region Growing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7014–7023.
54. Ahn, J.; Cho, S.; Kwak, S. Weakly Supervised Learning of Instance Segmentation with Inter-pixel Relations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2209–2218.
55. Chang, Y.T.; Wang, Q.; Hung, W.C.; Piramuthu, R.; Tsai, Y.H.; Yang, M.H. Weakly-supervised Semantic Segmentation via Sub-category Exploration. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 8991–9000.
56. Fu, K.; Lu, W.; Diao, W.; Yan, M.; Sun, H.; Zhang, Y.; Sun, X. WSF-NET: Weakly Supervised Feature-fusion Network for Binary Segmentation in Remote Sensing Image. *Remote Sens.* **2018**, *10*, 1970. [CrossRef]
57. 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.
58. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
59. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
60. Su, H.; Jampani, V.; Sun, D.; Gallo, O.; Learned-Miller, E.; Kautz, J. Pixel-adaptive Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11166–11175.
61. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Can Semantic Labeling Methods Generalize to Any City? The Inria Aerial Image Labeling Benchmark. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium, Fort Worth, TX, USA, 23–28 July 2017; pp. 3226–3229.

62. Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
63. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. SLIC Superpixels Compared to State-of-the-art Superpixel Methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2282. [CrossRef] [PubMed]