




Article

A Hybrid Attention-Aware Fusion Network (HAFNet) for Building Extraction from High-Resolution Imagery and LiDAR Data

Peng Zhang ^{1,2}, Peijun Du ^{1,2,*} , Cong Lin ^{1,2}, Xin Wang ^{1,2} , Erzhu Li ³, Zhaohui Xue ⁴  and Xuyu Bai ^{1,2}

¹ Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, Key Laboratory for Land Satellite Remote Sensing Applications of Ministry of Natural Resources, School of Geography and Ocean Science, Nanjing University, Nanjing 210023, China; pzhangrs@smail.nju.edu.cn (P.Z.); dg1727017@smail.nju.edu.cn (C.L.); wang.xin@smail.nju.edu.cn (X.W.); dg1727001@smail.nju.edu.cn (X.B.)

² Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China

³ School of Geography, Geomatics and Planning, Jiangsu Normal University, Xuzhou 221116, China; liezrs2018@jsnu.edu.cn

⁴ School of Earth Sciences and Engineering, Hohai University, Nanjing 211100, China; zhaohui.xue@hhu.edu.cn

* Correspondence: peijun@nju.edu.cn; Tel.: +86-159-0515-9291

Received: 24 September 2020; Accepted: 13 November 2020; Published: 16 November 2020



Abstract: Automated extraction of buildings from earth observation (EO) data has long been a fundamental but challenging research topic. Combining data from different modalities (e.g., high-resolution imagery (HRI) and light detection and ranging (LiDAR) data) has shown great potential in building extraction. Recent studies have examined the role that deep learning (DL) could play in both multimodal data fusion and urban object extraction. However, DL-based multimodal fusion networks may encounter the following limitations: (1) the individual modal and cross-modal features, which we consider both useful and important for final prediction, cannot be sufficiently learned and utilized and (2) the multimodal features are fused by a simple summation or concatenation, which appears ambiguous in selecting cross-modal complementary information. In this paper, we address these two limitations by proposing a hybrid attention-aware fusion network (HAFNet) for building extraction. It consists of RGB-specific, digital surface model (DSM)-specific, and cross-modal streams to sufficiently learn and utilize both individual modal and cross-modal features. Furthermore, an attention-aware multimodal fusion block (Att-MFBlock) was introduced to overcome the fusion problem by adaptively selecting and combining complementary features from each modality. Extensive experiments conducted on two publicly available datasets demonstrated the effectiveness of the proposed HAFNet for building extraction.

Keywords: building extraction; high-resolution imagery (HRI); light detection and ranging (LiDAR); multimodal data fusion; deep learning; attention mechanism

1. Introduction

Accurate building information extracted from earth observation (EO) data is essential for a wide range of urban applications, such as three-dimensional modeling, infrastructure planning, and urban expansion analysis. Since high-resolution imagery (HRI) became more accessible and affordable, extracting buildings from HRI has been of great interest. HRI provides valuable spectral, geometric, and texture information that are useful to distinguish buildings from non-building objects. However,

building extraction from HRI is still challenging due to the large intra-class and low inter-class variation of building objects [1], shadow effect, and relief displacement of high buildings [2]. Airborne light detection and ranging (LiDAR) technology provides a promising alternative for extracting buildings. Compared with optical sensors, LiDAR measurements are not influenced by shadows, and offer height information of the land surface which can help to separate buildings from other manmade objects (e.g., roads and squares). Nevertheless, LiDAR-based building extraction methods are limited due to the lack of texture and boundary information [3].

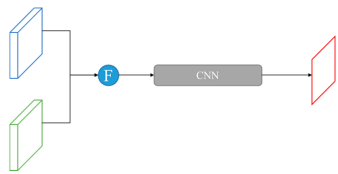
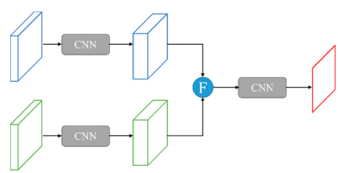
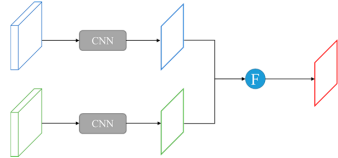
A lot of fusion techniques have been developed to integrate HRI and LiDAR data for building extraction and have been shown to perform better than using a single modality. In the early stage, the methods were relatively simple, in which elevation and intensity information derived from LiDAR data were stacked with HRI bands to obtain more informative images [4,5]. However, a simple concatenation of raw data may not be powerful enough to separate the classes of interest [6]. Pixel-wise prediction requires more discriminative feature representation. To address this issue, a set of hand-crafted features (e.g., spectral, shape, height, and textural features) [7,8] were extracted and fed into the supervised classifier(s) using advanced machine learning algorithms (e.g., support vector machine, ensemble learning, or active and semi-supervised learning) [9]. Although these feature-level fusion methods perform much better than using simply stacked raw data, they still have some limitations. Firstly, these methods require careful engineering and remarkable expert knowledge to design feature extractors, which are always goal-specific and not allowed to be directly applied from one specific task to another. Meanwhile, due to the heterogeneity gap existing in multimodal data, hand-crafted low-level or middle-level features derived from HRI and LiDAR data are usually in unequal subspaces, making vector representations associated with similar semantics completely different. A higher level of abstract features will be helpful to narrow the heterogeneity gap.

In recent years, deep learning (DL) has become the fastest-growing trend in big data analysis. DL models, especially the convolutional neural networks (CNNs), achieved significant improvement in RS image analysis tasks including scene classification [10–12], land use and land cover (LULC) classification [13,14], and urban object extraction [15–17]. The advantage of CNNs is that hierarchical deep features from a low-level to high-level can be automatically extracted via a common end-to-end learning process, instead of manually designing or handcrafting features [18]. These high-level features can help to bridge the heterogeneity gap between different data modalities at the feature level. The fully convolutional network (FCN) [19] has been shown to perform well on urban object extraction due to the ability of pixel-wise labeling [1,20,21]. However, these advanced CNNs and FCNs are mostly restrained to three-channel RGB images, which cannot be directly adopted to multimodal RS data. It is necessary to build a comprehensive DL-based model to integrate HRI and LiDAR data for more accurate and robust building extraction.

Building a DL-based multimodal fusion network has two architectural design choices: “where” and “how” to fuse different modalities. Based on “where” to fuse different modalities, current fusion methods can be categorized into data-level, feature-level, and decision-level fusion. Data-level fusion methods combine data from each modality as a single input of the model. A common way of data-level fusion in a deep network is to concatenate different data sources into a single data cube to be processed [1,22]. Feature-level fusion methods integrate features obtained from each modality at the feature-learning stage: two networks are trained in parallel to learn features of different modalities, and their activations are then fused into one stream, e.g., by feature concatenation or element-wise summation [23–27]. Decision-level fusion methods perform integrations of different outputs. In the case of classification, different models predict the classes, and their predictions are then fused, e.g., by averaging or majority-voting [28,29]. “How” to fuse different modalities refers to how to construct a fusion operation that can merge representations of different modalities, thereby forcing the network to learn a joint representation. Feature concatenation and element-wise summation are the most commonly used fusion operations.

Table 1 shows the architectures, advantages, and limitations of current DL-based fusion methods. Especially, there are several limitations that should be properly addressed. Firstly, these methods cannot fully leverage all kinds of useful features (e.g., individual modal and cross-modal), which may lead to insufficient feature learning and unsatisfied building extraction results. Individual modal features refer to the features derived from a single modality, and the cross-modal features are newly learned by fusing features from different modalities. Benefiting from the complementary information provided by all modalities, cross-modal features are considered to be more discriminative and representative than individual features. However, some useful individual modal information which is helpful for building extraction will be inevitably lost after the fusion operations. Therefore, a reliable and robust building extraction result requires sufficient learning and utilizing not only cross-modal features but also individual modal features. As a matter of fact, neither of the current fusion methods can sufficiently leverage the two types of features due to their architecture designs (cf. Table 1). Secondly, the most commonly used fusion operations, feature concatenation and element-wise summation may yield fusion ambiguity in the scenarios of the feature redundancy and noises among different modalities. When only one modality carries discriminative information while the counterpart one provides only some useful or even misleading information, negative features will be inevitably introduced via simple feature concatenation or summation operation. For example, when distinguishing buildings from trees with similar height, the features derived from LiDAR data may carry misleading information. A simple concatenation or summation of the two modalities will inevitably introduce negative height information. In this case, the features of HRI should be highlighted and the features of LiDAR data should be suppressed. A selection module is required to adaptively highlight the discriminative features and suppress the irrelevant features for more effective multimodal fusion.

Table 1. Current DL-based fusion methods.

Fusion Level	Architectures	Advantages	Limitations
Data-level		Relatively simple and easily implemented	Ignore the qualitative distinction of different modalities; cannot be initialized with the pre-trained CNN
Feature-level		Be capable of learning cross-modal features	Fail to fully exploit individual modal features
Decision-level		Perform well on learning individual modal features	Lack of sufficient learning of cross-modal features

To fully leverage both individual modal and cross-modal features, a novel hybrid fusion architecture is proposed to make full use of cross-modal and individual modal features. In this hybrid fusion architecture, two streams are employed to learn individual features from HRI and LiDAR data, respectively. Another stream is specially designed to explore cross-modal complements by feature-level fusion of the other two streams. At the decision stage, the predictions from the three streams are also combined to produce a comprehensive building extraction result. Thus, cross-modal features can be sufficiently learned by the third stream and individual features are preserved and

contribute to the final prediction of buildings. Compared with traditional fusion networks, the proposed hybrid network can fully leverage the benefits from both individual modal and cross-modal features, aiming to obtain more robust and reliable building extraction results.

To overcome the cross-modal fusion ambiguity resulting from simple concatenation or summation, an attention-aware multimodal fusion block (Att-MFBlock) is introduced to adaptively highlight the discriminative features and suppress the irrelevant features. Attention mechanism refers to the strategy of highlighting the most pertinent piece of information instead of paying equal attention to all available information. Channel-wise attention [30] was demonstrated to improve the representational power by explicitly modeling the interdependencies between the channels of its convolutional features. Recently some researchers [31,32] attempted to exploit the potential of channel-wise attention in combining multimodal features and demonstrate its ability to select complements from each modality. Motivated by these works, we proposed an attention-aware multimodal fusion block (Att-AFBlock) that extends the core idea of channel-wise attention to boost the fusion efficiency and sufficiency of HRI and LiDAR data.

The main contributions of this study are: (1) proposing a novel hybrid attention-aware fusion network (HAFNet) that could be used to extract buildings from HRI and LiDAR data, (2) analyzing how the proposed hybrid fusion architecture and attention-aware multimodal fusion block (Att-AFBlock) affect multimodal data fusion and building extraction results, and (3) comparing the proposed HAFNet with other classical fusion models in two public urban datasets.

This paper is organized as follows. Section 2 discusses related work. Section 3 describes the details of the proposed HAFNet. Experiment design is presented in Section 4. Results and discussions are presented in Section 5. Finally, the main conclusions of the study are summarized in Section 6.

2. Related Work

2.1. Fully Convolutional Networks (FCNs) for Semantic Labeling

Semantic labeling of RS images relates to the pixel-wise classification of images, which is called “semantic segmentation” in the computer vision community. Building extraction from RS data is essentially a semantic labeling task that assigns a label of “building” or “non-building” to every pixel of images. Traditional CNNs were generated for image-level classification, which requires fully connected layers at the end of the network to output a vector of the class score. By replacing the fully connected layers with convolutional layers, the fully convolutional network (FCN) was introduced to achieve a dense classification result at 1:8 resolution. Based on the FCN architecture, many models (e.g., SegNet, U-Net, and DeepLab) were proposed to perform pixel-wise predictions at 1:1 resolution by using a bottleneck architecture in which the feature maps are up-sampled to match the original input resolution. SegNet [33] used pooling indices computed in the max-pooling step to perform non-linear up-sampling, which was demonstrated as effective for precise relocalization of features [34]. U-Net [35] combined feature maps from different levels to produce finer segmentation results by skip-connections. The DeepLab networks, including DeepLab v1 [36], DeepLab v2 [37], DeepLab v3 [38], and DeepLab v3+ [39], used atrous convolutions to increase the receptive field and an atrous spatial pyramid pooling (ASPP) module to learn multi-scale contextual information. On the other hand, these semantic segmentation models have also been applied in remote sensing image processing tasks, such as pixel-wise classification and object extraction. Audebert et al. [23,34] proposed a SegNet-based fusion network for semantic labeling using high-resolution aerial images and LiDAR data and achieved state-of-the-art performance. Based on U-Net, Guo et al. [40] added an attention block to improve extraction building accuracy; Wagner et al. [41] proposed an instance segmentation model for building extraction, which consisted of identifying each instance of each object at the pixel level. Lin et al. [42] introduced a nested SE-DeepLab model for road extraction from very-high-resolution remote sensing images.

These FCNs often have an encoder-decoder architecture. The encoder part aims to exploit multi-scale deep features by using a deep CNN (e.g., VGG-16, ResNet-50) that consists of a series of convolutional layers, non-linear activation layers, and pooling layers. By using up-sampling layers, the decoder part reconstructs the spatial resolution of features to match the original input resolution. Take SegNet as an example. As illustrated in Figure 1, the encoder part of SegNet is composed of five convolutional blocks, each of which consists of two or three basic convolutional units and one pooling layer of stride two in the end. A basic convolutional unit includes a convolutional layer of kernel 3×3 with a padding of one, a batch normalization layer, and a non-linear activation layer (ReLU). These 13 convolutional layers correspond to the first 13 convolutional layers in the VGG16 network [43]. For the encoder part, its structure is symmetrical with respect to the encoder. Pooling layers are replaced by up-sampling layers to restore the full resolution of the input image. It is worth noting that the up-sampling operations reuse pooling indices restored in the pooling layers of the corresponding encoder part. This eliminates the need for learning to up-sample and is particularly effective on misplaced small objects [23,33].

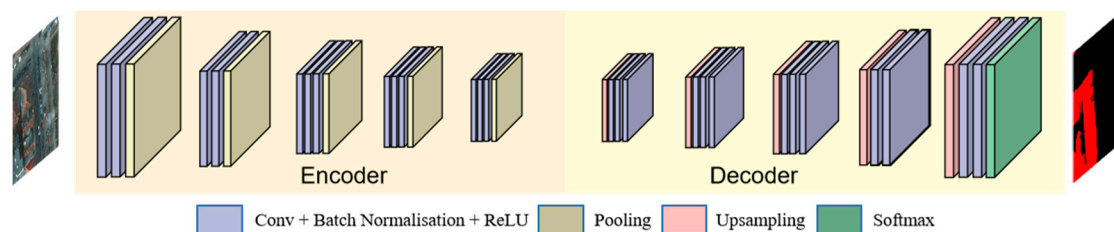


Figure 1. SegNet architecture for building extraction of EO data [23,33].

2.2. Attention Mechanism

The original idea of attention mechanisms used in natural language processing (NLP) was to generate a context vector that can assign weights on the input sequence, thus highlighting the salient feature of the sequence while suppressing the irrelevant counter-parts, obtaining a more contextualized prediction [44]. Since attention mechanisms achieved significant progress in the field of NLP [45], it has become an effective tool to extract the most useful information of the input signal by weighting the activations channel-wisely or spatially [46–51]. Hu et al. [30] proposed a channel-wise attention unit, namely the Squeeze-and-Excitation (SE) block, that adaptively highlights informative features and suppresses irrelevant features among the channels. SE blocks is a computationally efficient unit and can produce significant performance improvements for existing state-of-the-art deep architectures (e.g., Inception and ResNet). Wang et al. [49] argued that capturing long-range dependencies is of great importance in deep neural networks, while convolutional and recurrent operations only process a local neighborhood. To overcome the limitations that exist in neural networks, they proposed non-local operations that compute the response at a position as a weighted sum of the features at all positions. In the RS community, attention mechanisms have also been used in various image analysis tasks. Jin et al. [52] used a global-spatial-context attention module to detect shadows in aerial imagery. Tian et al. [53] introduced an attention mechanism for object detection and found that attention mechanism was capable of enhancing the features of the object regions while reducing the influence of the background. Li et al. [54] proposed a multiscale self-adaptive attention network for scene classification and the attention mechanism was used to adaptively select useful information by learning the weights in different channels.

The above-mentioned attention-based methods are mostly based on individual modality. Recently some researchers attempted to exploit the potential of channel-wise attention in combining multimodal features. Chen and Li. [31] introduced the channel-wise attention mechanism to handle the cross-modal cross-level fusion problem for RGB-D salient object detection. Mohla et al. [32] proposed a dual attention-based fusion network for hyperspectral and LiDAR classification and achieved state-of-the-art

classification performance. Our study aims to handle the HRI-LiDAR fusion ambiguity by using a modality and channel-wise attention module.

3. Hybrid Attention-Aware Fusion Network (HAFNet)

Here we develop a hybrid attention-aware fusion network (HAFNet) for building extraction from HRI and LiDAR data. The proposed network is based on a novel hybrid fusion architecture that can make full use of both cross-modal and individual modal features from multimodal data. A new attention-aware multimodal fusion block (Att-AFBlock) is introduced to boost fusion efficiency and sufficiency. We first introduce the overall architecture of HAFNet, and then describe Att-AFBlock in details.

3.1. Network Architecture

The overall architecture of the proposed HAFNet is shown in Figure 2. Red (R), green (G), and blue (B) bands from HRI and the LiDAR-derived digital surface model (DSM) are used as the input data of this network. Among many available FCN-based networks, the SegNet was selected as the basic network due to the following points. Firstly, SegNet has been demonstrated as effective for precise relocalization of features [34], which is of critical importance for building extraction. Secondly, SegNet can provide a good balance between accuracy and computational cost. Therefore, other state-of-the-art semantic segmentation models are not demonstrated in this paper. It should be noted that our contribution can be easily adapted to other networks rather than limited to SegNet.

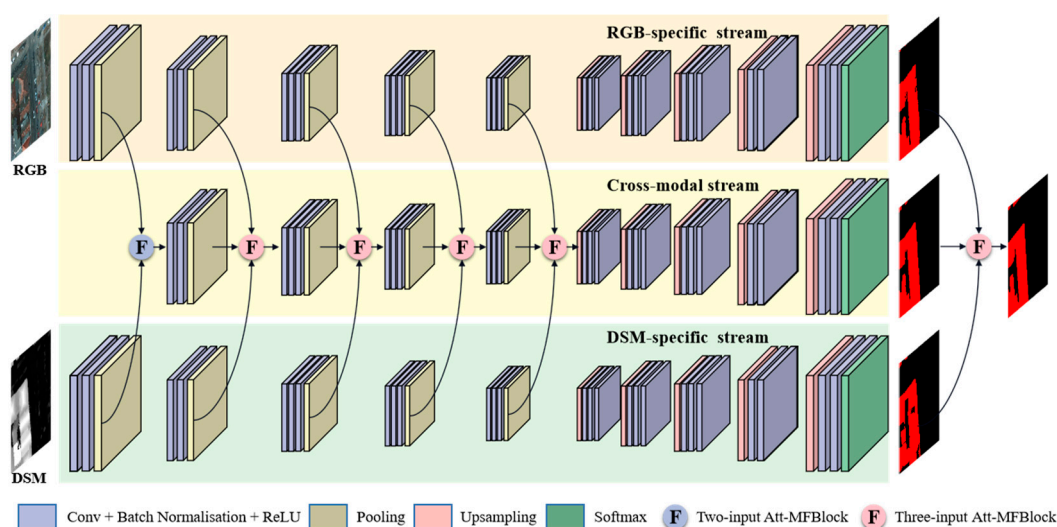


Figure 2. An overview of the hybrid attention-aware fusion network (HAFNet).

The HAFNet consists of three streams, i.e., RGB-specific, DSM-specific, and cross-modal streams. RGB-specific and DSM-specific streams are based on two parallel SegNet, aiming to learn individual modal features of RGB data and DSM data, respectively. The cross-modal stream is similar to the other two streams, while it cuts off the first convolution block. This stream is used to combine the activations from RGB-specific and DSM-specific streams for exploring cross-modal complements. Given that the current fusion networks are mostly based on two streams, the reasons for designing the three-stream architecture should be explained. First, (Chen et al.) [31] demonstrated that adding a specific stream to combine different modalities at the early stage helps to learn more discriminative cross-modal features. Second, in the decision stage, the predictions of three streams will be combined for a comprehensive building extraction result. Only using this three-stream architecture can ensure that both the individual modal features and cross-modal features have been sufficiently learned and utilized for final prediction.

3.2. Attention-Aware Multimodal Fusion Block

Note that the fusion operation used in this architecture is a newly designed fusion module, rather than commonly used feature concatenation or element-wise summation. The fusion module is based on an attention-aware multimodal fusion block (Att-MFBlock) (Figure 3). The purpose of the proposed Att-MFBlock is to adaptively reweight feature channels from different modalities to highlight the discriminative features and suppress the irrelevant features. Inspired by the channel-wise attention mechanism (e.g., SE block) that produce significant performance improvements for existing state-of-the-art deep architectures, we aim to explore how the individual modality-based channel-wise attention mechanism can be extended to multimodal data fusion. As illustrated in Figure 2, most of the fusion blocks have three inputs, including contributions of RGB-specific, DSM-specific, and cross-modal streams, while the first fusion block has only two inputs. For the convenience of expression, we take the three-input fusion block as an example. The Att-MFBlock is a computational unit that can be formulated as:

$$X_F = F_{\text{Att-MFB}}(X_1, X_2, X_3) = \sum_{m=1}^3 X_m \cdot \theta_m. \quad (1)$$

here, $X_1, X_2, X_3 \in \mathbf{R}^{H \times W \times C}$ denotes input features from RGB-specific, DSM-specific, and cross-modal streams, respectively. $X_F \in \mathbf{R}^{H \times W \times C}$ denotes the output fused features. The Att-MFBlock aims to learn a 1D channel-wise weight $\theta_m \in \mathbf{R}^{C \times 1 \times 1}$ of m -th modality, and then multiply it to the corresponding modality. As illustrated in Figure 3, the Att-MFBlock consists of two main components: channel-wise global pooling and modality and channel-wise relationship modeling.

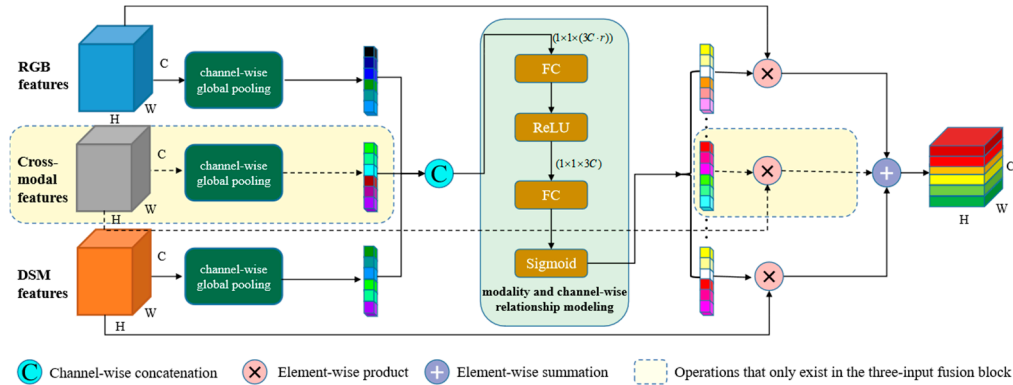


Figure 3. Att-MFBlock: attention-aware multimodal Fusion Block. Different colors of the vector elements indicate the different values of the channel-wise statistics or learned weights of each modality.

(1) Channel-wise global pooling

In order to capture channel dependencies within different modalities, it is necessary to obtain the feature descriptors of different channels first, and then model the relationship between them. The traditional convolution method can only perform operations on the features from different channels in the limited receptive field, while ignoring the contextual information outside the local region. This issue becomes more severe in the lower layers of a network which have relatively small receptive field sizes. To tackle this issue, a global average pooling layer is employed to abstract global spatial information of each channel. Formally, the channel-wise statistics $P \in \mathbf{R}^C$ are calculated by:

$$P_m^c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W X_m^c(i, j). \quad (2)$$

here X_m^c denotes input features of c -th channel of m -th modality, and the W and H are the width and height of X_m^c .

(2) Modality and channel-wise relationship modeling

After capturing the global channel-wise information, a second operation is needed to fully model the relationships between each channel of different modalities. To achieve the desired performance, the function must be flexible and capable of learning a nonlinear interaction between channels. A gating mechanism with a sigmoid activation proposed by [30] was employed:

$$\theta = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 P)). \quad (3)$$

here $P_{con} = [[P_1^1, P_1^2, \dots, P_1^C], [P_2^1, P_2^2, \dots, P_2^C], [P_3^1, P_3^2, \dots, P_3^C]] \in \mathbf{R}^{3C}$ denotes the concatenated channel-wise statistics of the three modalities, where P_m^c denotes the statistic of the c -th channel of the m -th modality. $\theta = [[\theta_1^1, \theta_1^2, \dots, \theta_1^C], [\theta_2^1, \theta_2^2, \dots, \theta_2^C], [\theta_3^1, \theta_3^2, \dots, \theta_3^C]] \in \mathbf{R}^{3C}$, where θ_m^c denotes the learned weights of the c -th channel of the m -th modality. σ , δ refer to the Sigmoid, ReLU function, respectively. $\mathbf{W}_1 \in \mathbf{R}^{(3C \cdot r) \times 3C}$ and $\mathbf{W}_2 \in \mathbf{R}^{3C \times (3C \cdot r)}$ refer to the parameters of two fully connected layers. r is a manually setting ratio which allows us to vary the capacity and computational cost of the model. For example, if r is set too high, the model will gain enough capacity to learn the relationship between channels while requiring massive computational costs. In order to find the best parameter, we performed preliminary experiments with different values of r . At the feature-level fusion stage, given the dimension of concatenated input channels was too high, r was set to 1/4, 1/8, 1/16, and 1/32 in order to reduce the model complexity; while at the decision-level fusion stage, r was set to 4, 8, 16, and 32 to enable the model to fully explore the relationships between decisions of different modalities. Results reported that the model achieved a good tradeoff between accuracy and complexity when r was set to 16 and 1/16.

Through the modality and channel-wise relationship modeling block, the channel-wise weights of each modality are adaptively learned and multiplied by the corresponding input features. At last, final fused features are obtained by an element-wise summation of all recalibrated features from different modalities:

$$X_F = \sum_{m=1}^3 X_m \cdot \theta_m, \quad (4)$$

where θ_m denotes the channel-wise weights of m -th modality, respectively.

4. Experiment Design

4.1. Datasets

Experimental evaluations were conducted on the two datasets of the ISPRS 2D Semantic Labeling Challenge, which can be freely downloaded from the ISPRS official website (<http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html>). Both datasets are comprised of high-resolution true orthophoto (TOP) tiles and corresponding LiDAR-derived DSM data with a resolution of 5 cm that cover two cities in Germany: Potsdam and Vaihingen. The ground truth (GT) contains the major classes of impervious surfaces, buildings, low vegetation, trees, cars, and clutter/background. In this paper, only the ground truth of buildings is needed. The two datasets were manually divided into training and testing parts by the contest organizers. The Potsdam dataset contains 38 images in total, of which 24 are used as the training set and the remaining as the test set. There are 33 images in Vaihingen dataset. Sixteen of them are used as the training set and the remaining for the test.

4.2. Accuracy Assessment

Three commonly used metrics, namely the overall accuracy (OA), F1 score, and mean intersect over union (IoU), were used to access the quantitative performance of each method. OA is the ratio of

the number of the correctly labeled pixels to the total number of the whole image. F1 score is computed as the harmonic mean between precision and recall:

$$precision = \frac{tp}{tp + fp} \quad (5)$$

$$recall = \frac{tp}{tp + fn} \quad (6)$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (7)$$

where tp , fp , and fn refer to the true positive, false positive, and false negative, respectively, which can be calculated by the confusion matrix.

IoU was used to measure the overlap rate of the detected building pixels and labeled building pixels:

$$IoU = \frac{target \cap detected}{target \cup detected} \quad (8)$$

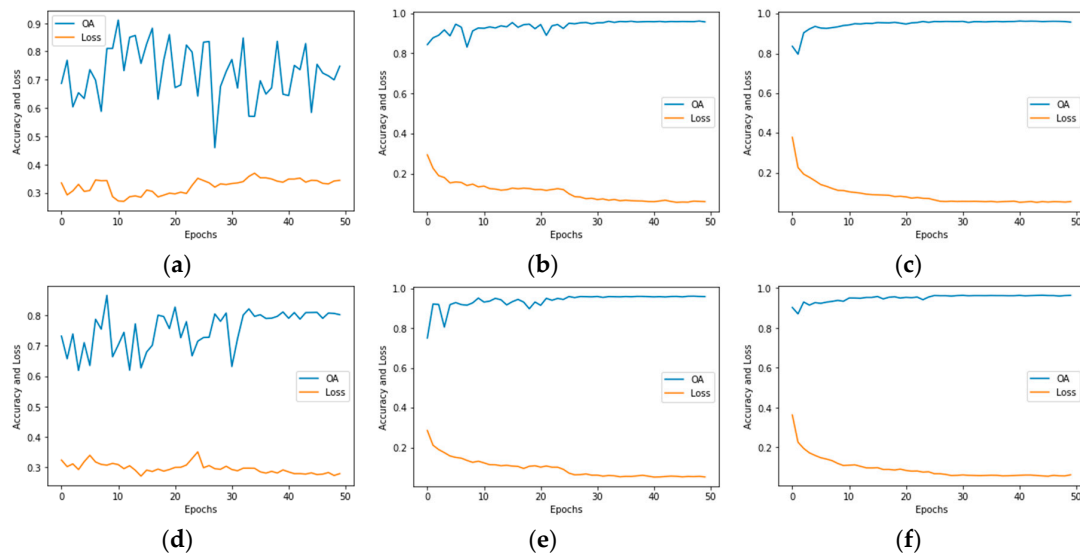
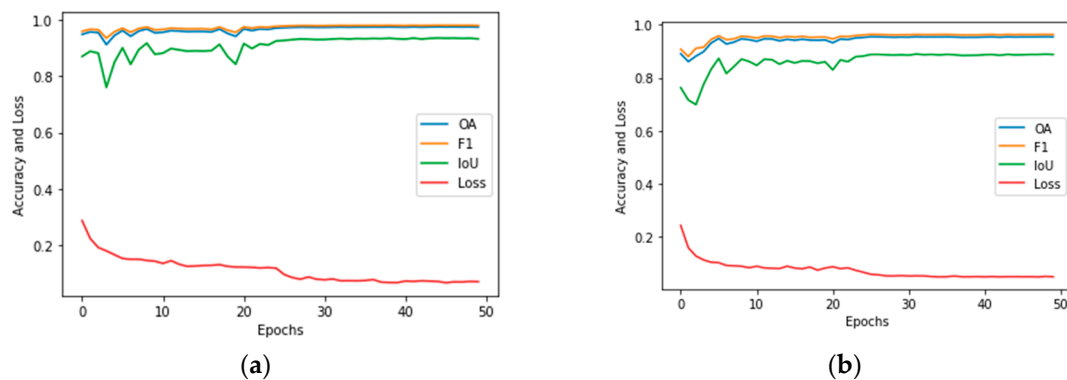
4.3. Method Implementation

Due to the limitations of GPU memory, the original images cannot be directly processed in our deep network. In this work, a sliding window approach was used to clip the original images into 128×128 patches and test the patches one by one from the top left to the bottom right. The stride of the sliding window refers to the size of the overlapping region between two consecutive patches. A smaller stride 64 was used to generate more training patches. Besides, each image patch was rotated every 90 degrees and flipped in both the horizontal and vertical directions to produce eight augmented patches. These data augmentation methods were employed to avoid over-fitting.

The proposed HAFNet was implemented using the PyTorch framework with an NVIDIA GeForce RTX 2080Ti GPU (11 GB memory). The model was trained with the cross-entropy loss function and optimized with the Stochastic Gradient Descent (SGD) optimizer and a batch size of 10. For the encoder of RGB-specific streams, the weights were initialized with the pre-trained VGG-16 model. For any other part of the network, the weights were randomly initialized using the policy from [55]. Audebert et al. [34] proposed a SegNet-based fusion network for semantic segmentation of HRI and LiDAR data, and experimental results showed that the model achieved the highest accuracy when the learning rate of the pre-trained part was one-tenth of that of the randomly initialized part. In order to search the optimal parameters, we experimented with different learning rates for the encoder of RGB-specific stream (l_{e_RGB}) and other parts (l_{other}) with two strategies (i.e., $\frac{l_{other}}{l_{e_RGB}} = 1$ or $\frac{l_{other}}{l_{e_RGB}} = 10$) in a validation set (thirty percent of the training set were randomly selected as the validation set). The experimental results (Table 2) showed that the model performed better when $\frac{l_{other}}{l_{e_RGB}}$ was set to 10. It indicated that a lower learning rate on the pre-trained part could enable the model to achieve higher accuracy. Otherwise, when the learning rates were set too high (i.e., $l_{other} = 0.1$), the model achieved the lowest accuracy on the validation set. The loss and accuracy of the model experienced sharp fluctuations and failed to converge (see Figure 4a,d). When the learning rates were set low (i.e., $l_{other} = 0.01$ or $l_{other} = 0.001$), the model possessed better accuracy, convergence, and stability. The model achieved the highest accuracy when $l_{e_RGB} = 0.001$ and $l_{other} = 0.01$. Therefore, we finally determined l_{e_RGB} , l_{other} as 0.001 and 0.01 to train the proposed MAFNet in the whole training sets. There were 50 epochs during the training and each epoch had 1000 iterations. A multi-step method was applied here to adjust the learning rate, which can be divided by a factor of 10 after 25, 35, and 45 epochs [23]. The speed of training was about 4 min per epoch, so it needed 200 min to complete the whole training process (50 epochs). During the test time, the speed of producing building extraction results was about 0.07 s per patch (128×128 size image). The changing accuracies and losses of the two datasets with the increasing epochs are illustrated in Figure 5.

Table 2. Statistical results on the validation set with different learning rates.

$\frac{l_{other}}{l_{e_{RGB}}}$	$l_{e_{RGB}}$	l_{other}	OA (%)	F1 Score (%)	IoU (%)
1	0.1	0.1	90.15	92.51	74.85
10	0.01	0.1	92.33	94.03	80.65
1	0.01	0.01	96.90	97.61	91.54
10	0.001	0.01	97.15	97.79	92.27
1	0.001	0.001	96.60	97.39	90.71
10	0.0001	0.001	96.83	97.55	91.35

**Figure 4.** Plots showing the overall accuracy (OA) and the loss of the proposed MAFNet using different learning rates. (a) $(l_{e_{RGB}}, l_{other}) = (0.1, 0.1)$, (b) $(l_{e_{RGB}}, l_{other}) = (0.01, 0.01)$, (c) $(l_{e_{RGB}}, l_{other}) = (0.001, 0.001)$, (d) $(l_{e_{RGB}}, l_{other}) = (0.01, 0.1)$, (e) $(l_{e_{RGB}}, l_{other}) = (0.001, 0.01)$, and (f) $(l_{e_{RGB}}, l_{other}) = (0.0001, 0.001)$.**Figure 5.** Plots showing the accuracy (including OA, F1 score, and IoU) and loss of the proposed MAFNet for training the Potsdam dataset (a) and Vaihingen dataset (b).

5. Results and Discussions

In Section 5, the effectiveness of the proposed HAFNet for building extraction is evaluated in several aspects. Firstly, we analyze whether the hybrid fusion architecture has the ability to learn and utilize both individual modal and cross-modal features, and how the different kinds of features

affect the building extraction results. We also compare the hybrid fusion architecture with other fusion methods (e.g., feature-level and decision-level fusion). Secondly, we investigate whether the proposed Att-MFBlock can be used to handle the cross-modal fusion ambiguity and learn more discriminative and representative cross-modal features, and how it gains an advantage over other fusion methods at the decision stage. Thirdly, the proposed HAFNet was compared with three classical fusion networks in the task of building extraction. Finally, we compared different semantic segmentation networks as the basic network of the proposed MAFNet.

5.1. Effect of the Hybrid Fusion Architecture

As mentioned in the introduction, we believe that our proposed hybrid fusion architecture can outperform others due to its ability in sufficient learning and utilizing both individual modal and cross-modal features. To prove this ability of the hybrid fusion architecture, we extracted and visualized feature maps in different levels and building extraction results from different streams (Figure 6). The results can be summarized in three aspects. Firstly, in the same stream, deeper level features appeared to be more abstract and discriminative than the shallower ones. Secondly, the same level features derived from different streams were highly different from each other, indicating that the individual modal and cross-modal features learned from the hybrid fusion network are capable of providing comprehensive multi-view information about building objects. Interestingly, compared with RGB-specific and DSM-specific streams, the features learned from the cross-modal stream carried more discriminative information which favored better object localization and spatial refinement. It indicated that using a whole stream to gradually fuse different level features from RGB-specific and DSM-specific streams performs well on exploring more informative cross-modal complements. The statistical results obtained from different streams (Table 3) also confirmed that the cross-modal stream performed better than either RGB-specific or DSM-specific stream.

Table 3. Statistical results obtained from different streams. “Cross-modal stream” can also be seen as a feature-level fusion method. “RGB + DSM” denotes the decision-level fusion of RGB and DSM data. “RGB + DSM + cross-modal” denotes the proposed hybrid fusion method.

Streams	Potsdam			Vaihingen		
	OA (%)	F1 Score (%)	IoU (%)	OA (%)	F1 Score (%)	IoU (%)
RGB	96.06	97.43	88.78	96.11	97.40	86.12
DSM	93.88	96.06	86.85	92.17	94.83	82.48
Cross-modal	97.60	98.43	89.65	96.60	97.71	86.89
RGB + DSM	96.87	97.95	89.12	96.32	97.58	86.67
RGB + DSM + cross-modal	97.96	98.78	90.10	97.04	98.17	87.32

The good performance of the cross-modal stream brings a new question: is it necessary to utilize individual modal features for final building extraction? To better understand the advantage of our proposed hybrid fusion architecture over other fusion methods, the statistical results obtained from different fusion methods (i.e., feature-level fusion, decision-level fusion, and hybrid fusion) were compared and are displayed in Table 3. Using only the cross-modal stream can be seen as a feature-level fusion method. Data-level fusion methods are not demonstrated in this paper because their architectures cannot be initialized with the pre-trained VGG16, which makes it unfair to compare with other fusion methods. As shown in Table 3, although the cross-modal stream (i.e., the feature-level fusion) outperformed the decision-level fusion of RGB and DSM streams, it cannot yet compare with the fusion of three streams (i.e., the hybrid fusion architecture). The results confirmed our original hypothesis that a reliable and robust building extraction result requires a combination of individual modal and cross-modal features. Although the cross-modal features performed better than individual modal features, some useful individual modal information which is helpful for building extraction would be inevitably lost after the fusion operations. The hybrid fusion architecture is designed for a

comprehensive feature engineering that can maximize the use of available information and achieve the most reliable building extraction results.

Streams	Inputs	Feat_conv2	Feat_conv3	Feat_conv4	Feat_conv5	Predictions	GT
RGB							
DSM							
Cross							
RGB							
DSM							
Cross							
RGB							
DSM							
Cross							

Figure 6. Visualized features in different layers and output predictions of RGB-specific, DSM-specific, and cross-modal streams. Feat_ n refers to the feature map derived from the n -th convolutional block.

5.2. Effect of Att-MFBlock

The purpose of Section 5.2 is to verify the effectiveness of the proposed attention-aware multimodal fusion block (Att-MFBlock). For comparison, the attention module was removed in fusion blocks, and multimodal features would be directly combined without modality and channel-wise weighting. As illustrated in Figure 7, cross-modal features in different layers with/without an attention module are visually distinguished from each other. Specifically, the features with an attention module contain more discriminative information about buildings and focus on exploring more spatial cues to refine the boundaries of buildings.

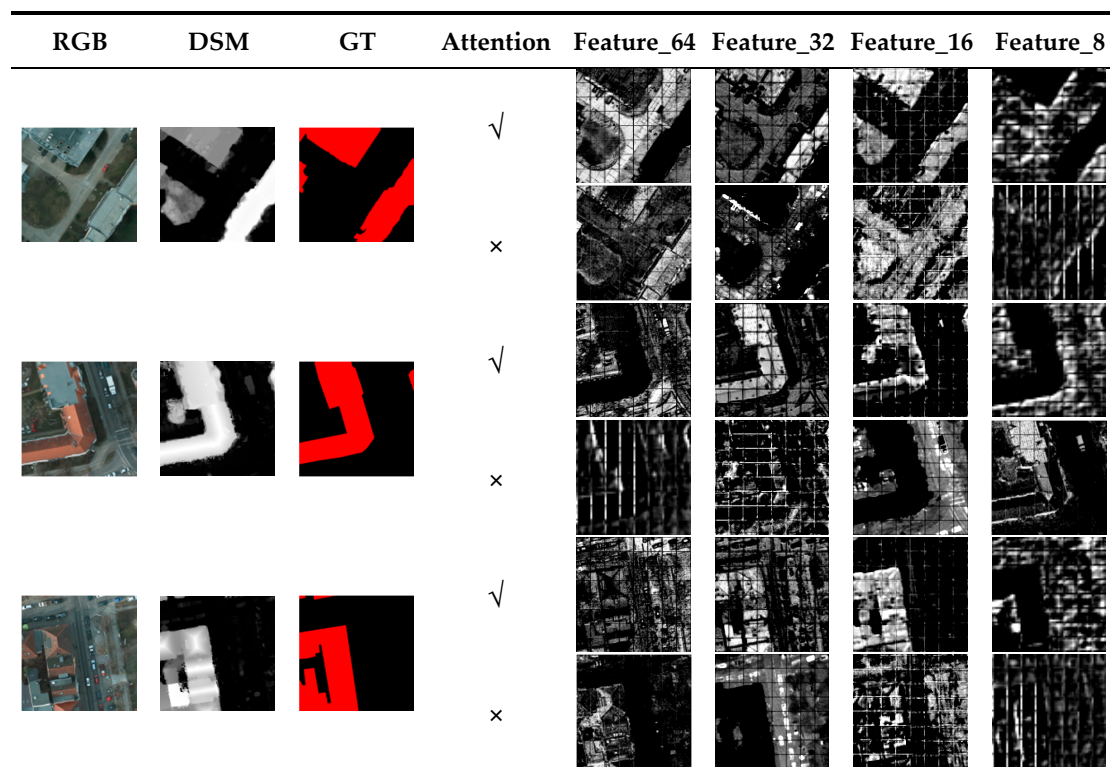


Figure 7. Visualized cross-modal features in different layers with/without an attention module.

On the contrary, the features without the attention module fail to achieve the desired progressive enhancement. In each level of fusion, the combined features are easily affected by noises and cannot reflect the discriminative information of buildings. It suggests that fusion methods without attention may lead to vague and uninformative cross-modal combination, making it hard to select desired cross-modal complements for gainful multimodal fusion. Instead, some negative features and noisy backgrounds from any modality may be preferred for predictions. The quantitative results of fusion networks with/without attention module shown in Table 4 demonstrate the significant boost by involving the attention module into Att-MFBlock.

Table 4. Quantitative results of fusion networks with/without an attention module.

Attention	Potsdam			Vaihingen		
	OA (%)	F1 Score (%)	IoU (%)	OA (%)	F1 Score (%)	IoU (%)
✓	97.96	98.78	90.10	97.04	98.17	87.32
×	96.56	97.48	89.36	96.32	97.46	86.44

At the decision stage, how to combine RGB-specific, DSM-specific, and cross-modal streams into a final fused prediction were investigated. There are two commonly used strategies: averaging and majority-voting. The former calculates the average of softmax activations of each stream and use the average value to make a final prediction. The latter calculates the mode of each prediction as the final prediction. For a fair comparison, different methods (i.e., averaging, majority-voting, and Att-MFBlock) were individually used to produce the final prediction and the rest of the network were kept unchanged. Experimental results showed that the proposed Att-MFBlock produced a more accurate and reliable fused prediction than averaging and majority-voting methods, especially in some hard pixels misclassified by most of the streams. For example, as illustrated in Figure 8, the pixels in the green rectangular box were misclassified as buildings in the RGB-specific stream (Figure 8c) and cross-modal stream (Figure 8d). As shown in Figure 8f,g, using averaging or majority-voting methods

failed to achieve good performance, while the result of using Att-MFBlock was highly desirable (Figure 8h). This may owe to the attention mechanism adaptively highlighting the positive information and suppressing the negative information. For the pixels in the green rectangular box, the Att-MFBlock was able to attribute more weight to the prediction coming out of the RGB-specific stream. The similar results that Att-MFBlock outperforms other methods were observed elsewhere as highlighted by other rectangular boxes displayed in Figure 8. The experimental results indicated that the Att-MFBlock not only favored better feature selection and fusion at the feature-level but also performed well at highlighting positive predictions at the decision-level.

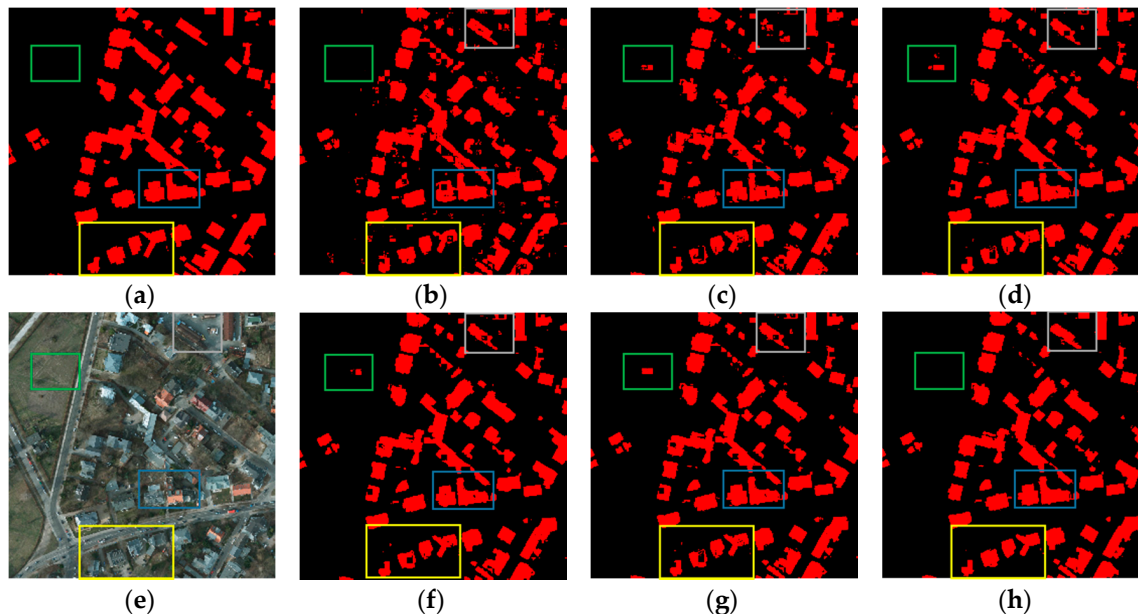


Figure 8. Predictions from different streams and the fused results using different decision-level fusion methods. (a) GT, (b) DSM prediction, (c) RGB prediction, (d) cross-modal prediction, (e) RGB image, (f) majority-voting, (g) averaging, and (h) Att-MFBlock.

On the other hand, the quantitative comparison in Table 5 also demonstrated that our proposed Att-MFBlock outperformed other methods. Otherwise, the averaging and majority-voting methods performed similarly in both two datasets. Unexpectedly, compared with the results from RGB-specific, DSM-specific, or cross-modal streams (Table 3), the two fusion methods failed to achieve the desired progressive enhancement: they performed better than RGB-specific and DSM-specific streams, while performing worse than the cross-modal stream. This may be due to the fact that the averaging and majority-voting methods tend to produce solid and smooth fused results rather than adaptively selecting the most important ones.

Table 5. Quantitative results of different decision-level fusion methods.

Decision-Level Fusion Methods	Potsdam			Vaihingen		
	OA (%)	F1 score (%)	IoU (%)	OA (%)	F1 Score (%)	IoU (%)
Averaging	97.43	98.32	89.44	96.48	97.65	86.79
Majority-voting	97.44	98.32	89.42	96.48	97.64	86.83
Att-MFBlock	97.96	98.78	90.10	97.04	98.17	87.32

5.3. Comparisons to Other Classical Fusion Networks

Three classical fusion networks, including FuseNet [56], RC-SegNet [34], and V-FuseNet [23] were used for comparisons. These networks were selected because all of them have already been proven as

effective in fusing RGB image and depth/LiDAR data. Otherwise, all of them are based on SegNet, which could make a fair comparison with our proposed network.

Table 6 lists the quantitative results of FuseNet, V-FuseNet, RC-SegNet, and HAFNet. HAFNet outperformed other models in both Potsdam and Vaihingen datasets. The visualized building extraction maps also verified the effectiveness of HAFNet. Specifically, when building objects were composed of different materials and appeared in varied shapes, sizes, and spectrum reflectance, FuseNet, V-FuseNet, and RC-SegNet failed to detect all kinds of buildings and tended to miss some spatial details. This will lead to empty holes inside extracted building objects, which are illustrated in Figure 9. In contrast, benefiting from the proposed hybrid fusion scheme and Att-MFBlock, HAFNet can adaptively select and combine cross-modal and individual modal features to perform more solid and accurate inference.

Table 6. Quantitative results of FuseNet, V-FuseNet, RC-SegNet, and HAFNet.

Fusion Network	Potsdam			Vaihingen		
	OA (%)	F1 Score (%)	IoU (%)	OA (%)	F1 Score (%)	IoU (%)
FuseNet	97.39	98.29	89.32	96.12	97.41	86.45
V-FuseNet	97.50	98.36	89.78	96.62	97.74	86.73
RC-SegNet	97.42	98.31	89.63	96.71	97.79	86.92
HAFNet	97.96	98.78	90.10	97.04	98.17	87.32

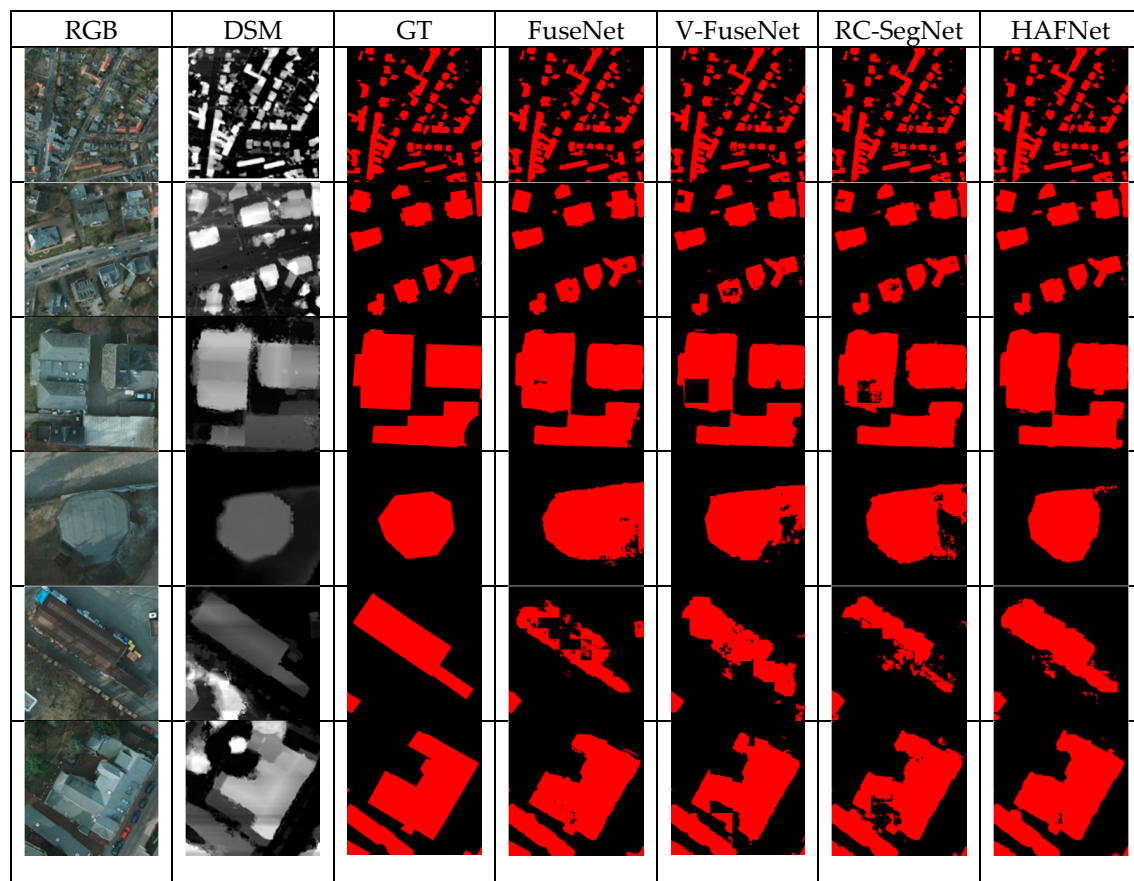


Figure 9. Building extraction results using FuseNet, V-FuseNet, RC-SegNet, and HAFNet.

5.4. Comparisons between Different Basic Networks of the MAFNet

As described in Section 3.1, we claimed that our proposed MAFNet can be easily adopted to other basic semantic segmentation networks. To prove that the contribution of our work is not limited to SegNet, several classical and popular semantic segmentation models, FCN-8s, U-Net, and DeepLab

v3+, were compared with SegNet as the basic networks of MAFNet. For a fair comparison, all basic networks used pretrained VGG-16 network as the encoder. Table 7 lists the quantitative results of SegNet, FCN-8s, U-Net, and DeepLab v3+ based MAFNet. SegNet and U-Net performed better than DeepLab v3+ and FCN-8s. This may be due to the notion that decoders of DeepLab v3+ and FCN 8s are relatively simple, while SegNet and U-Net both have elegant architectures with well-designed decoders. SegNet uses indices computed in the max-pooling step to perform non-linear up-sampling, and U-Net uses skip connection to help decoders gradually recover the object details. The final building extraction results using different basic networks also demonstrated the superiority of SegNet and U-Net in feature relocation and detail recovery (Figure 10). Otherwise, the overall performance of U-Net is similar to SegNet, while the latter performed slightly better in some areas (see the second columns of Figure 10). This might be because reusing low-level features by skip connection brings not only spatial details but also image noise.

Table 7. Quantitative results of SegNet, FCN-8s, U-Net, and DeepLab v3+ based MAFNet.

Basic Networks	Potsdam			Vaihingen		
	OA (%)	F1 Score (%)	IoU	OA (%)	F1 Score (%)	IoU
SegNet	97.96	98.78	90.10	97.04	98.17	87.32
FCN-8s	97.29	98.21	89.36	95.99	92.25	86.21
U-Net	97.83	98.68	90.17	96.92	98.07	87.14
DeepLab v3+	97.60	98.42	90.49	96.10	97.38	86.45

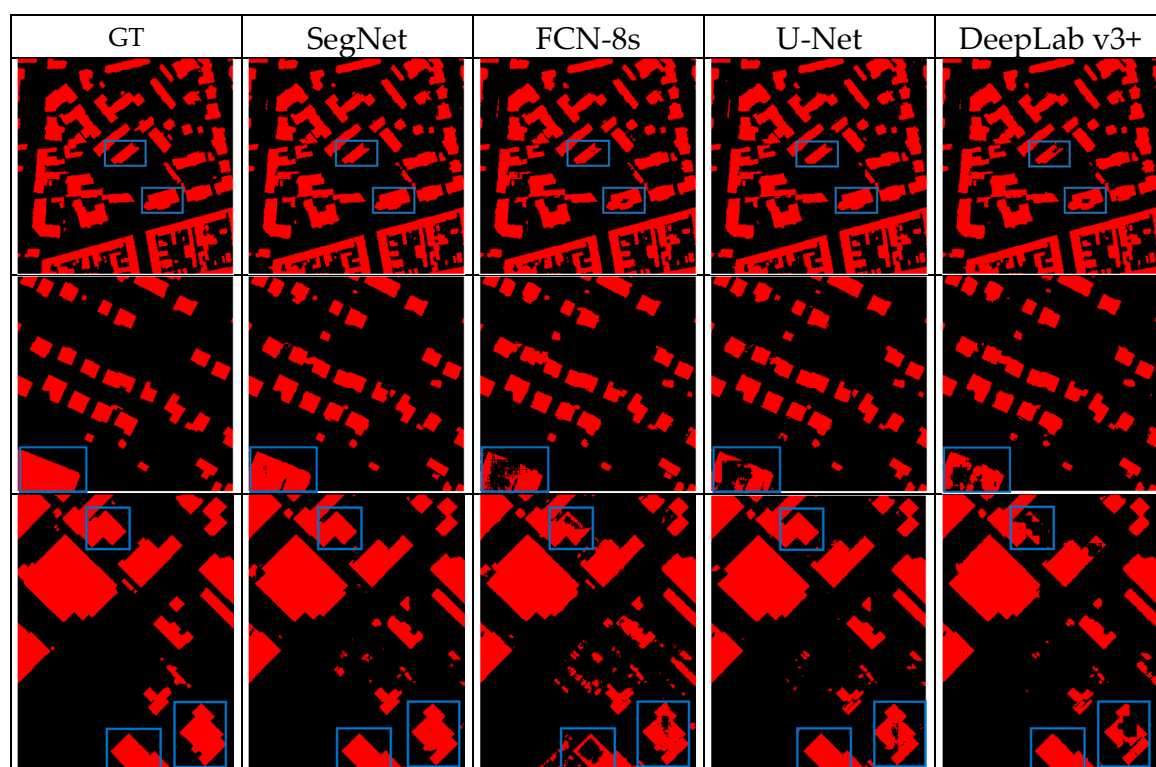


Figure 10. Building extraction results using SegNet, FCN-8s, U-Net, and DeepLab v3+ based MAFNet.

6. Conclusions

A multimodal attention fusion network (HAFNet) was proposed for automatic building extraction from HRI and LiDAR data. The network used a novel hybrid fusion architecture to sufficiently learn and utilize both individual modal and cross-modal features and employed a new attention-aware multimodal fusion block (Att-MFBlock) to overcome the cross-modal fusion problems. Experimental

results obtained from Potsdam and Vaihingen datasets confirmed the effectiveness of HAFNet. Results demonstrated that only the combining of both individual modal and cross-modal features can achieve the best building extraction results. Besides, the Att-MFBlock not only favored better feature selection and fusion at the feature-level but also performed well on highlighting positive predictions at the decision-level. Furthermore, compared with the three classic fusion models, HAFNet produced more solid and accurate building extraction results when the building objects were in high diversity. The results demonstrated that our contribution can be easily adapted to other segmentation model-based fusion architectures and extended to support a varied range of modalities. In the future, we plan to use other state-of-the-art models as basic networks and extend the network to support more than two modalities.

Author Contributions: Conceptualization, P.Z. and P.D.; methodology, P.Z.; software, P.Z.; writing—original draft preparation, P.Z., C.L., and X.W.; writing—review and editing, P.Z., P.D., E.L., Z.X., and X.B.; visualization, P.Z.; funding acquisition, P.D. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the National Natural Science Foundation of China (No. 41631176).

Acknowledgments: The authors thank the ISPRS for making the Vaihingen and Potsdam datasets available.

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

References

- Huang, J.; Zhang, X.; Xin, Q.; Sun, Y.; Zhang, P. Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network. *ISPRS J. Photogramm. Remote Sens.* **2019**, *151*, 91–105. [\[CrossRef\]](#)
- Zhang, K.; Yan, J.; Chen, S.C. Automatic Construction of Building Footprints From Airborne LIDAR Data. *IEEE Trans. Geosci. Remote Sens.* **2006**, *44*, 2523–2533. [\[CrossRef\]](#)
- Zhou, G.; Zhou, X. Seamless fusion of LiDAR and aerial imagery for building extraction. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 7393–7407. [\[CrossRef\]](#)
- Dalponte, M.; Bruzzone, L.; Gianelle, D. Fusion of Hyperspectral and LIDAR Remote Sensing Data for Classification of Complex Forest Areas. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 1416–1427. [\[CrossRef\]](#)
- Lee, D.S.; Shan, J. Combining Lidar Elevation Data and IKONOS Multispectral Imagery for Coastal Classification Mapping. *Mar. Geod.* **2003**, *26*, 117–127. [\[CrossRef\]](#)
- Chen, Y.; Li, C.; Ghamisi, P.; Jia, X.; Gu, Y. Deep Fusion of Remote Sensing Data for Accurate Classification. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1253–1257. [\[CrossRef\]](#)
- Karsli, F.; Dihkan, M.; Acar, H.; Ozturk, A. Automatic building extraction from very high-resolution image and LiDAR data with SVM algorithm. *Arabian J. Geosci.* **2016**, *9*. [\[CrossRef\]](#)
- Zarea, A.; Mohammadzadeh, A. A Novel Building and Tree Detection Method From LiDAR Data and Aerial Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 1864–1875. [\[CrossRef\]](#)
- Du, P.; Bai, X.; Tan, K.; Xue, Z.; Samat, A.; Xia, J.; Li, E.; Su, H.; Liu, W. Advances of Four Machine Learning Methods for Spatial Data Handling: A Review. *J. Geovis. Spat. Anal.* **2020**, *4*. [\[CrossRef\]](#)
- Li, E.; Xia, J.; Du, P.; Lin, C.; Samat, A. Integrating Multilayer Features of Convolutional Neural Networks for Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 5653–5665. [\[CrossRef\]](#)
- Zhong, Y.; Zhu, Q.; Zhang, L. Scene Classification Based on the Multifeature Fusion Probabilistic Topic Model for High Spatial Resolution Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 6207–6222. [\[CrossRef\]](#)
- Zou, Q.; Ni, L.; Zhang, T.; Wang, Q. Deep Learning Based Feature Selection for Remote Sensing Scene Classification. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 2321–2325. [\[CrossRef\]](#)
- Ienco, D.; Interdonato, R.; Gaetano, R.; Minh, H.T.D. Combining Sentinel-1 and Sentinel-2 Satellite Image Time Series for land cover mapping via a multi-source deep learning architecture. *ISPRS J. Photogramm. Remote Sens.* **2019**, *158*, 11–22. [\[CrossRef\]](#)
- Storie, C.D.; Henry, C.J. Deep Learning Neural Networks for Land Use Land Cover Mapping. In Proceedings of the Igarss 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; IEEE: New York, NY, USA, 2018; pp. 3445–3448.

15. Abdollahi, A.; Pradhan, B.; Shukla, N.; Chakraborty, S.; Alamri, A. Deep Learning Approaches Applied to Remote Sensing Datasets for Road Extraction: A State-Of-The-Art Review. *Remote Sens.* **2020**, *12*, 1444. [\[CrossRef\]](#)
16. Pan, X.; Yang, F.; Gao, L.; Chen, Z.; Zhang, B.; Fan, H.; Ren, J. Building Extraction from High-Resolution Aerial Imagery Using a Generative Adversarial Network with Spatial and Channel Attention Mechanisms. *Remote Sens.* **2019**, *11*, 917. [\[CrossRef\]](#)
17. Sun, G.; Huang, H.; Zhang, A.; Li, F.; Zhao, H.; Fu, H. Fusion of Multiscale Convolutional Neural Networks for Building Extraction in Very High-Resolution Images. *Remote Sens.* **2019**, *11*, 227. [\[CrossRef\]](#)
18. Du, L.; You, X.; Li, K.; Meng, L.; Cheng, G.; Xiong, L.; Wang, G. Multi-modal deep learning for landform recognition. *J. Photogramm. Remote Sens.* **2019**, *158*, 63–75. [\[CrossRef\]](#)
19. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 640–651.
20. Xin, J.; Zhang, X.; Zhang, Z.; Fang, W. Road Extraction of High-Resolution Remote Sensing Images Derived from DenseUNet. *Remote Sens.* **2019**, *11*, 2499. [\[CrossRef\]](#)
21. Yang, H.; Wu, P.; Yao, X.; Wu, Y.; Wang, B.; Xu, Y. Building Extraction in Very High Resolution Imagery by Dense-Attention Networks. *Remote Sens.* **2018**, *10*, 1768. [\[CrossRef\]](#)
22. Liu, W.; Yang, M.; Xie, M.; Guo, Z.; Li, E.; Zhang, L.; Pei, T.; Wang, D. Accurate Building Extraction from Fused DSM and UAV Images Using a Chain Fully Convolutional Neural Network. *Remote Sens.* **2019**, *11*, 2912. [\[CrossRef\]](#)
23. 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. [\[CrossRef\]](#)
24. Sun, Y.; Zhang, X.; Xin, Q.; Huang, J. Developing a multi-filter convolutional neural network for semantic segmentation using high-resolution aerial imagery and LiDAR data. *ISPRS J. Photogramm. Remote Sens.* **2018**, *143*, 3–14. [\[CrossRef\]](#)
25. Xu, Y.; Du, B.; Zhang, L. Multi-source remote sensing data classification via fully convolutional networks and post-classification processing. In Proceedings of the IGARSS 2018–2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 3852–3855.
26. Hazirbas, C.; Ma, L.; Domokos, C.; Cremers, D. FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture. In *Computer Vision—Accv 2016, Pt I*; Lai, S.H., Lepetit, V., Nishino, K., Sato, Y., Eds.; Springer: Cham, Switzerland, 2017; Volume 10111, pp. 213–228.
27. Zhang, W.; Huang, H.; Schmitz, M.; Sun, X.; Wang, H.; Mayer, H. Effective Fusion of Multi-Modal Remote Sensing Data in a Fully Convolutional Network for Semantic Labeling. *Remote Sens.* **2018**, *10*, 52. [\[CrossRef\]](#)
28. 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. [\[CrossRef\]](#)
29. Marcos, D.; Hamid, R.; Tuia, D. Geospatial Correspondences for Multimodal Registration. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; IEEE: New York, NY, USA, 2016; pp. 5091–5100. [\[CrossRef\]](#)
30. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/Cvf Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; IEEE: New York, NY, USA, 2018; pp. 7132–7141. [\[CrossRef\]](#)
31. Chen, H.; Li, Y. Three-stream Attention-aware Network for RGB-D Salient Object Detection. *IEEE Trans. Image Process.* **2019**. [\[CrossRef\]](#)
32. Mohla, S.; Pande, S.; Banerjee, B.; Chaudhuri, S. FusAtNet: Dual Attention based SpectroSpatial Multimodal Fusion Network for Hyperspectral and LiDAR Classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 92–93.
33. Badrinarayanan, V.; Handa, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Robust Semantic Pixel-Wise Labelling. *arXiv* **2015**, arXiv:1505.07293.
34. Audebert, N.; Le Saux, B.; Lefèvre, S. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In Proceedings of the Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; pp. 180–196.

35. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention, Pt III*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing Ag: Cham, Switzerland, 2015; Volume 9351, pp. 234–241.
36. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.
37. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [[CrossRef](#)]
38. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
39. 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.
40. Guo, M.; Liu, H.; Xu, Y.; Huang, Y. Building Extraction Based on U-Net with an Attention Block and Multiple Losses. *Remote Sens.* **2020**, *12*, 1400. [[CrossRef](#)]
41. Wagner, F.H.; Dalagnol, R.; Tarabalka, Y.; Segantini, T.Y.; Thomé, R.; Hirye, M. U-Net-Id, an Instance Segmentation Model for Building Extraction from Satellite Images—Case Study in the Joanópolis City, Brazil. *Remote Sens.* **2020**, *12*, 1544. [[CrossRef](#)]
42. Lin, Y.; Xu, D.; Wang, N.; Shi, Z.; Chen, Q. Road Extraction from Very-High-Resolution Remote Sensing Images via a Nested SE-Deeplab Model. *Remote Sens.* **2020**, *12*, 2985. [[CrossRef](#)]
43. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
44. Schlemper, J.; Oktay, O.; Schaap, M.; Heinrich, M.; Kainz, B.; Glocker, B.; Rueckert, D. Attention gated networks: Learning to leverage salient regions in medical images. *Med. Image Anal.* **2019**, *53*, 197–207. [[CrossRef](#)]
45. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; Mit Press: Cambridge, MA, USA, 1996; pp. 5998–6008.
46. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T.-S. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5659–5667.
47. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
48. Lin, G.; Shen, C.; van den Hengel, A.; Reid, I. Efficient Piecewise Training of Deep Structured Models for Semantic Segmentation. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3194–3203.
49. 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, 18–23 June 2018; pp. 7794–7803.
50. Yuan, Y.; Wang, J. Ocnet: Object context network for scene parsing. *arXiv* **2018**, arXiv:1809.00916.
51. Zhao, H.; Zhang, Y.; Liu, S.; Shi, J.; Loy, C.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.
52. Jin, Y.; Xu, W.; Hu, Z.; Jia, H.; Luo, X.; Shao, D. GSCA-UNet: Towards Automatic Shadow Detection in Urban Aerial Imagery with Global-Spatial-Context Attention Module. *Remote Sens.* **2020**, *12*, 2864. [[CrossRef](#)]
53. Tian, Z.; Zhan, R.; Hu, J.; Wang, W.; He, Z.; Zhuang, Z. Generating Anchor Boxes Based on Attention Mechanism for Object Detection in Remote Sensing Images. *Remote Sens.* **2020**, *12*, 2416. [[CrossRef](#)]
54. Li, L.; Liang, P.; Ma, J.; Jiao, L.; Guo, X.; Liu, F.; Sun, C. A Multiscale Self-Adaptive Attention Network for Remote Sensing Scene Classification. *Remote Sens.* **2020**, *12*, 2209. [[CrossRef](#)]

55. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1026–1034.
56. Hazirbas, C.; Ma, L.; Domokos, C.; Cremers, D. Fusetnet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In Proceedings of the Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; pp. 213–228.

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



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).