remote sensing

MDPI

Article

FusionHeightNet: A Multi-Level Cross-Fusion Method from Multi-Source Remote Sensing Images for Urban Building Height Estimation

Chao Ma [1,2,3], Yueting Zhang [1,2,*], Jiayi Guo [1,2], Guangyao Zhou [1,2] and Xiurui Geng [1,2]

[1] Key Laboratory of Technology in Geo-Spatial Information Processing and Application Systems, Beijing 100190, China; machao191@mails.ucas.ac.cn (C.M.); guojy@aircas.ac.cn (J.G.); zhougy@aircas.ac.cn (G.Z.); gengxr@sina.com.cn (X.G.)
[2] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
[3] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: zhangyueting06@mails.ucas.ac.cn

Abstract: Extracting buildings in urban scenes from remote sensing images is crucial for the construction of digital cities, urban monitoring, urban planning, and autonomous driving. Traditional methods generally rely on shadow detection or stereo matching from multi-view high-resolution remote sensing images, which is cost-intensive. Recently, machine learning has provided solutions for the estimation of building heights from remote sensing images, but challenges remain due to the limited observation angles and image quality. The inherent lack of information in a single modality greatly limits the extraction precision. This article proposes an advanced method using multi-source remote sensing images for urban building height estimation, which is characterized by multi-level cross-fusion, the multi-task joint learning of footprint extraction and height estimation, and semantic information to refine the height estimation results. The complementary and effective features of synthetic aperture radar (SAR) and electro-optical (EO) images are transferred through multi-level cross-fusion. We use the semantic information of the footprint extraction branch to refine the height estimation results, enhancing the height results from coarse to fine. Finally, We evaluate our model on the SpaceNet 6 dataset and achieve 0.3849 and 0.7231 in the height estimation metric δ_1 and footprint extraction metric *Dice*, respectively, which indicate effective improvements in the results compared to other methods.

Keywords: building height estimation; synthetic aperture radar (SAR); electro-optical (EO); multi-level cross-fusion; semantic information to refine height results

Citation: Ma, C.; Zhang, Y.; Guo, J.; Zhou, G.; Geng, X. FusionHeightNet: A Multi-Level Cross-Fusion Method from Multi-Source Remote Sensing Images for Urban Building Height Estimation. *Remote Sens.* **2024**, *16*, 958. https://doi.org/10.3390/rs16060958

Academic Editors: Abdul Bais, Keshav D Singh and Sajid Saleem

Received: 22 January 2024
Revised: 3 March 2024
Accepted: 7 March 2024
Published: 8 March 2024

1. Introduction

With the rapid development of global urbanization, urban management and planning [1,2] have become important issues in various countries. Accurate and efficient urban building extraction and height estimation are crucial for applications such as the construction of digital cities [3], the study of human activities [4], urban monitoring and planning [5], and autonomous driving [6]. Height information can reflect the structural indicators of a building. In many emergency situations, the building structure as well as height estimation are key to aid information retrieval. However, the dense distribution of buildings in urban areas, complex structures, and the lack of sufficient high-precision observation data make instance-level building height estimation [7] a critical and challenging task.

In recent years, remote sensing technology has experienced rapid development. Compared with other sensors, such as LiDAR [8], the acquisition cost of remote sensing images is obviously lower. Therefore, estimating building heights from remote sensing images has become a credible idea [9]. However, since a single image may correspond to countless

height structures, estimating the height of buildings from a single remote sensing image is a challenging problem. In addition, high-precision building height estimation often requires high-resolution remote sensing images. Blurred images greatly increase the difficulty of estimating building heights, which makes estimating building heights from single-view remote sensing images [10] a challenging issue.

EO and SAR [11] are the two main types of sensors for remote sensing applications. EO sensors such as Quickbird, Worldview, etc., usually have meter-level or even sub-meter-level resolution. They are passive optical systems and acquire images by illuminating the ground with sunlight. SAR is an active imaging system that actively emits electromagnetic waves and receives echoes to achieve imaging, which allows the acquisition of SAR images not restricted by the climatic conditions and time. With the vigorous development of SAR sensors, more high-resolution SAR sensors are being deployed, such as TerraSAR-X, COSMO-SkyMed, etc. In optical images, the building structure, color, texture information, and edges are clearer, generally achieving better building extraction results compared with SAR images. However, the imaging mechanism of SAR allows the sensor to observe the target at an oblique angle. The two dimensions of SAR images are the azimuth and range, which means that SAR images contain the three-dimensional structure information of the target. SAR images and optical images have different observation angles and imaging dimensions, which makes the two modalities complementary and can provide more target information. With the development of remote sensing detection technology, it has become increasingly popular to detect the same area with different types of sensors. In urban areas with both optical images and SAR images, improving the precision of building height estimation by fusing optical images and SAR images has begun to attract attention. Although there have been some studies [9,12] aiming to fuse SAR images and optical images to estimate building heights, these methods ignore the exploration of fusion methods adopting SAR images and optical images, which simply combine the features of the two modalities. This simple combination of heterogeneous features often leads to a decline in results and fails to achieve the purpose of complementing the feature information of the two modalities. This article studies a more effective fusion method to improve the precision of building height estimation.

Deep learning has seen rapid progress in the fields of computer vision and natural language processing in recent years. The basic composition of the network has also evolved from a CNN-based network to a Transformer-based network [13]. The Transformer network was originally applied in the field of natural language processing, mainly in processing sequence-to-sequence information reasoning tasks, due to its ability to realize information interactions between long-distance patches. With the proposal of ViT [14] and DETR [15], the Transformer network was introduced into the field of computer vision. Utilizing the global self-attention mechanism provided by Transformer networks, there has been a significant enhancement in the performance of various downstream tasks within the field of computer vision. Compared with CNNs, the main advantage of the Transformer network is that it can model long-distance attention, while a CNN pays more attention to local information. Therefore, both have their own advantages when processing images. In addition, the Transformer network does not need to stack too many layers and can focus on global and local pixel information at the same time, eliminating the need for depth. The complementary relationship between the global modeling ability of the Transformer network and the local modeling ability of the CNN network provides a new idea for the fusion of the heterogeneous features of optical images and SAR images.

Building height estimation from a single remote sensing image is similar to the monocular depth estimation task in the field of computer vision. Many studies [16,17] have introduced depth estimation algorithms into the building height estimation task of remote sensing images. Typically, the proposed network focuses on completing the specific task of depth estimation or height estimation. Some researchers [18,19] have introduced other different tasks to perform joint learning with the depth estimation task recently. This idea is also applicable to the problem of building height estimation from a single remote

sensing image. Building footprint extraction and height estimation are essentially two downstream tasks with a strong correlation. The two tasks are interdependent, and the introduction of the building footprint extraction task as an auxiliary task contributes to the better generation of height estimation boundaries. Therefore, we explore ways to introduce the building footprint extraction task to improve the precision of height estimation.

This study proposes an advanced method using multi-source remote sensing images for urban building height estimation, called FusionHeightNet. It is characterized by the multi-level cross-fusion of multi-source remote sensing data, the multi-task joint learning of footprint extraction and height estimation, and the use of semantic information to refine the height estimation results. The complementary and effective features of multi-source remote sensing data are transferred through multi-level cross-fusion. We perform joint learning on the building footprint extraction task and the height estimation task simultaneously under the same framework. The semantic information of the footprint extraction branch is used to refine the height estimation results, which helps to elevate the height estimation results from coarse to fine. Our contributions can be summarized as follows.

1 We propose an advanced method for building height estimation by fusing multi-source remote sensing images and refining the height results through building footprint semantic information, achieving robust and effective instance-level building height reconstruction in urban scenes.
2 We propose a multi-level cross-fusion module for multi-source remote sensing data. Images of different modalities use independent backbone network parts and we apply multi-level cross-fusion at different locations to achieve the transmission and fusion of complementary and effective features.
3 The height results' refinement by the semantic information framework adopts a joint learning method for the height estimation task and the footprint extraction task. Through joint learning, the hidden constraints of the building footprint semantic information embedded in the height estimation task are mined, and the height estimation precision is effectively improved by refining the height estimation results from coarse to fine.
4 Experiments are conducted on the SpaceNet 6 dataset. The ablation results show the effectiveness of the proposed method. The comparison results show that the proposed method achieves higher height estimation metrics.

The remainder of this article is organized as follows. In Section 2, we introduce related works. In Section 3, we explain the framework details of the proposed method. In Section 4, we introduce and discuss the extensive experiments and results. Finally, Section 5 summarizes this work.

2. Related Works

2.1. Building Footprint Extraction

Building footprint extraction is a pixel-level classification problem of remote sensing images, dividing each pixel into two categories: building and non-building. Traditional methods mainly rely on manually designed features, including methods based on classifiers, segmentation, original geometry, and specific indices. Classifier-based methods, such as that developed by Caglar Senaras et al. [20], combine the detection results of multiple classifiers in a hierarchical structure, namely FSG, to improve the performance by fusing the decisions of multiple classifiers. Konstantinos Karantzalos et al. [21] proposed a region-based level set segmentation method. The essence of this method is to optimize the position and geometry of the evolution curve based on its statistical description by measuring the information in the area that makes up a specific image partition. Melissa Cote et al. [22] exploited geometric information to generate roof contours based on candidate points and further refined them through level set curve evolution enhanced by mean square error maps. Xin Huang et al. [23] established a relationship between the implicit characteristics of buildings (e.g., brightness, size, and contrast) and the properties of morphological operators (e.g., reconstruction, granularity, and directionality), and they proposed a multi-scale mor-

phological index using the automatic extraction of buildings from high-resolution remote sensing images. Traditional methods generally have the problem of poor generalization, and the extracted results are greatly limited by the hand-designed features.

With the outstanding performance of deep learning in the field of computer vision, it has become a better solution to automatically extract image features through deep learning to complete downstream segmentation tasks. Methods using semantic segmentation networks to extract building footprints are beginning to emerge. The mainstream architecture of semantic segmentation is based on the structure of an encoder and decoder. The classic algorithm starts from FCN [24], which considers the process of extracting features and making the feature map smaller as the encoder and considers the subsequent upsampling process as the decoder. Only the convolutional layer is used to achieve a large rise in precision. Another commonly used model is UNet [25], which is a U-shaped symmetrical network structure. The left side is an encoder composed of convolutional layers, and the right side is an upsampling layer. In addition, the feature map of each level in the UNet encoder will be cascaded with the corresponding decoder input. This jump connection method achieves a great improvement in precision. The DeepLab series methods include DeepLab v1 [26], which introduces the probabilistic graphical model, DeepLab v2 [27], which uses a dilated convolution algorithm to expand the receptive field, DeepLab v3 [28], which improves the ASPP module, and DeepLab v3+ [29], which is designed based on the V3 decoder module and modified Xception [30]. PSPNet [31] improves on FCN, introduces pyramid pooling, and makes full use of context information. Later, some variations based on the above network structure began to appear.

2.2. Building Height Estimation

The task of building height estimation is to perform pixel-by-pixel height estimation from remote sensing images. Some researchers have utilized shadow information (SFS) [32] to estimate the height. The shadow information contains two parts, which are formed due to occlusion and the unilluminated parts of the building itself. Some shadow detection algorithms have been proposed. Jiahang Liu et al. [33] used a series of transformations to separate shadow and non-shadow areas. Zhang Hong-ya et al. [34] excluded shadows based on object attributes and the spatial relationships between objects. Mustafa Teke et al. [35] proposed a shadow detection algorithm that combined near-infrared information with RGB bands. Xiran Zhou et al. [36] proposed a shadow pattern classification system that summarized different shadow shapes into many pattern categories and classified the extracted shadows into patterns to automatically determine the edges of building shadows.

By calculating the set mapping relationship between detected shadows and building heights, the height of the building can be estimated. Alexis Comber et al. [37] classified building shadows according to their relative characteristics and the spatial background within the scene, and they calculated the building heights through shadows. O. Benarchid et al. [38] used invariant color features to extract shadow information and implemented building extraction in ultra-high-resolution multi-spectral images. P. L. N. Raju et al. [39] used instance-based and rule-based methods to manually or automatically extract roofs and shadow areas, and then they used the ratio method and sun–satellite geometric relationship to associate roofs with shadows to estimate the height of the building. These shadow detection methods have many restrictions regarding application scenarios. They achieve better results in scenes with complete shadows and isolated building scenes, but have poor results in densely distributed building areas. In addition, there are some studies using statistical models and remote sensing data attribute methods. Chen Yang et al. [40] used spatial information Gaussian process regression (Si-GPR). Xuecao Li et al. [41] proposed a VVH index that integrates the dual polarization information (i.e., VV and VH) of Sentinel-1 SAR data. These methods are greatly affected by the characteristics of building targets in remote sensing images.

With the advancement of deep learning, the majority of research has begun utilizing deep learning methods to estimate the height of buildings. There are two main architectures

of deep neural networks; one is based on the encoder–decoder architecture, and the other is based on the GAN architecture [42]. The purpose of the network based on the encoder–decoder architecture is to estimate the height value of each pixel in the image. Hamed Amini Amirkolaee et al. [43] proposed an encoder–decoder architecture based on a CNN to estimate the height values from a single aerial image. IM2HEIGHT [7] proposed a complete convolution–deconvolution network architecture to model the fuzzy mapping between monocular remote sensing images and height maps. The purpose of the architecture based on the GAN network is to simulate and generate a height map of the scene. IMG2DSM [44] used conditional generative adversarial networks to achieve image to DSM conversion. U-IMG2DSM [45] used VAEs and GAN to simulate DSMs from optical images.

2.3. Multi-Source Remote Sensing Data Fusion

In the current field of remote sensing, various countries have begun to use various types of sensors, bringing rich data sources to remote sensing data. Multi-modal data are becoming more and more popular in remote sensing observations. Compared with single-modal data, multi-modality provides observation data from a variety of sensors, enriches the target information, and is beneficial to the extraction of complex targets. Multi-modal data fusion combines a variety of different modal remote sensing data to achieve complementary information between different types of remote sensing data, integrate multiple types of information to achieve high-precision predictions, and obtain better results than single-modal images. Current fusion methods for heterogeneous multi-modal remote sensing data mainly fuse at the feature level and decision level. The simplest method is the feature stacking method. Mattia Pedergnana et al. [46] superimposed the height and intensity features extracted from LiDAR data to the spectral bands of multi-spectral images. There are also some methods [47–49] that utilize morphological profiles, attribute profiles, and extinction profiles to provide high-quality fusion results. Yuanyuan Wang et al. [50] combined InSAR and optical images through 3D geometric fusion and semantic texturing. Although this simple stacking of features takes into account multi-modal data, it increases the dimensionality of features and does not always perform better in more complex downstream tasks. Subspace-based methods project features into low-dimensional subspaces to simplify the feature representation of downstream tasks and improve the computational efficiency. Behnood Rasti et al. [51] used IHS transformation for fusion, and some studies [52,53] have used the PCA method. In recent years, more and more fusion methods based on deep learning [54–56] have been proposed. Deep learning models are used to reason about the relationships between high-order features of heterogeneous data. This fusion method has better robustness and generalization. For example, Audebert et al. [57] used a dual-stream deep network to fuse optical and OpenStreetMap data.

3. Methodology

In an area with both an EO image $x_{opt} \in \mathbb{R}^{H \times W \times C_{opt}}$ and SAR image $x_{sar} \in \mathbb{R}^{H \times W \times C_{sar}}$, and the two modal remote sensing images have the same spatial resolution of $H \times W$, our goal is to extract the footprint results of the building $Pred_{footprint} \in \mathbb{R}^{H \times W \times 2}$ and estimate the pixel-by-pixel height value results of the building $Pred_{height} \in \mathbb{R}^{H \times W \times 1}$ under the same framework. We propose an end-to-end framework, called FusionHeightNet, which applies independent encoders to remote sensing images of two modalities. The overall structure of the proposed network is shown in Figure 1. The encoders have the same structure but do not share parameters. First, we apply the basic feature extractor based on a CNN and then perform the first cross-fusion of the two modal features. Next, we introduce a Transformer-based encoder module to infer and encode the high-level features of each of the two modalities. The output features are used as input for subsequent footprint extraction and height estimation after a second cross-fusion. The two tasks of footprint extraction and height estimation use independent decoder structures to achieve pixel-level footprint segmentation and height estimation, respectively. Then, the semantic

information of footprint extraction is input to the height estimation branch to guide the refinement of the final height estimation results. Under the entire network framework, joint learning is implemented for images of two modalities and two output tasks. The experimental results prove that this joint learning method achieves better results than a single modality and single task. In the remainder of this section, we first introduce the multi-level encoder architecture based on CNN–Transformer in Section 3.1, and we then introduce the multi-level cross-fusion module of multi-modal features at different levels of the encoder in Section 3.2. Next, the two-task decoder module of footprint extraction and height estimation is introduced in Section 3.3, and the design of height estimation refined by footprint semantic information is introduced in Section 3.4. Finally, in Section 3.5, we introduce the details of the loss function.

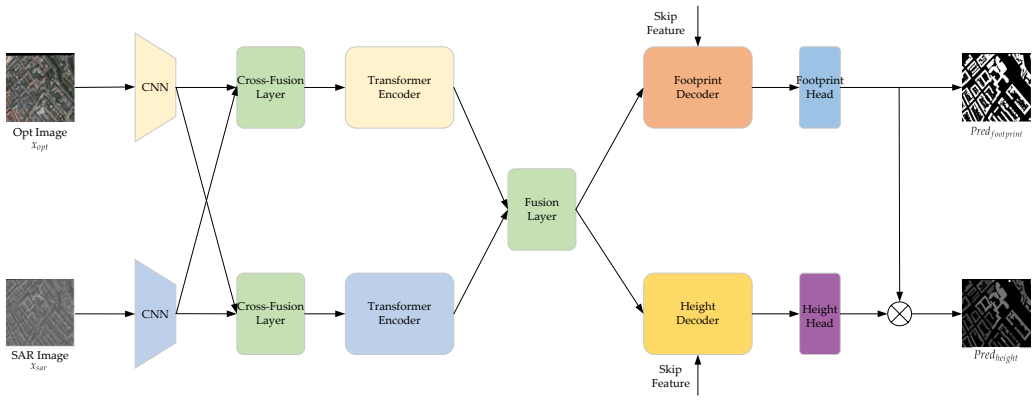


Figure 1. Detailed architecture of the proposed network.

3.1. Encoder Module Based on CNN–Transformer

CNNs are mainly used to extract the local features of targets, and it is more difficult to extract global receptive fields. The Vision Transformer network [14] is adept at reasoning about the relationships between long-distance patches, especially the capture of global features. However, the extraction of local features is ignored, and the foreground and background are more difficult to distinguish. Before the Transformer network [58] was applied to the field of computer vision, the most common method to improve the feature representations of CNNs was to expand the receptive field, but this approach would damage the pooling layer of the network.

Inspired by methods such as DETR [15] and TransUNet [50], the CNN and Transformer networks have advantages in extracting the local adjacent features and global features of the target, respectively. The CNN and Transformer networks can be combined to make full use of their complementary effects. The method that we propose performs feature extraction on optical images and SAR images through independent encoders, including the basic feature extraction module based on the CNN and the high-level feature encoder based on the Transformer network.

The basic feature extraction module is based on the CNN and uses skip connections to extract high-level features while retaining the details of low-level features, fully mining the local context information of the input images. The details of the basic feature extraction module are shown in Figure 2. The input image $x_{opt}/x_{sar} \in \mathbb{R}^{H\times W\times C}$ first goes through the convolution layer and a maximum pooling operation is used to obtain the feature map $F_0 \in \mathbb{R}^{H/2\times W/2\times C_0}$. Next, we perform three levels of downsampling operations. At each level, multiple convolutional layers and residual connection operations are used. The output features $F_1 \in \mathbb{R}^{H/4\times W/4\times C_1}$, $F_2 \in \mathbb{R}^{H/8\times W/8\times C_2}$ are input into the decoders of height estimation and footprint extraction and are the same as F_0, and the cascade of feature maps corresponding to the same resolution in the decoder is jointly used as the input of the next decoder block. The final output $F_{out} \in \mathbb{R}^{H/16\times W/16\times C_{out}}$ of the basic feature extraction module is used as the input of the next encoder block.

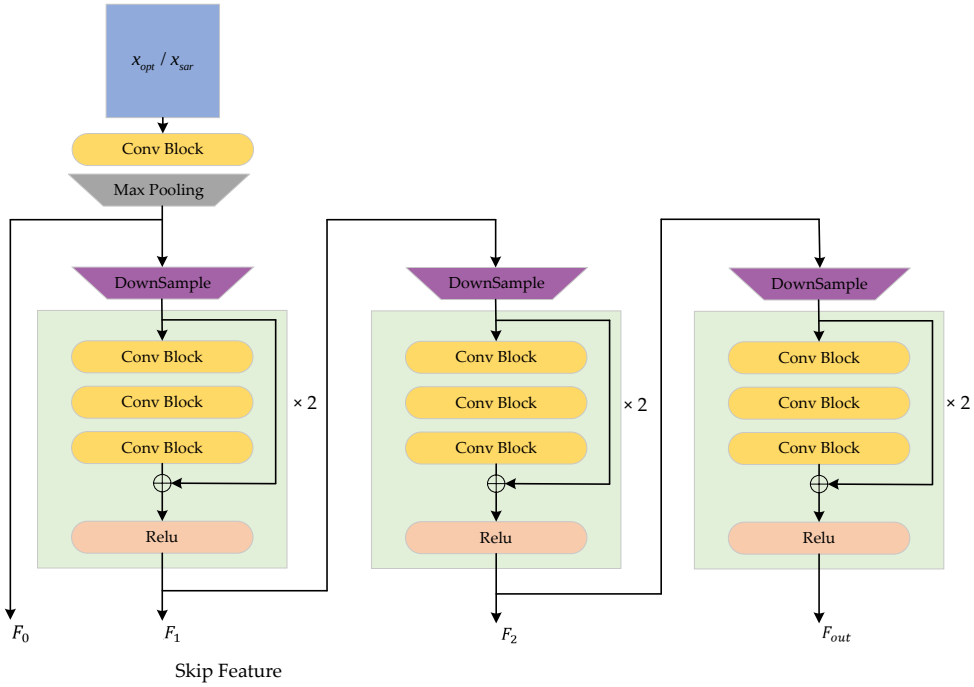


Figure 2. Structural details of the basic feature extraction module. The intermediate-level extracted features F_0, F_1, and F_2 are cascaded with the features of the same resolution in the decoder for height estimation and footprint extraction.

The high-level feature inference module of the encoder is based on a Transformer block. The structural details of the module are shown in Figure 3. The previous output features of the encoder $x_{patch} \in \mathbb{R}^{P\times P\times C\times N}$ are first flattened and divided into N patches with the size of $P \times P$, where $N = \frac{H\times W}{P\times P}$. Next, we perform a patch embedding operation on x_{patch} and encode and map it to the computable embedding space $E_{patch} \in \mathbb{R}^{P\times P\times D\times N}$ of the Transformer encoder block. At the same time, like ViT [14], the position embedding $E_{pos} \in \mathbb{R}^{D\times N}$ calculated for the input N patches is also added to the input of the Transformer block. The input of the Transformer block is as follows:

$$F_{encoder} = E_{patch} + E_{pos} \tag{1}$$

The Transformer encoder block is based on multi-head self-attention (MSA). The input is divided into multiple heads, and each head performs separate self-attention calculations. The calculation process of the self-attention of the i^{th} head is as follows:

$$Q_i = W_i^Q E_{patch}, K_i = W_i^K E_{patch}, V_i = W_i^V E_{patch} \tag{2}$$

$$Head_i(Q_i, K_i, V_i) = Softmax(\frac{Q_i K_i^T}{\sqrt{d}})V_i \tag{3}$$

where Q_i, K_i, V_i, respectively, represent the query matrix, key matrix, and value matrix calculated by the self-attention of the i^{th} head, and W_i^Q, W_i^K and W_i^V, respectively, represent the learnable weights of the query matrix, key matrix, and value matrix.

The output is obtained by the cascading of multiple self-attention heads:

$$MSA(Q, K, V) = Concat(head_1 W_1, head_2 W_2, \cdots, head_i W_i) \tag{4}$$

where W_i represents the i^{th} learnable weight. The calculation process of the Transformer encoder block is as follows:

$$x_k^1 = x_{k-1} + MSA(LN(x_{k-1}))$$
$$x_k^2 = x_k^1 + MLP(LN(x_k^1)) \quad (5)$$

where LN represents the layer norm and MLP represents the multi-layer perceptron.

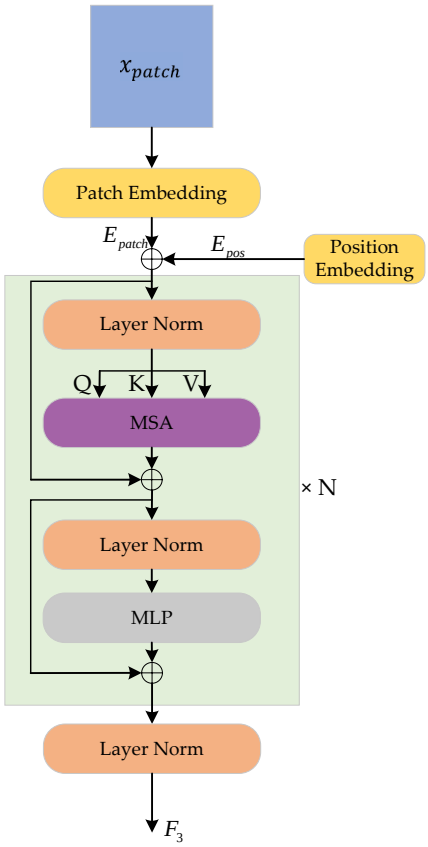


Figure 3. The high-level feature inference module of the encoder. All processing is performed on the feature with the lowest resolution of $H/16 \times W/16$.

3.2. Multi-Level Cross-Fusion Module

Due to the different imaging mechanisms of the sensors, the features of optical images and SAR images are heterogeneous. The direct cascade features of two modalities will cause mutual interference, whether in the data domain or feature domain. Some key features of targets will be buried beneath data from another modality. On the other hand, for the same target, different sensors provide data with different angles, and the characteristics of optical images and SAR images can complement each other. Therefore, inspired by the self-attention calculation method of Transformer, we use multi-head cross-attention calculation for the features of the two modalities, instead of using direct cascade or addition methods to mine the complementary information between two modal features, while ignoring the parts that interfere with each other.

The structure of the cross-fusion module is shown in Figure 4. The features of the SAR image and optical image extracted by the encoder are first fed into the Layer Norm layer, respectively, and then the multi-head cross-attention is calculated. The cross-attention calculation of the i^{th} head of the SAR encoder branch is as follows:

$$Q_i^{SAR} = W_i^{Q^{SAR}} E_{Optical}, K_i^{SAR} = W_i^{K^{SAR}} E_{SAR}, V_i^{SAR} = W_i^{V^{SAR}} E_{SAR} \quad (6)$$

$$CrossAttnHead_i^{SAR}(Q_i^{SAR}, K_i^{SAR}, V_i^{SAR}) = Softmax(\frac{Q_i^{SAR} K_i^{SAR^T}}{\sqrt{d}}) V_i^{SAR} \quad (7)$$

where Q_i^{SAR}, K_i^{SAR}, and V_i^{SAR}, respectively, represent the query matrix, key matrix, and value matrix of the cross-attention of the i^{th} head of the SAR encoder branch; $W_i^{Q^{SAR}}$, $W_i^{K^{SAR}}$, $W_i^{V^{SAR}}$, respectively, represent the learning weights of the query matrix, key matrix, and value matrix learned by the SAR encoder branch.

The output is obtained by the cascading of multiple cross-attention heads:

$$MCA^{SAR}(Q,K,V) = Concat(CrossAttnHead_1^{SAR}W_1^{SAR}, CrossAttnHead_2^{SAR}W_2^{SAR}, \cdots, CrossAttnHead_i^{SAR}W_i^{SAR}) \quad (8)$$

where W_i^{SAR} represents the learnable weight of the SAR encoder branch.

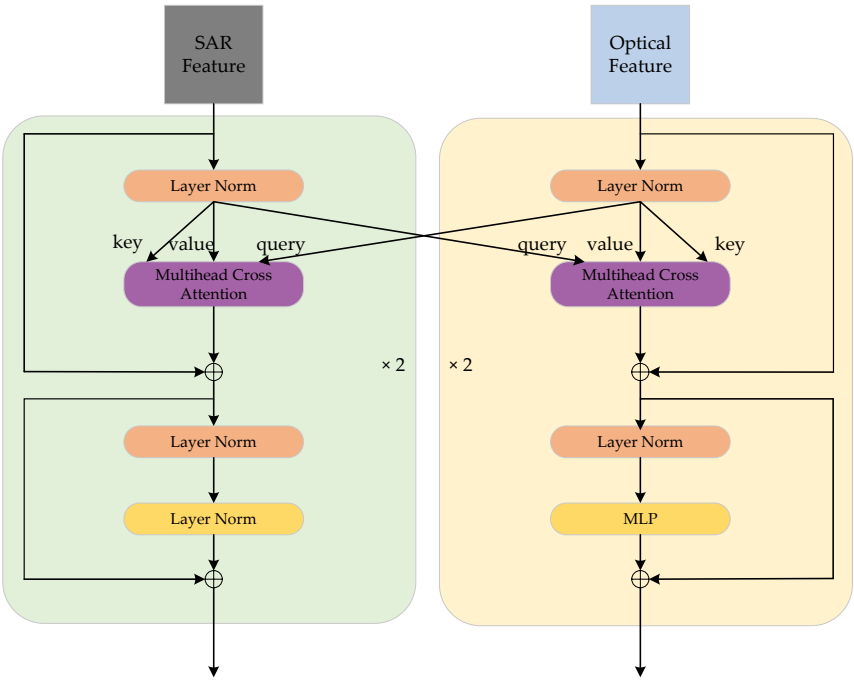


Figure 4. Structural details of the cross-fusion module. In the encoder branch of each modality, the features of this modality itself are used as keys and values in the cross-attention operation, while the features of the other modality are used as queries.

The calculation process of the cross-fusion module of the SAR encoder branch is as follows:

$$\begin{aligned} x_k^1 &= x_{k-1} + MSA^{SAR}(LN(x_{k-1})) \\ x_k^2 &= x_k^1 + MLP(LN(x_k^1)) \end{aligned} \quad (9)$$

In the same way, the cross-attention of the i^{th} head of the optical encoder branch is as follows:

$$Q_i^{Optical} = W_i^{Q^{Optical}} E_{SAR}, K_i^{Optical} = W_i^{K^{Optical}} E_{Optical}, V_i^{Optical} = W_i^{V^{Optical}} E_{Optical} \quad (10)$$

$$CrossAttnHead_i^{Optical}(Q_i^{Optical}, K_i^{Optical}, V_i^{Optical}) = Softmax(\frac{Q_i^{Optical} K_i^{Optical^T}}{\sqrt{d}}) V_i^{Optical} \quad (11)$$

where $Q_i^{Optical}$, $K_i^{Optical}$, and $V_i^{Optical}$, respectively, represent the query matrix, key matrix, and value matrix of the cross-attention of the i^{th} head of the optical encoder branch; $W_i^{Q^{Optical}}$, $W_i^{K^{Optical}}$, $W_i^{V^{Optical}}$, respectively, represent the learning weights of the query matrix, key

matrix, and value matrix that can be learned by the optical encoder branch. The output is obtained by the cascading of multiple cross-attention heads:

$$MCA^{Optical}(Q,K,V) = Concat(CrossAttnHead_1^{Optical}W_1^{Optical}, CrossAttnHead_2^{Optical}W_2^{Optical}, \cdots, CrossAttnHead_i^{Optical}W_i^{Optical}) \quad (12)$$

where $W_i^{Optical}$ represents the learnable weight of the optical encoder branch. The calculation process of the cross-fusion module of the optical encoder branch is as follows:

$$\begin{aligned} x_k^1 &= x_{k-1} + MSA^{Optical}(LN(x_{k-1})) \\ x_k^2 &= x_k^1 + MLP(LN(x_k^1)) \end{aligned} \quad (13)$$

3.3. Two-Task Decoder Module for Footprint Extraction and Height Estimation

In order to improve the estimation precision of the building height, our framework also introduces a branch of footprint extraction while estimating the height map. The structure of the decoder is shown in Figure 5. The two task branches share the output of the previous encoder and perform inference from the lowest-resolution feature map $F_{fusion} \in \mathbb{R}^{H/16\times W/16\times C_3}$. In the decoder stage, the feature map F_{fusion} from the encoder is gradually restored to the original resolution through three decoder blocks. The structure of the decoder block is shown in Figure 6. In the inference process of the intermediate-resolution feature map, the skip connection method is used. The feature map output by the previous decoder block is upsampled and cascaded with the feature map of the same resolution output by the basic feature extraction module in the encoder, which is the input of the next decoder block. The processing procedure of the i^{th} decoder block can be expressed by the following:

$$F_{d_i} = Conv(Conv(Concat(Upsample(F_{d_{i-1}}), F_{e_i}))) \quad (14)$$

where $F_{d_i}, F_{d_{i-1}}$ represent the output of the i^{th} and $i-1^{th}$ decoder block, respectively. F_{e_i} represents the feature map output by the encoder's basic feature extraction module. *Conv* represents the convolution block, which consists of 3×3 convolution and *ReLU* layers. *Upsample* represents a $2\times$ upsampling operator. Through this method, different resolutions of local and global features are simultaneously considered during the decoding inference process. The architectures of the height estimation branch and the footprint extraction branch in the decoder stage are the same, and the two branches do not share parameters.

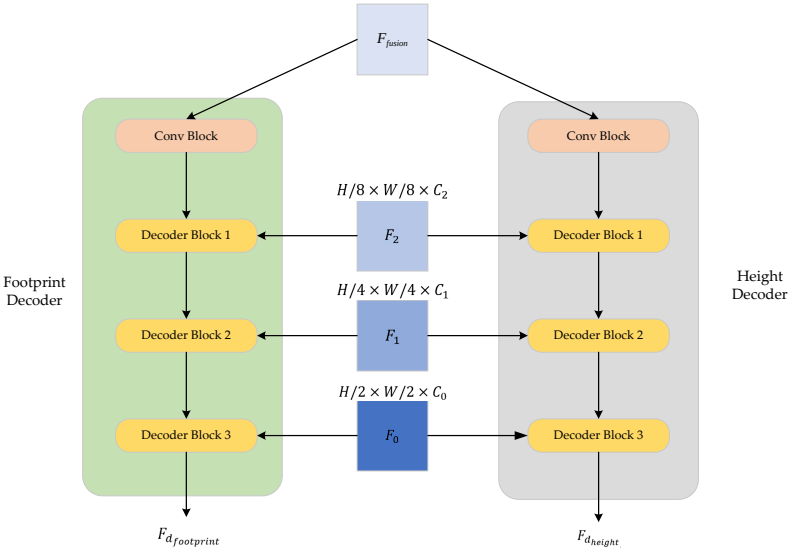


Figure 5. The structure of the two-task branch decoder. The footprint extraction branch and the height estimation branch share the features extracted by the encoder. Each branch contains three decoder blocks, and skip connections are employed at different decoder stages.

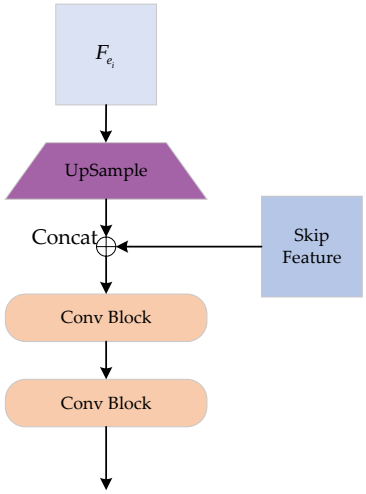


Figure 6. Structural details of decoder block.

3.4. Semantic Information to Refine Height Estimation

In order to generate the footprint map and height map, the outputs of the two decoder branches are respectively processed by the footprint extraction head and the height estimation head, as shown in Figure 7. The footprint extraction head can be expressed by the following formula:

$$Pred_{footprint} = Conv(Conv(F_{d_{footprint}})) \tag{15}$$

where $F_{d_{footprint}} \in \mathbb{R}^{H \times W \times C_{footprint}}$ represents the output of the footprint extraction branch of the decoder, which is input into the footprint extraction head, composed of two convolution blocks. The first convolution block consists of a 3×3 convolution layer and a BN layer. The dimension of the feature map remains unchanged after processing. The second convolution block consists of a 1×1 convolution layer and a BN layer. After processing, the dimension of the output is 2, which represents the two categories of footprint and background. The output probability is used to perform loss calculations with the ground truth of the footprint. The height estimation is expressed by the following:

$$Pred_{height} = Sigmoid(Conv(Conv(F_{d_{height}}))) \times \arg\max(Pred_{footprint}) \tag{16}$$

where $F_{d_{height}} \in \mathbb{R}^{H \times W \times C_{height}}$ represents the output of the height estimation branch of the decoder. Compared with the footprint extraction head, a sigmoid layer is added at the output to limit the final height prediction to between 0 and 1.

In order to improve the precision of height estimation, we propose the use of semantic information to refine the height estimation results. The building footprint semantic information is processed by arg max and used as the correction information of the auxiliary branch, which is pixel-wise multiplied with the height results to refine them and avoid interference from background areas.

3.5. Loss Function

The proposed method consists of two tasks: height estimation and footprint extraction. In the height estimation task, we use the Smooth L1 Loss, which is as follows:

$$L_{height} = SmoothL1Loss(Pred_{height}, GT_{height}) \tag{17}$$

$$SmoothL1Loss(x, y) = \begin{cases} 0.5(x-y)^2 & if|x-y| < 1 \\ |x-y| - 0.5 & otherwise \end{cases} \tag{18}$$

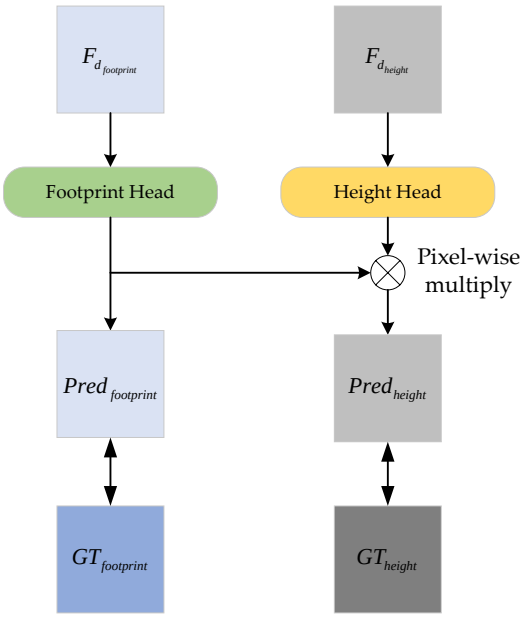


Figure 7. Semantic information used to refine height estimation.

The Smooth L1 Loss is smoother for zero points than the L1 Loss. Compared with the L2 Loss, it responds more gently to outliers and has better robustness in regression problems. The footprint extraction branch, as an auxiliary branch, is used to refine the height estimation. The loss of footprint extraction consists of the Dice loss and cross-entropy loss:

$$L_{footprint} = L_{cross-entropy}(Pred_{height}, GT_{height}) + L_{dice}(Pred_{height}, GT_{height}) \quad (19)$$

$$L_{cross-entropy}(x, y) = -[(y \times log(x)) + (1 - y)log(1 - x)] \quad (20)$$

$$L_{dice}(x, y) = 1 - \frac{2\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i} \quad (21)$$

where N represents the number of pixels. The Dice loss is derived from optimizing the Dice evaluation matrix, which is the abstract representation of the F1 score. Since the proposed method involves the joint learning of two tasks, the final loss function is obtained as the weighted sum of the losses of the two tasks. We assign different weights to the two tasks, which improves the precision of the height estimation results while maintaining the training of the footprint extraction auxiliary task. $\lambda_{footprint}$ is set to 1 and λ_{height} is set to 0.5 in the experimental implementation. The final loss function is as follows:

$$Loss_{final} = \lambda_{footprint} \times L_{footprint} + \lambda_{height} \times L_{height} \quad (22)$$

4. Results and Discussion

4.1. Experimental Setup and Evaluation Metrics

All experiments in this article were conducted under the PyTorch 1.12.1 framework on the Ubuntu 18.04 system. The CUDA version of the experimental platform is 11.3. The hardware platform was equipped with an AMD Ryzen 7 5800X 8 core CPU processor and an Nvidia GeForce RTX 3090 graphics card. The training optimizer was the SGD optimizer, where the weight decay rate was set to 10^{-4} and the momentum weight was set to 0.9. In our proposed framework, the inputs contain two modal data: EO images and SAR image. The final output tasks include height estimation and footprint extraction. All operations are jointly trained from scratch under the framework of joint learning.

For the two tasks of building height estimation and footprint extraction, we apply a series of evaluation metrics to evaluate the performance of each method. For the building

height estimation task, the evaluation metrics include MAE (mean absolute error), MSE (mean square error), $RMSE$ (root mean square error), and threshold precision indicators δ_1, δ_2, δ_3, which are used to measure the proportion of pixels that maintain error control within a specified range and prioritize overall error stability. The evaluation metrics are as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Pred_{height} - GT_{height}\right| \tag{23}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Pred_{height} - GT_{height})^2 \tag{24}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Pred_{height} - GT_{height})^2} \tag{25}$$

$$\delta_1 = \max\left(\frac{Pred_{height}}{GT_{height}}, \frac{GT_{height}}{Pred_{height}}\right) < 1.25^1 \tag{26}$$

$$\delta_2 = \max\left(\frac{Pred_{height}}{GT_{height}}, \frac{GT_{height}}{Pred_{height}}\right) < 1.25^2 \tag{27}$$

$$\delta_3 = \max\left(\frac{Pred_{height}}{GT_{height}}, \frac{GT_{height}}{Pred_{height}}\right) < 1.25^3 \tag{28}$$

For the building footprint extraction task, the evaluation metrics include the $Dice$ score, IoU (intersection over union), $Precision$, $Recall$, and $Accuracy$, which are defined as follows:

$$Dice = \frac{2TP}{2TP + FP + TN} \tag{29}$$

$$IoU = \frac{TP}{TP + FP + TN} \tag{30}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{31}$$

$$Precision = \frac{TP}{TP + FP} \tag{32}$$

$$Recall = \frac{TP}{TP + FN} \tag{33}$$

where TP represents the number of positive samples that are correctly classified as positive samples, TN represents the number of negative samples that are correctly classified as negative samples, FP represents the number of negative samples that are incorrectly classified as positive samples, and FN represents the number of positive samples that are incorrectly classified as positive samples.

4.2. Datasets

All our experiments were conducted on the SpaceNet 6 [59] dataset, which is a multi-sensor all-weather remote sensing dataset, including EO images as well as SAR images. The EO images were collected by the WorldView2 satellite of Maxar with a resolution of 0.5 m. The SAR images were collected by Capella Space, with the same resolution of 0.5 m, and contained four types of polarization data (HH, HV, VH, VV). The area of the SpaceNet 6 dataset was the city of Amsterdam, the Netherlands. The footprints and heights of the buildings in the area were marked. The footprints were marked through the public 3D Basisregistratie Adressen en Gebouwen (3DBAG) dataset, and we removed incorrectly marked, lost, and destroyed buildings. The height annotation was derived from the digital elevation model measured by LiDAR.

In order to evaluate the performance of the methods, all images were uniformly cropped to 512 × 512 pixels. At the same time, we split the dataset according to the longitude and latitude of the area where the images were located, of which 2654 samples were used as the train set and 747 samples were used as the validation set. We ensured that there were no overlapping regions between the train and validation sets to evaluate our method more accurately. In terms of data enhancement, for each sample, we only used random horizontal inversion and did not perform other data enhancement methods.

4.3. Comparison Experiments

In order to evaluate our proposed method, FusionHeightNet, we compare it with other deep learning-based SOTA methods, including ResUNet [60], DeepLabv3+ [29], UperNet [61], PSPNet [31], SegNet [62], and DUC-HDC [63]. The results of all comparison experiments are shown in Table 1. In the tables of this article, an upward arrow (↑) signifies that a larger value corresponds to a better performance metric for the model. Conversely, a downward arrow (↓) indicates that a smaller value is associated with a better performance metric for the model. All methods include two tasks, footprint extraction and height estimation. All methods use the same data enhancement and preprocessing methods and we use the same hyperparameters to ensure the fairness of the comparison experiments. The height maps estimated by the model are normalized to be between 0 and 1. The experiments show that our proposed method has advantages over other SOTA methods.

Table 1. Comparison experiment results with EO and SAR images from the SpaceNet 6 dataset.

Method	***Dice*** ↑	***IoU*** ↑	***Accuracy*** ↑	***Precision*** ↑	***Recall*** ↑	δ_1 ↑	δ_2 ↑	δ_3 ↑	***MAE*** ↓	***MSE*** ↓	***RMSE*** ↓
DeepLabv3+	0.6996	0.4703	0.9635	0.6154	0.5601	0.3179	0.5069	0.5806	0.0596	0.0069	0.0805
ResUNet	0.5527	0.3102	0.9467	0.6149	0.3496	0.2053	0.3266	0.3821	0.0790	0.0107	0.0996
UPerNet	0.7029	0.4663	0.9622	0.6095	0.5628	0.3556	0.5226	0.5708	0.0578	0.0067	0.0789
PSPNet	0.6411	0.3978	0.9501	0.5632	0.4970	0.2156	0.3683	0.4469	0.0645	0.0070	0.0819
SegNet	0.4581	0.2165	0.9333	0.5291	0.2425	0.0784	0.1444	0.1999	0.0881	0.0112	0.1027
DUC-HDC	0.6660	0.4111	0.9562	0.6131	0.4837	0.2708	0.4180	0.4671	0.0671	0.0083	0.0879
FusionHeightNet	**0.7231**	**0.4932**	**0.9660**	**0.6385**	**0.5844**	**0.3849**	**0.5531**	**0.6051**	**0.0540**	**0.0063**	**0.0763**

Our method has more advanced performance in both footprint extraction and height estimation, while the other SOTA pixel-level inference methods all exhibit certain performance limitations. DeepLabv3+ and PSPNet are based on the encoder and decoder architecture, using atrous convolution and pyramid pooling to capture long-distance context information. However, there are disadvantages in fine-grained detail estimation, and the pooling operation reduces the spatial resolution. ResUNet uses a U-shaped network structure, which has more advantages in terms of the spatial resolution, but the performance of the encoder backbone will greatly limit the final inference results. SegNet adopts a more lightweight structure based on the U-shaped architecture. Although this simple structure will improve the computational efficiency, it has weak fine-grained target performance and performs poorly in pixel-level height estimation. DUC-HDC uses dense convolution and pyramid pooling to improve the precision by stacking more complex convolutional layers. This approach creates redundancies in computational efficiency and limits the performance in height estimation. UPerNet achieves relatively good performance among the other methods. It fuses multi-scale information through top-down and bottom-up paths. The overall structure is relatively simple and it avoids redundancies in computational efficiency. It achieves a score of 0.356 in δ_1 for height estimation and 0.703 in *Dice* for footprint extraction.

Our proposed FusionHeightNet achieved the highest δ_1 metric of 0.385 and the highest Dice metric of 0.723, which shows that our method has better performance in both height estimation and footprint extraction tasks. At the same time, our method achieved the lowest errors of 0.0540, 0.0064, and 0.0763 in the MAE, MSE, and $RMSE$ metrics for height estimation, respectively. UPerNet and DeepLabv3+ are the methods closest to ours. Compared with these two methods, our method has 6.9% and 10% improvements in MAE, 4.3% and 7.6% improvements in MSE, and 3.2% and 5.2% improvements in $RMSE$, respectively, which shows that the height estimated by our proposed method is more accurate than that of others. In the footprint extraction auxiliary task, compared with UPerNet and DeepLabv3+, our method has 2% and 2.3% improvements in the *Dice* metric and 2.7% and 2.3% improvements in the IoU, which shows that our proposed method can also extract building footprints more accurately while ensuring high-precision height estimation. It is proven that the two tasks of height estimation and footprint extraction are mutually reinforcing, and the two tasks have similar feature representations in the high-dimensional feature space.

In summary, our proposed method effectively captures the correlation between height estimation and footprint extraction tasks and demonstrates the ability to fuse the complementary feature information of EO images and SAR images, avoiding the interaction of the heterogeneous feature information of the two modalities to the greatest extent. It achieved better performance compared to other state-of-the-art pixel-level methods.

We show the visualization results of all methods in Figures 8 and 9 to evaluate all methods from a quality perspective. It can be clearly seen from the visualization results that the height results predicted by other methods have major flaws compared to our method. Due to the limitation of the encoder backbone structure, ResUNet has many missed detection problems, which will affect the height estimation results. There are many jagged edges in the results of DUC-HDC, and the precision of edge details is insufficient. Overly dense convolution will greatly reduce the efficiency of gradient backpropagation. There are many holes in the results generated by SegNet, the estimated height results are discontinuous, and the lightweight network structure will bring about insufficient receptive fields. The height map structure generated by PSPNet has relatively smooth edges, but the precision of the details is poor, and the height predictions of some structures will be lost. Compared with the above-mentioned methods, DeepLabv3+ and UPerNet obtain better footprint extraction precision, but the difference between the height predictions and the ground truth is obvious, and buildings of different heights are not distinguished. Compared with the above-mentioned SOTA methods, our proposed method is more accurate in estimating the edge details of footprint extraction, and it can also clearly distinguish buildings of different heights in height estimation. This reflects the ability of our method to extract important features from EO images and SAR images, as well as the effectiveness of semantic information to refine the height estimation results. In particular, the joint learning of the two tasks is conducive to mining high-level features to make the final decision.

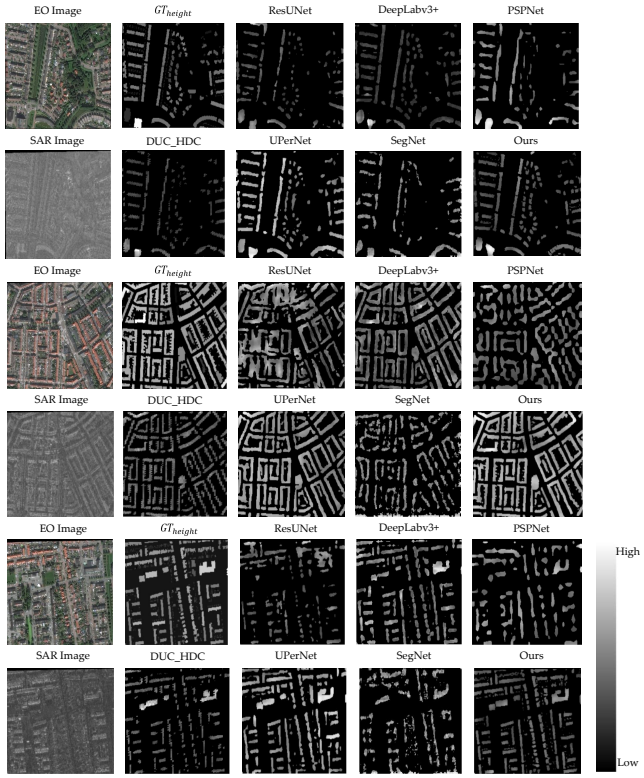


Figure 8. Visualization results of height estimation on SpaceNet 6 dataset (Part 1). In the height result map, brighter pixel colors correspond to greater heights.

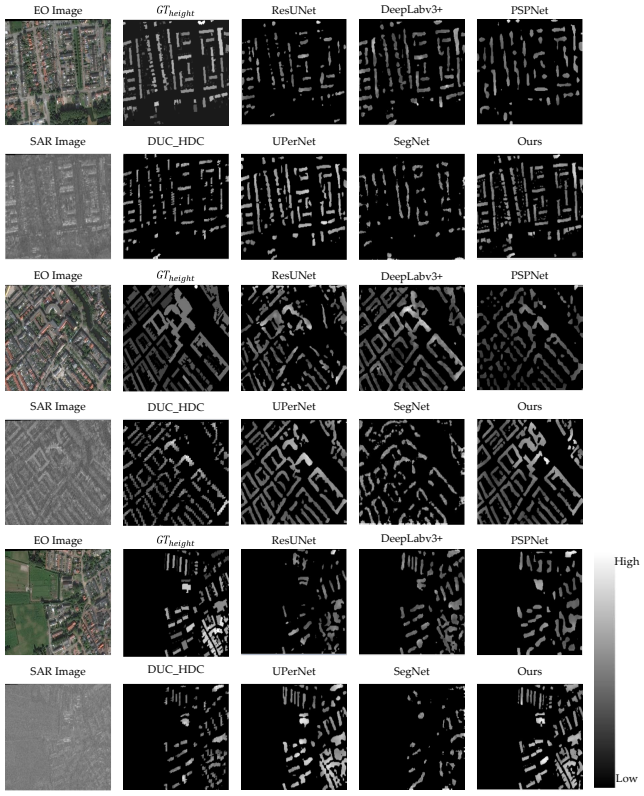


Figure 9. Visualization results of height estimation on SpaceNet 6 dataset (Part 2). In the height result map, brighter pixel colors correspond to greater heights.

4.4. Ablation Experiment and Discussion

In the ablation experiments, we discuss the effectiveness of the multi-level cross-fusion, independent encoders, and multi-task joint learning for semantic information in refining the height estimation results. The experimental results are shown in Tables 2–4. In the experiments, we applied different input combinations of remote sensing data modalities, including only input EO images, only input SAR images, and the joint input of optical images and SAR images. At the same time, we also performed ablation validation on different tasks, including footprint extraction only and the joint learning of footprint extraction and height estimation. Finally, we discuss the contribution of SAR images to the model's performance and the spatial variability of the model's predictability.

Table 2. Ablation experiment results (Part 1). Here, ablation experiments were conducted on different input types and different tasks to evaluate the effectiveness of our proposed method.

Method	EO	SAR	Footprint	Height	*Dice* ↑	*IoU* ↑	*Precision* ↑	*Recall* ↑	δ_1 ↑	*MAE* ↓	*MSE* ↓
EXP-A1	√		√		0.7105	0.4813	0.6249	0.5771			
EXP-A2		√	√		0.4736	0.2445	0.4530	0.3150			
EXP-A3	√	√	√		0.7208	0.4885	0.6269	0.5785			
EXP-A4	√		√	√	0.7148	0.4863	0.6254	0.5804	0.3524	0.0542	0.0065
EXP-A5		√	√	√	0.4789	0.2501	0.4595	0.3193	0.1975	0.0801	0.0104
EXP-A6	√	√	√	√	0.6944	0.4844	0.6229	0.5790	0.3762	0.0549	0.0066
EXP-A7	√	√	√	√	**0.7231**	**0.4932**	**0.6385**	**0.5844**	**0.3849**	**0.0540**	**0.0063**

Table 3. Ablation experiment results (Part 2). Two different encoder designs are verified here, i.e., the two modal images share the same encoder or adopt independent encoders.

Method	EO	SAR	Footprint	Height	*Dice* ↑	*IoU* ↑	*Precision* ↑	*Recall* ↑	δ_1 ↓	*MAE* ↓	*MSE* ↓
EXP-B1	√	√	√	√	0.7093	0.4928	0.6293	0.5795	0.3741	0.0550	0.0065
EXP-B2	√	√	√	√	**0.7231**	**0.4932**	**0.6385**	**0.5844**	**0.3849**	**0.0540**	**0.0063**

Table 4. Comparison experiment results with only EO images from the SpaceNet 6 dataset.

Method	*Dice* ↑	*IoU* ↑	*Accuracy* ↑	*Precision* ↑	*Recall* ↑	δ_1 ↑	δ_2 ↑	δ_3 ↑	*MAE* ↓	*MSE* ↓	*RMSE* ↓
DeepLabv3+	0.6744	0.4605	0.9614	0.6126	0.5556	0.3112	0.4941	0.5605	0.0613	0.0073	0.0827
ResUNet	0.7090	0.4670	0.9634	0.6273	0.5536	0.3053	0.4952	0.5683	0.0621	0.0077	0.0839
UPerNet	0.6943	0.4544	0.9604	0.6108	0.5504	0.3220	0.4985	0.5614	0.0573	0.0065	0.0784
PSPNet	0.6286	0.3770	0.9470	0.5445	0.4828	0.2353	0.3873	0.4490	0.0672	0.0079	0.0861
SegNet	0.4873	0.2605	0.9422	0.6177	0.2825	0.0663	0.1470	0.2166	0.0862	0.0112	0.1018
DUC-HDC	0.6651	0.4102	0.9564	**0.6381**	0.4717	0.2382	0.3861	0.4371	0.0716	0.0091	0.0920
FusionHeightNet	**0.7148**	**0.4863**	**0.9653**	0.6254	**0.5804**	**0.3524**	**0.5409**	**0.6009**	**0.0542**	**0.0065**	**0.0764**

4.4.1. Effectiveness of Multi-Level Cross-Fusion Module

In the ablation experiment shown in Table 2, EXP-A1 only inputs optical images for footprint extraction. Similarly, EXP-A2 only inputs SAR images. EXP-A3 adopts the multi-level cross-fusion method that we propose, and it uses independent encoders for optical images and SAR images. It can be seen from the results that the *Dice* and *IoU* of footprint extraction, obtained by inputting only optical images, are 0.7105 and 0.4813, respectively, while only inputting SAR images results in 0.4736 and 0.2445, respectively. Optical images have better footprint extraction precision than SAR images. This is mainly because the boundaries of targets in SAR images are more difficult to identify due to its imaging mechanism. In addition, there is a large gap in the image quality of SAR images compared with optical images, such as coherence speckle noise. These factors bring great challenges to the footprint extraction of building targets in SAR images.

In EXP-A3, using multi-level cross-fusion, values of 0.7208 and 0.4885 are obtained in *Dice* and *IoU*, respectively, which are 1% and 0.7% higher than in the method of only

inputting optical images. This shows that our method can effectively fuse the features of the two modalities. Next, we added the height estimation branch. In EXP-A4, only optical images were input to predict the height maps and building footprints simultaneously, achieving 0.3524 and 0.7148 in δ_1 and *Dice*, respectively. In EXP-A5, only SAR images were input to simultaneously predict the height maps and building footprints, achieving 0.1975 and 0.4789 in δ_1 and *Dice*, respectively. In EXP-A6, a simple method of cascading two modal feature maps was used to obtain 0.3762 and 0.6944 in δ_1 and *Dice*. Although there is an improvement in height estimation, the accuracy of footprint extraction will decrease. This is because the direct cascading of two heterogeneous features will cause some redundant features to interfere with each other, and complementary effective information will be obscured by this interference.

In EXP-A7, a multi-level cross-fusion method is used to simultaneously predict height maps and building footprints, achieving 0.3849 and 0.7231 in δ_1 and *Dice*, respectively, which are 3.25% and 0.8% higher than in the method of only inputting optical images. This shows that our method effectively integrates the additional effective three-dimensional information provided by SAR images. The three-dimensional information contained in SAR images can provide a gain for height estimation, but when SAR images are input only for height estimation, the poor footprint extraction results limit the precision of height estimation. By fusing optical images, we take advantage of the optical image to more accurately extract footprints and the SAR image to supplement three-dimensional information, effectively improving the height estimation results.

4.4.2. Effectiveness of Multi-Task Joint Learning

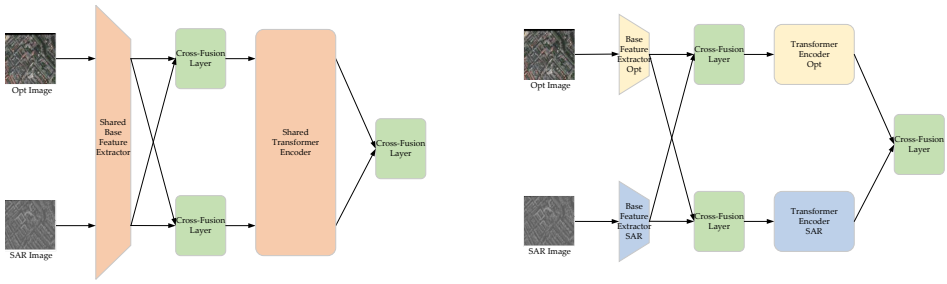
We evaluate the effectiveness of joint learning for the two tasks of height estimation and footprint extraction. In the ablation experiments provided in Table 2, both EXP-A4 and EXP-A1 only input optical images. Compared with EXP-A1, EXP-A4 adds a height estimation branch and adopts a joint learning method for the two tasks. It achieves improvements of 0.43% and 0.5% in the evaluation metrics *Dice* and *IoU* for footprint extraction. Similarly, EXP-A5 and EXP-A2 only input SAR images. Compared with EXP-A1, EXP-A5 adds the height estimation branch and adopts a joint learning method between the two tasks, achieving improvements of 0.53% and 0.56% in the evaluation metrics *Dice* and *IoU*. In EXP-A7 and EXP-A3, which use multi-level cross-fusion, EXP-A7 adopts a joint learning method for the two tasks, and it achieves improvements of 0.23% and 0.47% in the evaluation metrics *Dice* and *IoU* for footprint extraction. Thus shows that our multi-task joint learning method and the use of semantic information to refine the height results achieve the mutual promotion of the two tasks.

The footprint extraction and height estimation of buildings in a single image are two different but closely related tasks. Although the final output results of the two tasks are different, they complement each other in understanding the image content and scene structure. Both tasks share feature representations and mutually promote each other in the final output results. More accurate semantic information of footprints is beneficial in inferring height results, and the provision of depth information can also help with more accurate footprint extraction. Therefore, there is a strong correlation between the footprint extraction task and the height estimation task in the high-level feature representation space. The method designed in this article involves the end-to-end joint learning of the two tasks, where they learn and optimize each other during the training process to achieve better results. In the output of the height estimation branch, we employed footprint semantic information to refine the height estimation results by exploring the relevance between the two tasks in terms of understanding the image content and building structure, which can effectively mine the feature representation shared by the two tasks.

4.4.3. Effectiveness of Independent Encoders

We evaluate the design of encoders for images from two modalities. In EXP-B1, optical images and SAR images share the same encoder, including sharing the same basic

feature extraction module and high-level feature inference module. In EXP-B2, independent encoders are used for images of the two modalities. The comparison of the two methods is shown in Figure 10. In Table 3, the evaluation results of the two designs are shown. The design of two modalities sharing an encoder achieved scores of 0.3741 and 0.7093 in δ_1 and *Dice*, respectively. The design of two modalities using independent encoders achieved scores of 0.3849 and 0.7231 in δ_1 and *Dice*, with improvements of 1.08% and 1.38%, respectively, which shows that optical images and SAR images have obvious differences in the feature space and independent encoders are conducive to the extraction of effective features for their respective feature spaces.



(a) EXP-B1 (b) EXP-B2

Figure 10. Two different designs of encoders for two modalities. The left is the design of EXP-B1, where the input images of two modalities share the same basic feature extraction module and high-level feature inference module in the encoder. The right is the design of EXP-B2, where the input images of the two modalities use independent basic feature extraction and high-level feature inference module in the encoder.

4.4.4. Contribution of SAR Images to the Model's Performance

This section discusses the contribution of SAR images to model performance. The results of only inputting SAR images are generally poor, mainly due to the poor readability and coherent speckle of SAR images, obvious geometric distortion, target overlap, and unclear edges of targets. These factors greatly increase the difficulty in extracting effective features from SAR images by the model.

As two types of remote sensing images obtained by different sensors, EO images and SAR images observe targets from different angles. EO images provide color and texture details of building targets, imaged at an approximate top-down angle. SAR images are different from EO images in that they are imaged from a side view angle. The two coordinate axes of the image represent the azimuth and range directions, respectively, and the pixel values represent the radar echo intensity rather than color information. These unique characteristics enable SAR images to provide information that cannot be provided by EO images, such as the lateral structure information of targets and distinguishing different targets based on the scattering characteristics of different surface materials, which is beneficial in improving the estimation results of the building height.

The information between SAR images and EO images can complement each other. Remote sensing images with different modalities have some heterogeneous features, sometimes even overwhelming the effective homogeneous features to be extracted. Therefore, this article explores an effective fusion method to avoid heterogeneous feature interference as much as possible, extract effective homogeneous features between SAR images and EO images, and improve the performance of downstream tasks. Table 4 shows the metric results of different models that only input EO images. Comparing the results of inputting EO and SAR images in Table 1, different models perform differently under different inputs. Most models perform better when inputting both EO and SAR images, while the ResUNet and PSPNet methods perform worse. The difference in results depends on the different feature extraction structures and multi-source data fusion methods. This indicates that SAR does not always reduce the performance of the model. On the contrary, SAR images can provide supplementary information for EO images.

The results in Tables 1 and 4 show that the FusionHeightNet proposed in this article designs effective multi-source remote sensing data fusion algorithms to avoid redundant heterogeneous features as much as possible, extract effective homogeneous features of the target, and improve the network model's understanding of the target structure.

4.4.5. Spatial Variability of the Model's Predictability

To further evaluate the spatial variability of the proposed model, we present the visual prediction results and performance metrics of the model in different scenarios in Figures 11 and 12. Figure 11 shows scenes where the model performs well, with δ_1 values exceeding 0.5. Figure 12 shows scenes with unsatisfactory model performance, with δ_1 values below 0.2.

When comparing the two scenarios, the model generally performs better in scenarios where the height distribution of the buildings is relatively average. However, in some scenarios with significant height differences, the model may experience significant errors. This is mainly because the loss of a single scene image is the accumulation of all pixel errors, with higher buildings contributing more to the total loss, while lower buildings contribute less. Therefore, the focus of model optimization will turn to higher buildings and it will reduce the focus on lower buildings.

In addition, when estimating buildings with larger footprint areas composed of multiple substructures, the model may also encounter the problem of uneven height estimation and predict height values with significant deviations for individual buildings. This is mainly due to some structural interference at the top of the building, which causes the model to incorrectly identify a single building as multiple structures of different heights.

At the same time, we also found that good footprint extraction results can improve the height estimation results, even in some scenarios with large height variance, which is consistent with the correlation between footprint extraction and height estimation mentioned earlier.

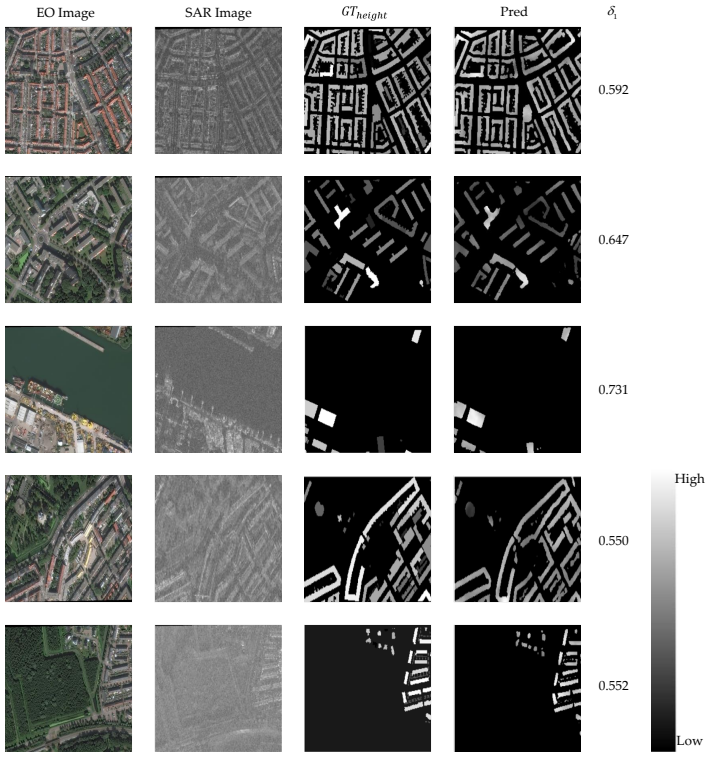


Figure 11. The scenes where the model performs well, with δ_1 values exceeding 0.5. In the height result map, brighter pixel colors correspond to greater heights.

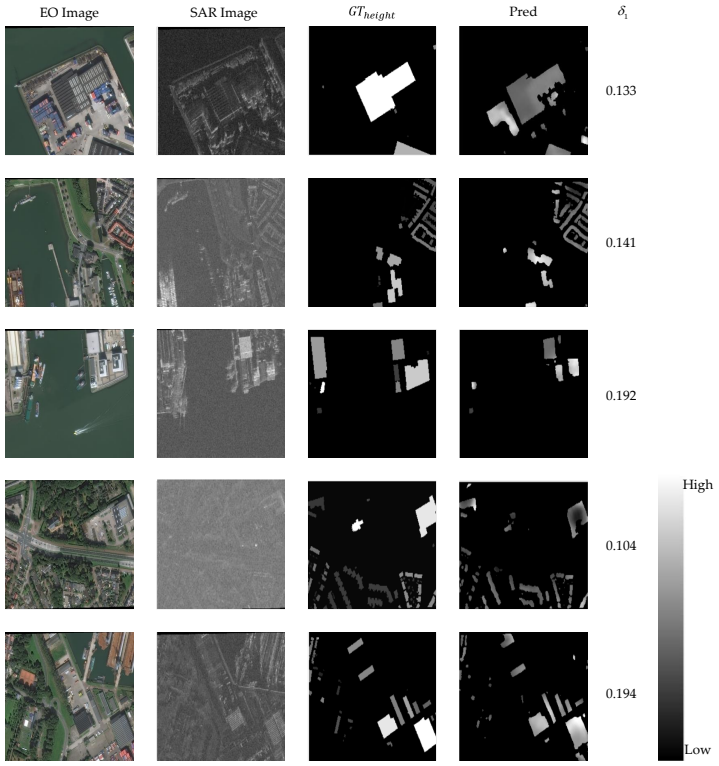


Figure 12. The scenes where the model performs unsatisfactorily, with δ_1 values below 0.2. In the height result map, brighter pixel colors correspond to greater heights.

5. Conclusions

In this paper, we propose an advanced method based on multi-source remote sensing image fusion for building height estimation, named FusionHeightNet. The proposed method introduces the auxiliary task of footprint extraction, aiming to achieve the joint learning of building height estimation and footprint extraction under one framework. In order to solve the problem of differences in the feature representations of the two modalities, we propose a multi-level cross-fusion method. The images of the two modalities use independent encoders; cross-fusion attention modules are designed at different positions of the encoders to extract complementary and effective information of the two modalities and avoid mutual interference between redundant features, preventing the flooding of effective information. By fusing multi-source remote sensing images, we make full use of the advantages of optical images to more accurately extract footprints and the supplementary three-dimensional information from SAR images, effectively improving the height estimation results. In addition, we also propose a two-task joint learning architecture using semantic information to refine the height estimation, fully exploring the strong correlation between the footprint extraction task and the height estimation task in the high-level feature space. The experimental results show that our method outperforms other deep learning-based methods in the evaluation metrics of height estimation and footprint extraction. We also evaluate the effectiveness of the multi-level cross-fusion, independent encoders, and multi-task joint learning of semantic information to refine the height results through ablation experiments. Our study provides a new solution for the application of multi-source remote sensing fusion in downstream tasks. In the future, we will explore more effective fusion methods for the mining of multi-source remote sensing images to improve the performance in downstream tasks, as well as better methods for the multi-task allocation of footprint extraction and height estimation.

Author Contributions: Conceptualization, methodology, C.M. and Y.Z.; software, validation, C.M. and J.G.; writing—original draft preparation, C.M. and X.G.; writing—review and editing, C.M. and

Y.Z.; visualization, supervision, G.Z. and Y.Z.; funding acquisition, Y.Z. and G.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant numbers 61991421 and 61991420, and the Key Research and Development Program of the Aerospace Information Research Institute Chinese Academy of Sciences, grant number E1Z208010F.

Data Availability Statement: The datasets presented in this paper are available through https://spacenet.ai/sn6-challenge/, accessed on 11 February 2020.

Acknowledgments: The authors would like to thank the Aerospace Information Research Institute under the Chinese Academy of Sciences.

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

Abbreviations

The following abbreviations are used in this manuscript:

SAR	synthetic aperture radar
EO	electro-optical
CNN	convolutional neural network
ViT	Vision Transformer
DETR	Detection Transformer
GAN	generative adversarial network
DSM	digital surface model
FCN	fully convolutional network
ASPP	atrous spatial pyramid pooling
InSAR	interferometric synthetic aperture radar
IHS	inverse hyperbolic sine
PCA	principal component analysis
MSA	multi-head self-attention
MLP	multi-layer perceptron
LN	layer norm
ReLU	rectified linear unit
MAE	mean absolute error
MSE	mean square error
IoU	intersection over union
RMSE	root mean square error
LiDAR	light detection and ranging
SOTA	state of the art
FSG	fuzzy stacked generalization
VAEs	variational autoencoders

References

1. Poister, T.H.; Streib, G. Elements of strategic planning and management in municipal government: Status after two decades. *Public Adm. Rev.* **2005**, *65*, 45–56. [CrossRef]
2. Mäntysalo, R.; Jarenko, K.; Nilsson, K.L.; Saglie, I.L. Legitimacy of informal strategic urban planning—Observations from Finland, Sweden and Norway. *Eur. Plan. Stud.* **2015**, *23*, 349–366. [CrossRef]
3. Couclelis, H. The construction of the digital city. *Environ. Plan. B Plan. Des.* **2004**, *31*, 5–19. [CrossRef]
4. Turaga, P.; Chellappa, R.; Subrahmanian, V.S.; Udrea, O. Machine recognition of human activities: A survey. *IEEE Trans. Circuits Syst. Video Technol.* **2008**, *18*, 1473–1488. [CrossRef]
5. Durieux, L.; Lagabrielle, E.; Nelson, A. A method for monitoring building construction in urban sprawl areas using object-based analysis of Spot 5 images and existing GIS data. *ISPRS J. Photogramm. Remote Sens.* **2008**, *63*, 399–408. [CrossRef]
6. Hsu, L.T.; Gu, Y.; Kamijo, S. Autonomous driving positioning using building model and DGNSS. In Proceedings of the IEEE 2016 European Navigation Conference (ENC), Helsinki, Finland, 30 May–2 June 2016; pp. 1–7.
7. Mou, L.; Zhu, X.X. IM2HEIGHT: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network. *arXiv* **2018**, arXiv:1802.10249.
8. Collis, R. Lidar. *Appl. Opt.* **1970**, *9*, 1782–1788. [CrossRef] [PubMed]
9. Chen, Y.; Yan, Q.; Huang, W. MFTSC: A Semantically Constrained Method for Urban Building Height Estimation Using Multiple Source Images. *Remote Sens.* **2023**, *15*, 5552. [CrossRef]

10. Liu, W.; Sun, X.; Zhang, W.; Guo, Z.; Fu, K. Associatively segmenting semantics and estimating height from monocular remote-sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–17. [CrossRef]
11. Moreira, A.; Prats-Iraola, P.; Younis, M.; Krieger, G.; Hajnsek, I.; Papathanassiou, K.P. A tutorial on synthetic aperture radar. *IEEE Geosci. Remote Sens. Mag.* **2013**, *1*, 6–43. [CrossRef]
12. Cai, B.; Shao, Z.; Huang, X.; Zhou, X.; Fang, S. Deep learning-based building height mapping using Sentinel-1 and Sentienl-2 data. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *122*, 103399.
13. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
14. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
15. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; Springer: Heidelberg, Germany, 2020; pp. 213–229.
16. Chen, Z.; Zhang, Y.; Qi, X.; Mao, Y.; Zhou, X.; Wang, L.; Ge, Y. HeightFormer: A Multilevel Interaction and Image-Adaptive Classification–Regression Network for Monocular Height Estimation with Aerial Images. *Remote Sens.* **2024**, *16*, 295. [CrossRef]
17. Liu, C.J.; Krylov, V.A.; Kane, P.; Kavanagh, G.; Dahyot, R. IM2ELEVATION: Building height estimation from single-view aerial imagery. *Remote Sens.* **2020**, *12*, 2719. [CrossRef]
18. Chen, P.Y.; Liu, A.H.; Liu, Y.C.; Wang, Y.C.F. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2624–2632.
19. Jiao, J.; Cao, Y.; Song, Y.; Lau, R. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 53–69.
20. Senaras, C.; Ozay, M.; Vural, F.T.Y. Building detection with decision fusion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2013**, *6*, 1295–1304. [CrossRef]
21. Karantzalos, K.; Argialas, D. A region-based level set segmentation for automatic detection of man-made objects from aerial and satellite images. *Photogramm. Eng. Remote Sens.* **2009**, *75*, 667–677. [CrossRef]
22. Cote, M.; Saeedi, P. Automatic rooftop extraction in nadir aerial imagery of suburban regions using corners and variational level set evolution. *IEEE Trans. Geosci. Remote Sens.* **2012**, *51*, 313–328. [CrossRef]
23. Huang, X.; Zhang, L. A multidirectional and multiscale morphological index for automatic building extraction from multispectral GeoEye-1 imagery. *Photogramm. Eng. Remote Sens.* **2011**, *77*, 721–732. [CrossRef]
24. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
25. 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—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18; Springer: Cham, Switzerland, 2015; pp. 234–241.
26. 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.
27. 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] [PubMed]
28. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
29. 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.
30. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
31. 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.
32. Pentland, A. Shape information from shading: A theory about human perception. In Proceedings of the Second International Conference on Computer Vision, Tampa, FL, USA, 5–8 December 1988; IEEE: Toulouse, France, 1988; pp. 404–413.
33. Liu, J.; Fang, T.; Li, D. Shadow detection in remotely sensed images based on self-adaptive feature selection. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 5092–5103.
34. Zhang, H.; Sun, K.; Li, W. Object-oriented shadow detection and removal from urban high-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 6972–6982. [CrossRef]
35. Teke, M.; Başeski, E.; Ok, A.Ö.; Yüksel, B.; Şenaras, Ç. Multi-spectral false color shadow detection. In Proceedings of the ISPRS Conference on Photogrammetric Image Analysis, Munich, Germany, 5–7 October 2011; Springer: Cham, Switzerland, 2011; pp. 109–119.
36. Zhou, X.; Myint, S.W. Shadow Pattern-Enhanced Building Height Extraction Using Very-High-Resolution Image. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *16*, 180–190. [CrossRef]

37. Comber, A.; Umezaki, M.; Zhou, R.; Ding, Y.; Li, Y.; Fu, H.; Jiang, H.; Tewkesbury, A. Using shadows in high-resolution imagery to determine building height. *Remote Sens. Lett.* **2012**, *3*, 551–556. [CrossRef]
38. Benarchid, O.; Raissouni, N.; El Adib, S.; Abbous, A.; Azyat, A.; Achhab, N.B.; Lahraoua, M.; Chahboun, A. Building extraction using object-based classification and shadow information in very high resolution multispectral images, a case study: Tetuan, Morocco. *Can. J. Image Process. Comput. Vis.* **2013**, *4*, 1–8.
39. Raju, P.; Chaudhary, H.; Jha, A. Shadow analysis technique for extraction of building height using high resolution satellite single image and accuracy assessment. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2014**, *40*, 1185–1192. [CrossRef]
40. Yang, C.; Zhao, S. A building height dataset across China in 2017 estimated by the spatially-informed approach. *Sci. Data* **2022**, *9*, 76. [CrossRef]
41. Li, X.; Zhou, Y.; Gong, P.; Seto, K.C.; Clinton, N. Developing a method to estimate building height from Sentinel-1 data. *Remote Sens. Environ.* **2020**, *240*, 111705. [CrossRef]
42. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*.
43. Amirkolaee, H.A.; Arefi, H. Height estimation from single aerial images using a deep convolutional encoder-decoder network. *ISPRS J. Photogramm. Remote Sens.* **2019**, *149*, 50–66. [CrossRef]
44. Ghamisi, P.; Yokoya, N. IMG2DSM: Height simulation from single imagery using conditional generative adversarial net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 794–798. [CrossRef]
45. Paoletti, M.E.; Haut, J.M.; Ghamisi, P.; Yokoya, N.; Plaza, J.; Plaza, A. U-IMG2DSM: Unpaired simulation of digital surface models with generative adversarial networks. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 1288–1292. [CrossRef]
46. Pedergnana, M.; Marpu, P.R.; Dalla Mura, M.; Benediktsson, J.A.; Bruzzone, L. Classification of remote sensing optical and LiDAR data using extended attribute profiles. *IEEE J. Sel. Top. Signal Process.* **2012**, *6*, 856–865. [CrossRef]
47. Chini, M.; Pierdicca, N.; Emery, W.J. Exploiting SAR and VHR optical images to quantify damage caused by the 2003 Bam earthquake. *IEEE Trans. Geosci. Remote Sens.* **2008**, *47*, 145–152. [CrossRef]
48. Ghamisi, P.; Benediktsson, J.A.; Phinn, S. Land-cover classification using both hyperspectral and LiDAR data. *Int. J. Image Data Fusion* **2015**, *6*, 189–215. [CrossRef]
49. Pedergnana, M.; Marpu, P.R.; Dalla Mura, M.; Benediktsson, J.A.; Bruzzone, L. A novel technique for optimal feature selection in attribute profiles based on genetic algorithms. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 3514–3528. [CrossRef]
50. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
51. Rasti, B.; Ghamisi, P. Remote sensing image classification using subspace sensor fusion. *Inf. Fusion* **2020**, *64*, 121–130. [CrossRef]
52. Rasti, B.; Ulfarsson, M.O.; Sveinsson, J.R. Hyperspectral feature extraction using total variation component analysis. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 6976–6985. [CrossRef]
53. Rasti, B.; Ghamisi, P.; Gloaguen, R. Hyperspectral and LiDAR fusion using extinction profiles and total variation component analysis. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3997–4007. [CrossRef]
54. Moosavi, V.; Talebi, A.; Mokhtari, M.H.; Shamsi, S.R.F.; Niazi, Y. A wavelet-artificial intelligence fusion approach (WAIFA) for blending Landsat and MODIS surface temperature. *Remote Sens. Environ.* **2015**, *169*, 243–254. [CrossRef]
55. 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]
56. Li, H.; Ghamisi, P.; Soergel, U.; Zhu, X.X. Hyperspectral and LiDAR fusion using deep three-stream convolutional neural networks. *Remote Sens.* **2018**, *10*, 1649. [CrossRef]
57. Audebert, N.; Le Saux, B.; Lefèvre, S. Joint learning from earth observation and openstreetmap data to get faster better semantic maps. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 67–75.
58. Ming, Y.; Meng, X.; Fan, C.; Yu, H. Deep learning for monocular depth estimation: A review. *Neurocomputing* **2021**, *438*, 14–33. [CrossRef]
59. Shermeyer, J.; Hogan, D.; Brown, J.; Van Etten, A.; Weir, N.; Pacifici, F.; Hansch, R.; Bastidas, A.; Soenen, S.; Bacastow, T.; et al. SpaceNet 6: Multi-sensor all weather mapping dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 196–197.
60. Xiao, X.; Lian, S.; Luo, Z.; Li, S. Weighted res-unet for high-quality retina vessel segmentation. In Proceedings of the 2018 9th International Conference on Information Technology in Medicine and Education (ITME), Hangzhou, China, 19–21 October 2018; IEEE: Toulouse, France, 2018; pp. 327–331.
61. Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; Sun, J. Unified perceptual parsing for scene understanding. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 418–434.

62. Badrinarayanan, V.; Kendall, A.; SegNet, R.C. A deep convolutional encoder-decoder architecture for image segmentation. *arXiv* **2015**, *5*, 2481–2495. [CrossRef]
63. Wang, P.; Chen, P.; Yuan, Y.; Liu, D.; Huang, Z.; Hou, X.; Cottrell, G. Understanding convolution for semantic segmentation. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; IEEE: Toulouse, France, 2018; pp. 1451–1460.