remote sensing

MDPI

Article

MFTSC: A Semantically Constrained Method for Urban Building Height Estimation Using Multiple Source Images

Yuhan Chen [1,2], Qingyun Yan [1,*] and Weimin Huang [3]

[1] School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China; yhchen@hrbeu.edu.cn
[2] Qingdao Innovation and Development Base (Centre), Harbin Engineering University, Qingdao 266400, China
[3] Faculty of Engineering and Applied Science, Memorial University, St. John's, NL A1B 3X5, Canada; weimin@mun.ca
* Correspondence: 003257@nuist.edu.cn

Abstract: The use of remote sensing imagery has significantly enhanced the efficiency of building extraction; however, the precise estimation of building height remains a formidable challenge. In light of ongoing advancements in computer vision, numerous techniques leveraging convolutional neural networks and Transformers have been applied to remote sensing imagery, yielding promising outcomes. Nevertheless, most existing approaches directly estimate height without considering the intrinsic relationship between semantic building segmentation and building height estimation. In this study, we present a unified architectural framework that integrates the tasks of building semantic segmentation and building height estimation. We introduce a Transformer model that systematically merges multi-level features with semantic constraints and leverages shallow spatial detail feature cues in the encoder. Our approach excels in both height estimation and semantic segmentation tasks. Specifically, the coefficient of determination (R^2) in the height estimation task attains a remarkable 0.9671, with a root mean square error (RMSE) of 1.1733 m. The mean intersection over union (mIoU) for building semantic segmentation reaches 0.7855. These findings underscore the efficacy of multi-task learning by integrating semantic segmentation with height estimation, thereby enhancing the precision of height estimation.

Keywords: height estimation; multi-task learning; Vision Transformer; remote sensing; synthetic aperture radar

Citation: 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. https://doi.org/10.3390/rs15235552

Academic Editors: Kuishuang Feng, Xiaowei Jia and Yiqun Xie

Received: 24 October 2023
Revised: 24 November 2023
Accepted: 27 November 2023
Published: 29 November 2023

1. Introduction

Buildings play a pivotal role in urban areas, and the analysis of their distribution holds substantial value for a variety of applications, including the assessments of urban livability [1,2] and urban planning [3]. Consequently, continuous monitoring of building changes remains an essential task. Furthermore, the precise determination of relative building heights is of paramount importance in the domains of urban planning and development.

Traditional methods of updating building data are burdened with substantial costs in terms of labor and resources, rendering comprehensive coverage and standardized information a challenging endeavor [4]. Thankfully, remote sensing technology provides a highly accurate means of obtaining a wide range of data related to building heights. These data can be effectively harnessed for the formulation of comprehensive urban planning schemes, the evaluation of urban volume and floor area ratios [5], and its utilization as fundamental data for urban disaster prevention and mitigation [6]. In practice, the increasing utilization of multi-source high-resolution satellite data offers a promising avenue for efficiently extracting building information over expansive areas through remote sensing techniques [7].

The remote sensing data used for estimating the height of surface objects can be broadly categorized into three groups: optical images [8,9], synthetic aperture radar (SAR) images [10–14], and the fusion of these two data sources [15,16].

Optical remote sensing images offer a rich source of visual information, encompassing attributes such as building size, shape, and relative positioning. By integrating the analysis of visual cues like perspective relationships, shadows, and textures, along with the application of image measurement principles and feature extraction algorithms, it becomes possible to deduce relative height differences between buildings [17,18]. SAR serves as an active ground detection technique, providing robust penetrability that allows it to effectively penetrate through clouds, smoke, and vegetation. This capability yields valuable information about terrain and ground objects [19]. The estimation of the height of ground objects in SAR images relies on the analysis of phase information, particularly examining the phase differences between adjacent pixels. Consequently, SAR is widely utilized for the height estimation of ground objects [11–13]. The fusion of SAR and optical images for building height extraction capitalizes on the distinctive imaging characteristics of both modalities. By combining the respective strengths of SAR and optical data through image fusion, more accurate building height data can be extracted [20].

The rapid progress in computer vision has enabled the estimation of relative height from a single image. This achievement is realized through data-driven methods that learn implicit mapping relationships [21], which are not explicitly derived from mathematical modeling. Unlike conventional mathematical modeling approaches, this data-driven method does not require precise modeling of physical parameters like depth of field, and the camera's internal and external characteristics. Instead, it leverages extensive image datasets for training, facilitating the acquisition of more intricate representations of height-related features. Consequently, significant advancements have been made in monocular depth estimation (MDE) tasks [22–25]. MDE involves the estimation of object depths in a scene from a single 2D image, a task closely related to building height estimation. Several methods based on Vision Transformers (ViTs) [25,26] have been introduced. ViT offers superior feature extraction capabilities, robustness, interpretability, and generalization abilities in comparison to convolutional neural networks (CNNs). It can adapt to images of various sizes and shapes, allowing the learning of comprehensive feature representations from extensive image data.

ViT [27] has made significant strides in the past three years and has found extensive applications in semantic segmentation [28–30] and depth estimation [25,26]. In the realm of semantic segmentation, ViT restores the feature map to the original image size and conducts pixel-wise classification by incorporating an upsampling layer or a transposed convolutional layer into the network architecture. This approach allows for efficient processing and precise prediction of large-scale image data, providing robust support for a variety of computer vision tasks. In the context of depth estimation, ViT facilitates the reconstruction of 3D scenes by estimating depth information from a single image. This data-driven approach learns implicit mapping relationships, enabling the prediction of scene depth information from the image.

Currently, there is a paucity of research on height estimation using multi-source remote sensing images, especially within the context of multi-task learning with semantic constraints to enhance height estimation. Existing studies primarily concentrate on analyzing remote sensing mechanisms or utilizing multi-view remote sensing images for relative height estimation through dense matching [17,31,32]. Recent endeavors have explored the utilization of SAR or optical remote sensing data for multi-task learning [7,33–35]. Additionally, some studies have integrated ground object height and RGB images to perform semantic segmentation tasks [36]. These studies have showcased promising results, signifying that the joint processing of SAR and high-resolution remote sensing data can bolster the accuracy of building extraction and height estimation tasks. Moreover, they underscore the intrinsic relationship between semantic information and ground object height, highlighting the effectiveness and necessity of simultaneously conducting semantic

segmentation and height estimation tasks. In recent years, deep learning methods have been employed for relative height estimation through generative techniques [37–39], as well as end-to-end approaches [40,41]. For semantic segmentation, regressing the height of the building area using a height estimation model necessitates the effective separation of the building from the background while estimating its height. The continuity of the regression model presents challenges in distinguishing the foreground and background in the height estimation task. Traditionally, a threshold is set for post-processing, but semantic segmentation tasks are adept at learning to differentiate the foreground and background, offering significant assistance in this regard.

In this study, our objective is to enhance the accuracy of height estimation by incorporating building semantic information constraints into the relative height estimation task. To accomplish this, we introduce a novel approach named the multi-level feature fusion Transformer with semantic constraint(s) (MFTSC), designed to jointly address building extraction and height estimation. Our methodology leverages multimodal optical and SAR satellite imagery as the input data. In the context of single-view satellite imagery, the process of extracting buildings and estimating their heights heavily relies on semantic features derived from the imagery. By integrating multiple tasks and capitalizing on feature reuse, our approach presents a potentially superior solution when compared to conventional individual implementations, enabling the establishment of implicit constraints across these tasks.

Our study makes several noteworthy contributions, which are summarized as follows:

(1) We introduce a high-precision height estimation method that employs the Swin Transformer as the backbone of the encoder and incorporates the use of prompts for shallow information. In the decoder, we employ a straightforward cross-connection design and utilize the Transformer to fuse shallow and deep features. This approach significantly enhances the model's ability to capture ground details.

(2) We devise a unified architecture that integrates building height estimation and building semantic segmentation tasks, with both sharing the same encoder. This joint training approach establishes an implicit constraint, thereby improving the accuracy of both tasks.

(3) To bridge the existing gap in the field of combining building extraction and height estimation tasks from multi-source remote sensing images, we conducted a comprehensive series of comparative experiments. The aim was to provide detailed experimental results that contribute to the current body of knowledge in this domain.

2. Related Work

2.1. MDE

Estimating building height is conceptually similar to MDE, a well-explored field in computer vision. MDE focuses on estimating the depth of objects within a scene from a single 2D image [24]. This task shares common challenges with the estimation of ground object height from remote sensing images. Both involve the complexity of recovering depth information from a 2D image projection of a 3D scene, where depth information is inherently lost, and its retrieval from a single image is challenging. MDE has diverse applications, including 3D reconstruction [42], autonomous navigation [43], augmented reality [24], and virtual reality [24]. Recent years have witnessed significant progress in MDE, primarily driven by advancements in deep learning techniques and the availability of extensive datasets for training depth estimation models. The prevalent approach in MDE is to train deep neural networks to directly predict depth maps from 2D images. These networks are typically trained on large-scale image datasets that include corresponding depth maps, employing techniques such as supervised learning [25,26,44,45], unsupervised learning [46], or self-supervised learning [47].

2.2. Semantic Segmentation

Semantic segmentation is a pixel-level classification task, and many semantic segmentation models adopt the encoder–decoder architecture, exemplified by models like Unet [48,49], LinkNet [50,51], PSPNET [52], and more. Various studies utilizing Unet-based approaches have been instrumental in automatically extracting buildings from remote sensing imagery [53,54]. In recent times, there has been a surge of interest in directly integrating semantic segmentation with the task of height estimation, all from a single remote sensing image [55–57]. These studies have consistently demonstrated that the incorporation of semantic information can significantly enhance the accuracy of height estimation. Nonetheless, the manual annotation of tags can be a cumbersome process, necessitating exploration into methods to streamline the semantic tagging procedure. Given this imperative, there is an urgent need to investigate the feasibility and efficacy of employing building tags exclusively for this purpose.

2.3. ViT

The advent of the Vision Transformer (ViT) [27] has captured the interest of computer vision researchers. However, pure Transformers exhibit high computational complexity and involve a substantial number of model parameters, demanding extensive optimization efforts for ViT. A promising development in this regard is the Swin Transformer [29], which represents a hierarchical Transformer and offers a versatile backbone for various computer vision tasks. By implementing shifted window computations, self-attention is constrained within non-overlapping local windows while also allowing for cross-window connections, leading to enhanced computational efficiency. This layered architecture excels in modeling across different scales and maintains linear computational complexity concerning image size. The Swin Transformer has found wide applications in remote sensing, including hyperspectral classification [58] where a multi-scale mixed spectral attention model based on the Swin Transformer achieved top-class performance across multiple datasets. Additionally, the work of Wang et al. [28] introduced BuildFormer, a novel Vision Transformer featuring a dual-path structure. This innovative design accommodates the use of a large window for capturing global context, substantially enhancing its capabilities for processing extensive remote sensing imagery.

2.4. Multi-Modal Fusion and Joint Learning for Remote Sensing

SAR offers the capability to retrieve height information of ground objects by analyzing the phase and amplitude information of radar echoes. However, the accurate retrieval of height information using SAR data is a complex process, as it is influenced by various factors, including terrain, vegetation, and buildings. This extraction process typically involves intricate signal processing and data analysis techniques. Nevertheless, deep learning has emerged as a promising approach to simplify the height extraction process and enable end-to-end elevation information extraction [40,41]. However, most existing research in this domain focuses on single data sources or single-task-based high-level information extraction, which may not generalize well to multi-source remote sensing data or multi-task joint learning. Researchers are actively exploring various methods, such as multi-modal fusion and multi-task learning, to enhance the accuracy and efficiency of height extraction from SAR data. Multi-task learning using both optical and SAR data is a complex endeavor that involves intricate processing and analysis. Acquiring suitable datasets that contain high-resolution optical and SAR data to support such tasks is also a challenging issue. Recent studies have started to investigate the use of SAR or optical remote sensing data for multi-task learning [33–35], demonstrating the potential of multi-task learning in remote sensing. However, numerous challenges remain, such as integrating multi-source data and developing effective algorithms for joint learning. Further research is essential to address these challenges and fully exploit the potential of multi-task learning in remote sensing applications. In recent remote sensing research, there is growing interest in utilizing combined ground object height and RGB images for semantic segmentation tasks.

For example, Xiong et al. [36] demonstrated a strong correlation between the geometric information in the normalized digital surface model (nDSM) and the semantic category of land cover. Jointly utilizing two modalities, RGB and nDSM (height), has the potential to significantly improve segmentation performance, underlining the reliability of Transformer-based networks for multimodal fusion. This research highlights the interplay between semantic information and feature height information. Additionally, recent studies have investigated the use of RGB images for joint height estimation and semantic segmentation tasks in deep learning for remote sensing.

2.5. Multi-Task Learning

Previous studies [15,36] have yielded promising results, underscoring that joint processing of SAR and high-resolution remote sensing data can significantly enhance the accuracy of building extraction and height estimation tasks. These investigations have emphasized the connection between semantic and height information of ground objects, highlighting the effectiveness and necessity of simultaneously performing semantic segmentation and height estimation tasks. Currently, many deep learning tasks predominantly rely on single-task learning, yet multi-task learning, which allows the simultaneous learning of multiple related tasks and the sharing of information between them, offers superior generalization abilities compared to single-task learning [59].

Srivastava et al. [60] employed joint height estimation and semantic labeling on monocular aerial images, utilizing a single decoder with a fully connected layer to perform both height estimation and semantic segmentation tasks. In contrast, Carvalho et al. [61] proposed a framework for joint semantics and local height, processing the two tasks separately in the middle part of the decoder. Gao et al. [62] harnessed contrastive learning with an encoder featuring shared parameters, alongside cross-task contrast loss and cross-pixel contrast loss for height estimation and semantic segmentation. The decoder employed contrastive learning to encourage the model to learn detailed features. Lu et al. [63] introduced a unified deep learning architecture that can generate both estimated relative height maps and semantically segmented maps from RGB images, allowing for end-to-end training while accomplishing relative height estimation and semantic segmentation simultaneously. However, they failed to consider the independent relationship between building texture details and building semantic information. According to the correlation between semantic segmentation and height estimation, Zhao et al. [64] investigate and propose a semantic-aware unsupervised domain adaptation method for height estimation. They found that incorporating semantic supervision improves the accuracy of height estimation for single-view orthophotos under unsupervised domain adaptation.

Collectively, these studies have demonstrated that the integration of multi-task learning can effectively enhance the model's capability to perform height estimation and semantic segmentation tasks, resulting in improved accuracy.

3. Method

The proposed MFTSC encoder is built upon the Swin Transformer, and the overall architecture of the model is depicted in Figure 1. To enhance feature extraction, we devised the texture feature-extraction module (TEM) for capturing local neighborhood information from the shallow layer of the image, which is then used to construct the feature map E_0. Subsequently, we harnessed the Swin Transformer's robust feature extraction capabilities to form feature maps $\mathbf{E_1}$, $\mathbf{E_2}$, $\mathbf{E_3}$, and $\mathbf{E_4}$. These feature maps operate at various resolution scales, corresponding to 1/2, 1/4, 1/8, 1/16, and 1/32.

The height estimation decoder in MFTSC follows a design inspired by [51,65,66], comprising four layers. However, within each layer of the height decoder, we integrate a multi-dimensional feature-aggregation Transformer (MFT) to consolidate diverse dimensional information from both the encoder and the height decoder. Subsequently, we employ an advanced upsampling algorithm (details provided below) to obtain pointwise features. Toward the end of the height decoder, we introduce the height head, which combines

shallow local neighborhood features from the encoder with semantic features from the height decoder, thereby fine-tuning the height information prediction for each point.

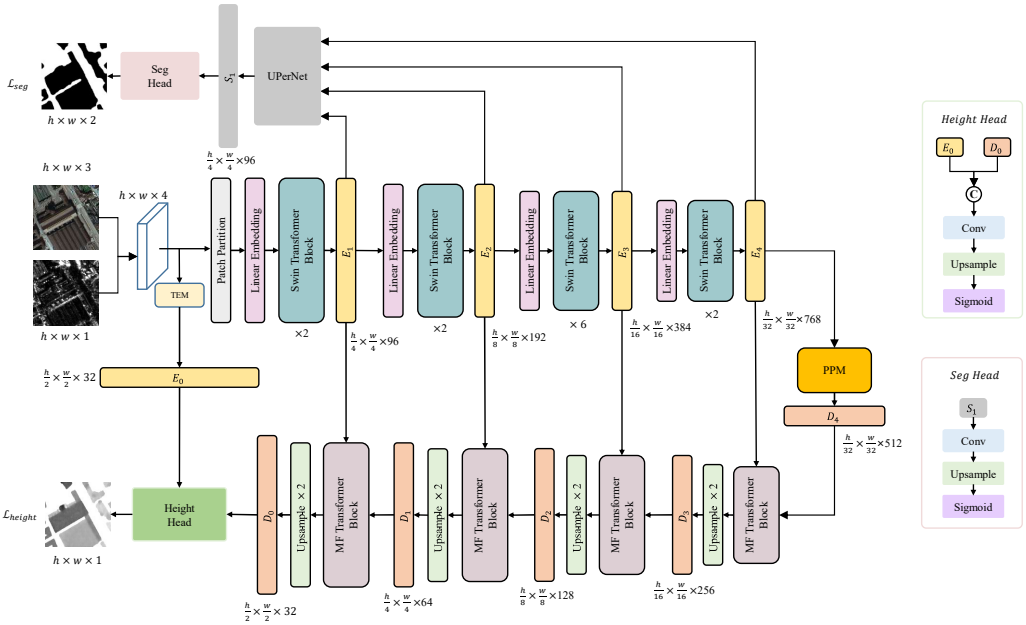

Figure 1. Detailed architecture of MFTSC: Given an input image, a visual Transformer-based encoder (Swin Transformer) and a TEM module extract multi-scale feature maps. The input of the feature pyramid pooling module (PPM) with the coarsest resolution ($\mathbf{E}_4$), and then the multi-scale feature $\mathbf{E}_1 \cdots \mathbf{E}_4$ are, respectively, passed into the height estimation decoder and semantic segmentation decoder. Finally, the height head is used for height estimation and Seg Head is used for semantic segmentation.

For the semantic segmentation decoder, we employ UPerNet [67], which capitalizes on multi-scale information fusion to enhance its segmentation performance. Finally, at the conclusion of the semantic segmentation decoder, we utilize convolution and upsampling techniques to refine the prediction of semantic labels for individual points.

3.1. TEM

We adopt a convolution module with a kernel size of 3 × 3 and max pooling to create TEM. Concretely, TEM comprises two 3 × 3 convolutional layers, each of which is accompanied by a batch normalization operation and a ReLU activation function. To retain rich spatial details, the shallow spatial detail features are extracted alongside the high-resolution feature representation, with a downsampling factor set to 2.

The TEM in our approach operates in a manner akin to 'Prompt' technology, which relies on language models to extract text-relevant features. In a similar vein, our TEM is responsible for extracting shallow information from the image, effectively constraining the area for ground object regression. This process significantly improves the accuracy of height estimation. As shown in Figure 2, the feature map generated by TEM clearly highlights the building area, effectively confining the regression area of ground objects within its scope.

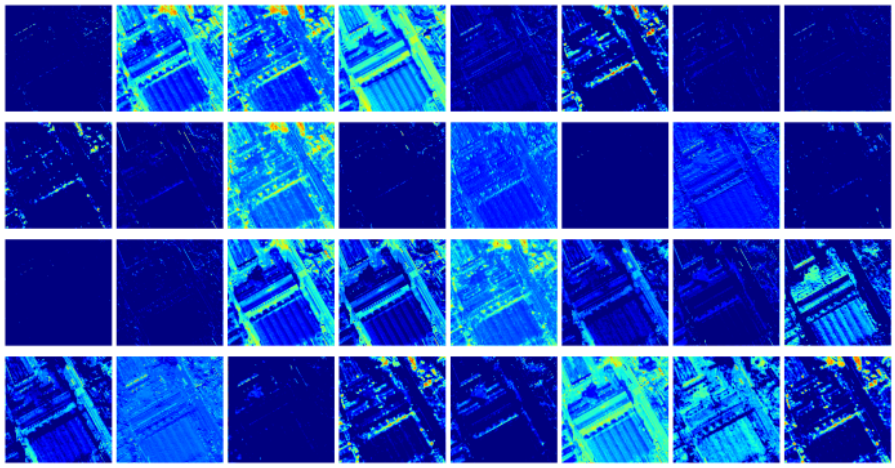

Figure 2. Features from the TEM of MFTSC, expanded by channel dimensions.

3.2. Swin Transformer

Figure 3 illustrates the Swin Transformer block, which is a notable variant of the Transformer architecture. Transformer encoders are a fundamental component of the Transformer architecture, a widely-used deep learning model in various natural language processing tasks. The encoder typically contains multi-head self-attention (MSA) and multi-layer perceptron (MLP) modules, along with layer normalization (LN) and residual connections positioned before and after each module. The MLP in the standard Transformer is a two-layer fully connected neural network with a Gaussian error linear unit (GELU) nonlinearity.

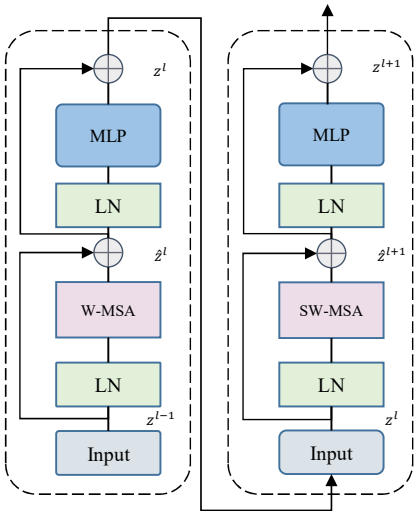

Figure 3. Swin Transformer block.

In the Swin Transformer block, the MSA block is replaced by two windows-based self-attention mechanisms: window multi-head self-attention (W-MSA) and shifted window multi-head self-attention (SW-MSA). Two layers of Swin Transformer Blocks are connected together as a basic unit, with the first layer comprising MLP and W-MSA, and the following Swin Transformer Block consisting of MLP and SW-MSA. LN is applied before the W-MSA, SW-MSA, and MLP modules, while residual connections are applied after these modules. LN normalizes the input of each module, while residual connections facilitate the flow of gradients through the network, mitigating the problem of vanishing gradients.

The above process can be expressed by the following formulas:

$$\hat{\mathbf{z}}^l = \text{W-MSA}(\text{LN}s(\mathbf{z}^{l-1})) + \mathbf{z}^{l-1}, \quad (1)$$

$$\mathbf{z}^{l} = \mathrm{MLP}(\mathrm{LN}(\hat{\mathbf{z}}^{l})) + \hat{\mathbf{z}}^{l}, \tag{2}$$

$$\hat{\mathbf{z}}^{l+1} = \text{SW-MSA}(\mathrm{LN}(\mathbf{z}^{l})) + \mathbf{z}^{l}, \tag{3}$$

$$\mathbf{z}^{l+1} = \mathrm{MLP}(\mathrm{LN}(\hat{\mathbf{z}}^{l+1})) + \hat{\mathbf{z}}^{l+1}, \tag{4}$$

where $\hat{\mathbf{z}}^{l}$ and $\mathbf{z}^{l}$ represent the output features of the (S)W-MSA module and the MLP module in the Swin Transformer block l, respectively. W-MSA and SW-MSA denote windowed multi-head self-attention and shifted-window multi-head self-attention, respectively.

3.3. PPM

For tasks that demand high regression accuracy, it is vital to capture spatial features at multiple scales to enhance the model's robustness. One approach is to use pooling modules of varying sizes to extract spatial information at different scales. Additionally, to address the challenge of losing context information across different sub-regions, models like PSPNET propose a hierarchical global prior structure. This involves integrating information from different scales and subregions to create a global scene on the final feature map of a deep neural network. This significantly improves the accuracy of height regression.

The pyramid spatial pooling (PSP) module employs a pyramid pooling strategy to capture context at multiple scales. This enables the model to better understand the spatial relationships between objects and their surroundings and to more accurately differentiate objects that may appear similar but have different contexts.

The PPM (see Figure 4) transforms the input feature map $\mathbf{D}_4 \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times 512}$ into four feature maps of varying spatial dimensions through pooling operations. These four different feature maps are then dimensionally reduced using 1×1 convolutions and subsequently resized to match the size of the input feature map through linear interpolation. After this, the input feature map is concatenated with the four interpolated feature maps. Finally, feature fusion is accomplished through the use of a ConvModule as:

$$\mathbf{D}_4 = \mathrm{ConvModule}(\mathrm{Concat}(Pool_i(\mathbf{E}_4), \mathbf{E}_4)), i \in [1, 2, 3, 6], \tag{5}$$

In the above formula, ConvModule is composed of 3 × 3 convolution, regularization, and ReLU activation function, Concat is the splicing operation, $Pool_i$ represents the mean pooling layer of different sizes i, and $\mathbf{E}_4$ is the input feature map.

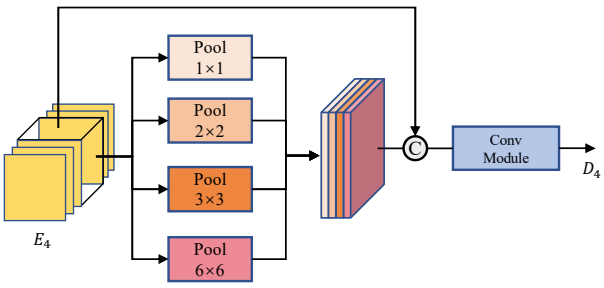

Figure 4. PPM. The PPM performs pyramid pooling to extract multi-scale global context from an input image.

3.4. MFT

In height estimation tasks, the utilization of both shallow spatial detail features and deep semantic features is essential in the decoder section. While many previous conventional methods have made use of both shallow and deep features, they often emphasize local information. To overcome this limitation and promote a more effective fusion of shallow and deep features, we introduce the MFT, as depicted in Figure 5.

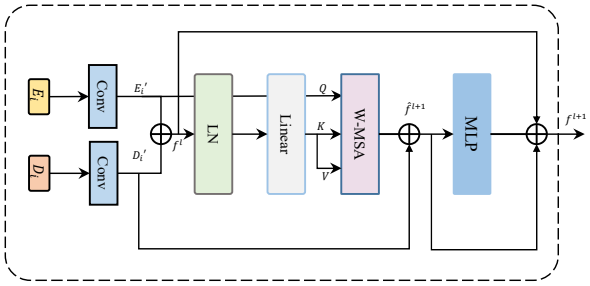

Figure 5. MFT. The MFT model employs a two-step approach to align and fuse features. Firstly, convolutional layers are used for feature alignment, followed by the use of Transformers for feature fusion. This approach allows the model to extract informative representations by capturing spatial relationships between features and integrating information from different levels of the network.

The MFT aims to capitalize on the advantages of MSA in computer vision, as it has been shown to enhance model accuracy, reduce loss, and improve generalization. However, the main challenge associated with MSA is its high computational complexity. To address this issue, we employ Window-based W-MSA, which reduces computational complexity. Specifically, we process the deep and shallow features through a series of steps:

(1) Convolution: The features undergo convolution to extract and enhance their representations.

(2) LN: We apply LN to normalize the feature maps, improving their stability.

(3) Window partition: The features are divided into windows to enable parallelized computations.

(4) Linear layer: The feature map passes through a linear layer to calculate window self-attention scores.

(5) Window merge: The window self-attention scores are employed to merge information from different windows.

(6) Residual connection: A residual connection is used to ensure smooth information flow, and the feature map proceeds through the MLP.

This process allows for effective feature aggregation and interaction between deep and shallow features, contributing to improved height estimation.

Simultaneously, we project the feature $\mathbf{E}_i$ from the encoder as $\mathbf{Q}$ in the W-MSA, while the feature $\mathbf{D}_i$ from the deep layer of the decoder is projected as $\mathbf{K}$ and $\mathbf{V}$. This design facilitates effective interaction between the encoder and decoder features, enhancing the model's efficiency in handling multiple tasks. Furthermore, this design establishes direct communication between the height estimation branch and the encoder, leading to improved coordination between the semantic segmentation branch and the height estimation branch during gradient propagation.

Given the input feature vectors $\mathbf{E}_i$ and $\mathbf{D}_i$, this approach employs a convolution module to project the features into the same dimension and perform feature superposition and fusion to obtain the fused feature $\mathbf{f}^l$ of size $\mathbb{R}^{H\times W\times C}$. This fused feature is then reshaped to $\mathbb{R}^{N\times C}$, where N is the sequence length and C is the feature dimension. After applying LN, the window partition technique is used to convert $\mathbf{f}^l$ and $\mathbf{E}_i^{'}$ into $w = \frac{H}{M} \times \frac{W}{M}$ feature vectors $\mathbf{f}^p \in \mathbb{R}^{w\times M\times M\times C}$ and $\mathbf{E}^p \in \mathbb{R}^{w\times M\times M\times C}$, respectively. Within each window, the window k features are linearly projected to obtain $\mathbf{Q}_k = \mathbf{W}_k^Q \mathbf{E}_k^p$, $\mathbf{K}_k = \mathbf{W}_k^Q \mathbf{f}^p$, and $\mathbf{V}_k = \mathbf{W}_k^Q \mathbf{f}^p$. The multi-head attention mechanism is then applied to these projections, following which, the window reverse technique is used to retrieve the $\mathbf{f}_{att}$ feature from the resultant vector. The $\mathbf{D}_i^{'}$ feature of the decoder is then connected to the resultant feature using residual connections. Finally, the MLP module is used to perform residual connections with previous features.

The specific forward process of MFT is as follows:

$$\mathbf{f}^l = \mathrm{LN}\left(\mathbf{M}^{E_i}\mathbf{E}_i + \mathbf{M}^{D_i}\mathbf{D}_i\right), \tag{6}$$

$$\mathbf{f}_{att} = \text{W-MSA}(\mathbf{f}^l, \mathbf{E}_i^{'}), \tag{7}$$

$$\hat{\mathbf{f}}^{l+1} = \mathbf{f}_{att} + \mathbf{D}_i^{'}, \tag{8}$$

$$\mathbf{f}^{l+1} = \mathrm{MLP}\left(\hat{\mathbf{f}}^{l+1}\right) + \mathbf{f}^l + \hat{\mathbf{f}}^{l+1}, \tag{9}$$

where $\mathbf{M}^{E_i}$ and $\mathbf{M}^{D_i}$ represent learnable matrices, and MLP represents multilayer perceptron. $\mathbf{E}_i$ and $\mathbf{D}_i$ represent features from the encoder and decoder, respectively.

The W-MSA algorithm computes MSA within a window of size $M \times M$, and a total of $\left\lceil \frac{H}{M} \times \frac{W}{M} \right\rceil$ windows are used for this purpose. Specifically, the features $\mathbf{f}^l$ and $\mathbf{E}_i^{'}$ are passed through the W-MSA module for MSA computation, and $\mathbf{f}^l$ and $\mathbf{E}_i^{'}$ are divided into windows to obtain $\mathbf{f}^p$ and $\mathbf{E}^p$ for each window. The self-attention calculation is performed as follows:

$$\mathbf{Q}_k = \mathbf{W}_k^Q \mathbf{E}_k^p, \mathbf{K}_k = \mathbf{W}_k^K \mathbf{f}^p, \mathbf{V}_k = \mathbf{W}_k^V \mathbf{f}^p, \tag{10}$$

$$\text{Attention}(\mathbf{Q}_k, \mathbf{K}_k, \mathbf{V}_k)_k = \text{SoftMax}(\frac{\mathbf{Q}_k {\mathbf{K}_k}^T}{\sqrt{d}})\mathbf{V}_k, \tag{11}$$

where $\mathbf{Q}_k, \mathbf{K}_k, \mathbf{V}_k \in \mathbb{R}^{M^2 \times C}$ denote the *query*, *key*, and *value* matrices, respectively, where M represents the size of the window and C is the dimensionality of the query and key matrices.

The attention calculation involves a multi-head attention mechanism, which partitions the input into multiple heads. Each head independently performs an attention calculation, and the results from all the heads are concatenated to obtain the final output. The multi-head self-attention calculation for the k-th window is expressed as follows:

$$\text{MSA}_k(\mathbf{Q}_k, \mathbf{K}_k, \mathbf{V}_k) = \text{Concat}(\mathbf{head}_1, \ldots, \mathbf{head}_i){\mathbf{W}_k}^O, \tag{12}$$

where ${\mathbf{W}_k}^O$ represents the learned parameter for the k-th head. Additionally, each head is defined as:

$$\mathbf{head}_i = \text{Attention}(\mathbf{Q}_k \mathbf{W}_i^{Q_k}, \mathbf{K}_k \mathbf{W}_i^{K_k}, \mathbf{V}_k \mathbf{W}_i^{V_k}), \tag{13}$$

where $\mathbf{W}_i^{Q_k} \in \mathbb{R}^{M^2 \times C'}$, $\mathbf{W}_i^{K_k} \in \mathbb{R}^{M^2 \times C'}$, and $\mathbf{W}_i^{V_k} \in \mathbb{R}^{M^2 \times C'}$, C' represents the dimension of each head. If there are N heads, then $C' = C/N$.

For MSA, the computational complexity of each window in W-MSA and the overall computational complexity of W-MSA are given by Equations (14)–(16).

$$\Omega(\text{MSA}) = 4HWC^2 + 2(HW)^2C, \tag{14}$$

$$\Omega(\text{Window}) = 4M^2C^2 + 2M^4C, \tag{15}$$

$$\Omega(\text{W-MSA}) = 4HWC^2 + 2M^2HWC, \tag{16}$$

where $H \times W$ is the total number of patches; C represents the feature dimension; M represents the size of each window.

3.5. Decoder

To ensure accurate height estimation, it is crucial to combine deep semantic and shallow spatial features. Our approach follows a bottom-up strategy similar to classic methods like Unet and LinkNet. We start with the lowest-resolution feature map and employ skip connections to capture detailed features by upsampling the encoder's feature map. These encoder–decoder features are then connected and processed through convolutional layers. However, during the fusion of encoder–decoder features via convolution, there is an issue of

information loss. This loss is primarily due to the local nature of the convolution operation, which compresses and mixes information and is further exacerbated by the upsampling operation in the decoder. Consequently, the model has a limited receptive field, making it challenging to acquire global features effectively. To overcome this limitation, we utilize the MFT module, which facilitates the fusion of global and local features. The MFT module employs window self-attention and significantly reduces the model's memory usage.

The height estimation decoder is depicted on the right side of Figure 1. Starting from the output $\mathbf{D}_4$ of the PPM module, we input the encoder features $\mathbf{E}_i (i \in 0 \cdots 4)$ and decoder features $\mathbf{D}_i (i \in 0 \cdots 4)$ into the MFT module for feature fusion. Following the MFT module, we perform upsampling to obtain the feature map $\mathbf{D}_{i-1}$. For a given encoder feature $\mathbf{E}_i$ and decoder feature $\mathbf{D}_i$, the process can be expressed as:

$$\mathbf{D}_{i-1} = \text{Upsample}(\text{MFT}(\mathbf{E}_i, \mathbf{D}_i)), \tag{17}$$

To generate the final feature map $\mathbf{D}_0$, we use the height head for processing. The height head combines the shallow features $\mathbf{E}_0$ extracted from the shallow feature extraction module with the deep features. This combination is achieved by concatenating the two sets of features and passing them through convolutional and upsampling modules. Finally, we apply the Sigmoid activation function to ensure that the output values fall within the [0,1] interval. The height head process can be expressed as:

$$\mathbf{H}_{Height} = \text{Sigmoid}\Big(\text{Upsample}\Big(\mathbf{M}^H(\text{Concat}(\mathbf{E}_0, \mathbf{D}_0))\Big)\Big), \tag{18}$$

where Sigmoid represents the activation function, Upsample denotes upsampling, Concat means tensor splicing, $\mathbf{M}^H$ represents the learnable matrix, $\mathbf{E}_0$ and $\mathbf{D}_0$ are shallow features from the encoder and decoder, respectively.

Given the relative simplicity of the task of semantic segmentation for building masks, we have opted for a streamlined semantic segmentation module. In our approach, we leverage the UPerNet model [67] as the central component for our segmentation tasks, specifically aimed at distinguishing the foreground from the background within building areas. UPerNet is a well-established decoder model renowned for semantic segmentation, and it effectively enhances segmentation performance through the amalgamation of multi-scale information. This is achieved by configuring sub-networks for segmentation at varying scales, allowing UPerNet to process images across multiple scales and amalgamate data from different scales to substantially bolster segmentation accuracy.

In our height estimation task, the challenge lies in simultaneously regressing the height of the building area while distinguishing the building from the background. This simultaneous operation complicates the separation of foreground and background, often necessitating the introduction of post-processing thresholds. Nevertheless, this challenge of foreground–background separation can be effectively addressed through semantic segmentation tasks. Recognizing the rich interplay between height estimation and semantic segmentation, we have adopted a multi-task learning approach to enhance the precision of our height estimation process. This approach enables UPerNet to be represented by the following formula:

$$\mathbf{S}_1 = \text{ConvModule}(\text{Concat}(\mathbf{P}_i)), i \in [2, 5], \tag{19}$$

$$\mathbf{P}_i = \mathbf{M}^{P_i}(\mathbf{E}_i) + \text{Upsample}(\mathbf{P}_{i+1}), i \in [2, 4], \tag{20}$$

$$\mathbf{H}_{Seg} = \text{Sigmoid}\Big(\text{Upsample}\Big(\mathbf{M}^S(\mathbf{S}_1)\Big)\Big), \tag{21}$$

In the above formula, ConvModule consists of 3×3 convolution, batch regularization, and ReLU activation function. The fusion block combines the four feature maps P2, P3, P4, and P5, then performs dimensionality reduction and scales the result to the size of the input image. $\mathbf{M}^{P_i}$ and $\mathbf{M}^S$ represent the learnable matrix, and Upsample means an upsampling

operation using bilinear interpolation. The final module produces the fusion feature $\mathbf{S}_1$, from which the result is obtained through the segmentation head.

3.6. Loss Function

In the height estimation task, we utilize a hybrid loss function comprising the *MSE* loss and L_1 loss functions, with the associated loss weights being fine-tuned to enhance the model's performance. As for the loss function in the semantic segmentation task, we opt for the standard binary cross-entropy loss function. The exact formulation of the loss function is given by:

$$\mathcal{L}_{height}(\mathbf{x_h}, \mathbf{y_h}) = \frac{1}{N}\sum_{i}^{N}\Big[\lambda_1 \times (\mathbf{x_h}_i - \mathbf{y_h}_i)^2 + (1-\lambda_1) \times |\mathbf{x_h}_i - \mathbf{y_h}_i|\Big], \tag{22}$$

$$\mathcal{L}_{seg}(\mathbf{x_s}, \mathbf{y_s}) = -[(\mathbf{y_s} * \log(\mathbf{x_s})) + (1-\mathbf{y_s})\log(1-\mathbf{x_s})], \tag{23}$$

where $\mathcal{L}_{height}$ represents the loss function for height estimation in the model. $\mathbf{x_h}$ and $\mathbf{y_h}$ denote the predicted value and the ground truth value of the building height, respectively. $\mathcal{L}_{seg}$ represents the binary cross-entropy loss for building segmentation, and $\mathbf{x_s}$ and $\mathbf{y_s}$ represent the predicted value and the ground truth segmentation mask for the building, respectively. According to the literature [63,68], we set λ_1 to 0.85.

In the pursuit of multi-task learning and concurrent optimization of two distinct tasks, we deliberately allocate distinct weights to the respective heads. The primary objective is to attain exceptional accuracy in height estimation, a notably intricate task, while semantic segmentation is comparatively more straightforward. Consequently, we assign a greater weight to the height estimation head and a relatively smaller weight to the semantic segmentation head. This strategic weight assignment ensures that the model prioritizes the height estimation task, demanding elevated precision and accuracy without compromising the overall performance of the semantic segmentation task. Therefore we set λ_2 to 0.8.

$$\mathcal{L}_{totol} = \lambda_2 \times \mathcal{L}_{height} + (1-\lambda_2) \times \mathcal{L}_{seg}, \tag{24}$$

4. Experiments

4.1. Experimental Platform Parameter Settings

All experiments were conducted using the PyTorch framework on a Windows 11 system, which is equipped with an Intel(R) Core(TM) i5 10400 CPU @ 2.90 GHz processor and an Nvidia GeForce RTX 3060 graphics card. For optimization, we employed the AdamW optimizer with a consistent learning rate of 0.0001 across all experiments. Given the simple nature of the semantic segmentation task and our objective to improve the accuracy of the height estimation task, we propose to balance the weight assigned to each task. Specifically, we assign smaller weights to the semantic segmentation task. To achieve this, we begin with a warm-up training of 10 epochs solely focused on the height estimation task. This strategy helps the model to learn the optimal parameters for the height estimation task without being distracted by the semantic segmentation task. Subsequently, we introduce a semantic segmentation head for multi-task learning. Building upon the approaches presented in [63,68], we set the weight for the semantic segmentation task to 0.2. We empirically found that this value leads to the best performance of our method compared to other values.

To assess the performance of each method, we employed a range of standard evaluation metrics, including mean Intersection over Union (mIoU), mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), and the coefficient of determination (R^2).

$$\text{mIoU} = \frac{1}{N}\sum_{k=1}^{N}\frac{\text{TP}_k}{\text{TP}_k + \text{FP}_k + \text{FN}_k}, \tag{25}$$

where TP_k, FP_k, TN_k, and FN_k indicate the true positive, false positive, true negative, and false negative, respectively, for the specific object indexed as class k.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{\mathbf{y}}_i - \mathbf{y}_i|, \tag{26}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbf{y}}_i - \mathbf{y}_i)^2, \tag{27}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2}, \tag{28}$$

$$\text{R}^2 = 1 - \frac{\sum_{i=1}^{n}(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2}{\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})^2}, \tag{29}$$

where $\mathbf{y}_i$ represents the ith true values of the variable being predicted, and $\hat{\mathbf{y}}_i$ is the i-th predicted values of the variable. $\bar{\mathbf{y}}$ represents the mean of the true values.

4.2. Datasets

The evaluation of our proposed method was conducted using the Data Fusion Competition 2023 (DFC2023) dataset [69]. This dataset comprises optical images (RGB) acquired from Gaofen-2 and SuperView satellites, with spatial resolutions of 0.8 m and 0.5 m, respectively. The SAR image, which is co-registered with the optical imagery, is obtained from the Gaofen-3 satellite, with a spatial resolution of 1 m. A nDSM for reference was generated from stereoscopic images captured by Gaofen-7 and WorldView, featuring a ground sampling distance of approximately 2 m. To ensure uniformity, all images were resampled to a common 0.5 m resolution.

The dataset is diverse and spans seventeen cities across six continents, excluding Antarctica. It provides a broad representation of various landforms, architectural styles, and building types. To facilitate consistent evaluation, all images were cropped into image patches of size 512 × 512 pixels. The dataset was divided into 1330 image patches for training and 443 for testing, totaling 1773 image patches. For each image block, we conducted data augmentation operations, which encompassed random horizontal and vertical flipping. Additionally, we applied a 2% linear stretch to the SAR image to reduce its inherent noise.

4.3. Compare Experiment

To verify the effectiveness of the proposed method in this paper, we conducted a comprehensive comparative experiment with several state-of-the-art methods, including DeepLabv3+ [70], Res-Unet [71], PSPNET [52], Res-LinkNet [50], VGG-Unet [72], Unet++ [73], VGG-LinkNet [51], and PAN [74], with all using ResNet50 or VGG13 as their backbone networks. We also compared our method with Swin-Unet [65], which uses the Swin Transformer (Swin-T) as the backbone network, and Swin-UPerNet [29], which combines the encoder and decoder in one network using the Swin Transformer. Additionally compared with the generative adversarial learning method Pix2Pix [39], which uses Res-Unet as the generator. Finally, we compare with recent excellent deep learning methods for MDE tasks, including NeWCRFs [44], GLPDepth [45], and PixelFormer [26].

The experimental results presented in Table 1 highlight the significant advantages of our proposed method over other state-of-the-art techniques. In contrast, other methods exhibit certain limitations. For instance, DeepLabv3+ is effective in capturing long-distance contextual information, which is beneficial for understanding the global context of an image. However, it tends to struggle to preserve fine-grained details, which are crucial for capturing local intricacies and small-scale features. Similarly, methods such as PSPNET and PAN excel at capturing multi-scale information by employing pooling operations that help

incorporate context at different levels. While this multi-scale approach is advantageous for capturing a wide range of object scales, it often comes at the cost of sacrificing fine-grained details, as pooling operations can lead to a loss of spatial resolution. Models like Res-Unet, Res-LinkNet, VGG-Unet, and VGG-LinkNet employ U-shaped architectures with ResNet50 and VGG13 backbones. While they use skip connections to combine low-level encoder features with high-level decoder features to enhance regression accuracy, their performance is constrained by the limitations of their backbones. Unet++, which outperforms Unet by incorporating additional skip connections and dense connection mechanisms, achieves superior performance. Nevertheless, it faces challenges during training due to high computational requirements, measured in terms of floating point operations (FLOPs). Swin-Unet can capture multi-scale features and reduce computational costs, but it may sacrifice fine-grained details due to downsampling operations. Swin-UPerNet, which adopts the Swin-T backbone and combines the encoder and decoder in one network using the UPerNet architecture, struggles to capture fine-grained details effectively.

Table 1. Comparative experimental results of the DFC23 dataset.

Method	Backbone	mIoU ↑	MAE (m) ↓	MSE (m^2) ↓	RMSE (m) ↓	R^2 ↑	Params ↓	FLOPs ↓
DeepLabv3+	ResNet50	0.628	0.7488	2.4398	1.4834	0.9064	106.72 M	1.369 G
Res-Unet	ResNet50	0.718	0.7289	2.4882	1.4895	0.9129	130.10 M	1.325 G
PSPNET	ResNet50	0.648	0.7785	2.5575	1.5089	0.9127	**8.96 M**	**0.743 G**
Res-LinkNet	ResNet50	0.745	0.7763	2.7441	1.5660	0.9047	124.72 M	1.620 G
VGG-Unet	VGG13	0.713	0.7193	2.4169	1.4644	0.9187	73.76 M	0.856 G
VGG-LinkNet	VGG13	0.631	0.8039	2.6927	1.5545	0.8952	42.68 M	1.245 G
Unet++	ResNet50	0.782	0.6809	2.4893	1.4652	0.9218	195.96 M	2.130 G
PAN	ResNet50	0.666	0.8102	2.7240	1.5700	0.8778	97.05 M	1.174 G
Swin-Unet	Swin-T	0.671	0.8608	3.5535	1.7884	0.9177	168.89 M	1.087 G
Swin-UPerNet	Swin-T	0.521	1.2082	6.4657	2.3984	0.8616	80.79 M	0.565 G
Pix2Pix	ResNet50	0.749	0.7676	2.9890	1.6394	0.9398	141.17 M	1.434 G
NeWCRFs	Swin-T	0.678	0.9701	3.8080	1.8679	0.8896	353.65 M	1.897 G
GLPDepth	MiT-b4	0.723	0.7679	2.8529	1.5950	0.9058	244.90 M	2.282 G
PixelFormer	Swin-T	0.682	0.8396	3.0784	1.6694	0.9044	305.35 M	1.620 G
MFTSC	Swin-T	**0.785**	**0.5390**	**1.5167**	**1.1733**	**0.9671**	302.38 M	1.686 G

Pix2Pix relies on GAN for building height generation, but it often generates overly smooth images and exhibits sensitivity and instability, which pose challenges during training and result in suboptimal RMSE indicators. The adversarial training process, which involves a generator and a discriminator network, aims to achieve a balance between generating realistic images and fooling the discriminator. However, this balance can be difficult to achieve, resulting in overly smooth outputs that lack fine details. Similarly, the MDE models, namely NeWCRFs, GLPDepth, and PixelFormer, deliver unsatisfactory performance across multiple evaluation metrics. This can be attributed to their original design for MDE tasks, primarily dealing with RGB street view images, which exhibit lower complexity compared to remote sensing images. The disparity between the specific task of building height estimation in remote sensing images and the broader scope of MDE tasks contributes to the subpar performance of these models. Remote sensing images often encompass unique challenges such as variations in resolution, sensor characteristics, and environmental conditions. Thus, it becomes imperative to develop a distinct model exclusively tailored for accurate height estimation in remote sensing applications.

Our proposed MFTSC architecture achieves the highest mIoU of 0.7855, indicating superior performance in semantic segmentation. Additionally, our method demonstrates significantly better results in terms of MAE (0.5390 m) compared to other methods, while the results of PSPNET and Swin-UPerNet are relatively poorer. This suggests that MFTSC is more sensitive to differences between true and predicted values and is able to predict target values more accurately. Furthermore, MFTSC achieves the best results in terms of MSE and RMSE (1.5167 square m and 1.1733 m, respectively), while the results of Swin-UPerNet are relatively poorer. This indicates that MFTSC predicts target values more accurately and with greater concentration. Finally, MFTSC achieves the highest R^2 of 0.9671, while the result of Swin-UPerNet is relatively poorer. This shows that the correlation between

predicted results of MFTSC and ground-truth values is stronger, allowing it to better capture the change trend of the target value.

In summary, our proposed method effectively captures the correlation between predicted and ground-truth height values, demonstrating the ability to capture multi-scale features and fine-grained details while reducing computational costs and mitigating the vanishing gradient problem. However, it is important to note that the method's disadvantage is its complex network architecture, necessitating training for both height estimation and semantic segmentation tasks simultaneously.

Figures 6 and 7 present the experimental results. It is evident from the figures that the model utilizing the Swin Transformer as the backbone network emphasizes global features, resulting in smoother results compared to models using VGG and ResNet as the backbone network. In models with ResNet as the backbone network, such as Unet-Res50, PSPNET-Res50, DeepLabv3+, PAN, and LinkNet-Res50, there are many missing holes in the generated results, and the estimated ground object height values display discontinuities. This could be attributed to the inadequate receptive field of the backbone network, hindering the full extraction of large-scale spatial information. Models employing VGG as the backbone network, such as Unet-VGG13 and LinkNet-VGG13, also exhibit the problem of insufficient receptive field, leading to discontinuous height information generation and the loss of height information in certain critical areas. Although Unet++ effectively alleviates the issue of information loss by employing multi-level feature fusion and dense skip connections within its decoder module, there still exists a discrepancy between the predicted height and the ground truth height in localized regions.

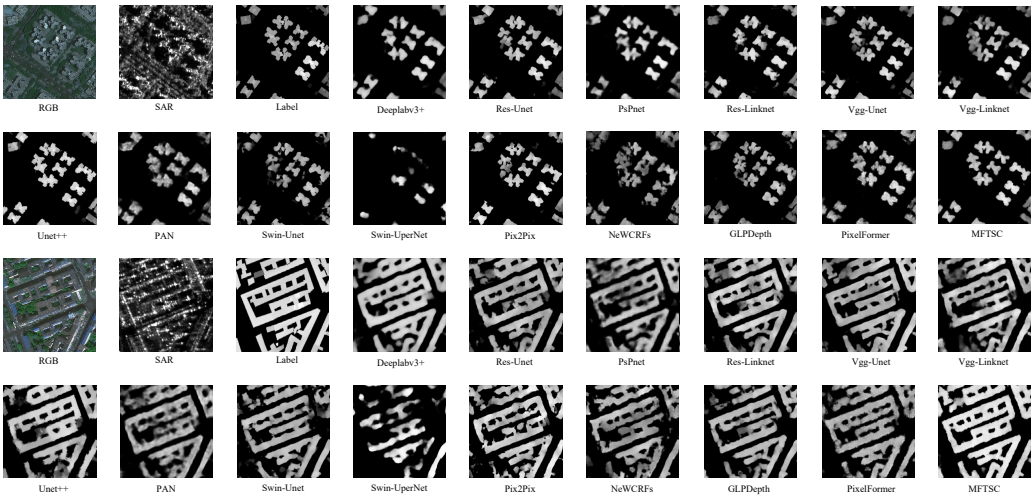

Figure 6. Comparative experimental results on the DFC23 dataset (part 1).

In contrast to models using VGG or ResNet as the backbone network, Swin Transformer-based models, such as Swin-Unet and MFTSC, excel at extracting spatial information and modeling global features. Consequently, the generated results exhibit relatively continuous regional height information. However, Swin-UPerNet underperforms due to its sole use of quadruple upsampling and the UPerNet module. This limitation hampers the preservation of texture information from the original remote sensing image, leading to issues such as large gaps and a lack of sufficient texture.

On the other hand, Swin-Unet harnesses the Swin Transformer as its backbone network, allowing the model to adeptly extract spatial information and possess long-distance dependence capabilities. Additionally, it employs a skip connection in the decoder segment, facilitating the acquisition of detailed features from the encoder feature map. However, Swin-Unet still necessitates a four-fold upsampling process in the decoder, which may cause the model to lose its capacity to extract and model the texture information present in the original remote sensing image.

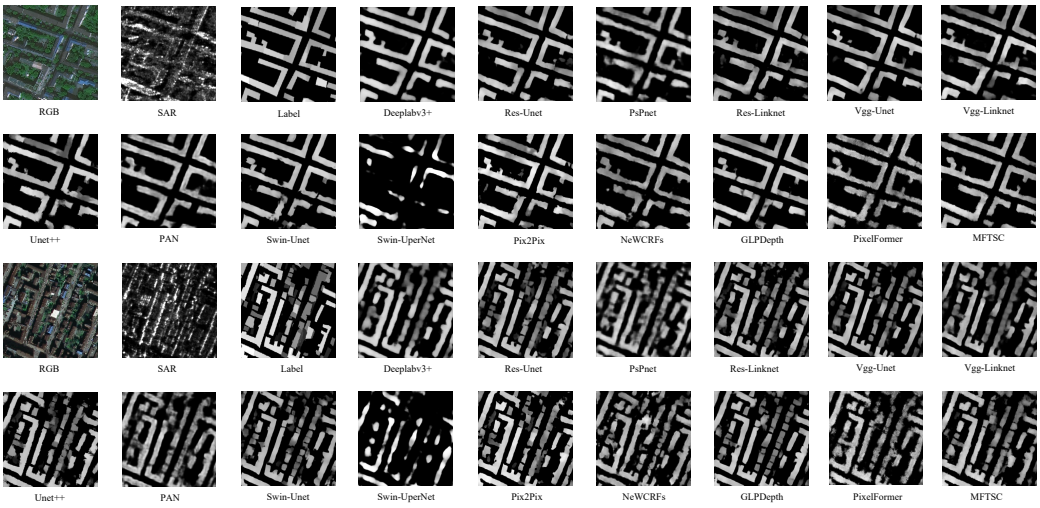

Figure 7. Comparative experimental results on the DFC23 dataset (part 2).

Regarding models designed for MDE tasks, while they leverage the Transformer as a robust backbone for feature extraction, their decoder architecture is not ideally suited for remote sensing images. As a result, issues like blurring artifacts and incomplete area coverage emerge in models like NeWCRFs, GLPDepth, and PixelFormer.

In contrast, MFTSC capitalizes on the texture cues within the original remote sensing image and preserves the Swin Transformer's robust spatial modeling ability. Moreover, MFTSC only requires double the upsampling in the decoder, thus effectively preserving the texture information and spatial continuity produced by the model. Furthermore, MFTSC adeptly fuses shallow and deep features, leading to more precise regression. The incorporation of semantic information also enhances MFTSC's ability to approximate the actual building outlines, especially in the edge areas of the buildings.

4.4. Discussion

In our ablation experiments, we aimed to assess the impact of various components on the performance of MFTSC. Here are the specific experimental settings:

(1) The effect of different backbone models

In this experiment, we evaluated the performance impact of utilizing different backbone models in the MFTSC architecture. The results are presented in Table 2. When we employ Swin-T as the backbone model, it allows the MFTSC to capture a broader global receptive field in the encoder segment. This, in combination with the MFT module in the height decoder part and the local detail features from the TM module at the end, results in a comprehensive ability to model both global and local features effectively. In contrast, ResNet and VGG exhibit limitations in terms of their receptive fields and feature extraction capabilities. Consequently, models using these backbones cannot fully leverage both global and local features, leading to relatively lower overall performance.

Table 2. Impact of different backbone replacements on MFTSC.

Backbone	mIoU ↑	MAE (m) ↓	MSE (m^2) ↓	RMSE (m) ↓	R^2 ↑	Params ↓	FLOPs ↓
ResNet	0.7462	0.8119	3.0569	1.6610	0.9646	257.28 M	1.172 G
VGG	0.7633	0.7633	2.8127	1.5867	0.9610	202.93 M	2.541 G
Swin-T	0.7855	0.5390	1.5167	1.1733	0.9671	302.38 M	1.686 G

(2) The impact of different modules

We discussed the ablation experiments of different modules, and the specific results are shown in Table 3.

A: TEM module effectiveness.

In this experiment, we removed the TEM module from MFTSC. The results indicate that the TEM module plays a crucial role in enhancing both object height estimation and building segmentation. It effectively extracts shallow spatial detail features, which are essential for prompting the decoder.

B: Remove building semantic constraints.

Here, we excluded the Seg decoder and Seg head in MFTSC. This experiment demonstrates that semantic segmentation constraints significantly improve the performance of object height estimation. It underscores the rationale behind our method, which leverages semantic information to enhance object height estimation performance.

C: Using a simple decoder.

For this test, we replaced the MFT module in the decoder with convolution for contextual feature aggregation. The results clearly highlight the limitations of the convolution module, especially for tasks with high complexity, where the Transformer has a distinct advantage in feature aggregation.

D: Remove the PPM module.

In this experiment, the PPM module was excluded from the MFTSC model, resulting in a decrease of 0.0436 in mIoU and a decrease of 0.0046 in R^2. These findings emphasize the critical role of the PPM module in achieving optimal performance in MFTSC.

E: Remove the PPM module and TEM module.

Here, we removed both the PPM module and the TEM module from MFTSC. This simultaneous removal led to a significant decrease in R^2, highlighting the effectiveness of both the PPM and TEM modules.

Table 3. Ablation experiment: Impact of different modules on MFTSC. A: Remove the TEM module, B: remove semantic constraints, C: use the simple decoder, D: remove the PPM module, E: remove the PPM module and TEM module.

Method	mIoU ↑	MAE (m) ↓	MSE (m^2) ↓	RMSE (m) ↓	R^2 ↑	Params ↓	FLOPs ↓
A	0.7789	0.6085	1.6432	1.2449	0.9627	302.24 M	1.619 G
B	0.7387	0.7633	2.8127	1.5867	0.9481	298.03 M	1.651 G
C	0.6919	0.8608	3.5535	1.7884	0.9425	169.00 M	1.357 G
D	0.7419	0.7207	2.4858	1.5005	0.9625	258.32 M	1.640 G
E	0.7752	0.5689	1.6732	1.2304	0.9595	258.19 M	1.573 G

(3) The impact of soft and hard parameter sharing on the model

Multi-task learning can be divided into hard parameter sharing and soft parameter sharing in remote sensing tasks. For hard parameter sharing, the model shares some parameters across all tasks and only uses individual parameters at the final classification head or segmentation head. Shared layers across the network tend to learn feature representations that are better for all tasks, so models with hard parameter sharing are hard to fit. For soft parameter sharing, each task has its parameters, and parameters are shared among different tasks. Finally, combining the differences between the parameters of different tasks allows for holistic multi-task learning. We designed MFTSC in two modes: soft parameter sharing mode (Soft) and hard parameter sharing mode (Hard). The specific results are shown in Table 4. Notably, we observed that only encoder parameter sharing (Soft) outperforms encoder–decoder parameter sharing (hard). One can refer to Figure 8 for a visual representation of the hard parameter sharing mode.

Table 4. The impacts of parameter sharing strategies in different ways on the accuracy of the model.

Method	mIoU ↑	MAE (m) ↓	MSE (m^2) ↓	RMSE (m) ↓	R^2 ↑	Params ↓	FLOPs ↓
Soft	0.7855	0.5390	1.5167	1.1733	0.9671	302.38 M	1.686 G
Hard	0.7339	0.7386	2.4585	1.4963	0.9544	254.12 M	1.611 G

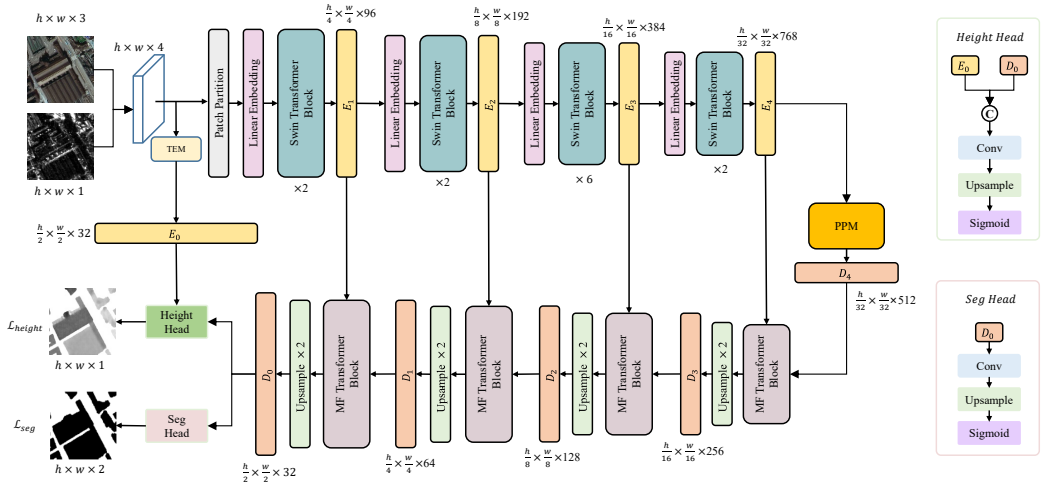

Figure 8. Hard. MFTSC architecture using hard parameter sharing.

(4) Model parameters and FLOPs

Table 1 provides comprehensive data on model parameters, calculations, and FLOPs for various models. Here is an analysis of the key findings:

CNN-based models: CNN-based models generally have a lower parameter count, with PSPNET having the lowest at 8.96 M parameters and a relatively low 0.743 G FLOPs. These models are computationally efficient but may not perform as well in intensive tasks.

Unet and LinkNet models: Unet and LinkNet models have larger parameter counts and FLOPs compared to PSPNET but tend to outperform PSPNET, PAN, and DeepLabv3+ in more demanding tasks. Unet++ demonstrates remarkable results but comes with a challenge during training due to its substantial number of parameters and FLOPs.

Pix2Pix: The application of the Pix2Pix training method to train Res-Unet yields favorable results, but optimizing GANs for model training remains a challenge.

Swin-T backbone models: Models that use Swin-T as a backbone typically have a high parameter count. However, when the number of parameters is small, the resulting effect is significantly diminished. Swin-UPerNet demonstrates the least effectiveness, primarily due to the unsuitability of the decoder for height estimation tasks.

MDE models: Models designed specifically for height estimation tasks have extraordinarily large numbers of parameters and FLOPs. For example, NeWCRFs have 353.65 M parameters and GLPDepth has 2.282 G FLOPs. In contrast, MFTSC uses a multi-task learning architecture to harness the global modeling capabilities of Swin-T while controlling parameters and FLOPs at a moderate level compared to NeWCRFs, GLPDepth, and PixelFormer.

These findings highlight the trade-offs between model complexity, computational requirements, and task performance. MFTSC stands out by intelligently introducing multi-task learning while maintaining moderate parameter and FLOP levels.

5. Conclusions

This study introduces a novel encoder–decoder architecture, referred to as MFTSC, aimed at jointly addressing the challenges of semantic segmentation and height estimation tasks. The proposed architecture leverages semantic constraints to enhance the accuracy of height estimation, avoiding the limitations associated with CNNs in dense prediction tasks for remote sensing images. Key aspects of the approach include the adoption of the Swin Transformer as a backbone network, the development of a multi-task learning architecture for height estimation and semantic segmentation, and the integration of modules like TEM and PPM to improve the model's ability to capture gradient texture and context at multiple scales. The MFT module effectively fuses features from the encoder and decoder,

facilitating the combination of global and local information. Experimental results indicate that our method outperforms traditional CNN-based methods across various model types.

In future research, we intend to investigate the potential benefits of integrating our approach with the rapidly evolving diffusion model. This integration has the potential to enhance the accuracy of height estimation tasks. These efforts are expected to contribute to the ongoing advancement of remote sensing applications.

Author Contributions: Conceptualization, methodology, software, Y.C., Q.Y. and W.H.; validation, Y.C. and Q.Y.; writing—original draft preparation, Y.C.; writing—review and editing, Q.Y. and W.H.; visualization, Y.C. and Q.Y.; supervision, Q.Y.; project administration, Q.Y.; funding acquisition, Q.Y. 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 (42001362) to Q. Yan.

Data Availability Statement: The datasets presented in this paper are available through https://dfc.geovisearth.com/en, accessed on 15 May 2023.

Acknowledgments: The authors would like to thank the Aerospace Information Research Institute under the Chinese Academy of Sciences, the Universität der Bundeswehr München, and GEOVIS Earth Technology Co., Ltd., for making the used remote sensing data freely available.

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

Abbreviations

R^2	coefficient of determination
RMSE	root mean square error
mIoU	mean intersection over union
MSE	mean square error
MAE	mean absolute error
SAR	synthetic aperture radar
MDE	monocular depth estimation
CNNs	convolutional neural networks
nDSM	normalized digital surface model
MFTSC	multi-level feature fusion Transformer with semantic constraint(s)
TEM	texture feature-extraction module
ViT	Vision Transformer
MFT	multi-dimensional feature-aggregation Transformer
PPM	pyramid pooling module
MSA	multi-head self-attention
LN	layer normalization
GELU	Gaussian error linear unit
W-MSA	window multi-head self-attention
SW-MSA	shifted window multi-head self-attention
PSP	pyramid spatial pooling
DFC2023	Data Fusion Competition 2023
FLOPs	floating point operations

References

1. Skalicky, V.; Čerpes, I. Comprehensive assessment methodology for liveable residential environment. *Cities* **2019**, *94*, 44–54. [CrossRef]
2. Chi, Y.L.; Mak, H.W.L. From comparative and statistical assessments of liveability and health conditions of districts in Hong Kong towards future city development. *Sustainability* **2021**, *13*, 8781. [CrossRef]
3. Dabous, S.A.; Shanableh, A.; Al-Ruzouq, R.; Hosny, F.; Khalil, M.A. A spatio-temporal framework for sustainable planning of buildings based on carbon emissions at the city scale. *Sustain. Cities Soc.* **2022**, *82*, 103890. [CrossRef]
4. Li, Z.; Shi, W.; Wang, Q.; Miao, Z. Extracting man-made objects from high spatial resolution remote sensing images via fast level set evolutions. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 883–899.
5. Han, K.; Bao, S.; She, M.; Pan, Q.; Liu, Y.; Chen, B. Exploration of intelligent building planning for urban renewal. *Sustainability* **2023**, *15*, 4565. [CrossRef]

6. Cao, Y.; Xu, C.; Aziz, N.M.; Kamaruzzaman, S.N. BIM–GIS integrated utilization in urban disaster management: The contributions, challenges, and future directions. *Remote Sens.* **2023**, *15*, 1331. [CrossRef]
7. Guo, H.; Shi, Q.; Du, B.; Zhang, L.; Wang, D.; Ding, H. Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4287–4306. [CrossRef]
8. Lee, T.; Kim, T. Automatic building height extraction by volumetric shadow analysis of monoscopic imagery. *Int. J. Remote Sens.* **2013**, *34*, 5834–5850. [CrossRef]
9. Licciardi, G.A.; Villa, A.; Dalla Mura, M.; Bruzzone, L.; Chanussot, J.; Benediktsson, J.A. Retrieval of the height of buildings from WorldView-2 multi-angular imagery using attribute filters and geometric invariant moments. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2012**, *5*, 71–79. [CrossRef]
10. Brunner, D.; Lemoine, G.; Bruzzone, L.; Greidanus, H. Building height retrieval from VHR SAR imagery based on an iterative simulation and matching technique. *IEEE Trans. Geosci. Remote Sens.* **2009**, *48*, 1487–1504. [CrossRef]
11. Elkhrachy, I. Flash flood water depth estimation using SAR images, digital elevation models, and machine learning algorithms. *Remote Sens.* **2022**, *14*, 440. [CrossRef]
12. Moya, L.; Mas, E.; Koshimura, S. Sparse representation-based inundation depth estimation using sAR data and digital elevation model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 9062–9072. [CrossRef]
13. Parida, B.R.; Tripathi, G.; Pandey, A.C.; Kumar, A. Estimating floodwater depth using SAR-derived flood inundation maps and geomorphic model in kosi river basin (India). *Geocarto Int.* **2022**, *37*, 4336–4360. [CrossRef]
14. 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]
15. Fieuzal, R.; Baup, F. Estimation of leaf area index and crop height of sunflowers using multi-temporal optical and SAR satellite data. *Int. J. Remote Sens.* **2016**, *37*, 2780–2809. [CrossRef]
16. Sportouche, H.; Tupin, F.; Denise, L. Building detection by fusion of optical and SAR features in metric resolution data. In Proceedings of the 2009 IEEE International Geoscience and Remote Sensing Symposium, Cape Town, South Africa, 12–17 July 2009; IEEE: Piscataway, NJ, USA, 2009; Volume 4.
17. Liasis, G.; Stavrou, S. Satellite images analysis for shadow detection and building height estimation. *ISPRS J. Photogramm. Remote Sens.* **2016**, *119*, 437–450. [CrossRef]
18. Qi, F.; Zhai, J.Z.; Dang, G. Building height estimation using Google Earth. *Energy Build.* **2016**, *118*, 123–132. [CrossRef]
19. Kulkarni, S.C.; Rege, P.P. Pixel level fusion techniques for SAR and optical images: A review. *Inf. Fusion* **2020**, *59*, 13–29. [CrossRef]
20. Sportouche, H.; Tupin, F.; Denise, L. Extraction and three-dimensional reconstruction of isolated buildings in urban scenes from high-resolution optical and SAR spaceborne images. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 3932–3946. [CrossRef]
21. Gao, J.; O'Neill, B.C. Mapping global urban land for the 21st century with data-driven simulations and Shared Socioeconomic Pathways. *Nat. Commun.* **2020**, *11*, 2302. [CrossRef]
22. Eigen, D.; Fergus, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2650–2658.
23. Xu, D.; Ricci, E.; Ouyang, W.; Wang, X.; Sebe, N. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5354–5362.
24. Ming, Y.; Meng, X.; Fan, C.; Yu, H. Deep learning for monocular depth estimation: A review. *Neurocomputing* **2021**, *438*, 14–33. [CrossRef]
25. Agarwal, A.; Arora, C. Depthformer: Multiscale vision transformer for monocular depth estimation with global local information fusion. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 3873–3877.
26. Agarwal, A.; Arora, C. Attention attention everywhere: Monocular depth prediction with skip attention. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 5861–5870.
27. 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.
28. Wang, L.; Fang, S.; Meng, X.; Li, R. Building extraction with vision transformer. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef]
29. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10–17 October 2021; pp. 10012–10022.
30. Chen, Y.; Liu, P.; Zhao, J.; Huang, K.; Yan, Q. Shallow-Guided Transformer for Semantic Segmentation of Hyperspectral Remote Sensing Imagery. *Remote Sens.* **2023**, *15*, 3366. [CrossRef]
31. Xie, Y.; Feng, D.; Xiong, S.; Zhu, J.; Liu, Y. Multi-scene building height estimation method based on shadow in high resolution imagery. *Remote Sens.* **2021**, *13*, 2862. [CrossRef]
32. Sun, Y.; Shahzad, M.; Zhu, X.X. Building height estimation in single SAR image using OSM building footprints. In Proceedings of the 2017 Joint Urban Remote Sensing Event (JURSE), Dubai, United Arab Emirates, 6–8 March 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–4.

33. Wang, C.; Pei, J.; Wang, Z.; Huang, Y.; Wu, J.; Yang, H.; Yang, J. When deep learning meets multi-task learning in SAR atr: Simultaneous target recognition and segmentation. *Remote Sens.* **2020**, *12*, 3863. [CrossRef]
34. Ma, X.; Ji, K.; Zhang, L.; Feng, S.; Xiong, B.; Kuang, G. An open set recognition method for SAR targets based on multitask learning. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
35. Heiselberg, P.; Sørensen, K.; Heiselberg, H. Ship velocity estimation in SAR images using multitask deep learning. *Remote Sens. Environ.* **2023**, *288*, 113492. [CrossRef]
36. Xiong, Z.; Chen, S.; Wang, Y.; Mou, L.; Zhu, X.X. GAMUS: A geometry-aware multi-modal semantic segmentation benchmark for remote sensing data. *arXiv* **2023**, arXiv:2305.14914.
37. Hambarde, P.; Dudhane, A.; Patil, P.W.; Murala, S.; Dhall, A. Depth estimation from single image and semantic prior. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1441–1445.
38. Hambarde, P.; Murala, S.; Dhall, A. UW-GAN: Single-image depth estimation and image enhancement for underwater images. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–12. [CrossRef]
39. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
40. Atteia, G.; Collins, M.J.; Algarni, A.D.; Samee, N.A. Deep-Learning-Based Feature Extraction Approach for Significant Wave Height Prediction in SAR Mode Altimeter Data. *Remote Sens.* **2022**, *14*, 5569. [CrossRef]
41. Sun, Y.; Hua, Y.; Mou, L.; Zhu, X.X. Large-scale building height estimation from single VHR SAR image using fully convolutional network and GIS building footprints. In Proceedings of the 2019 Joint Urban Remote Sensing Event (JURSE), Vannes, France, 22–24 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–4.
42. Ding, Y.; Lin, L.; Wang, L.; Zhang, M.; Li, D. Digging into the multi-scale structure for a more refined depth map and 3D reconstruction. *Neural Comput. Appl.* **2020**, *32*, 11217–11228. [CrossRef]
43. Dong, X.; Garratt, M.A.; Anavatti, S.G.; Abbass, H.A. Towards real-time monocular depth estimation for robotics: A survey. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 16940–16961. [CrossRef]
44. Yuan, W.; Gu, X.; Dai, Z.; Zhu, S.; Tan, P. New crfs: Neural window fully-connected crfs for monocular depth estimation. *arXiv* **2022**, arXiv:2203.01502.
45. Kim, D.; Ka, W.; Ahn, P.; Joo, D.; Chun, S.; Kim, J. Global-local path networks for monocular depth estimation with vertical cutdepth. *arXiv* **2022**, arXiv:2201.07436.
46. 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.
47. Petrovai, A.; Nedevschi, S. Exploiting pseudo labels in a self-supervised learning framework for improved monocular depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1578–1588.
48. 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.
49. Chen, Y.; Yan, Q. Vision Transformer is required for hyperspectral semantic segmentation. In Proceedings of the 2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI), Chengdu, China, 19–21 August 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 36–40.
50. Chaurasia, A.; Culurciello, E. Linknet: Exploiting encoder representations for efficient semantic segmentation. In Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, USA, 10–13 December 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–4.
51. Yan, Q.; Chen, Y.; Jin, S.; Liu, S.; Jia, Y.; Zhen, Y.; Chen, T.; Huang, W. Inland water mapping based on GA-LinkNet from CyGNSS data. *IEEE Geosci. Remote Sens. Lett.* **2022**, *20*, 1–5. [CrossRef]
52. 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.
53. Shao, Z.; Tang, P.; Wang, Z.; Saleem, N.; Yam, S.; Sommai, C. BRRNet: A fully convolutional neural network for automatic building extraction from high-resolution remote sensing images. *Remote Sens.* **2020**, *12*, 1050. [CrossRef]
54. Deng, W.; Shi, Q.; Li, J. Attention-gate-based encoder–decoder network for automatical building extraction. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2611–2620. [CrossRef]
55. Zheng, Z.; Zhong, Y.; Wang, J. Pop-Net: Encoder-dual decoder for semantic segmentation and single-view height estimation. In Proceedings of the IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 4963–4966.
56. Xing, S.; Dong, Q.; Hu, Z. SCE-Net: Self-and cross-enhancement network for single-view height estimation and semantic segmentation. *Remote Sens.* **2022**, *14*, 2252. [CrossRef]
57. Zhang, B.; Wan, Y.; Zhang, Y.; Li, Y. JSH-Net: Joint semantic segmentation and height estimation using deep convolutional networks from single high-resolution remote sensing imagery. *Int. J. Remote Sens.* **2022**, *43*, 6307–6332. [CrossRef]

58. Chen, Y.; Wang, B.; Yan, Q.; Huang, B.; Jia, T.; Xue, B. Hyperspectral Remote-Sensing Classification Combining Transformer and Multiscale Residual Mechanisms. *Laser Optoelectron. Prog.* **2023**, *60*, 1228002. [CrossRef]
59. Bhattacharjee, D.; Zhang, T.; Süsstrunk, S.; Salzmann, M. Mult: An end-to-end multitask learning transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12031–12041.
60. Srivastava, S.; Volpi, M.; Tuia, D. Joint height estimation and semantic labeling of monocular aerial images with CNNs. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 5173–5176.
61. Carvalho, M.; Le Saux, B.; Trouvé-Peloux, P.; Champagnat, F.; Almansa, A. Multitask learning of height and semantics from aerial images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 1391–1395. [CrossRef]
62. Gao, Z.; Sun, W.; Lu, Y.; Zhang, Y.; Song, W.; Zhang, Y.; Zhai, R. Joint learning of semantic segmentation and height estimation for remote sensing image leveraging contrastive learning. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5614015. [CrossRef]
63. Lu, M.; Liu, J.; Wang, F.; Xiang, Y. Multi-Task learning of relative height estimation and semantic segmentation from single airborne rgb images. *Remote Sens.* **2022**, *14*, 3450. [CrossRef]
64. Zhao, W.; Persello, C.; Stein, A. Semantic-aware unsupervised domain adaptation for height estimation from single-view aerial images. *ISPRS J. Photogramm. Remote Sens.* **2023**, *196*, 372–385. [CrossRef]
65. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 205–218.
66. Yan, Q.; Huang, W. Sea Ice Sensing From GNSS-R Data Using Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 18160835.
67. 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.
68. Bhat, S.F.; Alhashim, I.; Wonka, P. Adabins: Depth estimation using adaptive bins. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4009–4018.
69. Persello, C.; Hänsch, R.; Vivone, G.; Chen, K.; Yan, Z.; Tang, D.; Huang, H.; Schmitt, M.; Sun, X. 2023 IEEE GRSS Data Fusion Contest: Large-scale fine-grained building classification for semantic urban reconstruction [Technical Committees]. *IEEE Geosci. Remote Sens. Mag.* **2023**, *11*, 94–97. [CrossRef]
70. 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.
71. 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: Piscataway, NJ, USA, 2018; pp. 327–331.
72. Iglovikov, V.; Shvets, A. Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. *arXiv* **2018**, arXiv:1801.05746.
73. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested U-net architecture for medical image segmentation. In Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, 20 September 2018; Proceedings 4; Springer: Cham, Switzerland, 2018; pp. 3–11.
74. Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid attention network for semantic segmentation. *arXiv* **2018**, arXiv:1805.10180.