remote sensing

MDPI

Article

Knowledge-Guided Multi-Task Network for Remote Sensing Imagery

Meixuan Li [1], Guoqing Wang [1,*], Tianyu Li [1], Yang Yang [1], Wei Li [2], Xun Liu [2] and Ying Liu [2]

[1] The Center for Future Media, University of Electronic Science and Technology of China, Chengdu 611731, China; 202221080132@std.uestc.edu.cn (M.L.); tylisky@uestc.edu.cn (T.L.); yang.yang@uestc.edu.cn (Y.Y.)
[2] The Department of Imaging Technology, Beijing Institute of Space Mechanics and Electricity, Beijing 100094, China
* Correspondence: gqwang0420@uestc.edu.cn

Abstract: Semantic segmentation and height estimation tasks in remote sensing imagery exhibit distinctive characteristics, including scale sensitivity, category imbalance, and insufficient fine details. Recent approaches have leveraged multi-task learning methods to jointly predict these tasks along with auxiliary tasks, such as edge detection, to improve the accuracy of fine-grained details. However, most approaches only acquire knowledge from auxiliary tasks, disregarding the inter-task knowledge guidance across all tasks. To address these challenges, we propose KMNet, a novel architecture referred to as a knowledge-guided multi-task network, which can be applied to different primary and auxiliary task combinations. KMNet employs a multi-scale methodology to extract feature information from the input image. Subsequently, the architecture incorporates the multi-scale knowledge-guided fusion (MKF) module, which is designed to generate a comprehensive knowledge bank serving as a resource for guiding the feature fusion process. The knowledge-guided fusion feature is then utilized to generate the final predictions for the primary tasks. Comprehensive experiments conducted on two publicly available remote sensing datasets, namely the Potsdam dataset and the Vaihingen dataset, demonstrate the effectiveness of the proposed method in achieving impressive performance on both semantic segmentation and height estimation tasks. Codes, pre-trained models, and more results will be publicly available.

Keywords: multi-task learning; remote sensing; semantic segmentation; height estimation; edge detection

check for updates

Received: 30 October 2024
Revised: 4 January 2025
Accepted: 28 January 2025
Published: 31 January 2025

Citation: Li, M.; Wang, G.; Li, T.; Yang, Y.; Li, W.; Liu, X.; Liu, Y. Knowledge-Guided Multi-Task Network for Remote Sensing Imagery. *Remote Sens.* **2025**, *17*, 496. https://doi.org/10.3390/rs17030496

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the rapid advancement of aerospace technologies, including aircraft, satellites, optical cameras, LiDAR, and radar [1–4], remote sensing analysis has received significant attention as a means to comprehend the geographic and ecological environment in recent years. Semantic segmentation and height estimation in a single image are two important topics in remote sensing imagery, since the learned semantic and height information is crucial for the task of accurately measuring and analyzing the 3D shapes and positions of buildings in ground scenes, and such a process has wide-ranging applications, e.g., urban planning [5,6], building change detection [7,8], and disaster monitoring [9,10].

Semantic segmentation aims to assign each pixel in an image to a specific class, such as building, car, and low vegetation. Height estimation can be approached as a pixel-level regression task designed for estimating the vertical elevation associated with each pixel in

a remote sensing image. Numerous studies have been proposed to address these two tasks independently [11–14]. However, single-task learning encounters inherent limitations as it lacks the ability to leverage image features from multiple perspectives, thereby leading to potential instances of false recognition. For instance, using single-task learning for height estimation may mistakenly assign similar heights to objects sharing the same material properties, such as trees and low vegetation. Likewise, semantic segmentation models often struggle to accurately classify small objects within extensive areas of the same class, such as low vegetation, trees, and cars, as shown in Figure 1c,d.

Compared to addressing semantic segmentation and height estimation as separate task, multi-task learning accomplishes them jointly by leveraging a unified framework. By employing a multi-task learning approach for height estimation and semantic segmentation, the integration of diverse task features becomes possible. This integration facilitates the acquisition of semantic information pertaining to objects within the image by the height estimation module while enabling the semantic segmentation module to leverage 3D height information [15,16]. Consequently, a synergistic effect is achieved, wherein the performance of both tasks is mutually enhanced.

Typical multi-task learning approaches have been successfully applied in various fields, including computer vision [17–20], natural language processing [21–23], and autonomous driving [24–26]. However, directly applying these approaches to remote sensing imagery encounters specific challenges: (i) Scale sensitivity [27], which refers to the variation in extracted features across different scales. Achieving a comprehensive understanding of the imaged area often requires consideration of multiple scales, posing a significant challenge for analysis and interpretation. (ii) Imbalanced class distribution [28], where the majority of remote sensing images consist of background, large buildings, and low vegetation. In comparison to these large objects, distributions of small objects cannot be easily modeled and learned by traditional network designs. (iii) Capturing the fine-grained details of objects proves challenging in various tasks, notably semantic segmentation and height estimation. To address the issue of scale sensitivity, Pop-Net [29] utilizes a pyramid network based on an encoder–dual decoder framework for semantic segmentation and height estimation. BAMTL [30] enhances the prediction ability by using edge detection as an auxiliary task while exploiting the correlation between the primary tasks. However, neither method simultaneously considers both multi-scale and auxiliary tasks. Additionally, BAMTL overlooks the crucial interaction of relevant knowledge across different tasks, leading to a deficiency in knowledge guidance.

To address the problems mentioned above, we propose KMNet, a novel knowledge-guided multi-task network to predict semantic segmentation and height estimation simultaneously for remote sensing imagery, which comprises three components: preliminary feature extraction, multi-scale knowledge-guided fusion, and final prediction. Specifically, the preliminary feature extraction stage utilizes HRNet V2 as an encoder to extract features from multiple scales in parallel and transfer these abundant features to the next stage. Following this, the multi-scale knowledge-guided fusion (MKF) module generates a knowledge bank consisting of multi-scale multi-task feature maps and utilizes a cross-knowledge propagation (CKP) module to fuse the knowledge features from the knowledge bank. Subsequently, the final prediction module utilizes the fusion features to generate the ultimate predictions of semantic segmentation and height estimation. To demonstrate the effectiveness of our proposed method, we conducted experiments on two public remote sensing datasets: the Potsdam dataset and the Vaihingen dataset [31]. The experimental results indicate that the proposed method enhances the performance of both semantic segmentation and height estimation tasks while exhibiting exceptional capability in preserving

fine shapes, particularly in the context of object boundaries and contours, as shown in Figure 1.

To summarize, the main contributions of this paper are as follows:

(1) We propose a novel multi-scale knowledge-guided multi-task learning architecture for remote sensing imagery, which can be applied to various primary tasks (semantic segmentation, height estimation) and also allows for the incorporation of different auxiliary tasks (edge detection, semantic edge detection).
(2) We introduce the MKF module, which builds a multi-scale multi-task knowledge bank using task-specific feature maps and leverages the CKP module to transfer complementary information from this bank through cross-knowledge affinity, enhancing final predictions.
(3) We demonstrate our proposed method improves the performance of both semantic segmentation and height estimation against several state-of-the-art approaches on two classical datasets, namely, Potsdam and Vaihingen.

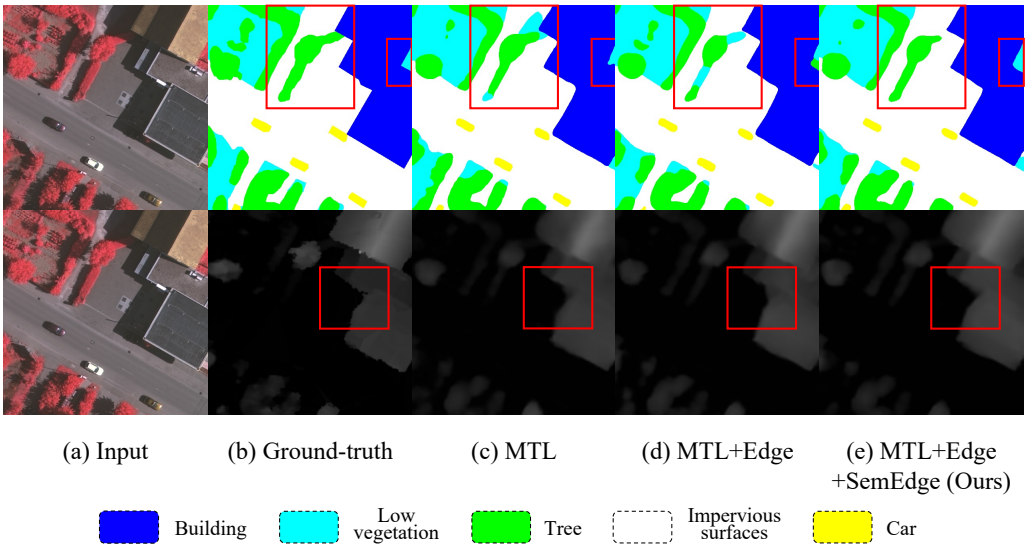


Figure 1. Prediction results by different methods, the data are sourced from the test set of the public remote sensing dataset Vaihingen. (**a**) input, (**b**) ground truth, (**c**) results by multi-task learning (MTL) network, (**d**) results by BAMTL, and (**e**) results by KMNet (ours), which receive the knowledge guidance of all other tasks.

2. Related Work

In this section, we introduce the related works of this study, including height estimation, semantic segmentation, and multi-task learning.

2.1. Semantic Segmentation

Semantic segmentation [32–37] is a classical task in computer vision and it involves assigning a categorical label to each pixel in an image, thereby facilitating the perception of the underlying environment at a pixel level. With advancements in neural networks, the performance of semantic segmentation has witnessed significant improvements, especially with the introduction of Fully Convolutional Networks (FCNs) [33], which utilizes convolutional layers instead of fully connected layers. Subsequently, several studies have been dedicated to the task of scene parsing in an end-to-end fashion, such as UNet [34], SegNet [38], and RefineNet [39], which have been developed specifically for semantic segmentation. Considering that semantic segmentation can be interpreted as a scale-sensitive dense prediction task, some works have explored approaches that leverage multi-scale information and use dilated convolution/atrous convolution to enlarge the receptive field,

improving the detailed prediction of segmentation [40–42]. To effectively utilize long-range relationships, Xu et al. [43] proposed a method that combines multi-scale feature extraction with the transformer architecture. Wang et al. [28] introduced an HRNet-based [44] approach for adaptive semantic segmentation of remote sensing scenes, and thus producing impressive results.

2.2. Height Estimation

Height estimation is a pixel-level regression task that aims to predict the elevation of each pixel in a remote sensing image captured from a single viewpoint. Analogous to depth estimation in autonomous driving [45–49], height estimation involves predicting distance information in three dimensions based on a two-dimensional image. Deep learning techniques have demonstrated notable achievements in regression tasks, including depth estimation, as evidenced by prior works [50–52]. Likewise, deep neural networks have exhibited significant benefits in addressing height estimation tasks for remote sensing scenes [53–56]. PLNet [57] utilizes progressive learning networks to predict the elevation information of objects in images in a coarse-to-fine manner. IMG2DSM [14] exploits generative adversarial networks with skip connections and penalty structure at the image patch scale to translate image-to-DSM, enabling the prediction of elevation information for remote scenes.

Although the single-task learning methods discussed above have achieved notable performance in semantic segmentation and height estimation, they inherently have limitations as they cannot leverage image features from multiple perspectives, which can potentially lead to misidentification.

2.3. Multi-Task Learning

Multi-task learning [58] is a prevalent framework in computer vision that enables simultaneous learning of multiple related tasks through a unified network. Numerous approaches [19,59,60] have been introduced to investigate the fusion of task-specific features in multi-task learning. In the PAD-Net [18] approach, an auxiliary prediction stage is proposed to augment the main task with comprehensive information. Building upon these advancements, MTI-Net [20] emphasizes the significance of multi-scale information in demanding tasks and employs HRNet as a backbone network for parallel feature extraction at different scales. InvPT [61] incorporates Transformer into multi-task learning, while TaskExpert [62] decomposes the backbone features and employs gating networks to decode task-specific features. Notably, recent studies [63–65] have made significant strides in exploring multi-task learning methods in the domain of remote sensing. According to studies [15,66,67], there exists a mutually beneficial relationship between semantic segmentation and height estimation, which has been demonstrated by using a unified network to learn both tasks simultaneously. To enhance the performance of semantic segmentation or height estimation, a common approach is utilizing edge detection as an auxiliary task to provide additional edge information, as discussed in [30,68,69]. However, these methods only focus on transferring knowledge from auxiliary information to the primary task and ignore the interaction of relevant knowledge across different tasks. Additionally, beyond low-level edge auxiliary information, high-level semantic information is equally crucial. This draws attention to the task of semantic edge detection, which is based on the recognition of edges and the classification of edge pixels into categories, while studies [70–72] have investigated various deep learning approaches for semantic edge detection, there is currently no utilization of semantic edge information for multi-task learning in the field of remote sensing. Hence, we propose KMNet, which adequately considers the interaction of relevant knowledge across different tasks on different scales and also allows the incorpo-

ration of different auxiliary tasks in a convenient way. Furthermore, the benefit of taking semantic edge detection as an auxiliary task is explored under our KMNet framework.

3. Methodology

As illustrated in Figure 2, the proposed KMNet model is a comprehensive network that predicts both semantic segmentation and height estimation. The end-to-end architecture mainly comprises three parts: preliminary feature extraction at multiple scales by a backbone encoder, multi-scale knowledge-guided fusion, and final prediction that leverages the propagated knowledge to generate the semantic segmentation and height estimation maps. In this section, we discuss each of the three parts first, and then present the loss function used for our model.

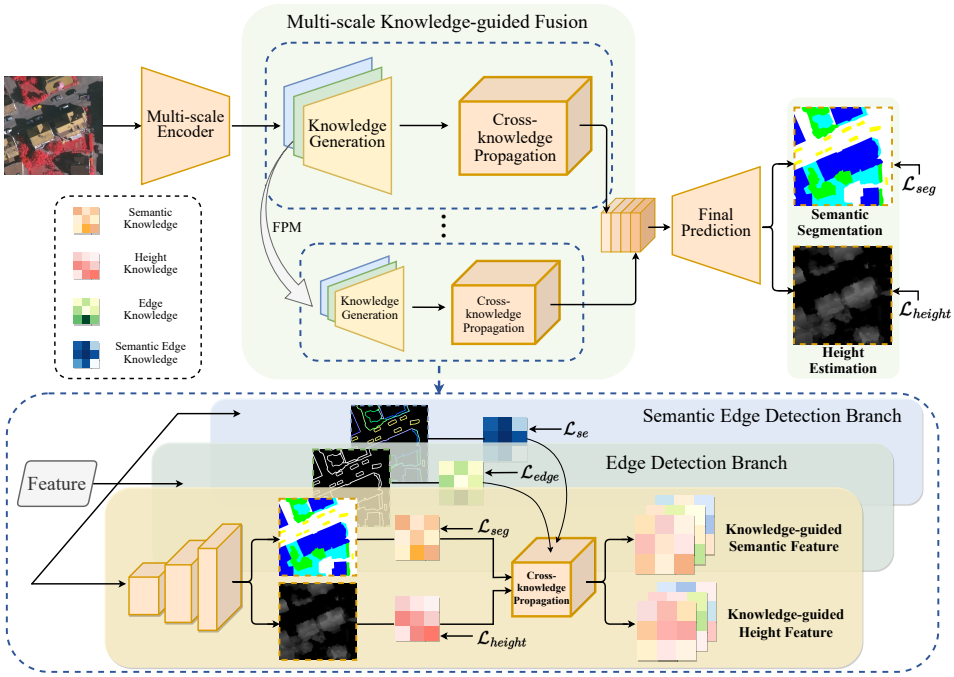


Figure 2. The overview of our Knowledge-guided Multi-task Network (KMNet). The multi-scale knowledge-guided fusion module (see Section 3.2) takes the preliminary features extracted by the encoder (see Section 3.1), generates multi-scale multi-task knowledge as a knowledge bank, and the CKP module propagates the useful and essential parts to the primary task. The fusion feature is transmitted to the final prediction stage (see Section 3.3) to produce the results of semantic segmentation and height estimation.

3.1. Front-End Network Structure

Multi-scale information plays a crucial role in remote sensing image processing due to the high resolution of the images and the presence of objects with varying sizes, including vehicles and building roofs. To be able to obtain knowledge at multiple scales, we utilize HRNet V2 [73] as the backbone of our KMNet, which can extract features from multiple scales for the next knowledge generation stage. Specifically, the KMNet framework begins with high-resolution convolutional feature extraction and gradually adds three lower-resolution feature extraction branches by downsampling during the feature extraction process. As in the common approaches [44,73], the resolution of each branch is reduced by half sequentially. By carefully formulating the parallel structure for joint learning of multiple resolutions and feature interactions across branches, we are able to obtain accurate semantic and location information simultaneously, which is particularly important for dense prediction tasks, such as semantic segmentation and height estimation.

3.2. Multi-Scale Knowledge-Guided Fusion

Following the front-end encoder, the multi-scale knowledge-guided fusion stage of KMNet is introduced. At this stage, KMNet first generates a knowledge bank consisting of multi-scale, multi-task feature maps. These knowledge features are then fused using cross-knowledge propagation (CKP) modules, as shown in Figure 3.

The knowledge bank is created through multiple task prediction branches, each generating task-specific feature maps at various scales. Each branch consists of three components: a feature propagation mechanism, a refinement module, and a decoder module. To facilitate interaction between different scales, a multi-scale feature propagation mechanism (FPM) is applied, following [20], to combine features from previous and current resolutions at each scale branch. After this, the refinement module, using two residual blocks, further enhances the combined features, and the decoder module generates task-specific feature maps via a single convolutional layer. This process yields rich, multi-scale knowledge that provides comprehensive auxiliary information for final predictions.

To effectively fuse this knowledge, we propose the CKP module, designed to capture the correlations between different knowledge features. The CKP process focuses on two primary tasks, semantic segmentation and height estimation, referred to as knowledge-demanders, which require complementary information from the knowledge bank. The other tasks (including the other primary task and two auxiliary tasks) act as knowledge-providers, offering the necessary supplementary information.

The CKP module takes as input the features of the knowledge-demander (x^t) and the knowledge-provider (x^s), both derived from the knowledge generation stage. To reduce memory usage, the feature x^s is first downsampled via depth-wise convolution, shrinking its size to $1/8$ of the original. To compute the relationship between x^t and x^s, the module generates a query (q^t) from x^t and a key (k^s) from x^s using linear layers. These are multiplied, scaled, and passed through a softmax function to produce a knowledge-guided map ($w^{t,s}$), representing the affinity between the two features.

To extract useful information from x^s related to x^t, a value (v^s) is computed from x^s. The CKP module multiplies v^s with $w^{t,s}$ and combines it with the upsampled feature of x^s (using pixel-shuffle) to recover lost information. Finally, the module concatenates multi-head features and applies a linear layer to generate $y^{t,s}$, representing the knowledge-guided features for the knowledge-demander task.

This process can be formulated as:

$$A_s^i = softmax\left(\frac{q^t \cdot k^s}{\sqrt{d_k}}\right)v^s + Up(v^s) \tag{1}$$

$$CKP(x^t, x^s) = Linear(Concat(A_s^1, \ldots, A_s^i)) \tag{2}$$

where d_k represents the head dimension.

3.3. Final Predictions

To achieve the final prediction and effectively leverage the multi-scale information, we employ upsampling operation to align the feature maps of different scales to the same size and concatenate them as an integrated feature map. The feature map is then refined through a convolutional layer, followed by batch normalization and rectified linear unit (ReLU) activation. Subsequently, the dimension of the refined features is reduced through one convolutional layer and upsampling operation to obtain the final prediction.

3.4. Optimization Method

To generate higher quality knowledge and learn the task-related knowledge, in addition to supervision on the final prediction, deep supervision is also performed on each task and at each scale of the knowledge generation process. This approach facilitates more effective utilization of the knowledge bank and improves the overall performance of the proposed framework.

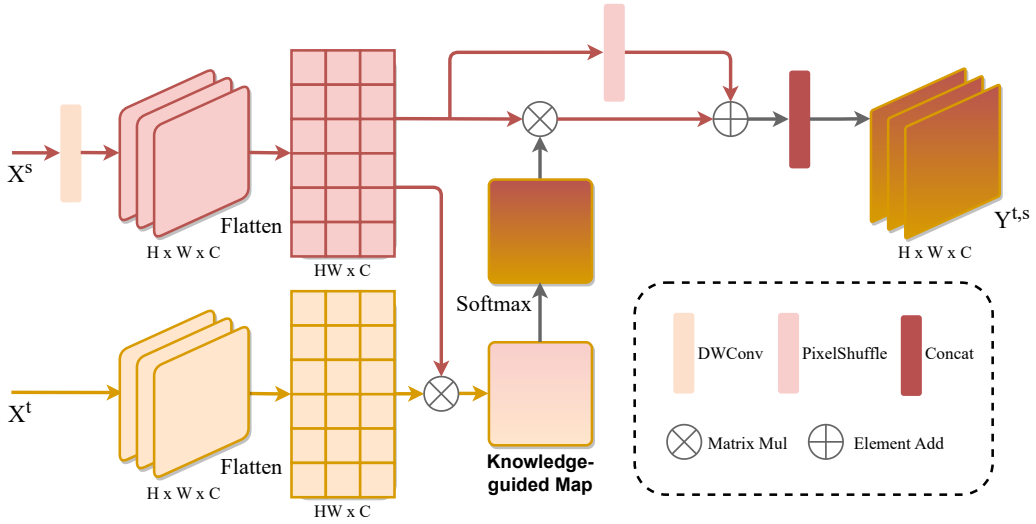


Figure 3. Illustration of cross-knowledge propagation module. The CKP module transforms the knowledge-demander feature map x^t with the knowledge-provider feature map x^s and produces the refined output feature $y^{t,s}$.

We customize the loss functions accordingly for each task. Following [18,20,30], we employ cross-entropy loss for semantic segmentation, which can be formulated as:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{h \times w} \sum_{u=1}^{h} \sum_{v=1}^{w} \sum_{c=1}^{|\mathcal{C}|} y_{u,v,c} \cdot \log(\hat{p}_{u,v,c}) \tag{3}$$

where $\hat{p} = \mathrm{softmax}(\hat{y})$ is applied over the channel dimension.

Similarly, we adopt balanced cross-entropy loss for edge detection. The BCE loss can be defined as follows:

$$\begin{aligned} \mathcal{L}_{\mathrm{BCE}} = -\frac{1}{h \times w} \sum_{u=1}^{h} \sum_{v=1}^{w} & (\beta * y_{u,v} \log(\hat{y}_{u,v}) \\ & + (1-\beta) * (1-y_{u,v}) \log(1-\hat{y}_{u,v})) \end{aligned} \tag{4}$$

where β is defined as $1 - \frac{y}{h \times w}$.

For height estimation, L1 loss is calculated to guide the optimization of height estimation branch.

$$\mathcal{L}_1 = \frac{1}{h \times w} \sum_{u=1}^{h} \sum_{v=1}^{w} |\hat{y}_{u,v,c} - y_{u,v,c}| \tag{5}$$

Since semantic edge segmentation in remote sensing images is an unbalanced multi-classification problem, we use a weighted cross-entropy loss function to constrain the prediction of object edge subcategories. In addition, we constrain all points belonging to

the edge with those not belonging to the edge using balanced cross-entropy loss. The loss function of semantic edge detection can be formulated as:

$$\mathcal{L}_{\mathrm{WCE}} = -\frac{1}{h \times w} \sum_{u=1}^{h} \sum_{v=1}^{w} \sum_{c=1}^{|\mathcal{C}|} w_c \cdot y_{u,v,c} \cdot \log(\hat{p}_{u,v,c}) \tag{6}$$

$$\mathcal{L}_{\mathrm{SED}} = \mathcal{L}_{\mathrm{WCE}}(\hat{y}, y) + 2.0 \cdot \mathcal{L}_{\mathrm{BCE}}(\hat{y}, y) \tag{7}$$

where w_c represents the weights of class c, which is the inverse of the weight of the class in the image.

Again, following [18,20], we optimize the objective of multi-task learning through a linearly combined loss function, which can be formulated as

$$\mathcal{L}_{MTL} = \sum_{i} w_i \cdot \mathcal{L}_i + \sum_{k=1}^{4} \sum_{i} w_i \cdot \mathcal{L}_i \tag{8}$$

where w_i is task-specific weight and $\mathcal{L}_i$ is task-specific loss function. For the preliminary prediction stage supervision, we compute $w_i \cdot \mathcal{L}_i$ for i-th task in the k-th scale.

4. Experiments Results and Analysis

In this section, we first describe the two public datasets used in our experiments. We then provide the implementation details and the evaluation metrics used in the experiments. Subsequently, we compare our proposed method with state-of-the-art multi-task algorithms on the aforementioned datasets, demonstrating its effectiveness. Finally, an ablation study is conducted to further analyze the proposed MKF module.

4.1. Dataset Description

Our experiments are conducted on two remote sensing datasets, namely Potsdam and Vaihingen [31], both of which comprise aerial digital images, corresponding semantic segmentation labels, and DSM data. In particular, the semantic segmentation annotation includes six classes, namely impervious surfaces, buildings, low vegetation, trees, cars, and clutter/background, where the last class is disregarded in our experiments. We obtain edge annotations by normalizing the difference between semantic segmentation annotations with edges and those without edges. Semantic edge annotations are obtained by multiplying edge annotations with semantic segmentation annotations, yielding edge information enriched with categories. For height estimation, we use the normalized DSM (nDSM) data, which is normalized to the range of [0, 1] in our experiments, following the approach used in a previous study [30].

(1) Potsdam Dataset: The Potsdam dataset comprises 38 aerial images of identical dimensions with a resolution of 6000 × 6000, obtained from ISPRS in Potsdam, Germany. The digital aerial images feature four bands, namely red, green, blue, and near-infrared bands, with a ground sampling distance of 9 cm. The DSM is generated via dense image matching with Trimble INPHO software (https://geospatial.trimble.com/en/products/software/trimble-inpho, accessed on 27 January 2025) and has a ground sampling distance of 5 cm. Following previous research [30,65], the RGB image data is utilized as the input modality, and the 38 image tiles are partitioned into 24 for training and 14 for testing.

(2) Vaihingen Dataset: The Vaihingen dataset comprises 33 aerial images with different sizes, averaging at 2494 × 2064 pixels per image. The aerial images contain IRRG (red, green, and near-infrared) bands, with 8-bit radiometric resolution and 9 cm ground sampling distance. Additionally, the DSM is obtained through dense image matching with Trimble INPHO software, with a 9 cm ground sampling distance. As mentioned above, the dataset

includes six categories for semantic segmentation, where each pixel is labeled accordingly. Following [74], 16 images are selected for model training and the remaining 17 images for model testing.

4.2. Implementation Details and Evaluation Metrics

(1) Implementation Details: During the model training phase, we employed the Adam optimizer with a weight decay of 0.0001. To enhance the model's performance, we utilized the "CosineAnnealingWarmUpRestarts" strategy to reset the learning rate every 20 epochs for a total of 80 epochs. The two datasets were assigned different initial learning rates, with 0.0001 used for the Vaihingen dataset and 0.0005 for the Potsdam dataset. Similar to the comparative approaches [20], the backbone networks utilized by our method were pre-trained on ImageNet. The weight parameters w_i for each task are determined by the Tree-structured Parzen Estimator (TPE) [75] hyperparameter search method.

To enable efficient training of the neural network and to optimize memory utilization, we employ a cropping technique for the original high-resolution remote sensing images. Specifically, we utilize a sliding window approach to extract patches of size 512 × 512 from the original image. In addition, we apply the following operations to each cropped patch, inspired by previous work [76]:

(i) Random rotation;
(ii) Resize with a random scale selected from the interval [0.75, 1.5];
(iii) Color jitter for brightness, contrast, saturation, and hue with a adjust factor of 0.25, for data augmentation.

(2) Evaluation Metrics: In this paper, following [30], we employ three metrics to evaluate the performance of semantic segmentation and six metrics to evaluate the performance of height estimation. The evaluation metrics for semantic segmentation are overall pixel accuracy (pixAcc), which measures the accuracy of the overall semantic segmentation, per-class pixel accuracy (clsAcc), which calculates the average accuracy of segmentation for different classes, and mean F1 score (mF1), which is defined as the harmonic mean of precision and recall. For the height estimation evaluation, the metrics include absolute relative error (*absRel*), mean absolute error (MAE), root mean square error ($RMSE$), and accuracy with thresholds (δ_i). The formulas are as follows:

$$absRel = \frac{1}{N}\sum_{i=1}^{N}\left|h_i - \hat{h}_i\right|/h_i \tag{9}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|h_i - \hat{h}_i\right| \tag{10}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(h_i - \hat{h}_i\right)^2} \tag{11}$$

$$\delta_i = \max\left(\frac{h_i}{\hat{h}_i}, \frac{\hat{h}_i}{h_i}\right) < 1.25^i, i \in 1,2,3 \tag{12}$$

where h represents the height ground truth, $\hat{h}$ denotes the predicted height estimation value, i is the index of the pixel in the image, and the total number of pixels in the image is represented by the N.

4.3. Comparisons with State-of-the-Art Methods

We compare the performance of our approach with state-of-the-art methods on the Potsdam dataset and the Vaihingen dataset, and the results are presented in Table 1 and Table 2, respectively.

Table 1. Quantitative Results on The ISPRS Potsdam Dataset. Seg denotes the semantic segmentation model, Height denotes the height estimation model, MTL denotes the multi-task learning approach, and bold indicates the best performance.

Method		Segmentation			Height					
Name	Task	pixAcc ↑	clsAcc ↑	mF1 ↑	absRel ↓	MAE ↓	RMSE ↓	$\delta_i < 1.25$ ↑	$\delta_i < 1.25^2$ ↑	$\delta_i < 1.25^3$ ↑
SVL_V1 [77]	Seg	77.8	-	79.9	-	-	-	-	-	-
UFMG4 [78]	Seg	87.9	-	89.5	-	-	-	-	-	-
S-RA-FCN [79]	Seg	88.6	-	90.2	-	-	-	-	-	-
HSN [80]	Seg	89.4	-	87.9	-	-	-	-	-	-
UZ_1 [6]	Seg	89.9	88.8	88.0	-	-	-	-	-	-
HRNet V2 [73]	Seg	90.8	90.6	90.6	-	-	-	-	-	-
IMG2DSM [14]	Height	-	-	-	-	-	3.890	-	-	-
Zhang et al. [81]	Height	-	-	-	-	-	3.870	-	-	-
Amirkolaee et al. [54]	Height	-	-	-	0.571	-	3.468	0.342	0.601	0.782
D3Net [13]	Height	-	-	-	0.391	1.681	3.055	0.601	0.742	0.830
IM2ELEVATION [55]	Height	-	-	-	0.429	1.744	3.516	0.638	0.767	0.839
PLNet [57]	Height	-	-	-	0.318	1.201	2.354	0.639	0.833	0.912
Srivastava et al. [82]	MTL	80.1	79.2	79.9	0.624	2.224	3.740	0.412	0.597	0.720
Carvalho et al. [15]	MTL	83.2	80.9	82.2	0.441	1.838	3.281	0.575	0.720	0.808
MTI-Net [20]	MTL	90.3	89.8	89.9	0.296	1.173	2.367	0.691	0.836	0.905
BAMTL [30]	MTL	91.3	90.4	90.9	0.291	1.223	2.407	0.685	0.819	0.897
InvPT [61]	MTL	91.1	90.3	90.6	0.253	1.210	2.402	0.673	0.829	0.904
TaskExpert [62]	MTL	90.7	89.9	90.2	0.273	1.292	2.513	0.650	0.818	0.898
Ours	MTL	**91.6**	**91.6**	**91.4**	**0.242**	**1.027**	**2.107**	**0.728**	**0.863**	**0.922**

Table 2. Quantitative Results on The ISPRS Vaihingen Dataset. Seg denotes the semantic segmentation model, Height denotes the height estimation model, MTL denotes the multi-task learning approach, and bold indicates the best performance.

Method		Segmentation			Height					
Name	Task	pixAcc ↑	clsAcc ↑	mF1 ↑	absRel ↓	MAE ↓	RMSE ↓	$\delta_i < 1.25$ ↑	$\delta_i < 1.25^2$ ↑	$\delta_i < 1.25^3$ ↑
FCN [33]	Seg	83.2	-	79.2	-	-	-	-	-	-
FPL [6]	Seg	83.8	76.5	78.8	-	-	-	-	-	-
SegNet [38]	Seg	84.1	-	81.4	-	-	-	-	-	-
ERN [83]	Seg	85.6	-	84.8	-	-	-	-	-	-
UZ_1 [6]	Seg	87.8	81.4	83.6	-	-	-	-	-	-
Deeplab V3+ [42]	Seg	88.2	85.4	86.3	-	-	-	-	-	-
HRNet V2 [73]	Seg	88.7	88.0	87.2	-	-	-	-	-	-
Zhang et al. [81]	Height	-	-	-	-	2.420	3.900	-	-	-
Amirkolaee et al. [54]	Height	-	-	-	1.163	-	2.871	0.330	0.572	0.741
IMG2DSM [14]	Height	-	-	-	-	-	2.580	-	-	-
D3Net [13]	Height	-	-	-	2.016	1.314	2.123	0.369	0.533	0.644
IM2ELEVATION [55]	Height	-	-	-	0.956	1.226	1.882	0.399	0.587	0.671
PLNet [57]	Height	-	-	-	0.833	1.178	1.775	0.386	0.599	0.702
Srivastava et al. [82]	MTL	79.3	70.4	72.6	4.415	1.861	2.729	0.217	0.385	0.517
Carvalho et al. [15]	MTL	86.1	80.4	82.3	1.882	1.262	2.089	0.405	0.562	0.663
MTI-Net [20]	MTL	87.6	85.5	85.7	1.549	1.571	2.306	0.356	0.580	0.707
BAMTL [30]	MTL	88.4	85.9	86.9	1.064	**1.078**	**1.762**	0.451	0.617	0.714
InvPT [61]	MTL	88.7	86.5	86.0	0.830	1.334	2.009	0.379	0.638	0.768
TaskExpert [62]	MTL	88.8	87.0	86.3	1.037	1.338	1.989	0.428	0.647	0.760
Ours	MTL	**89.1**	**88.7**	**87.9**	**0.695**	1.230	1.927	**0.470**	**0.696**	**0.806**

(1) Potsdam Dataset: For the semantic segmentation task, we present the experimental results from SVL_V1 [77], UFMG4 [78], S-RA-FCN [79], HSN [80], UZ_1 [6], and HRNet

V2 [73]. Regarding the height estimation task, we report the evaluation results from IMG2DSM [14,54,81], D3Net [13], IM2ELEVATION [55] and PLNet [57]. Additionally, we report the results from other multi-task learning methods in remote sensing, including the methods used in [15,82], and BAMTL [30]. Moreover, we also present the results of MTI-Net [20], InvPT [61], and TaskExpert [62], which focus on natural scenarios. As shown in Table 1, the quantitative results demonstrate the superiority of our proposed method not only over single-task learning methods but also over other multi-task learning methods in both semantic segmentation and height estimation tasks. The effectiveness in comparison to single-task learning can be attributed to the multi-task learning approach's capacity to capture visual information from diverse perspectives and exhibit enhanced generalization capabilities. Moreover, the effectiveness in relation to other multi-task learning methods highlights the utilization of rich and comprehensive knowledge from the multi-scale knowledge bank in the knowledge-fusion stage, which enables the provision of multi-type task-related information for the tasks of semantic segmentation and height estimation.

The qualitative results of our proposed method are shown in Figure 4 for the semantic segmentation task and the height estimation task. The experimental findings indicate a significant improvement in the edge quality of both the semantic segmentation and height estimation prediction maps. Additionally, as shown in Figure 4, benefiting from the extensive auxiliary information we utilize as a knowledge bank, finer edge details can be predicted for both larger objects (such as buildings and backgrounds) in rows 1 and 3 and smaller objects (such as cars and low vegetation) in rows 5 and 7, thereby improving the accuracy of the segmentation task. Moreover, the proposed approach demonstrates improved precision in identifying small and narrow regions of impervious surfaces in height predictions as shown in rows 2, 4, and 8.

(2) Vaihingen Dataset: For the semantic segmentation task, we report the results from FCN [33], FPL [6], SegNet [38], ERN [83], UZ_1 [6], Deeplab V3+ [42], and HRNet V2 [73]. For the height estimation task, we present the experimental results from [54,81], IMG2DSM [14], D3Net [13], IM2ELEVATION [55], and PLNet [57], which have been evaluated on the Vaihingen test set. Regarding the multi-task learning methods, we report the results of the methods used in [15,82], MTI-Net [20], InvPT [61], TaskExpert [62], and present the results generated by BAMTL [30]. The experimental results on the Vaihingen dataset, as shown in Table 2, reaffirm the effectiveness of utilizing multi-scale knowledge from multiple tasks to guide predictive semantic segmentation and height estimation in the learning process.

We provide qualitative comparisons for different methods in Figure 5 on the Vaihingen dataset. Our proposed method exhibits more distinct boundaries and higher prediction accuracy in the output maps for both tasks when compared to single-task learning approaches and multi-task learning methods that solely integrate semantic segmentation and height estimation. In the segmentation results presented in Figure 5, our method demonstrates superior performance in identifying boundaries, as depicted in rows 3 and 7, as well as recognizing small objects, as illustrated in rows 1 and 5. For height estimation, our method is more accurate and exhibits clearer boundaries in the predictions, as clearly evidenced by the examples shown in the red boxes in rows 2, 4, 6, and 8 in Figure 5.

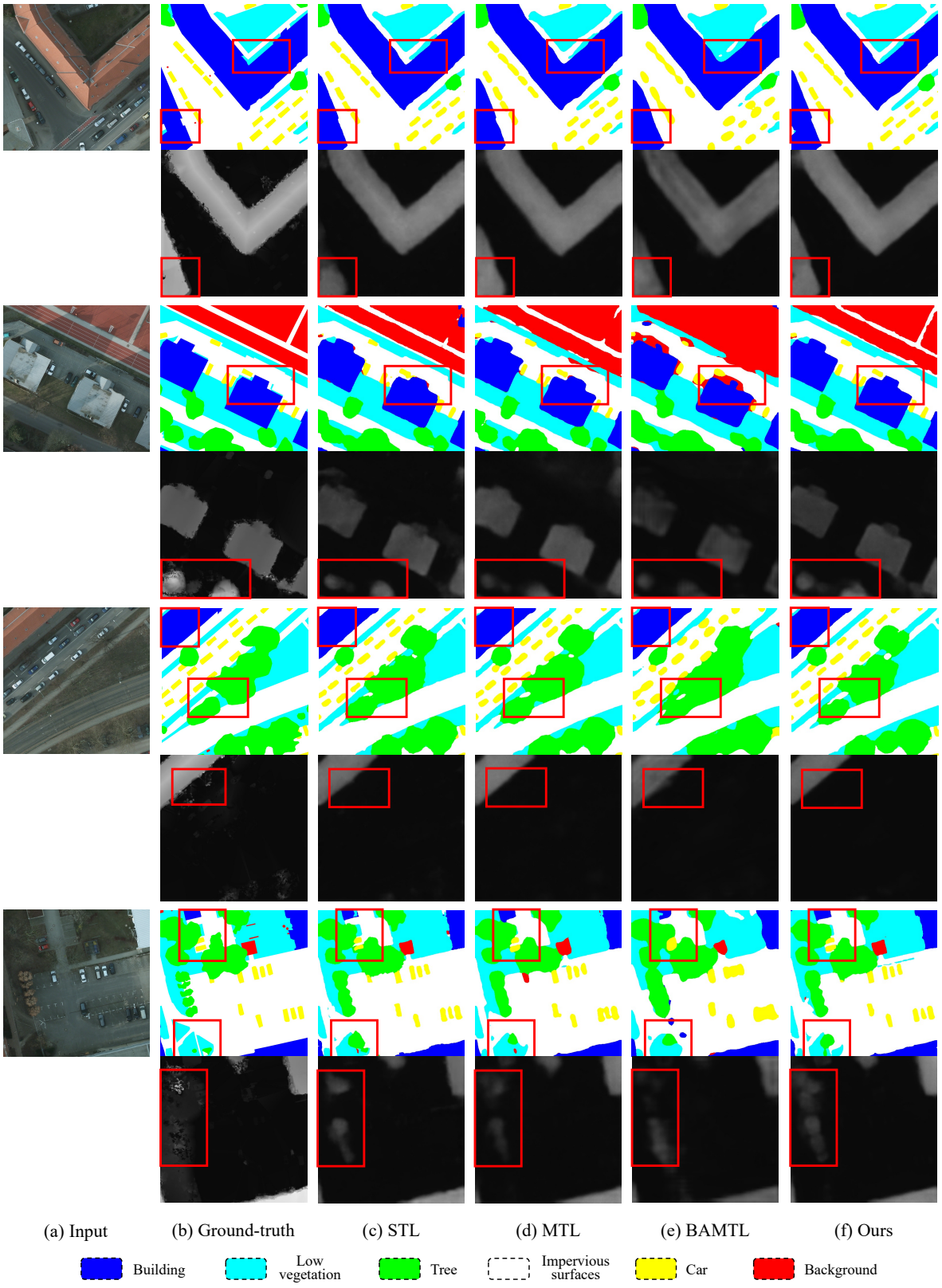


Figure 4. Qualitative comparisons for different methods on the Potsdam dataset. Column: (**a**) input image, (**b**) ground truth of segmentation and height, (**c**) predictions of the STL method, (**d**) predictions of the MTL method, (**e**) predictions of the BAMTL method, and (**f**) predictions of the KMNet method (ours). The visual results show that our proposed KMNet exhibits improvements in both semantic segmentation and height estimation performance, outperforming other approaches in terms of capturing intricate details.

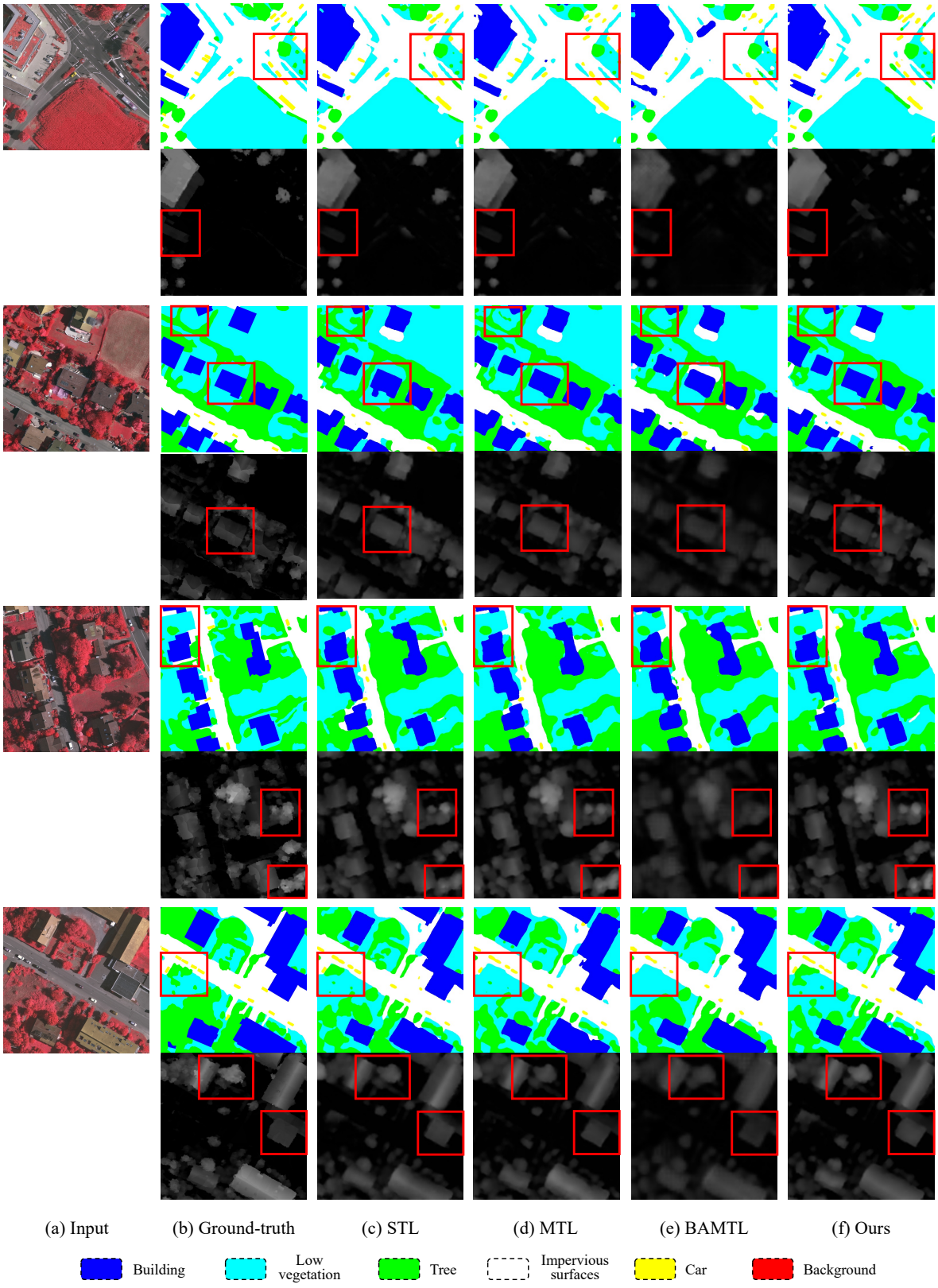


Figure 5. Qualitative comparisons for different methods on the Vaihingen dataset. Column: (**a**) input image, (**b**) ground-truth of segmentation and height, (**c**) predictions of the STL method, (**d**) predictions of the MTL method, (**e**) predictions of the BAMTL method, and (**f**) predictions of the KMNet method (ours). The visual results show that compare to the STL, MTL, and BAMTL methods, our proposed KMNet method has more clear boundaries and more accuracy in the predictions of segmentation and height.

4.4. Ablation Studies

To assess the efficacy of the proposed MKF module and evaluate the impact of the auxiliary tasks, ablation studies are conducted on the Vaihingen dataset.

Table 3 presents quantitative results of the effectiveness of the multi-scale knowledge-guided fusion module and the contribution of different auxiliary tasks on the Vaihingen dataset. The experimental results comparing row 1 and row 2 clearly demonstrate a significant improvement in accuracy for semantic segmentation and height estimation tasks with the incorporation of the MKF module. This observation underscores the efficacy of a knowledge-guided architecture, which generates a knowledge bank first and then utilizes the relative and useful knowledge within it to guide the final predictions. Consequently, this pivotal innovation serves as the fundamental cornerstone of our research endeavor. Row 2 also showcases the results of utilizing semantic segmentation and height estimation as exclusive knowledge-providers. Subsequently, row 3 introduces the incorporation of edge detection as an additional knowledge-provider, resulting in notable findings. The obtained results demonstrate a discernible improvement in semantic segmentation accuracy and height estimation evaluation metrics. This substantiates the notion that the guidance provided by edge knowledge effectively complements and enhances the tasks of semantic segmentation and height estimation. Furthermore, row 4 presents our exploration in the realm of semantic edge detection, demonstrating its potential for enhancing predictive outcomes. These improvements are consistently evident across a significant portion of the evaluation metrics, which confirm that semantic edge detection represents a valuable auxiliary information source.

Table 3. Ablation Experiments on The ISPRS Vaihingen Dataset.

MKF Module	Edge	SemEdge	Segmentation			Height					
			pixAcc ↑	clsAcc ↑	mF1 ↑	absRel ↓	MAE ↓	RMSE ↓	$\delta_i < 1.25$ ↑	$\delta_i < 1.25^2$ ↑	$\delta_i < 1.25^3$ ↑
			86.9	86.4	84.7	0.946	1.362	1.988	0.384	0.617	0.742
✓			88.7	87.3	86.9	0.914	1.245	1.912	0.409	0.641	0.764
✓	✓		88.9	88.2	87.5	0.714	1.231	1.908	0.446	0.693	0.805
✓	✓	✓	89.1	88.7	87.9	0.695	1.230	1.927	0.470	0.696	0.806

Table 4 presents the quantitative experimental results of different multi-task learning methods with various backbones on the Vaihingen dataset. The experimental results demonstrate that our approach outperforms methods utilizing the same HRNet48 backbone. Additionally, when comparing InvPT [61] and TaskExpert [62], both of which are large-parameter models using ViT as a backbone, our approach performs better while maintaining a smaller count of parameters. It is noteworthy that when using HRNet18 as the backbone, our method not only has a small parameter count but also surpasses the high-parameter BAMTL in most of the metrics. This further validates the effectiveness of our multi-scale knowledge-guided multi-task learning approach.

Table 4. Ablation Experiments of Backbone on The ISPRS Vaihingen Dataset.

Method Name	Backbone	Segmentation			Height						Params
		pixAcc ↑	clsAcc ↑	mF1 ↑	absRel ↓	MAE ↓	RMSE ↓	$\delta_i < 1.25$ ↑	$\delta_i < 1.25^2$ ↑	$\delta_i < 1.25^3$ ↑	
Srivastava et al. [82]	—	79.3	70.4	72.6	4.415	1.861	2.729	0.217	0.385	0.517	—
Carvalho et al. [15]	—	86.1	80.4	82.3	1.882	1.262	2.089	0.405	0.562	0.663	—
MTI-Net [20]	HRNet18	87.6	85.5	85.7	1.549	1.571	2.306	0.356	0.580	0.707	8.6
MTI-Net [20]	HRNet48	88.7	87.4	87.1	0.809	1.259	1.928	0.415	0.663	0.784	98.7
BAMTL [30]	ResNet101	88.4	85.9	86.9	1.064	**1.078**	**1.762**	0.451	0.617	0.714	59.1
BAMTL [30]	HRNet48	88.5	85.7	86.2	1.216	1.303	2.125	0.291	0.529	0.684	83.5
InvPT [61]	ViT-L	88.7	86.5	86.0	0.830	1.334	2.009	0.379	0.638	0.768	358.7
TaskExpert [62]	ViT-L	88.8	87.0	86.3	1.037	1.338	1.989	0.428	0.647	0.760	372.1

Table 4. *Cont.*

Method		Segmentation			Height						Params
Name	**Backbone**	**pixAcc ↑**	**clsAcc ↑**	**mF1 ↑**	**absRel ↓**	**MAE ↓**	**RMSE ↓**	$\delta_i < 1.25 \uparrow$	$\delta_i < 1.25^2 \uparrow$	$\delta_i < 1.25^3 \uparrow$	
Ours	HRNet18	88.7	87.9	87.2	0.802	1.321	2.058	0.414	0.660	0.785	14.8
Ours	HRNet48	**89.1**	**88.7**	**87.9**	**0.695**	1.230	1.927	**0.470**	**0.696**	**0.806**	137.7

5. Conclusions

In this paper, we proposed a novel knowledge-guided multi-task network named KMNet for remote sensing imagery. It leverages a multi-scale feature extraction process to extract preliminary features first. Subsequently, the MKF module is employed to generate a comprehensive multi-scale multi-task knowledge bank and employ it to guide the feature fusion process. The multi-knowledge fusion feature is then utilized to generate the final predictions for tasks such as semantic segmentation and height estimation. We evaluate our KMNet method on two publicly available remote sensing datasets and the results demonstrate that our method achieves comparable performance to several state-of-the-art approaches. The ablation study further highlights the importance of the MKF module and the integration of auxiliary tasks, while also demonstrating the performance superiority of our method, all achieved with a minimal number of parameters.

While we have focused on tasks such as semantic segmentation, height estimation, edge detection and semantic edge detection, it is essential to note that the architecture of KMNet holds potential for various primary and auxiliary tasks within the remote sensing domain, including image enhancement and object detection. However, similar to other complex multi-task learning models, applying this approach to non-pixel-level prediction tasks may necessitate a redesign of the task-specific head. As a multi-task framework for remote sensing image analysis, it empowers us to extract deeper insights from the data, conserve memory, and accelerate inference speed. In the future, we will investigate the application of this framework to other tasks, particularly focusing on its advantages in on-orbit deployment [84].

Author Contributions: Conceptualization, M.L. and G.W.; Data curation, M.L.; Investigation, T.L.; Methodology, M.L. and G.W.; Resources, G.W. and Y.Y.; Validation, M.L.; Visualization, M.L.; Writing—original draft, M.L.; Writing—review and editing, G.W., T.L., Y.Y., W.L., X.L. and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China (U23B2011, 62102069, U20B2063 and 62220106008), the Key Research and Development Program of Zhejiang Province (2024SSYS0091), and the Sichuan Provincial Science and Technology Support Program (2024NSFTD0034).

Data Availability Statement: No new data were generated in this study. The data used in this study are public-use data files prepared and disseminated to provide access to the full scope of the data. Data will be made available on request.

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

References

1. Ding, P.; Zhang, Y.; Deng, W.J.; Jia, P.; Kuijper, A. A light and faster regional convolutional neural network for object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2018**, *141*, 208–218. [CrossRef]
2. Kramer, H.J.; Cracknell, A.P. An overview of small satellites in remote sensing. *Int. J. Remote Sens.* **2008**, *29*, 4285–4337. [CrossRef]
3. Schowengerdt, R.A. *Remote Sensing: Models and Methods for Image Processing*; Elsevier: Amsterdam, The Netherlands, 2006.
4. Li, L.; Song, N.; Sun, F.; Liu, X.; Wang, R.; Yao, J.; Cao, S. Point2Roof: End-to-end 3D building roof modeling from airborne LiDAR point clouds. *ISPRS J. Photogramm. Remote Sens.* **2022**, *193*, 17–28. [CrossRef]

5. Beumier, C.; Idrissa, M. Digital terrain models derived from digital surface model uniform regions in urban areas. *Int. J. Remote Sens.* **2016**, *37*, 3477–3493. [CrossRef]
6. Volpi, M.; Tuia, D. Dense semantic labeling of sub-decimeter resolution images with convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2016**, *55*, 1–13.
7. Qin, R.; Tian, J.; Reinartz, P. 3D change detection–approaches and applications. *ISPRS J. Photogramm. Remote Sens.* **2016**, *122*, 41–56. [CrossRef]
8. Vega, P.J.S.; da Costa, G.A.O.P.; Feitosa, R.Q.; Adarme, M.X.O.; de Almeida, C.A.; Heipke, C.; Rottensteiner, F. An unsupervised domain adaptation approach for change detection and its application to deforestation mapping in tropical biomes. *ISPRS J. Photogramm. Remote Sens.* **2021**, *181*, 113–128. [CrossRef]
9. Tu, J.; Sui, H.; Feng, W.; Song, Z. Automatic building damage detection method using high-resolution remote sensing images and 3D GIS Model. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *3*, 43–50.
10. Vetrivel, A.; Gerke, M.; Kerle, N.; Nex, F.; Vosselman, G. Disaster damage detection through synergistic use of deep learning and 3D point cloud features derived from very high resolution oblique aerial images, and multiple-kernel-learning. *ISPRS J. Photogramm. Remote Sens.* **2018**, *140*, 45–59. [CrossRef]
11. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [CrossRef]
12. Peters, T.; Brenner, C.; Schindler, K. Semantic segmentation of mobile mapping point clouds via multi-view label transfer. *ISPRS J. Photogramm. Remote Sens.* **2023**, *202*, 30–39. [CrossRef]
13. Carvalho, M.; Le Saux, B.; Trouvé-Peloux, P.; Almansa, A.; Champagnat, F. On regression losses for deep depth estimation. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 2915–2919.
14. 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]
15. Carvalho, M.; Le Saux, B.; Trouvé-Peloux, P.; Champagnat, F.; Almansa, A. Multi-task learning of height and semantics from aerial images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 1391–1395. [CrossRef]
16. Standley, T.; Zamir, A.; Chen, D.; Guibas, L.; Malik, J.; Savarese, S. Which tasks should be learned together in multi-task learning? In Proceedings of the International Conference on Machine Learning, Online, 13–18 July 2020; pp. 9120–9132.
17. Misra, I.; Shrivastava, A.; Gupta, A.; Hebert, M. Cross-Stitch networks for multi-task learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3994–4003.
18. Xu, D.; Ouyang, W.; Wang, X.; Sebe, N. PAD-Net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 675–684.
19. Gao, Y.; Ma, J.; Zhao, M.; Liu, W.; Yuille, A.L. NDDR-CNN: Layerwise feature fusing in multi-task CNNs by neural discriminative dimensionality reduction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3205–3214.
20. Vandenhende, S.; Georgoulis, S.; Van Gool, L. MTI-Net: Multi-scale task interaction networks for multi-task learning. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part IV 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 527–543.
21. Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S.R. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv* **2018**, arXiv:1804.07461.
22. Liu, X.; He, P.; Chen, W.; Gao, J. Improving multi-task deep neural networks via knowledge distillation for natural language understanding. *arXiv* **2019**, arXiv:1904.09482.
23. Worsham, J.; Kalita, J. Multi-task learning for natural language processing in the 2020s: Where are we going? *Pattern Recognit. Lett.* **2020**, *136*, 120–126. [CrossRef]
24. Kargar, E.; Kyrki, V. Efficient latent representations using multiple tasks for autonomous driving. *arXiv* **2020**, arXiv:2003.00695.
25. Phillips, J.; Martinez, J.; Bârsan, I.A.; Casas, S.; Sadat, A.; Urtasun, R. Deep multi-task learning for joint localization, perception, and prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4679–4689.
26. Casas, S.; Sadat, A.; Urtasun, R. MP3: A unified model to map, perceive, predict and plan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14403–14412.
27. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
28. Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; Zhong, Y. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv* **2021**, arXiv:2110.08733.

29. 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; pp. 4963–4966.
30. Wang, Y.; Ding, W.; Zhang, R.; Li, H. Boundary-aware multitask learning for remote sensing imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *14*, 951–963. [CrossRef]
31. 2D Semantic Labeling Contest. 2014. Available online: https://www.isprs.org/documents/si/SI-2014/ISPRS_SI_report-website_WGIII4_Gerke_2014.pdf (accessed on 1 July 2020).
32. Ciresan, D.; Giusti, A.; Gambardella, L.; Schmidhuber, J. Deep neural networks segment neuronal membranes in electron microscopy images. *Adv. Neural Inf. Process. Syst.* **2012**, *25*. [CrossRef]
33. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 640–651.
34. 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: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
35. Xu, Y.; He, F.; Du, B.; Tao, D.; Zhang, L. Self-ensembling GAN for cross-domain semantic segmentation. *IEEE Trans. Multimed.* **2022**, *25*, 7837–7850. [CrossRef]
36. Ma, L.; Xie, H.; Liu, C.; Zhang, Y. Learning cross-channel representations for semantic segmentation. *IEEE Trans. Multimed.* **2022**, *25*, 2774–2787. [CrossRef]
37. Yin, C.; Tang, J.; Yuan, T.; Xu, Z.; Wang, Y. Bridging the gap between semantic segmentation and instance segmentation. *IEEE Trans. Multimed.* **2021**, *24*, 4183–4196. [CrossRef]
38. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
39. Lin, G.; Milan, A.; Shen, C.; Reid, I. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1925–1934.
40. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. *IEEE Comput. Soc.* **2016**. [CrossRef]
41. 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.* **2018**, *40*, 834–848. [CrossRef]
42. 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.
43. Xu, R.; Wang, C.; Zhang, J.; Xu, S.; Meng, W.; Zhang, X. RSSFormer: Foreground saliency enhancement for remote sensing land-cover segmentation. *IEEE Trans. Image Process.* **2023**, *32*, 1052–1064. [CrossRef]
44. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3349–3364. [CrossRef]
45. Watson, J.; Mac Aodha, O.; Prisacariu, V.; Brostow, G.; Firman, M. The temporal opportunist: Self-supervised multi-frame monocular depth. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 1164–1174.
46. Wang, L.; Zhang, J.; Wang, O.; Lin, Z.; Lu, H. SDC-Depth: Semantic divide-and-conquer network for monocular depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 541–550.
47. Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; Tao, D. Deep ordinal regression network for monocular depth estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2002–2011.
48. Li, R.; Xue, D.; Zhu, Y.; Wu, H.; Sun, J.; Zhang, Y. Self-supervised monocular depth estimation with frequency-based recurrent refinement. *IEEE Trans. Multimed.* **2022**, *25*, 5626–5637. [CrossRef]
49. Shao, S.; Li, R.; Pei, Z.; Liu, Z.; Chen, W.; Zhu, W.; Wu, X.; Zhang, B. Towards comprehensive monocular depth estimation: Multiple heads are better than one. *IEEE Trans. Multimed.* **2022**, *25*, 7660–7671. [CrossRef]
50. dos Santos Rosa, N.; Guizilini, V.; Grassi, V. Sparse-to-continuous: Enhancing monocular depth estimation using occupancy maps. In Proceedings of the 2019 19th International Conference on Advanced Robotics (ICAR), Belo Horizonte, Brazil, 2–6 December 2019; pp. 793–800.
51. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. GhostNet: More features from cheap operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1580–1589.
52. Zhang, Z.; Xu, C.; Yang, J.; Gao, J.; Cui, Z. Progressive hard-mining network for monocular depth estimation. *IEEE Trans. Image Process.* **2018**, *27*, 3691–3702. [CrossRef] [PubMed]

53. Mou, L.; Zhu, X.X. IM2HEIGHT: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network. *arXiv* **2018**, arXiv:1802.10249.
54. 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]
55. 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]
56. Mo, D.; Fan, C.; Shi, Y.; Zhang, Y.; Lu, R. Soft-aligned gradient-chaining network for height estimation from single aerial images. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 538–542. [CrossRef]
57. Xing, S.; Dong, Q.; Hu, Z. Gated feature aggregation for height estimation from single aerial images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
58. Caruana, R. Multitask learning. *Mach. Learn.* **1997**, *28*, 41–75. [CrossRef]
59. Ruder, S.; Bingel, J.; Augenstein, I.; Søgaard, A. Latent multi-task architecture learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Long Beach, CA, USA, 16–17 June 2019; Volume 33, pp. 4822–4829.
60. Zhang, Z.; Cui, Z.; Xu, C.; Yan, Y.; Sebe, N.; Yang, J. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4106–4115.
61. Ye, H.; Xu, D. Inverted pyramid multi-task transformer for dense scene understanding. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 514–530.
62. Ye, H. Taskexpert: Dynamically assembling multi-task representations with memorial mixture-of-experts. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 21828–21837.
63. Cheng, D.; Meng, G.; Xiang, S.; Pan, C. FusionNet: Edge aware deep convolutional networks for semantic segmentation of remote sensing harbor images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 5769–5783. [CrossRef]
64. Marmanis, D.; Schindler, K.; Wegner, J.D.; Galliani, S.; Datcu, M.; Stilla, U. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS J. Photogramm. Remote Sens.* **2018**, *135*, 158–172. [CrossRef]
65. Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114. [CrossRef]
66. Li, X.; Wen, C.; Wang, L.; Fang, Y. Geometry-aware segmentation of remote sensing images via joint height estimation. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
67. Feng, Y.; Sun, X.; Diao, W.; Li, J.; Niu, R.; Gao, X.; Fu, K. Height aware understanding of remote sensing images based on cross-task interaction. *ISPRS J. Photogramm. Remote Sens.* **2023**, *195*, 233–249. [CrossRef]
68. Cheng, D.; Meng, G.; Cheng, G.; Pan, C. SeNet: Structured edge network for sea-land segmentation. *IEEE Geosci. Remote Sens. Lett.* **2016**, *14*, 247–251. [CrossRef]
69. Volpi, M.; Tuia, D. Deep multi-task learning for a geographically-regularized semantic segmentation of aerial images. *ISPRS J. Photogramm. Remote Sens.* **2018**, *144*, 48–60. [CrossRef]
70. Hu, Y.; Chen, Y.; Li, X.; Feng, J. Dynamic feature fusion for semantic edge detection. *arXiv* **2019**, arXiv:1902.09104.
71. Xia, L.; Zhang, X.; Zhang, J.; Yang, H.; Chen, T. Building extraction from very-high-resolution remote sensing images using semi-supervised semantic edge detection. *Remote Sens.* **2021**, *13*, 2187. [CrossRef]
72. Liu, Y.; Cheng, M.M.; Fan, D.P.; Zhang, L.; Bian, J.W.; Tao, D. Semantic edge detection with diverse deep supervision. *Int. J. Comput. Vis.* **2022**, *130*, 179–198. [CrossRef]
73. Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Xiao, B.; Liu, D.; Mu, Y.; Wang, X.; Liu, W.; Wang, J. High-resolution representations for labeling pixels and regions. *arXiv* **2019**, arXiv:1904.04514.
74. Hou, J.; Guo, Z.; Wu, Y.; Diao, W.; Xu, T. BSNet: Dynamic hybrid gradient convolution based boundary-sensitive network for remote sensing image segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–22. [CrossRef]
75. Bergstra, J.; Bardenet, R.; Bengio, Y.; Kégl, B. Algorithms for hyper-parameter optimization. *Adv. Neural Inf. Process. Syst.* **2011**, *24*. [CrossRef]
76. Zhu, Y.; Sapra, K.; Reda, F.A.; Shih, K.J.; Newsam, S.; Tao, A.; Catanzaro, B. Improving semantic segmentation via video propagation and label relaxation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8856–8865.
77. Gerke, M. Use of the Stair Vision Library Within the ISPRS 2D Semantic Labeling Benchmark (Vaihingen). Available online: https://www.researchgate.net/publication/270104226_Use_of_the_Stair_Vision_Library_within_the_ISPRS_2D_Semantic_Labeling_Benchmark_Vaihingen (accessed on 1 January 2015).
78. Nogueira, K.; Dalla Mura, M.; Chanussot, J.; Schwartz, W.R.; Dos Santos, J.A. Dynamic multi-context segmentation of remote sensing images based on convolutional networks. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 7503–7520. [CrossRef]
79. Mou, L.; Hua, Y.; Zhu, X.X. Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7557–7569. [CrossRef]

80. Liu, Y.; Minh Nguyen, D.; Deligiannis, N.; Ding, W.; Munteanu, A. Hourglass-ShapeNetwork based semantic segmentation for high resolution aerial imagery. *Remote Sens.* **2017**, *9*, 522. [CrossRef]
81. Zhang, Y.; Chen, X. Multi-path fusion network for high-resolution height estimation from a single orthophoto. In Proceedings of the 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Shanghai, China, 8–12 July 2019; pp. 186–191.
82. 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; pp. 5173–5176.
83. Liu, S.; Ding, W.; Liu, C.; Liu, Y.; Wang, Y.; Li, H. ERN: Edge loss reinforced semantic segmentation network for remote sensing images. *Remote Sens.* **2018**, *10*, 1339. [CrossRef]
84. Shen, Y.; Liu, D.; Zhang, F.; Zhang, Q. Fast and accurate multi-class geospatial object detection with large-size remote sensing imagery using CNN and Truncated NMS. *ISPRS J. Photogramm. Remote Sens.* **2022**, *191*, 235–249. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.