Cross-Task Mamba Network for Building Extraction and Height Estimation from Single-View Remote Sensing Images

Yu He[1,2], Po Liu[2], Qingdong Wang[2], Nengcheng Chen[1], Liang Zhai[2]

[1]School of Future technology, China University of Geosciences (Wuhan), Wuhan 430074, China
Email: heyu666@cug.edu.cn
[2]Chinese Academy of Surveying and Mapping, Beijing 100036, China

Keywords: Building height estimation, Building extraction, Mamba, Multitask learning.

Abstract

Simultaneous building extraction and height estimation from single-view satellite imagery via multi-task learning presents a viable solution for large-scale urban 3D reconstruction. However, balancing the weights across different tasks and reducing conflicts between them remains a challenging problem. In this paper, we propose a Mamba-CNN based network to more effectively capture the spatial distribution of buildings. Additionally, we propose a cross-task feature fusion module to facilitate information exchange between tasks. Experiments conducted on the Vaihingen dataset demonstrate significant improvements achieved by the proposed method.

1. Introduction

Buildings constitute the primary components of urban areas, and analyzing their spatial distribution and attributes such as height holds significant importance for urban planning (Son et al., 2023), population estimation (Boo et al., 2022), and disaster management (Zheng et al., 2021). With advancements in remote sensing technology and deep learning, building extraction and height estimation from satellite imagery have facilitated large-scale 3D urban reconstruction.

Traditionally, building extraction and height estimation have been regarded as two separate tasks, and their results are combined to generate 3D building models. However, it is challenging to accomplish both tasks directly from a single image. Remote sensing images have inter-class similarity and intra-class difference, which bring difficulties to semantic segmentation. For height estimation, the same object may correspond to multiple different heights. Recent studies have shown that there is a strong correlation between the height changes and semantic changes of buildings. Jointly performing building extraction and height estimation tasks can promote both tasks simultaneously. D3-Net (Carvalho et al., 2020) proposed a single-encoder-dual-decoder architecture to jointly generate pixel-level building segmentation maps and height estimation maps. SCE-Net (Xing et al., 2022b) further improved cross-task feature interaction through dedicated modules. However, challenges remain due to task-specific semantic inconsistencies in remote sensing imagery (e.g., different objects sharing similar heights) and the difficulty in balancing inter-task weights.

To address the aforementioned issues, we propose BiMamba3D, a novel network architecture that integrates CNN and Mamba frameworks. BiMamba3D employs the encoder of MambaVision to extract features. After utilizing shallow decoders to separately learn features for building segmentation and height estimation, a Cross-task Mamba Module (CMM) is introduced to facilitate information exchange between the two tasks.

2. Related work

2.1 Single-task learning

2.1.1 Building segmentation: Building extraction has undergone a transformation from traditional methods based on geometry and regional segmentation to deep learning-driven approaches. Early methods relied on handcrafted features (Hao et al., 2019; Liasis & Stavrou, 2016) and had limited applicability. Deep learning methods treat building segmentation as a pixel classification problem, and by introducing Convolutional Neural Networks (CNN) to improve the model, the segmentation accuracy has been enhanced.

Fully Convolutional Networks (FCNs) (Long et al., 2015) were the first to replace the fully connected layers in traditional convolutional neural networks with convolutional layers, enabling pixel-wise prediction for inputs of arbitrary sizes. Since then, methods such as U-Net (‘U-Net’, 2015) and DeepLab (L.-C. Chen et al., 2018) have significantly advanced semantic segmentation technology. Recent studies often adopt specific strategies to improve network frameworks targeting specific issues in building extraction, such as multi-scale problems (R. Li et al., 2021; Xu et al., 2022; Zhou et al., 2022), uncertainty issues (J. Li et al., 2024), and the problem of small inter-class differences and large intra-class differences (Hamaguchi, 2024), thereby achieving more excellent performance.

The introduction of transformer has brought a new perspective to building segmentation. Through the self-attention mechanism, it can perform global context modeling and capture the spatial distribution patterns of building clusters, overcoming the limitations of CNN in terms of receptive fields. Chen et al. (2022) proposed combining Swin transformer with channel and spatial enhancement technologies to extract multi-scale features and achieve accurate building extraction. To improve the computational efficiency of transformer, STT (K. Chen et al., 2021) introduced a dual-path transformer framework,

representing buildings as sparse feature vectors in the feature space from both spatial and channel perspectives. In addition, since CNN is good at local feature extraction, most existing transformer-based building segmentation methods adopt a hybrid design, aiming to combine the advantages of convolutional neural networks and transformer. Zhang et al. (2024) designed a dual-branch structure that can capture both global dependencies and local features, and integrate them into a multi-scale global-local context representation with enhanced boundary features. Fu et al, (2024) proposed a parallel CNN-transformer architecture and introduced an interactive self-attention mechanism to fuse multi-level features from the two branches.

2.1.2 Building height estimation: Early monocular elevation estimation mainly includes random field models and CNN models. The core idea of random fields is to estimate building heights by modeling the spatial dependencies between pixels or regions and using the constraints of single-point features and neighborhood structures. Studies used Markov Random Field (MRF) and Conditional Random Field (CRF) to model the local and global structures of images. Considering that local features cannot provide sufficient information for predicting depth values, Batra (Batra & Saxena, 2012) modeled the relationships between adjacent regions. To obtain both local and global features, Saxena et al. (Saxena et al., 2008) calculated the features of four adjacent regions of a specified region and used the Laplacian model and MRF to estimate elevation.

Inspired by depth estimation tasks, researchers have introduced neural networks to estimate building heights. In terms of network frameworks, monocular elevation estimation mainly includes single-task learning and multi-task learning. Single-task learning directly uses neural networks for height value regression. IM2HEIGHT (Mou et al., 2018) first realized the mapping from single-view images to DSM using a fully convolutional network architecture. Amirkolaee et al. (2019) proposed a CNN with an encoder-decoder structure. In the encoder part, ResNet is used to model global and local features; in the decoder part, upsampling operations are utilized to increase the size of elevation results. Karatsiolis et al. (2021) proposed IMG2nDSM to estimate the heights of buildings and vegetation from a single aerial image. IM2ELEVATION (C.-J. Liu et al., 2020) improved the feature aggregation method by combining the features of the encoding layer with those of the final decoding layer, achieving good results. Xing et al. (2022a) proposed a gated feature aggregation method, which enhances the height estimation effect and preserves clearer object boundaries and contours by effectively combining low-level and high-level features.

LUMNet (Du et al., 2024) significantly improved the accuracy of height estimation by combining prior knowledge and multi-scale feature extraction. Mao et al. (Mao, Chen, et al., 2023) designed the SFFDE method, which integrates global and local information, and proposed a building modeling framework that combines building entity extraction and elevation estimation.

Due to the non-linear nature of surface object heights in remote sensing images and their extremely large dynamic range, some studies have attempted to convert the elevation estimation problem into a height interval classification problem. Li et al. (2022) divided height values into intervals with gradually increasing gaps, transformed the regression problem into an ordinal regression problem, and used ordinal loss for network training. Through post-processing techniques, the predicted height map of each image patch is converted into a seamless height map.

Feng et al. (2022) defined height discretization rules and introduced a distance penalty index, converting continuous ground height values into a soft probability distribution. In the inference stage, the predicted height is obtained by means of soft weighted summation. Wang et al. (2024) used the SAM model to filter out the background, and then proposed to use SAdabins to optimize the results of the regression task. Chen et al. (2023) proposed to adopt adaptive intervals in network design and clipped a small number of deviation values generated by the network.

2.2 Multi-task learnin

Multi-task learning attempts to make the building shapes in the obtained DSM more regular by simultaneously learning semantic segmentation information and building height information. How to balance the weight allocation among multi-task learning and how to promote feature exchange between different task branches are two key research directions in multi-task learning. LIGHT (Mao, Sun, et al., 2023) designed a unified multi-scale feature branch and proposed a gated cross-task interaction module to alleviate the feature gap between different tasks. Liu et al. (2022) realized feature interaction between semantic segmentation and elevation extraction tasks by designing a cross-task adaptive propagation module, while learning global context information and local geometric features. Feng et al. (2023) proposed a calibration and refinement attention module to filter inconsistent features between the two tasks, and introduced adjacent pixel affinity loss and soft-weighted ordinal loss for the two tasks to optimize the direction of task gradients.

3. Method

This paper presents a multi-task learning network, BiMamba3D, designed to jointly perform building segmentation and height estimation. The network architecture consists of encoder, decoder, a cross-task information interaction module CMM and two task-specific prediction heads. By incorporating the Mamba architecture, BiMamba3D can more effectively capture long-range spatial relationships in images, such as the global structure of buildings. Meanwhile, through the CMM, the semantic boundary information from the segmentation task can guide the edge refinement in height estimation. Conversely, the height gradient information can feedback to facilitate instance separation in the segmentation task.

3.1 Decoder

We use the encoder of MambaVision for feature extraction, following (Hatamizadeh & Kautz, 2025). We employ two U-shaped decoder structures for auxiliary semantic segmentation and height estimation decoding, respectively, which gradually reconstruct spatial resolution through hierarchical feature fusion.

Let the encoder extract four levels of features $\{X_1, X_2, X_3, X_4\}$. The number of channels in each feature map is denoted as $\{C_1, C_2, C_3, C_4\} = (196, 392, 784, 1568)$, and all features are spatially aligned through appropriate upsampling operations during decoding.

The decoder first projects the deepest feature map $X_4 \in \mathbb{R}^{C_4 \times H/8 \times W/8}$ into a unified channel space using a 3×3

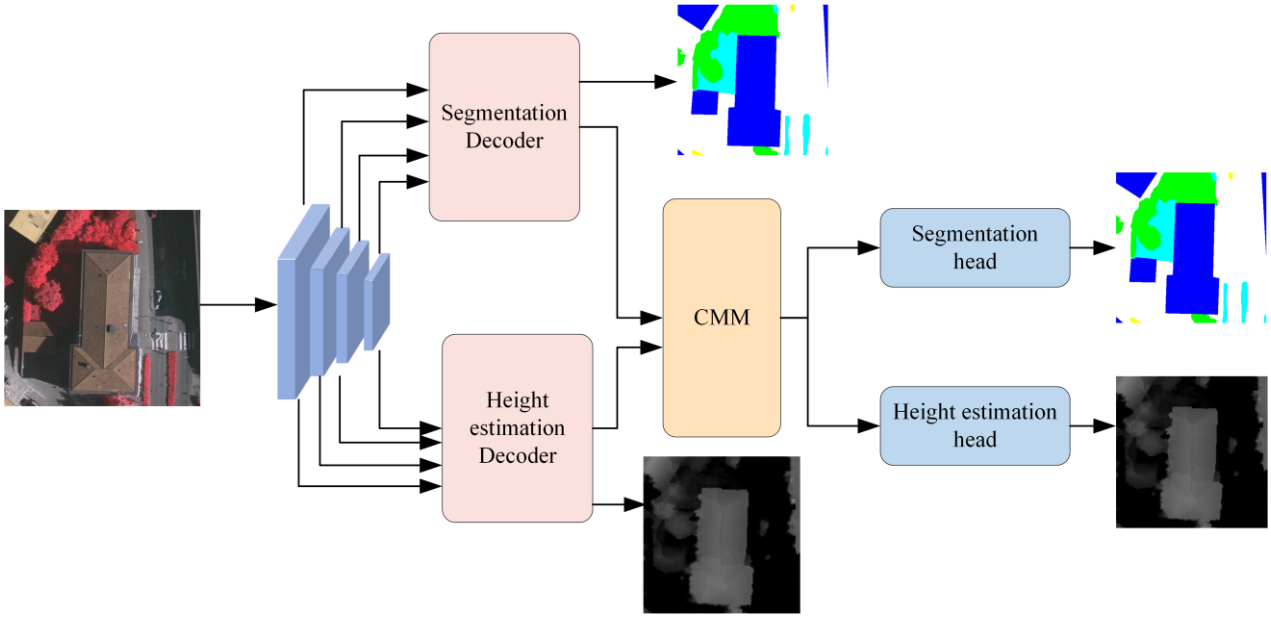

Figure 1. Pipeline of BiMamba3D.

convolution followed by batch normalization and ReLU activation, denoted as D_4. This representation is then iteratively upsampled and fused with the corresponding encoder features at each level via lateral connections:

$$\begin{aligned} D_4 &= \text{ConvBNReLU}(X_4) \\ D_3 &= \text{ConvBNReLU}(X_3) + \text{Upsample}(\text{ConvBNReLU}(D_4)) \\ D_2 &= \text{ConvBNReLU}(X_2) + \text{Upsample}(\text{ConvBNReLU}(D_3)) \\ D_1 &= \text{ConvBNReLU}(X_1) + \text{Upsample}(\text{ConvBNReLU}(D_2)) \end{aligned} \quad (1)$$

At each stage, feature maps are processed by a convolutional block before fusion, ensuring alignment in the feature domain. All upsampling operations are bilinear with a scale factor of 2, ensuring a gradual restoration of spatial resolution.

The fused high-resolution representation D_1 is refined through two additional convolutional layers:

$$F = \text{ConvBNReLU}(\text{ConvBNReLU}(D_1)), \quad (2)$$

To further enlarge the resolution to match the original input scale, the decoder applies two successive upsampling blocks, each consisting of a 3×3 convolution (with reduced channels), ReLU activation, and bilinear upsampling. Finally, a 1×1 convolution maps the feature to the desired number of output channels.

We instantiate this decoder architecture twice, forming a dual-head decoder design. Semantic segmentation branch outputs a *n*-channel per-pixel semantic label map:

$$f_{seg}, Y_{seg}^1 = \text{Decoder}_{seg}(X_1, X_2, X_3, X_4), \quad (3)$$

where *n* denotes the number of semantic segmentation classes.

Height estimation branch outputs a single-channel normalized elevation map:

$$f_{dsm}, Y_{dsm}^1 = \text{Decoder}_{dsm}(X_1, X_2, X_3, X_4), \quad (4)$$

3.2 CMM

To facilitate effective information exchange between the semantic segmentation and height estimation tasks, we propose a CMM module that performs both channel-wise and spatial-wise mutual enhancement. The CMM module leverages intermediate task-specific decoder features—denoted as $f_{seg} \in \mathbb{R}^{C\times H\times W}$ and $f_{dsm} \in \mathbb{R}^{C\times H\times W}$, and integrates them to produce a task-agnostic, enriched representation that benefits both downstream heads. These features are first projected into a unified latent space via 1×1 convolution:

$$\tilde{f}_{seg} = \text{Conv}_{1\times1}(f_{seg}), \quad \tilde{f}_{dsm} = \text{Conv}_{1\times1}(f_{dsm}), \quad (5)$$

To model cross-task dependencies at each spatial location, we treat the task features at a given spatial coordinate as a sequence across tasks. Specifically, we flatten the spatial dimensions and form a task-wise sequence at each location

$$\tilde{g}_{seg} = \text{flatten}(\tilde{f}_{seg}) \in \mathbb{R}^{N\times C}, \quad \tilde{g}_{dsm} = \text{flatten}(\tilde{f}_{dsm}) \in \mathbb{R}^{N\times C}, \quad (6)$$

$$f_{seq} = \text{Concat}(\tilde{g}_{seg}, \tilde{g}_{dsm}) \in \mathbb{R}^{N\times 2C}, \quad (7)$$

where $N=H\times W$. We then reshape this into a task sequence for each spatial location:

$$f'_{seq} = \text{Reshape}(f_{seq}) \in \mathbb{R}^{N\times 2\times C}, \quad (8)$$

Each 2-element sequence represents the task-specific features for a single spatial location. This formulation allows us to model the task interaction as a temporal sequence of length 2, with Mamba acting as the sequence encoder. The Mamba block is applied to the reshaped sequence:

$$f_{\text{encodee}} = \text{Mamba}(f'_{seq}) \in \mathbb{R}^{N\times 2\times C}, \quad (9)$$

We then reshape the encoded outputs back to task-specific feature maps:

$$f_{seg}^{out}, f_{dsm}^{out} \in \mathbb{R}^{C\times H\times W}, \quad (10)$$

These are obtained by slicing the sequence and reshaping accordingly. Finally, a residual connection is applied to preserve the original features:

$$f_{seg}^{final} = f_{seg} + f_{seg}^{out},\ f_{dsm}^{final} = f_{dsm} + f_{dsm}^{out}, \quad (11)$$

This residual fusion ensures that the learned interactions act as modulation rather than complete replacement, preserving task-specific knowledge while injecting cross-task awareness.

3.3 Task-specific Prediction Heads

After performing cross-task interaction using the CMM module, we obtain globally enhanced task-specific feature maps f_{seg}^{final} and f_{dsm}^{final} .To produce the final predictions, we employ lightweight, task-specific prediction heads that transform these enriched features into their respective output spaces.

The segmentation head is implemented as a simple yet effective convolutional classifier. It maps the feature f_{seg}^{final} to a multi-channel per-pixel class probability map. The structure consists of two convolutional layers with an intermediate non-linearity:

$$Y_{seg}^{2} = \mathrm{Conv}_{1\times1}(\mathrm{Relu}(\mathrm{Conv}_{3\times3}(f_{seg}^{final}))) \in \mathbb{R}^{B\times C_{seg}\times H\times W}, \quad (12)$$

where C_{seg} is the number of semantic classes. The 3×3 convolution captures local spatial context, while the 1×1 convolution projects the features to class logits. The head operates directly on the globally enhanced features without further upsampling or fusion, preserving the high-level consistency established by the interaction module.

The height estimation prediction head follows a similar structure, tailored for single-channel regression. Given f_{dsm}^{final}, the head produces a normalized height estimation map:

$$Y_{dsm}^{2} = \mathrm{Conv}_{1\times1}(\mathrm{Relu}(\mathrm{Conv}_{3\times3}(f_{seg}^{final}))) \in \mathbb{R}^{B\times1\times H\times W}, \quad (13)$$

3.4 Loss Function

To enhance the learning capacity of our multitask architecture, we propose a multi-stage supervision strategy that incorporates predictions from both early and late stages of the task heads. In particular, our framework produces two sets of predictions for each task: one from the initial decoder heads and another from the refined, post-interaction heads. These intermediate predictions act as auxiliary outputs that guide the learning process at different abstraction levels. To integrate all these outputs effectively during training, we design a joint loss formulation based on task uncertainty weighting, which allows the network to adaptively balance different loss components.

We adopt a hybrid loss for the height estimation prediction task and a cross-entropy loss for semantic segmentation. The height estimation loss includes pixel-wise L1 loss, gradient consistency loss, and structural similarity (SSIM) loss, combined as follows:

$$L_{dsm} = L_{l1} + \lambda_{grad}\cdot L_{grad} + \lambda_{ssim}\cdot(1-\mathrm{SSIM}), \quad (14)$$

In our experiments, we empirically set $\lambda_{grad} = 0.5$, $\lambda_{ssim} = 0.1$,following prior work.

For semantic segmentation, we use the standard pixel-wise cross-entropy loss:

$$L_{seg} = \mathrm{CrossEntropy}(Y_{seg}, Y_{seg\_gt}), \quad (15)$$

To combine losses from multiple prediction stages, we introduce learnable uncertainty weighting variables following the approach of (Cipolla et al., 2018). Each task-specific loss is weighted by an inverse-variance term that is learned during training, enabling the model to adjust the relative importance of different losses automatically. The total loss is formulated as:

$$L_{total} = \sum_{t\in\{seg,dsm\}}\sum_{i\in\{1,2\}} \frac{1}{2\exp(\log\sigma_t^{(i)})}\cdot L_t^{(i)} + \frac{1}{2}\log\sigma_t^{(i)}, \quad (16)$$

where $\sigma_t^{(i)}$ are learnable log-variance parameters corresponding to each task and stage. These values are optimized jointly with the network parameters.

4. Experiment

4.1 Implementation Details

All experiments in this paper were conducted on a DELL T5820 server running Windows 10. The hardware configuration includes an Intel(R) Core(TM) i9-10980XE CPU and a RTX 3090 GPU. The deep learning framework used is PyTorch. The number of training epochs is set to 150. The Adam optimizer is used with an initial learning rate of 0.0001, and a learning rate decay strategy is applied after 80 epochs. For parameter initialization, the feature extractor utilized weights MambaVision-L2-512-21K which pretrained on ImageNet-2K dataset.

4.2 Experimental results

In the experimental evaluation, we compare the performance of BiMamba3D against several methods on semantic segmentation and building height estimation tasks. The results are presented in the Table 1. BiMamba3D achieves the highest mIOU of 84.1, outperforming the second-best method ASSEH by 0.7. This indicates superior pixel-level classification accuracy in distinguishing building regions from background. For height estimation, BiMamba3D demonstrates the lowest RMSE of 1.03 meters and MAE of 1.02 meters, outperforming SCE-Net and ASSEH. These results highlight its capability to accurately predict building heights with finer details.

Method	mIOU	RMSE	MAE
D3-Net	-	2.08	1.26
BAMTL	-	1.76	1.07
SCE-Net	81.4	1.75	1.13
ASSEH	83.4	1.14	-
BiMamba3D	84.1	1.03	1.02

Table 1 Quantitative comparison on the Vaihingen dataset.

Figure 2 shows the height estimation results of Bimamba3D on the Vaihingen dataset, where Figure 2a) shows the height estimation results of Bimamba3D, and Figure 2b) shows the ground truth. it can be seen that Bimamba3D not only achieves more accurate height estimation results, but also has a clearer and more complete geometric structure. This indicates that

through cross-task learning, the height estimation task of Bimamba3D has learned the geometric information from the semantic segmentation task.

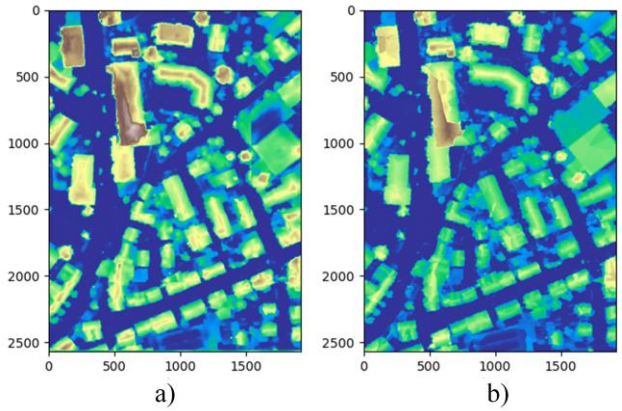

Figure 2. Height estimation on Vaihingen dataset.

Figure 3 shows the segmentation results of BiMamba3D on the Vaihingen dataset, where Figure 3a presents the model's prediction results and Figure 3b shows the ground truth. It can be observed that the segmentation results of BiMamba3D exhibit relatively regular and complete shapes.

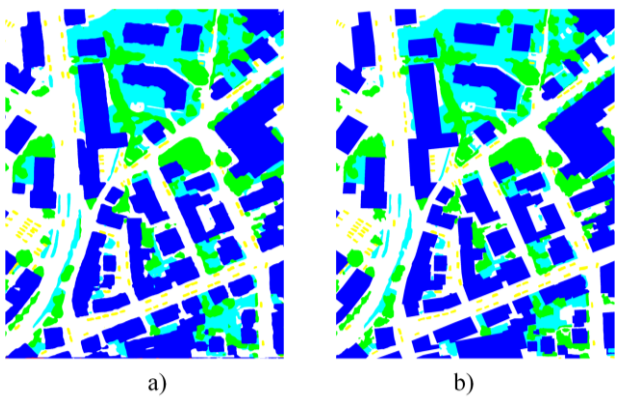

Figure 3. Segmentation on Vaihingen dataset.

4.3 Abolation Experiments

Table 2 shows that after adding the CMM module, both the mIOU and RMSE of the model have been improved. This improvement across both tasks validates the effectiveness of our CMM and the integration of CNN-Mamba architectures, which enable bidirectional knowledge transfer between segmentation and height estimation. Compared to MambaVision, which uses a single-task decoder, BiMamba3D shows significant gains in mIOU, underscoring the benefits of our multi-task learning framework.

Method	mIOU	RMSE	MAE
MambaVision	82.9	-	-
BiMamba3D (without CMM)	81.8	1.54	1.31
BiMamba3D	84.1	1.03	1.02

Table 2 Ablation experiments on the Vaihingen dataset.

Figures 4 and 5 illustrate the model results with and without the CMM module. Specifically, Figures 4(a) and 5(a) show the height estimation results and semantic segmentation results without the CMM module, while Figures 4(b) and 5(b) display the height estimation results and semantic segmentation results with the CMM module added. It can be seen that after adding the CMM module, the height estimation results have better geometric integrity and exhibit geometric consistency between different image patches. At the same time, the geometric appearance of the semantic segmentation results has also been significantly improved.

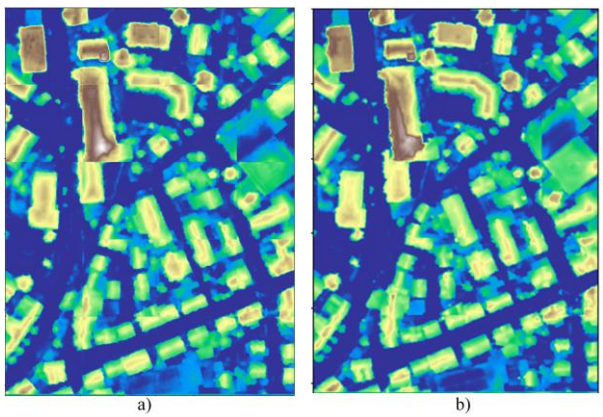

Figure 4. Height Estimation Results Before Adding CMM Module.

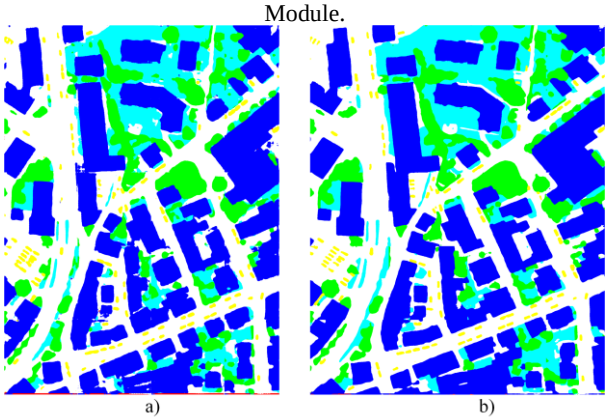

Figure 5. Segmentation Results Before Adding CMM Module.

5. Conclusion

This paper proposes a concise and efficient multi-task learning network, BiMamba3D, to address the issues of building segmentation and building height estimation in single-view remote sensing images. The network shares features between the semantic segmentation task and the height estimation task by designing a cross-task information interaction module, enabling implicit constraints between the two tasks and improving performance. In addition, the weights of the loss function are adjusted in an adaptive manner, allowing the model to jointly optimize the two tasks. Experiments on the Vaihingen dataset verify the effectiveness of the proposed method.

Acknowledgements

This work is funded by 3D Real Scene China Construction Project (A2505) and the Fundamental Research Funds of Chinese Academy of Surveying and Mapping (AR2414).

References

Amirkolaee, H. A., & Arefi, H. (2019). Height estimation from single aerial images using a deep convolutional encoder-decoder network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 149, 50–66. https://doi.org/10.1016/j.isprsjprs.2019.01.013

Batra, D., & Saxena, A. (2012). Learning the right model: Efficient max-margin learning in Laplacian CRFs. 2012 *IEEE Conference on Computer Vision and Pattern Recognition*, 2136–2143. https://doi.org/10.1109/cvpr.2012.6247920

Boo, G., Darin, E., Leasure, D. R., Dooley, C. A., Chamberlain, H. R., Lázár, A. N., Tschirhart, K., Sinai, C., Hoff, N. A., Fuller, T., Musene, K., Batumbo, A., Rimoin, A. W., & Tatem, A. J. (2022). High-resolution population estimation using household survey data and building footprints. *Nature Communications*, 13(1), 1330. https://doi.org/10.1038/s41467-022-29094-x

Carvalho, M., Le Saux, B., Trouve-Peloux, P., Champagnat, F., & Almansa, A. (2020). Multitask Learning of Height and Semantics From Aerial Images. *IEEE Geoscience and Remote Sensing Letters*, 17(8), Article 8. https://doi.org/10.1109/LGRS.2019.2947783

Chen, K., Zou, Z., & Shi, Z. (2021). Building extraction from remote sensing images with sparse token transformers. *Remote Sensing*, 13(21), 4441.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4), 834–848. https://doi.org/10.1109/tpami.2017.2699184

Chen, S., Shi, Y., Xiong, Z., & Zhu, X. X. (2023). Adaptive Bins for Monocular Height Estimation from Single Remote Sensing Images. *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, 7015–7018. https://doi.org/10.1109/IGARSS52108.2023.10281953

Chen, X., Qiu, C., Guo, W., Yu, A., Tong, X., & Schmitt, M. (2022). Multiscale Feature Learning by Transformer for Building Extraction From Satellite Images. *IEEE Geoscience and Remote Sensing Letters*, 19, 1–5. https://doi.org/10.1109/LGRS.2022.3142279

Du, S., Xing, J., Wang, S., Xiao, X., Li, J., & Liu, H. (2024). LUMNet: Land Use Knowledge Guided Multiscale Network for Height Estimation From Single Remote Sensing Images. *IEEE Geoscience and Remote Sensing Letters*, 21, 1–5. https://doi.org/10.1109/lgrs.2024.3374526

Feng, Y., Sun, X., Diao, W., Li, J., Niu, R., Gao, X., & Fu, K. (2023). Height aware understanding of remote sensing images based on cross-task interaction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 195, 233–249. https://doi.org/10.1016/j.isprsjprs.2022.11.014

Feng, Y., Sun, X., Diao, W., Li, J., Xu, T., Gao, X., & Fu, K. (2022). Soft Weighted Ordinal Classification for Monocular Height Estimation in Remote Sensing Image. *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 2750–2753. https://doi.org/10.1109/igarss46834.2022.9883187

Fu, W., Xie, K., & Fang, L. (2024). Complementarity-Aware Local–Global Feature Fusion Network for Building Extraction in Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–13. https://doi.org/10.1109/tgrs.2024.3370714

Hamaguchi, R. (2024). Multibranch Network for Addressing Intraclass Variation in Remote Sensing Building Detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17, 16710–16726. https://doi.org/10.1109/jstars.2024.3454110

Hao, L., Zhang, Y., & Cao, Z. (2019). Active Cues Collection and Integration for Building Extraction With High-Resolution Color Remote Sensing Imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(8), 2675–2694. https://doi.org/10.1109/JSTARS.2019.2926738

Hatamizadeh, A., & Kautz, J. (n.d.). MambaVision: A Hybrid Mamba-Transformer Vision Backbone.

Karatsiolis, S., Kamilaris, A., & Cole, I. (2021). IMG2nDSM: Height Estimation from Single Airborne RGB Images with Deep Learning. *Remote Sensing*, 13(12), 2417. https://doi.org/10.3390/rs13122417

Li, J., He, W., Cao, W., Zhang, L., & Zhang, H. (2024). UANet: An Uncertainty-Aware Network for Building Extraction From Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–13. https://doi.org/10.1109/TGRS.2024.3361211

Li, R., Zheng, S., Zhang, C., Duan, C., Wang, L., & Atkinson, P. M. (2021). ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of Fine-Resolution remotely sensed imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 181, 84–98. https://doi.org/10.1016/j.isprsjprs.2021.09.005

Li, X., Wang, M., & Fang, Y. (2022). Height Estimation From Single Aerial Images Using a Deep Ordinal Regression Network. *IEEE Geoscience and Remote Sensing Letters*, 19, 1–5. https://doi.org/10.1109/lgrs.2020.3019252

Liasis, G., & Stavrou, S. (2016). Satellite images analysis for shadow detection and building height estimation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 119, 437–450. https://doi.org/10.1016/j.isprsjprs.2016.07.006

Liu, C.-J., Krylov, V. A., Kane, P., Kavanagh, G., & Dahyot, R. (2020). IM2ELEVATION: Building Height Estimation from Single-View Aerial Imagery. *Remote Sensing*, 12(17), 2719. https://doi.org/10.3390/rs12172719

Liu, W., Sun, X., Zhang, W., Guo, Z., & Fu, K. (2022). Associatively Segmenting Semantics and Estimating Height From Monocular Remote-Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–17. https://doi.org/10.1109/TGRS.2022.3177796

Long, J., Shelhamer, E., & Darrell, T. (2015, June). Fully convolutional networks for semantic segmentation. 2015 IEEE Conference on Computer Vision and Pattern Recognition

(CVPR). *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA. https://doi.org/10.1109/cvpr.2015.7298965

Mao, Y., Chen, K., Zhao, L., Chen, W., Tang, D., Liu, W., Wang, Z., Diao, W., Sun, X., & Fu, K. (2023). Elevation estimation-driven building 3-D reconstruction from single-view remote sensing imagery. *IEEE Transactions on Geoscience Remote Sensing*, 61, 1–18.

Mao, Y., Sun, X., Huang, X., & Chen, K. (2023). Light: Joint Individual Building Extraction and Height Estimation from Satellite Images Through a Unified Multitask Learning Network. *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, 5320–5323. https://doi.org/10.1109/IGARSS52108.2023.10281565

Mou, L., & Zhu, X. X. (2018). IM2HEIGHT: Height Estimation from Single Monocular Imagery via Fully Residual Convolutional-Deconvolutional Network (Version 1). *arXiv*. https://doi.org/10.48550/ARXIV.1802.10249

Saxena, A., Chung, S. H., & Ng, A. Y. (2008). 3-D Depth Reconstruction from a Single Still Image. *International Journal of Computer Vision*, 76(1), 53–69. https://doi.org/10.1007/s11263-007-0071-y

Son, T. H., Weedon, Z., Yigitcanlar, T., Sanchez, T., Corchado, J. M., & Mehmood, R. (2023). Algorithmic urban planning for smart and sustainable development: Systematic review of the literature. *Sustainable Cities and Society*, 94, 104562. https://doi.org/10.1016/j.scs.2023.104562

U-Net: Convolutional Networks for Biomedical Image Segmentation. (2015). In O. Ronneberger, P. Fischer, & T. Brox, Lecture Notes in Computer Science (pp. 234–241). S*pringer International Publishing*. https://doi.org/10.1007/978-3-319-24574-4_28

Wang, Y., Tang, H., Zhou, Y., & Wang, G. (2024). HENet: A Height Estimation Network for Estimating the Height of Buildings from Single Optical Image. *2024 International Conference on New Trends in Computational Intelligence (NTCI)*, 307–311. https://doi.org/10.1109/NTCI64025.2024.10776318

Xing, S., Dong, Q., & Hu, Z. (2022a). Gated Feature Aggregation for Height Estimation From Single Aerial Images. *IEEE Geoscience and Remote Sensing Letters*, 19, 1–5. https://doi.org/10.1109/LGRS.2021.3090470

Xing, S., Dong, Q., & Hu, Z. (2022b). SCE-Net: Self- and Cross-Enhancement Network for Single-View Height Estimation and Semantic Segmentation. *Remote Sensing*, 14(9), 2252. https://doi.org/10.3390/rs14092252

Xu, H., Tang, X., Ai, B., Yang, F., Wen, Z., & Yang, X. (2022). Feature-Selection High-Resolution Network With Hypersphere Embedding for Semantic Segmentation of VHR Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–15. https://doi.org/10.1109/TGRS.2022.3183144

Zhang, H., Dou, H., Miao, Z., Zheng, N., Hao, M., & Shi, W. (2024). Extracting Building Footprint From Remote Sensing Images by an Enhanced Vision Transformer Network. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–14. https://doi.org/10.1109/TGRS.2024.3421651

Zheng, Z., Zhong, Y., Wang, J., Ma, A., & Zhang, L. (2021). Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters. *Remote Sensing of Environment*, 265, 112636. https://doi.org/10.1016/j.rse.2021.112636

Zhou, Y., Chen, Z., Wang, B., Li, S., Liu, H., Xu, D., & Ma, C. (2022). BOMSC-Net: Boundary Optimization and Multi-Scale Context Awareness Based Building Extraction From High-Resolution Remote Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–17. https://doi.org/10.1109/TGRS.2022.3152575