Zero-shot building footprint extraction and regularization based on Segment Anything model with Mesh Model

Jiachen Zhong[1], Yongjun Zhang[1, 2], Xinyi Liu[1, 2], Jinming Zhang[3], Liang Fei[4], Wang Xia[4], Bin Zhang[4], Weiwei Fan[1], Dongdong Yue[1]

[1] WHU, School of Remote Sensing and Information Engineering, 430079, Wuhan, Hubei, China – (zjc0813, zhangyj, liuxy0319, fanweiwei, yueyisui)@whu.edu.cn
[2] Technology Innovation Center for Collaborative Applications of Natural Resources Data in GBA, Ministry of Natural Resources, 510075, Guangzhou, Guangdong, China – (zhangyj, liuxy0319) @whu.edu.cn
[3] Aerospace Information Research Institute, Chinese Academy of Sciences, 100190, Beijing, China – nicnyzjm@whu.edu.cn
[4] China Railway Siyuan Survey and Design Group Co., Ltd., 430063, Wuhan, Hubei, China – feiliang3828912@126.com, 1548963280@qq.com, 007981@crfsdi.com

Keywords: Segment Anything Model (SAM), Building Footprint Extraction, Building Footprint Regularization, Triangle Meshs.

Abstract

With the advancement of urbanization, building footprint data plays an important role in urban planning, 3D Real Scene and smart cities. Traditional manual contouring methods are time-consuming and laborious, while deep learning-based building extraction methods often require a large amount of labeled data and have limited generalization ability. In this paper, a zero-shot framework based on Segment Anything Model (SAM) is proposed for extracting and regularing building footprints from 3D mesh data. The method mainly consists of three steps: 1) Coarse Prompt Generation, irrelevant element's masks such as ground and vegetation are eliminated by semi-global filtering and traditional classification method, and rough building mask is obtained as a boundary box prompt. 2) Fine mask generation: Using SAM's mask prompt capability, combined with logits map and grid elevation information with adaptive threshold to generate the fine mask prompt. Combine it with the updated bounding box to form hybrid prompt, and input SAM to generate a refined building mask. 3) Footprint regularization: Kinetic Partition, Markov random field, and Region Growth Algorithm are used to extract regularized building contours. Structural line segments from LSD guide the Kinetic Partitioning of the building. Markov random field matches building labels, while a region growth-based boundary reassignment refines the contours. The final regularized contour integrates the partitioned building zones. Our method achieved 78.31% AP50 on the Vaihingen dataset and obtained regular footprints that closely align with the true building contours on real Mesh data.

1. Introduction

Building regularized footprints extraction play a key role in practical applications of urban construction, such as urban planning, cadastral and topographic mapping, 3D Real Scene, and smart city. Different from rasterized building segmentation, building regularized footprints can express the geometric structure of buildings in a lighter and more accurate vector form, which contains geometric spatial information that can be effectively applied to data storage and relational analysis in Geographic Information System (GIS). However, due to the high accuracy requirements of building regularized footprint, the actual engineering applications at this stage are still based on manual sketching, which is expensive and time-consuming.

With the rapid development of deep learning technology, building extraction has become a hotspot of deep learning research for many years, and a large number of excellent works based on the architectures of convolutional neural network (CNN), recurrent neural network (RNN), graph neural network (GNN) and Transform have emerged. Most of the early researches were based on mask segmentation methods (Ronneberger, 2015; He, 2017), firstly obtaining instantiated masks of buildings through the network, and then regularizing and vectorizing the masks through a series of post-processing steps. Such methods are often trained with the ground-truth of pixel classification as a supervision, which makes the model ignore high-dimensional region information such as building corners and regular boundaries, thus obtaining rounded corners and irregular building contours. These errors continue to accumulate during post-processing, making it difficult to obtain correct results from the vectorization process.

In recent years, methods for directly extracting building polygons by learning geometric information such as points and edges of buildings have attracted attention. The methods for extracting polygonal buildings can be categorized into three main types: Vertex-based methods, which predict the building vertices from the image and then connect them to obtain the polygons; Contour-based methods, which predict and regress the sequence of polygonal vertices of a building directly from the contour of the building; and graphic primitives-based method, which extract the line primitives of a building and directly perform the construction of the building polygons.

Vertex-based methods usually include two-stage approaches: vertex prediction and vertex connection. The structural vertices of the building are first extracted from the image without prediction of the connectivity between the vertices. Therefore, the prediction of building polygons based on the extracted vertices is also needed. In this case, PolyWorld (Zorzi, 2022) employs a permutation matrix to represent the connectivity relationships between vertices. It utilizes a graph neural network to predict the connection strengths for each vertex pair, effectively transforming the problem of constructing building polygons into an optimization problem akin to a transportation problem. HiSup (Xu, 2023) simultaneously extracts masks, regional attraction field maps, and vertex features. It predicts connection strengths using a mask attraction strategy and constructs polygons by integrating the extracted masks and vertices according to this strategy. However, these methods heavily depend on the accuracy of vertex detection, making them prone to missed vertices and polygon construction errors, particularly in complex scenes or when obstructed by vegetation.

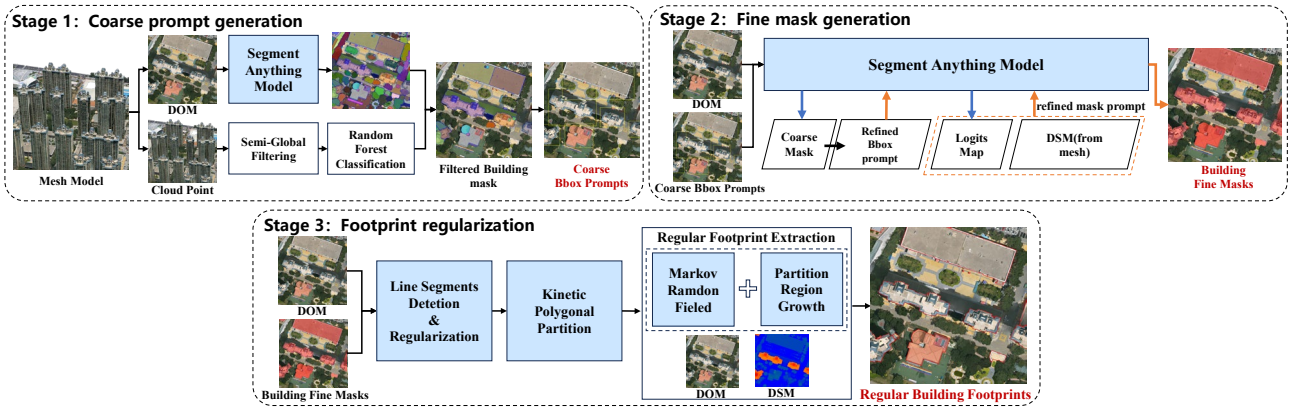


Figure 2. The proposed three stages of pipline. (a) Coarse Prompt Generation: SAM's automatic segmentation mode is used to extract filtered building masks from the orthophoto, which are then converted into coarse bounding box prompts; (b) Fine Mask Generation: The confidence feature map obtained from segmentation based on coarse prompts is overlaid with the DSM to incorporate height information, generating refined building masks; (c) Footprint Regularization: Line segments detected from the DOM and mask are used for kinetic partitioning, and the final regularized building footprints are obtained through Markov Random Field modeling and region growing.

Contour-based methods primarily obtain the building polygon by starting from the contours of the building instance mask and iteratively regressing the corner points to directly generate the vertex sequence. Early methods mainly utilized networks such as U-Net (Ronneberger, 2015) and Mask R-CNN (He, 2017) to detect building masks, followed by recurrent neural networks like ConvLSTM (Shi et al., 2015) and ConvGRU (Ballas et al., 2015) to detect building vertex sequences and perform coordinate regression. However, this approach is often performed under theassumption that the number of contour vertices is fixed, and thus can easily face the challenge of redundancy or insufficient detected building vertices. Therefore, BuilderMapper (Wei, 2023) proposes to perform classification filtering while detecting building vertices so as to eliminate redundancy. Compared to vertex-based methods, contour-based methods provide more direct access to building polygons. However, such methods often face greater training challenges (Girard, 2021) and struggle to handle buildings with holes (Wei, 2019).

Geometric primitive-based methods focus on the line segments rather than the vertices. Line2Poly (Wei, 2024) adopts a two-stage network architecture combining CNN and Transformer to propose a coarse-to-fine approach for extracting and optimizing building line segments, directly reconstructing building polygons from the discrete segments. P2Pformer (Zhang, 2024) directly extracts vertices, line segments, and corner points, predicting the connection order between these geometric primitives to generate building polygons, significantly streamlining the processing workflow.

However, the deep learning methods mentioned above usually require extensive labeled data and struggle to generalize effectively to complex scenarios, making it difficult to meet diverse practical needs. In recent years, Large Scale Model (LSM) have gained significant attention for their exceptional generalization and zero-shot learning capabilities. Among them, the Segment Anything Model (SAM) (Kirillov, 2023) demonstrates powerful zero-sample segmentation capabilities, which only needs to provide prompts such as points, boxes, and coarse-grained masks to generate fine masks.

The emergence of SAM makes it possible to transform an instance segmentation task into a target detection task. Therefore, inspired by SAM's related work (Chen, 2024), this paper proposes a novel zero-shot framework to extract regularized footprints of buildings from 3D mesh data based on Segment Anything Model. The methodological framework of this paper is divided into three main steps: Coarse prompt generation, Fine mask generation and Footprint regularization; To sum up, the contributions of this work includes three core point:

(i) a zero-shot building footprint extraction and regularization method;
(ii) a fine mask generation technique integrating image and height features;
(iii) a regularized building footprint extraction method based on Markov Random Field and Region Growing algorithms.

We tested our method on the Vaihingen dataset and mesh data from a city in China. The approach was validated using metrics such as Average Precision (AP) and Boundary F1 (denoted by B-F1) (Perazzi, 2016), demonstrating its strong capability to extract regularized building contours from mesh data with high precision and reliability.

2. Method

In this chapter, we introduce a zero-shot building footprint extraction method, which is a processing workflow based on the SAM framework, specifically designed for Mesh model data. This section covers the following aspects: a revisit of the SAM framework and a detailed introduction to the three stages of our method: coarse prompt generation, refined mask generation, and footprint regularization.

2.1 Preliminary: SAM

The SAM framework consists of three main components: an image encoder, a prompt encoder, and a mask decoder. It enables zero-shot interactive segmentation based on points, bounding boxes, and mask prompts. The image encoder in SAM is a pre-trained Masked Autoencoder (MAE) (He, 2022) based on a Vision Transformer (ViT) (Dosovitskiy, 2020), incorporating both global and local window attention mechanisms. It processes input images of size 1024×1024 and outputs image features with a resolution of 256×64×64. The prompt encoder can takes three types of prompts: points, boxes, and masks through applying

positional encoding to generate embedding tokens. The mask decoder, built on a Transformer architecture, interacts with the image features and prompt embeddings to generate the final mask.

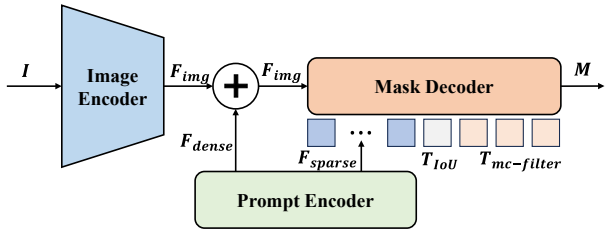
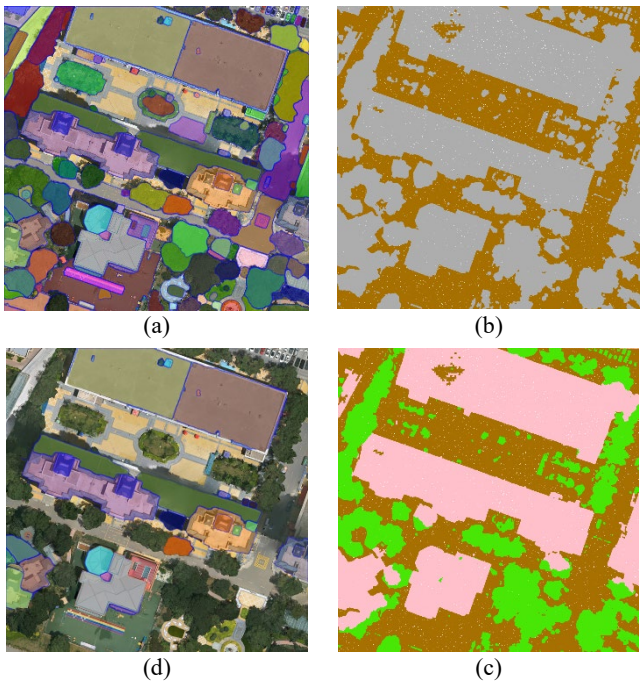


Figure 2. The architecture of SAM (Segment Anything Model)

2.2 Coarse Prompt Generation

Since SAM requires 2D images as input, we extract a Digital Orthophoto Map (DOM) from the Mesh data and use SAM's automatic segmentation mode to divide the entire image into multiple masks without semantic information. In an urban environment, these masks can be roughly categorized into ground (including small objects such as cars), vegetation, and buildings. By filtering out irrelevant masks such as ground and vegetation, we obtain coarse prompts for buildings.



(a) (b) (d) (c)

Figure 3. Effect demonstration of the mask filtering method. (a) Mask results automatically extracted by SAM; (b) Point cloud filtering results; (c) Point cloud classification results using a random forest; (d) Coarse building mask after filtering.

To distinguish ground masks, we first perform uniform sampling on the Mesh data to generate a point cloud and apply the Semi-Global Filtering (SGF) algorithm for filtering. The core idea of SGF is to create a discrete grid from the point cloud and determine the optimal height level for each grid through a semi-global optimization approach, minimizing the energy function to achieve ground point classification. For non-ground points, we use Random Forest method to further classify the data and remove irrelevant masks such as vegetation. The resulting coarse building masks are then converted into bounding boxes, which serve as prompts for subsequent refined mask generation.

2.3 Fine Mask Generation

When generating masks using point and box prompts, SAM also outputs a confidence distribution map corresponding to the mask. By feeding the confidence map back into SAM as a mask prompt, the quality of the generated mask can be effectively improved. However, ambiguous regions where building roof textures resemble the ground often lead to missing parts in the generated masks, as these regions are difficult to distinguish using only image information. To address this, we leverage SAM's ability to use confidence maps as prompts and design a method that utilizes height information to guide SAM in distinguishing ambiguous areas without requiring additional training.

First, based on the coarse bounding box prompts provided in Section 2.2, we input them into SAM to generate an initial coarse mask along with its logits map. We observe that ambiguous regions typically have positive logit values close to zero, while non-building areas have strictly negative values. Therefore, we normalize the DSM height values within the mask region to obtain a height distribution map H. For each pixel value h_ij in H, if its corresponding logit value is negative, we assign the pixel a value of $-h\_ij$; otherwise, it remains unchanged. The processed height distribution map is denoted as H', which is then multiplied by a threshold α and added to the logits map. This combined result is used as a refined mask prompt for SAM to generate a more precise building mask. The process can be expressed by the following equation:

$$(\mathcal{M}_0, \mathcal{L}_0) = \text{SAM}(\text{BBox}), \tag{1}$$

$$H_{ij} = \frac{H_{\text{raw},ij} - H_{min}}{H_{max} - H_{min}}, \forall (i,j) \in \Omega\,, \tag{2}$$

$$H'_{ij} = \begin{cases} -H_{ij}, & \text{if } \mathcal{L}_{0,ij} < 0 \\ H_{ij}, & \text{if } \mathcal{L}_{0,ij} \geq 0 \end{cases}, \tag{3}$$

$$\mathcal{L}_1 = \mathcal{L}_0 + \alpha H', \tag{4}$$

$$(\mathcal{M}_1, \mathcal{L}_1) = \text{SAM}(\mathcal{L}_1)\,, \tag{5}$$

where $\mathcal{M}_0$ = coarse mask
$\mathcal{L}_0$ = logits map of $\mathcal{M}_0$
Ω = the set of pixels in the $\mathcal{M}_0$ mask area
H_{raw} = original DSM
$H_{min} = \min_{(i,j)\in\Omega} H_{\text{raw},ij}$, $H_{max} = \max_{(i,j)\in\Omega} H_{\text{raw},ij}$
H = normalized DSM
H' = normalized DSM after process
α = the average score of logits in the $\mathcal{M}_0$ area
$\mathcal{L}_1$ = new logits map after DSM enhancement

2.4 Footprint Regularization

To address irregularities in the generated masks, such as rounded corners or aliasing boundaries, we employ Kinetic Partitioning (Bauchet, 2018), Markov Random Field (MRF) modeling, and region growing algorithms to extract regularized building footprints.

In the preprocessing stage, we use OpenCV's LSD method to detect structural line segments from the DOM image within the building mask region. These detected segments are then directionally and spatially corrected together with the mask contour lines to obtain a set of regularized line segments. We utilize the segment regularization algorithm integrated into CGAL to optimize both the direction and distance of these

Figure 4. Datasets used in the experiment. Real Mesh data from a city in China (left), Vaihingen dataset (right).

segments. Next, using these regularized line segments as input, we apply kinetic partitioning to divide the building mask region into multiple partitions. At this stage, we observe that the regularized building mask can be reconstructed by stitching together the contours of selected partitions. Therefore, obtaining the final regularized building footprint becomes a problem of selecting the partitions belonging to the building. This selection problem can be formulated as a binary label matching task, where each partition is assigned a label *l={in:1,out:0}*. This problem is modeled as an MRF optimization task, which is solved by constructing an energy function. The data term is defined as:

$$D_c = A_i \cdot \begin{cases} 1 - p_i, & \text{if } l_i^c = \text{in} \\ p_i, & \text{if } l_i^c = \text{out} \end{cases}, \tag{6}$$

where A_i = the area of sub-partition
p_i = IoU of partition and fine mask
l_i^c = sub-partition

The smoothing term is defined as:

$$V\left(l_i^c, l_j^c\right) = \text{len}_{i,j}^2 \cdot \left|1 - \left|p_i - p_j\right|\right| \cdot 1_{l_i^c \neq l_j^c}, \tag{7}$$

where $\text{len}_{i,j}$ = common side length of the two sub-partition
l_i^c, l_j^c = adjacent sub-partition

The energy function is defined as a linear combination of data terms, smoothing terms, and equilibrium coefficients λ (0.02):

$$E_c(l) = \sum_{i \in V_c} D_c(l_i^c) + \lambda \sum_{\{i,j\} \in E_c} V(l_i^c, l_j^c) \tag{8}$$

Considering the ambiguity in extracting mask boundaries, we set a strict confidence threshold β (0.95) for determining partitions as "in", only retaining those partitions fully within the building mask. Remaining partitions with non-zero confidence are considered "pending partitions." A region-growing-based boundary reassignment method is then introduced to reassign these "pending partitions." guided by grid height information: For each pending partition, its height approximation relative to neighboring "in partitions." N_p is computed using the following formula:

$$H_p = \frac{\sum_{i \in N_p} w_i H_i}{\sum_{i \in N_p} w_i}, \tag{9}$$

where H_p = height approximation
N_p = the set of neighboring "in partitions."
H_i = the height of the neighboring partition *i*
w_i = 1, weight factor

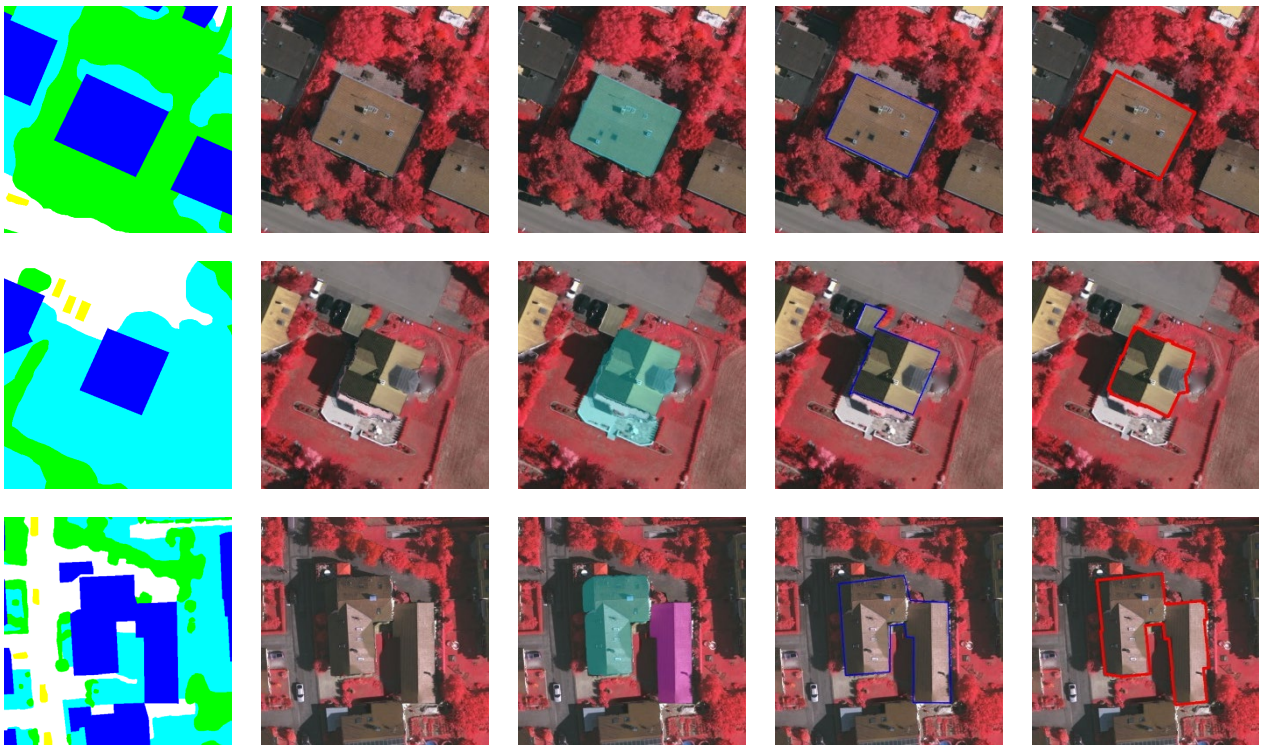
If the height difference is less than the threshold h (1.5), the partition is reassigned as "in"; otherwise, it is skipped. This process continues until no pending partitions remain. Finally, all partitions classified as "in" are merged to obtain the regularized building footprint.

3. Experiment

3.1 Dataset and Evaluation

To demonstrating the capability to extract regularized building footprints from Mesh data, we validate the proposed method using real Mesh data from a city in China. The dataset is in OSGB format and covers an area of 0.12 km². Additionally, to ensure the reliability of our method, we conduct quantitative and qualitative comparison experiments using the Vaihingen dataset. The dataset consists of 33 remote sensing images of varying sizes, each extracted from a larger top-level orthophoto image. The image selection process ensures that no areas without data are included. Both the top-level image and the DSM have a spatial resolution of 9 cm. The remote sensing images are in 8-bit TIFF format and consist of three bands: near-infrared, red, and green. for this experiment, we extract only the blue building masks with an RGB value of (0, 0, 255).

To evaluate the performance of our method on the Vaihingen dataset, we uses classical instance segmentation evaluation metrics: AP (averaged over intersection-over-union (IoU) thresholds of 0.50:0.05:0.95), AP50 (IoU threshold of 0.5), and AP75 (IoU threshold of 0.75). The IoU used for calculating APs in this study is based on masks rather than bounding boxes, with a larger AP indicating more accurate instance segmentation



(a) (b) (c) (d) (e)

Figure 5. Qualitative results on the Vaihingen dataset. (a) Ground truth; (b) Image; (c) Mask R-CNN; (d) HiSup; (e) Ours.

masks. Furthermore, to assess the degree of alignment between the predicted building boundaries and the ground truth, we use Boundary F1 (B-F1) (Perazzi, 2016) for evaluation. Compared to traditional region-level metrics, B-F1 is more sensitive and direct in evaluating boundary quality, making it more reliable for assessing fine-grained structure extraction tasks. Its calculation formula is as follows:

$$\mathrm{BF1} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}, \tag{10}$$

where Precision = predicted boundary points within tolerance
Recall = ground truth boundary points matched

3.2 Results and Comparisons

We employed two types of dataset to evaluate the proposed method. First, we conducted qualitative and quantitative comparison experiments with two learning-based methods: a pixel-based method Mask R-CNN, and a contour-based method HiSup. Second, we tested the zero-shot extraction capability of our method for regular footprints using real Mesh data.

Method	AP	AP50	AP75	Boundary F1
Mask RCNN	56.54	77.13	64.96	41.92
HiSup	60.97	**78.45**	68.21	**81.30**
our	**61.78**	78.31	**68.82**	62.73

Table 1. Result on Vaihingen Dataset

As shown in Table 1, compared to Mask R-CNN, the proposed method demonstrates a significant improvement in instance segmentation accuracy across AP, AP50, AP75, and boundary accuracy B-F1. Specifically, AP increased by 5.24%, AP50 by 1.18%, AP75 by 3.86%, and B-F1 by 20.81%. In comparison with HiSup, our method achieves an improvement of 0.81% in AP and 0.61% in AP75, though it performs slightly lower in AP50. Analyzing our building detection approach, this may be attributed to missed or misclassified small objects during the coarse prompt generation stage, leading to a relatively lower number of correctly segmented instances with IoU > 50%. Regarding boundary prediction, HiSup outperforms our method in B-F1 by 18.73%. This discrepancy is primarily due to the rule-based contour approach used in our method, which may produce more small protrusions when handling complex contours or textures, thereby affecting the B-F1 metric, which is based on the count of correctly predicted contour points.

Figure 5 compares the visualization results of Mask R-CNN, HiSup, and our method. The pixel-based method Mask R-CNN, can correctly detect buildings, but its mask results are irregular and do not fully cover the building areas. HiSup effectively extracts building polygons, and the predicted polygon outlines align well with the actual building boundaries. However, while HiSup generates polygons with relatively uniform vertices, it does not perfectly align with the structural points of buildings. Compared to Mask R-CNN, the proposed method better approximates building boundaries and accurately locates building corners. Additionally, for buildings with complex structures, our method outperforms HiSup. However, for buildings with regular structures, minor errors may occur, affecting the final accuracy.

Figure 6 presents the visualization results of HiSup and our method on real Mesh data. To HiSup, it can't extract building polygon correctly without training. To our method, most buildings are correctly segmented, and the extracted regularized footprints align well with the building structures, effectively demonstrates it's good generalization ability. However, minor errors may occasionally occur at corners. In terms of building

(a) Segment Anything results (b) HiSup (Without training) (c) Our mask results (d) Regular footprint results

Figure 6. Testing results on real Mesh data. Obviously, HiSup fails to produce correct results on untrained data, while our method effectively extracts refined building masks from SAM’s cluttered mask set and obtains regularized footprints that align with real building contours.

detection, some areas with dense surrounding vegetation or similar colors are easy to misclassification or omission.

4. Conclusion

In this paper, we propose a zero-shot framework based on the Segment Anything Model (SAM) for extracting and regularizing building footprints from 3D mesh data. Experimental results onthe Vaihingen dataset and real Mesh data validate the effectiveness of the proposed method, achieving performance comparable to classic learning-based approaches such as HiSup. However, our method still has certain limitations. Since each processing step operates independently, the overall processing speed is relatively affected. In future work, we aim to transform this workflow into a weakly supervised or self-supervised learning-based approach to enhance both processing speed and quality.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 42201474), the State Key Laboratory of Micro-Spacecraft Rapid Design and Intelligent Cluster (Grant No. MS01240125), and Hubei Provincial Natural Science Foundation of China(2024AFD8)

References

Ballas, N., Yao, L., Pal, C., & Courville, A., 2015. Delving deeper into convolutional networks for learning video representations. *arXiv preprint arXiv:1511.06432*. doi.org/10.48550/arXiv.1511.06432.

Bauchet J P., Lafarge F., 2018: Kippi: Kinetic polygonal partitioning of images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3146-3154. doi.org/10.1109/cvpr.2018.00332.

Chen, K., Liu, C., Chen, H., Zhang, H., Li, W., Zou, Z., & Shi, Z., 2024: RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, vol.62. doi.org/10.1109/tgrs.2024.3356074.

Dosovitskiy, A., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. doi.org/10.48550/arXiv.2010.11929.

Girard, N., Smirnov, D., Solomon, J., & Tarabalka, Y., 2021. Polygonal building extraction by frame field learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5891-5900. doi.org/10.1109/cvpr46437.2021.00583.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R., 2022. Masked autoencoders are scalable vision learners. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 16000-16009. doi.org/10.1109/cvpr52688.2022.01553.

He, K., Gkioxari, G., Dollár, P., & Girshick, R., 2017. Mask r-cnn. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2961-2969. doi.org/10.1109/ICCV.2017.322.

Kirillov A., Mintun E., Ravi N., et al., 2023: Segment anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4015-4026. doi.org/10.1109/ICCV51070.2023.00371.

Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., & Sorkine-Hornung, A., 2016. A benchmark dataset and evaluation methodology for video object segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 724-732. doi.org/10.1109/cvpr.2016.85.

Ronneberger O., Fischer P., Brox T., 2015: U-net: CoGnvolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention MICCAI International Conference*, 234-241. doi.org/10.1007/978-3-319-24574-4_28.

Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C., 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28.

Wei, S., Ji, S., & Lu, M., 2019. Toward automatic building footprint delineation from aerial images using CNN and regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 58 (3), 2178-2189. doi.org/10.1109/TGRS.2019.2954461.

Wei, S., Zhang, T., Ji, S., Luo, M., & Gong, J., 2023. BuildMapper: A fully learnable framework for vectorized building contour extraction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 197: 87-104. doi.org/10.1016/j.isprsjprs.2023.01.015.

Wei, S., Zhang, T., Yu, D., Ji, S., Zhang, Y., & Gong, J., 2024. From lines to Polygons: Polygonal building contour extraction from High-Resolution remote sensing imagery. *ISPRS Journal of*

Photogrammetry and Remote Sensing, 209: 213-232. doi.org/10.1016/j.isprsjprs.2024.02.001.

Zhang, T., Wei, S., Zhou, Y., Luo, M., Yu, W., & Ji, S., 2024. P2PFormer: A Primitive-to-polygon Method for Regular Building Contour Extraction from Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 62. doi.org/10.1109/TGRS.2024.3459011.

Zorzi, S., Bazrafkan, S., Habenschuss, S., & Fraundorfer, F., 2022: Polyworld: Polygonal building extraction with graph neural networks in satellite images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1848-1857. doi.org/10.1109/cvpr52688.2022.00189.