SingleRecon: Reconstructing Building 3D models of LoD1 from A Single Off-Nadir Remote Sensing Image

Ruizhe Shao, JiangJiang Wu, Jun Li, Shuang Peng, Hao Chen*

Abstract

3D building models are one of the most intuitive and widely used forms for understanding urban buildings. Generating 3D building models based on a single off-nadir satellite image is an economical and rapid method, particularly valuable in large-scale 3D reconstruction scenarios with limited time. In this paper, we propose a novel pipeline for automatically reconstructing LoD1 3D building models based on a single off-nadir satellite remote sensing image. Our pipeline is built upon a multi-task neural network called ONBuildingNet (Off-Nadir Building Reconstruction Network), which extracts building roof polygons and offsets from the image. Using this information, the pipeline computes the building footprint polygons and heights, constructs LoD1 building models, and then extract textures from the off-nadir image. ONBuildingNet introduces our proposed cross-field auxiliary task and multi-scale mask head to extract building roof polygons with accurate shapes. We have demonstrated through extensive experiments that our pipeline can automatically and rapidly construct LoD1 3D urban building models. Additionally, our proposed ONBuildingNet outperforms current state-of-the-art methods in extracting more shape accurate building roof polygons, thereby enhancing the accuracy of the final 3D models produced by our pipeline. Please visit our project page at https://shaoruizhe.github.io/building-stand-up.github.io/ to gain a more intuitive understanding of our method.

Keywords: 3D reconstruction, LoD1 3D building model, off-nadir satellite image, building roof polygon extraction

1 Introduction

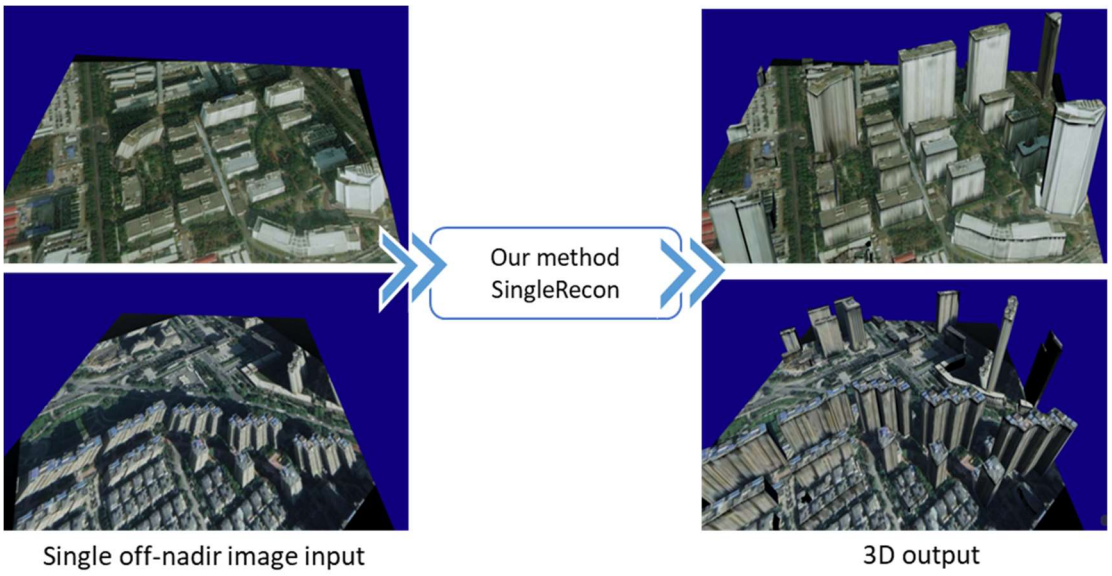

Figure 1 We propose a method that reconstructs 3d buildings form a single off-nadir remote sensing image input

Urban 3D building models are a data representation that can visually present and analysis various information about the shape, structure, and texture of urban buildings. They serve as a fundamental and widely used data format for various applications such as urban planning [1], urban spatial analysis [2, 3], disaster response [4], skyline analysis [5], etc.

There are several methods for reconstructing 3D urban buildings, including LiDAR -based methods [6, 7], multi-view imagery methods [8-11], and monocular image-based methods [12-16]. The first two classes of methods can achieve more detailed 3D reconstruction results, especially when combined with Unmanned Aerial Vehicles (UAVs) [11, 17]. However, monocular image-based methods also offer the advantage of being more economical and effective. Compared to the previous two methods, monocular image-based methods have lower data acquisition cost and shorter the production process. This approach is particularly suitable for applications where strict constraints exist on time and computational costs. As a result, monocular image-based methods have garnered significant attention in the academic community.

In current research, there are various forms of 3D building models, among which Level of Detail 1 (LoD1) [18] 3D buildings, as defined by SIG3D, have gained widespread application due to their high information efficiency and excellent visualization effects. Benefiting from the prior knowledge that introduces the 3D structure of buildings as prisms, LoD1 3D building models strike a balance between information efficiency and detail richness in the 3D model. LoD1 models require only building footprint polygons and building height information to construct the building model. This characteristic makes LoD1 models well-suited for monocular remote sensing image-based 3D reconstruction tasks, given the limited information available in monocular images.

To construct LoD1 3D building models, the most crucial step is to extract building footprint polygons and building heights from satellite remote sensing images. Currently, research on

methods for extracting building footprints and estimating heights includes approaches for nadir images and off-nadir images. The distinction between nadir and off-nadir images lies in the off-nadir angle at which the satellite captures the imagery. Nadir images are captured directly below the satellite, with an off-nadir angle (ONA) typically ranging from 0 to 25 degrees. In contrast, off-nadir images refer to those captured at ONA between 25 and 40 degrees [19], as shown on the left side of Figure 1. In these two types of images, the parallax effect in off-nadir images provides clues for estimating building heights, whereas nadir images are challenging for height determination. Consequently, some researchers utilize single off-nadir images for 3D reconstruction of buildings [20, 21]. In these studies, scholars focus on the offset between the building roof and the footprint caused by parallax, overcoming the challenge of invisible building footprints in off-nadir images. However, these methods still have two main limitations: First, the extracted building polygon shapes are not accurate enough, resulting in suboptimal performance when applied to 3D building reconstruction. Second, these methods only extract building footprints and heights, without further utilizing the building texture information from the image to create more visually appealing textured 3D building models.

To address these challenges, we propose a neural network that extracts accurate-shaped building roofs and building offsets from off-nadir satellite images. Additionally, we design a pipeline to utilize this information for reconstructing LoD1 3D building models with textures. In general, our main contributions are summarized as follows:

1. We propose a pipeline for textured LoD1 3D building reconstruction from a single off-nadir remote sensing image. Given an off-nadir image, this pipeline automatically constructs intuitive 3D models of urban buildings with texture information.
2. We propose a novel multi-task learning neural network called ONBuildingNet, which simultaneously extracts building roofs and building offsets. ONBuildingNet includes a novel multiscale mask head that customizes the extracted mask size based on different target sizes to enhance the accuracy of multiscale roof extraction.
3. We propose a cross-field auxiliary task that can be incorporated into the aforementioned ONBuildingNet multitask network. This auxiliary task predicts the orientation of the edges of the building roof polygon and is mutually constrained with the prediction of the building roof mask, enhancing the accuracy of polygon extraction.
4. We conduct extensive experiments to demonstrate the effectiveness of the proposed 3D building reconstruction pipeline, as well as the superior performance of the proposed ONBuildingNet compared to state-of-the-art (SOTA) methods.

2 Related work

2.1 Building Polygon Extraction from Remote Sensing Images

In this paper, extracting accuracy building polygons is a crucial step in constructing LoD1 3D building models. Therefore, building polygon extraction from remote sensing images is one of the key focuses of our research. This section provides a brief introduction to the research background of building polygon extraction methods.

The extraction methods for building polygons from remote sensing images have consistently been a topic of interest among researchers. Satellite remote sensing data has garnered widespread attention in academia due to its efficient acquisition and rich information content [22-28]. One research field closely related to building polygon extraction is the extraction of building masks. However, the task of extracting building polygons cannot be replaced by building mask extraction [20]. This discrepancy arises because mask-based methods primarily focus on pixel-level accuracy, neglecting the precision of polygon shapes, which includes the precision of vertex positions and edges. To achieve more accurate building polygons, scholars have proposed various effective methods. Frame field Learning[29] introduces a field called the Frame field, which indicates building edge directions in the image, and effectively extracts more accurate building shapes. [30] utilizes a signed distance function for building edges as an auxiliary task in neural networks to optimize building edge extraction. Li[31] extracts building corners and edges, using this information to enhance building polygon results. In PolyWorld [32], building corners are first detected, and then a graph neural network is constructed using these detected corners to form the building polygons. HEAT [33] encodes corners and edges as embeddings and designs a transformer to extract accurate vertex and edge positions. HiSup [34] proposes the Attraction Field Map (AFM) as an auxiliary task to assist the network in extracting more accurate building polygons.

For off-nadir imagery, the roofs of buildings in remote sensing images do not align perfectly with their footprints [20, 35]. This misalignment poses challenges for building polygon extraction. LOFT [20] is a multitask network that simultaneously extracts roofs from off-nadir images and predicts the offsets between the roofs and footprints. And then LOFT translates the roofs based on this distance to obtain footprint predictions. Compared to directly predicting footprints, this approach achieves better results. MTBR-Net [21, 36] creatively introduces the use of an HR-Net to extract building polygons from off-nadir remote sensing images. Inspired by these works, we also extract building offsets while extracting building roofs to obtain accurate building polygons.

2.2 Building 3D Reconstruction

Due to the impressive performance of 3D models in urban environments, methods for constructing 3D building models have been a focal point of academic research. Currently, based on input data, several mainstream approaches exist for 3D building construction. These include methods based on LiDAR data [37], multi-view stereo [8-10], oblique photogrammetry [11], and monocular image-based [12-16] techniques. The first three methods have mature technologies and can generate highly accurate 3D models. For instance, widely used oblique photogrammetry techniques can produce centimeter-level precision 3D mesh products [11] from high-resolution UAV images. However, the data acquisition for these methods is more complex and difficult [17, 38] compared to monocular image-based approaches. Additionally, methods based on oblique photogrammetry incur significant computational overhead, longer processing times, and complex reconstruction workflows as the cost of its high-precision[17]. On the other hand, the final category of monocular image-based 3D generation methods offers advantages such as easy data acquisition, fast 3D model production[14], and strong timeliness. However, these methods also come with challenges, including the need for substantial prior knowledge [39] and high technical complexity[12].

Monocular image-based 3D construction methods primarily follow these technical approaches: Firstly, depth estimation-based methods [30, 40, 41] estimate the height value for each pixel in the input image to obtain a 3D digital surface model (DSM) for the target area. Secondly, shadow information-based methods [8, 16, 42] extract shadows from input images and then calculating building heights using auxiliary information such as solar elevation angles to construct 3D representations. Thirdly, in off-nadir building-based methods [15, 21, 36], buildings are first extracted using object detection, and then the height of each building is calculated based on the parallax effect of buildings in remote sensing images. This approach results in 3D building models. Lastly, 3D generation-based methods. Recently, with the development of Neural Radiance Fields (NeRF) [43, 44] and 3D gaussian splatting[45], a new category of methods has emerged that directly generates 3D building models from images [46, 47]. However, these methods still face challenges related to high computational resource consumption [48]. Additionally, current 3D generative methods are not yet capable of rapidly producing large-scale 3D models [46, 47, 49].

3 Pipeline

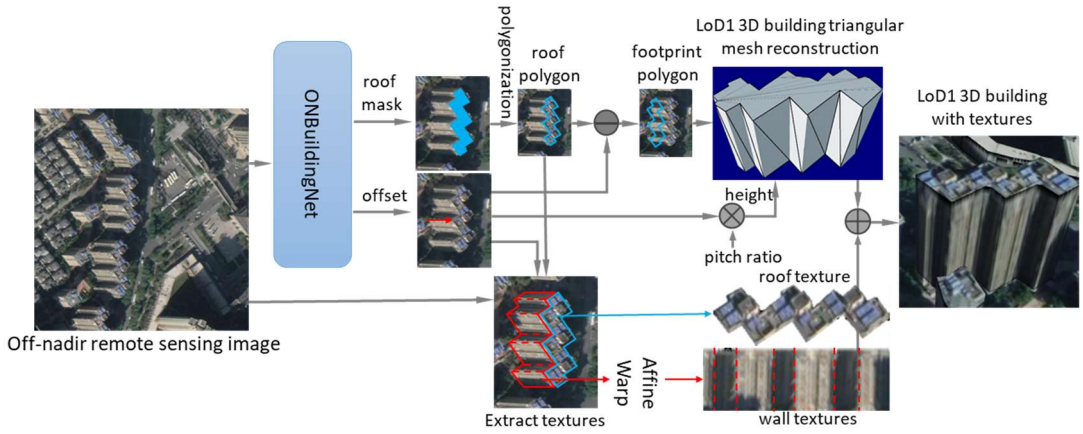

Figure 2 The pipeline of reconstructing 3D buildings from single off-nadir remote sensing image

In off-nadir remote sensing images, the footprint of buildings is often not visible, while the roof is generally more distinct. Based on this observation, instead of directly extracting the building's footprint, our pipeline extracts the building's roof and determines the offset of the roof relative to the footprint. Subsequently, it calculates the footprint polygon. After that, we use this building information to construct the LoD1 3D model.

Our method's pipeline is illustrated in

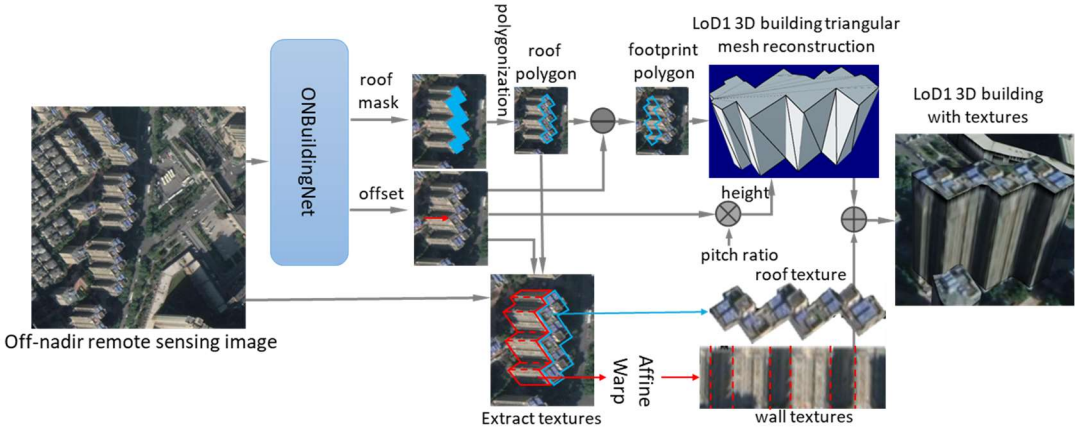

Figure 2. It consists of three main steps. First, we input an off-nadir remote sensing image and utilize the proposed multitask network ONBuildingNet to extract building roofs and the offset from each roof to its footprint. ONBuildingNet will be further detailed in Chapter 4. Second, we polygonise the extracted roof results to obtain the building polygons. Leveraging the offset and pitch angle information of the image, we calculate the height of each building. Finally, we construct LoD1 3D triangular mesh surface models for the buildings using their footprints and heights. To complete the model, we crop the roof and facade walls in the remote sensing image and perform an affine transformation to extract the roof and side textures for the 3D building. These textures are subsequently added to the triangular mesh. The resulting mesh model is stored in 3D Tile format. We provide a pseudocode in Algorithm 1 to show our approach more explicitly.

Algorithm 1: 3d building reconstruction for a single off-nadir remote sensing image

Input: remote sensing image I, direction angle θ, pitch angle p

$I' \leftarrow$Rotate$(I, -\theta)$ # get a rotated image I' of which direction angle is 0

$scores_{obj}, masks_{roof}, offsets \leftarrow$F.forward$(I')$# where F is the multitask neural network

$models_{building} \leftarrow$[ ]

for $score_{obj}, mask_{roof}, offset$ in zip$(scores_{obj}, masks_{roof}, offsets)$:

    # For each detected building:

    if $score_{obj}$>0.5:

        $polygon_{roof} \leftarrow$PolygonSimplify(FindContour$(mask_{roof})$)

        $polygon_{footprint} \leftarrow polygon_{roof} - offset$

        $height \leftarrow offset * tan(p)$

        $texture_{roof} \leftarrow I'$.PolygonCrop$(polygon_{roof})$

        $polygons_{FacadeWall} \leftarrow$GetFacade$(polygon_{roof}, offsets)$ # Facade walls are the walls that is visible on the off-nadir image

        $textures_{wall} \leftarrow$WarpAffine$(I'$.PolygonCrop$(polygons_{FacadeWall}))$ # extract facade texture form the image

        $models_{building}$.append(BuildModel$(polygon_{footprint}, height,$

$[texture_{roof},\ taxtures_{wall}]))$

Output: $models_{building}$

4 Off-Nadir Building Reconstruction Neural Network (ONBuildingNet)

4.1 Overview

To provide accurate results for the aforementioned 3D construction process, including building roof polygons and offset extraction, we designed a multi-task neural network, named ONBuildingNet, to extract precise roof and offset information. ONBuildingNet is designed based on the two following considerations: First, by sharing backbone weights across multiple tasks related to buildings, we encourage the backbone to extract generalized building features, thereby enhancing the performance of each task. Second, we introduce a cross-field auxiliary task that incorporates prior knowledge about building shapes as a new constraint within the network.

The architecture of ONBuildingNet is illustrated in Figure 3. Building upon the foundation of a general instance segmentation framework, our network incorporates an offset head and a cross-field head, and also replaces the standard FCN mask head with the proposed multi-scale mask head. This architecture is inspired by LOFT [20] and Mask R-CNN[50]. The cross-field prediction task (section 4.2) serves as an auxiliary task, which can make the shape of the extracted building polygons more accurate by mutually constraining the mask prediction task with the newly proposed align loss. With the multi-scale mask head (section 4.3), which enhances the precision of mask extraction across different scales, our method further improves the accuracy of building extraction.

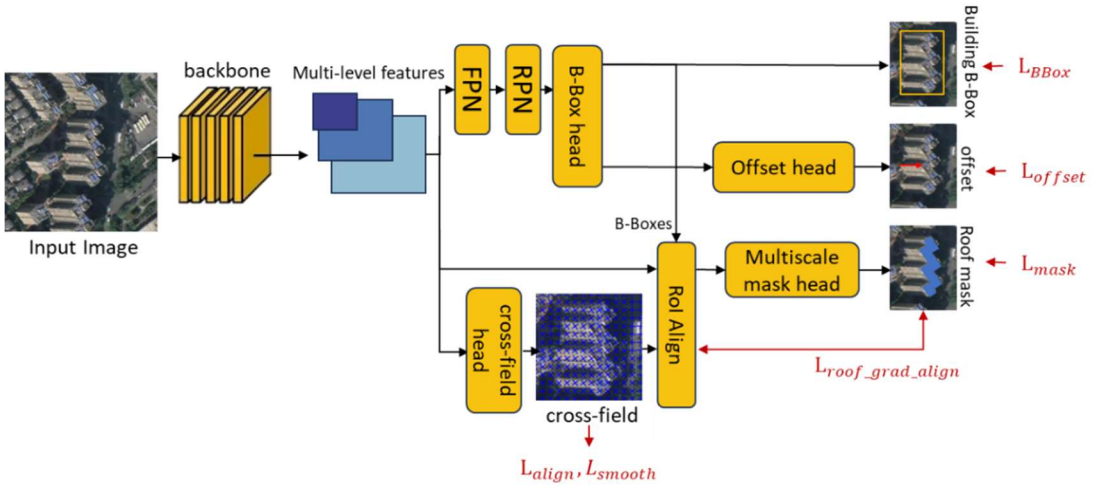

Figure 3 An overview of the proposed ONBuildingNet

4.2 Cross-field

To generate accurate 3D buildings, it is essential to extract roof polygons with precise shapes. However, due to the neural network's inherent tendency to emphasize low-frequency information while neglecting high-frequency details, the mask predictions produced by conventional networks trained with cross entropy mask loss tend to form clusters, losing the

sharp corners of building structures. After polygonization, the accuracy of vertex positions and edge directions is insufficient, resulting in inadequate shape accuracy.

To address the abovementioned issue, inspired by [29], we propose cross-field auxiliary task. A cross-field assigns a "cross" to each pixel, with each cross containing four mutually orthogonal directions, as illustrated in Figure 4.

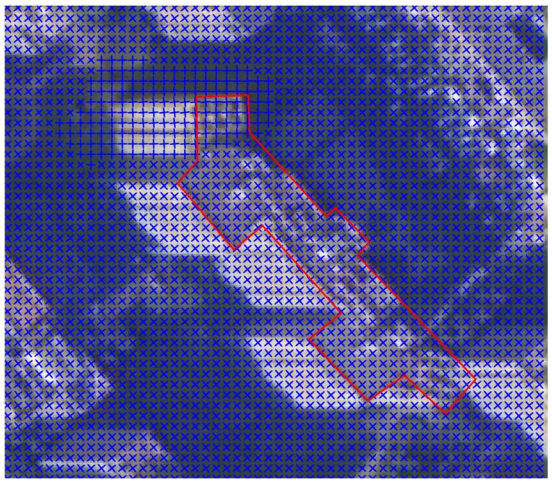

Figure 4 Illustration of cross-field on a building's image

4.2.1 Cross-field encoding and 2-level deviation function

To train the neural network for cross-field prediction, we need to encode the cross. Directly predicting the angles of the cross faces two drawbacks: First, angles do not have a one-to-one correspondence with crosses. Crosses with an angle θ are equivalent to those with an angle $\theta + \frac{\pi}{2}$. Second, nearby crosses may correspond to significantly different angles. For instance, the difference between angles θ and $\theta + \frac{\pi}{2} + \tau$ (where τ is a small value) is substantial, even though the corresponding crosses are similar. Therefore, we encode the crosses to satisfy both the one-to-one correspondence principle and the "closer cross, closer encoding" principle. We introduce complex functions and design a cross encoding scheme:

$$c^4 = e^{i\hat{\theta}\times 4} \tag{1}$$

where $\hat{\theta}$ represents the predicted angle. When computing the loss, we calculate the discrepancy between the encoding of true angle and the predicted angle encoding. This serves as the deviation function for the directions of the two crosses

$$f_1(\theta, c^4) = \left|e^{i\theta\times 4} - c^4\right| \tag{2}$$

The relationship between the angle difference and the deviation function is illustrated in Figure 5(a).

The function f_1 supervises neural networks for cross-field prediction training. However, when the angle differences are small, the values and gradients of f_1 are also small, which may lead to less accurate predictions. As an auxiliary task, imprecise cross-field predictions can cause incorrect edge directions in roof masks, especially for longer edges where the error is more pronounced. To address this issue, we propose a second-level encoding and discrepancy function:

$$c^8 = e^{i\hat{\theta}\times 8}, \qquad f_2(\theta, c^8) = \left|e^{i\theta\times 8} - c^8\right| \tag{3}$$

By encoding with higher orders of $e^{i\theta}$, the second-level discrepancy function obtains higher gradients (as shown in Figure 5(b)), thereby improving angle prediction accuracy.

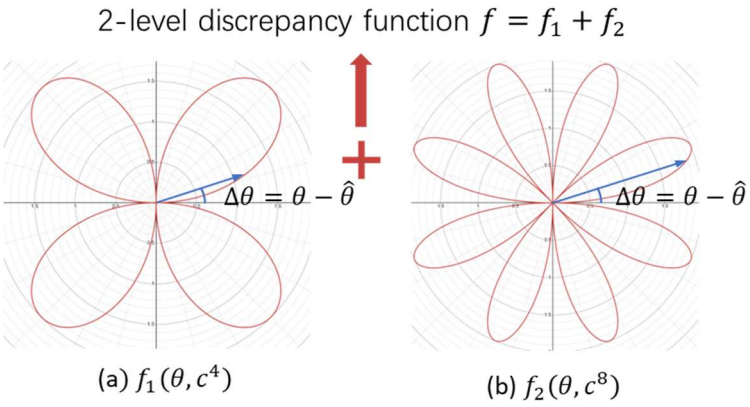

Figure 5 level 1 and level 2 discrepancy function vs. angle error $\Delta\theta$

We sum the discrepancy functions from both levels to obtain the joint discrepancy function for the 2-level approach:

$$f(\theta, c^4, c^8) = f_1(\theta, c^4) + f_2(\theta, c^8) = \left|e^{i\theta\times4} - c^4\right| + \left|e^{i\theta\times8} - c^8\right| \quad (4)$$

where f_1 and f_2 represent the angle deviation functions for the first and second levels of encoding, respectively. The first term maintains overall accuracy in predictions. Meanwhile, the last term ensures that even for small angle deviations, the loss function exhibits a significant gradient, thereby driving the two crosses toward precise alignment.

4.2.2 Cross-field auxiliary task in ONBuildingNet

We introduce a cross-field head into the multi-task network to extract cross-field information from the image, as shown in Figure 3. The cross-field head has a UNet decoder-like structure that combines semantic features from multiple levels of the backbone for predicting the cross-field.

To leverage the cross-field task to assist in optimizing the extracted roof shapes, we set the relation between the direction of the crosses and the direction of the roof edges as follows: At the edges of the roof, edge's two extension directions align with two opposing directions of the crosses, while the normal directions align with the other two opposing directions. Additionally, at the corners of the roof, considering that most roof corners are right angles, the two edges connecting the roof corner each align with two orthogonal directions of the crosses. An intuitive description of the cross-field on a building is shown in Figure 4.

Based on the above settings, we introduce three losses to train the model to predict the cross-field and optimize the mask shape with the cross-field. The first loss is L_{algin}, which ensures the consistency between predicted cross-fields and ground truth cross-fields:

$$L_{align} = f(dir(\nabla y), c^4, c^8) \cdot mask_{edge} \quad (5)$$

where y represents roof mask label, ∇y denotes the normal direction of the roof mask edge, $dir(\nabla y)$ is the direction angle of ∇y, and $mask_{edge}$ identifies the edge regions of the roof mask.

The second loss is L_{roof_align} which constrain the model's output cross-field and mask to satisfy above settings.

$$L_{roof\_grad\_align} = f(dir(\nabla \hat{y}), c^4, c^8) \quad (6)$$

where $\hat{y}$ is the predicted mask.

Furthermore, to ensure that the edges remain straight, we introduce a smooth loss L_{smoot} to the cross-field. This loss indirectly controls the direction of mask edge towards a fixed direction by constraining the spatial variations of the cross-field.

$$L_{smooth} = avg(\left|\left|\nabla c^4\right|\right| + \left|\left|\nabla c^8\right|\right|) \quad (7)$$

where $\left|\left|\nabla c^4\right|\right|$和$\left|\left|\nabla c^8\right|\right|$ are the gradient magnitude of c^4 and c^8, which can be approximately obtained by convolution with the Laplacian of Gaussian (LoG) operator in practical implementation.

4.3 Multiscale Mask Head

In Mask R-CNN, the design of the FCN mask head is mature and exhibits relatively stable performance[50]. However, in the context of building reconstruction in our study, the accuracy of mask prediction falls short of meeting the reconstruction requirements. Additionally, the FCN mask head predicts a fixed-size mask for all instances, which is then interpolated to match their actual sizes. However, due to significant variations in building sizes, the uniform-size mask predictions from the mask head result in poor detail fineness for large building roof predictions. Consequently, we propose the Multiscale mask Head as a replacement for the standard FCN mask head.

The architecture of our Multiscale mask Head is illustrated in Figure 6: For different scales of targets, we extract features from multiple levels of image features obtained from the backbone. Subsequently, these multi-level features are processed through a Multiscale mask Head, which is designed with a structure similar to a UNet decoder [51]. The output is mask predictions that align with the scale of the target.

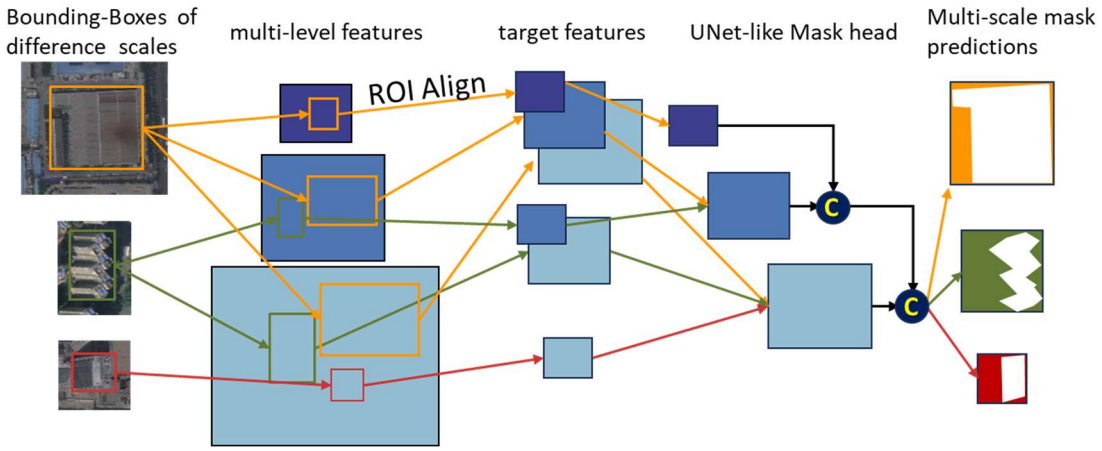

Figure 6 The proposed Multiscale mask Head. The first column represents proposal bounding-boxes obtained from the bounding-box head predictions, indicating the targets to be masked. The second column depicts multi-level features sourced from the backbone's multiple levels. The outputs have multi-scale that align with the scale of the target.

The benefits of designing the mask head in this manner are twofold: Firstly, it enables different-sized mask predictions for various target sizes. While ensuring that a significant number of small targets do not excessively consume memory, it also enhances the mask prediction resolution for larger targets. Secondly, by incorporating a UNet-like structure that fuses features from multiple levels, the predicting of mask output merges deep semantic information from deeper features with high-frequency information from shallower features.

Finally, the ONBuildingNet is optimized via minimizing the following joint loss function:

$$L = L_{rpn} + L_{BBox} + \lambda_1 L_{align} + \lambda_2 L_{smooth} + \lambda_3 L_{mask} + \lambda_4 L_{roof\_grad\_align} + L_{offset} \quad (8)$$

where L_{rpn}, L_{BBox}, and L_{mask} are the cross-entropy losses for RPN proposals, bounding-box prediction and mask prediction, respectively, which are the same as those in Mask R-CNN. L_{offset} is the loss for the offset head, where a standard smooth L1 Loss is used. L_{align}, $L_{roof\_grad\_align}$, and L_{smoo} are calculated according to formular (5), (6), and (7), respectively.

5 Experiments

5.1 Dataset

The BONAI dataset comprises 3,300 off-nadir remote sensing images, each with a size of 1024x1024 pixels. These images are distributed across six representative cities in China, including Shanghai, Beijing, Harbin, Jinan, Chengdu, and Xi'an. The dataset is split into a training set of 3,000 images and a test set of 300 images. In total, 268,958 building instances are annotated on, with each building instance labeled for both its roof and offset.

5.2 Implementation Details

We conducted training on a server equipped with an NVIDIA GeForce RTX 4090 GPU. The batch size was set to 2. We trained for 48 epochs, with the learning rate starting at 0.001 and decaying by a factor of 0.1 at the 32nd and 44th epochs. During testing in the 3D reconstruction pipeline, we used a confidence threshold of 0.5 for building instances, considering only those above this threshold.

Firstly, we evaluated the feature extraction performance of our ONBuildingNet. We used the following four evaluation metrics to assess network performance: roof AP_{50}[50], offset Mean Absolute Error (MAE), and Mean Tangent Angle Error (MTAE)[29]. Among them, MTAE is used to compare the shape of the extracted polygons.

$$MTAE = mean_{j \in V} \Delta\theta_j \quad (9)$$

where V is the collection of all the sample points on the extracted polygons. While $\Delta\theta_j$ is calculated as follows:

$$\Delta\theta_j = \cos^{-1}\left(\langle P_j, G_j \rangle\right) \quad (10)$$

where P_j is the edge dirction vector of the extracted polygons at j. And G_j is the edge dirction vector of the ground truth polygons at the point nearest to j on the ground truth polygons.

Secondly, we evaluate the accuracy of the generated 3D building models with height RMSE.

5.3 Experiment Results

For comparison, we chose three state-of-the-art instance segmentation methods as baselines for our ONBuildingNet: Geopose [40], LOFT [20], ViTAE [52]+offset head, and Cascade Mask-RCNN [53]+offset head (referred to as CMRCNN+offset in the subsequent discussion). Among these, Geopose is a DSM 3D reconstruction method for off-nadir remote sensing images, while the latter three are methods for extracting building features including roof and offset. We first compare the performance of different building feature extraction

methods. Then, we evaluate the 3D reconstruction results. Specifically, we compare the reconstruction performance of Geopose with the results obtained by incorporating different building feature extraction methods into our 3D reconstruction pipeline described in Section 3.

5.3.1 Roof polygon and offset extraction

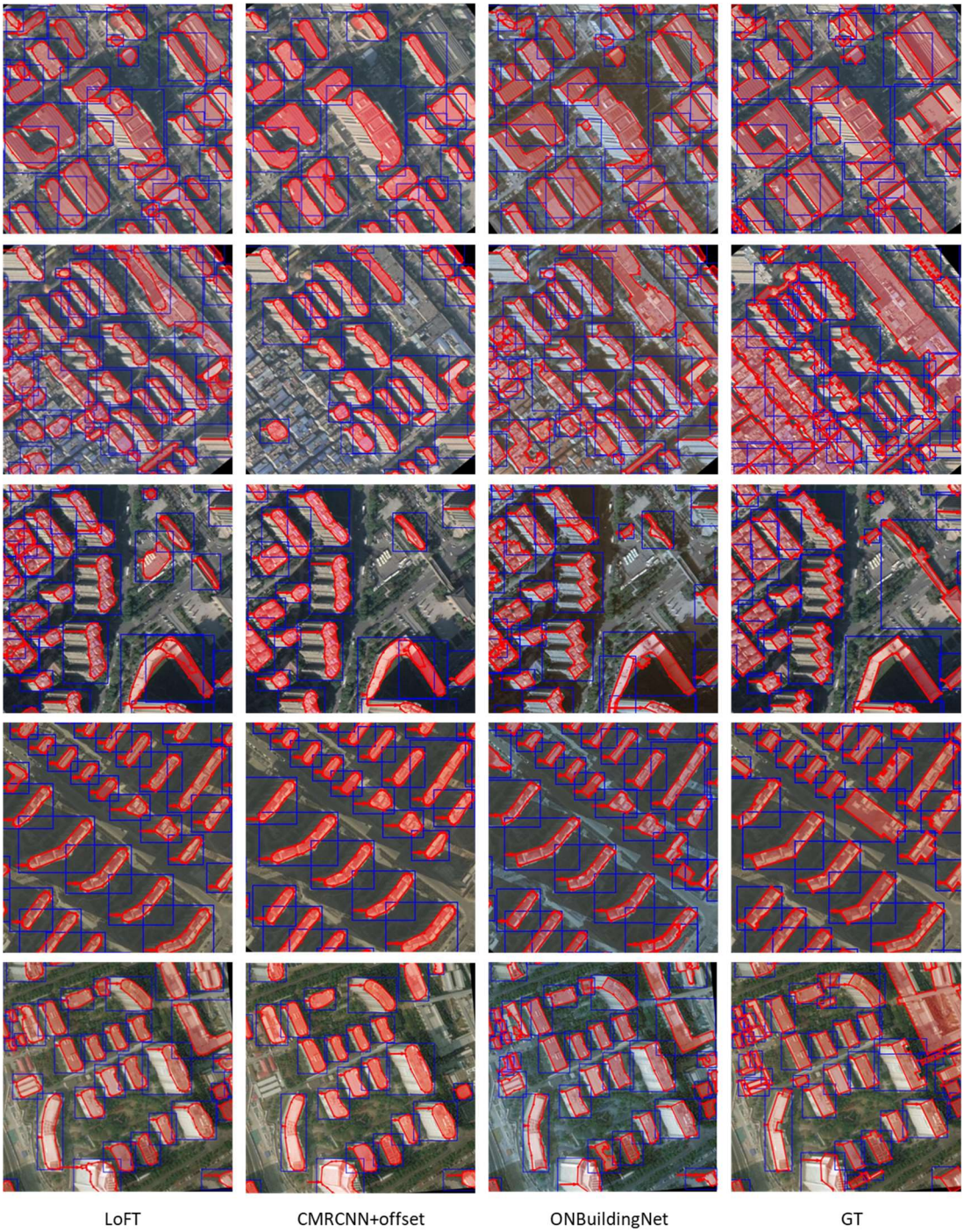

Figure 7 Building polygon extraction experiment illustration. In the figure, the red polygons represent the extracted building roofs, and the red arrows indicate the extracted building offsets.

It can be observed from Figure 7 that, as described in section 4.2, the building polygons extracted by existing methods appear as indistinct angular clusters. However, with the help of our proposed cross-field auxiliary task and multiscale mask head, the building roof polygons

extracted by our ONBuildingNet exhibit more accurate shapes. Our method aligns closely with the ground truth (GT) in terms of edge directions and corner positions.

Table 1 Roof extraction results

Method	AP_{50} ↑	MTAE/° ↓	Offset MAE/pixel ↓
LOFT	0.532	12.11	3.10
CMRCNN+offset	0.529	11.97	2.79
ViTAE+offset	0.570	11.92	2.94
ONBuildingNet	0.564	8.37	2.70

Table 1 reports the numerical evaluation of roof extraction results. Our method significantly better than the existing methos in terms of MTAE, indicating that our approach achieves excellent accuracy in extracted roof polygon's shape.

5.3.2 3D building reconstruction

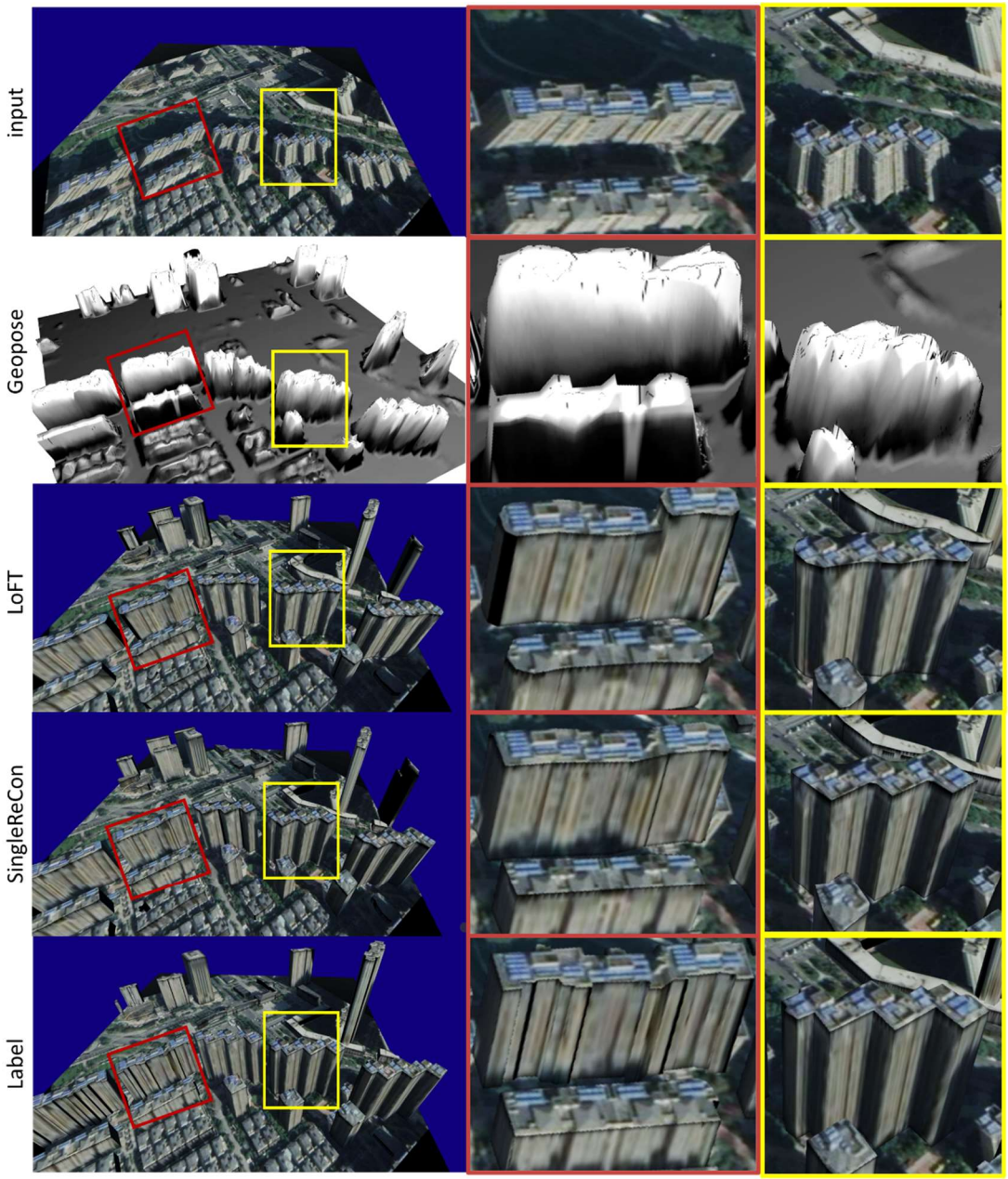

Figure 8 The 3D reconstruction results of different methods for building reconstruction. The output from Geopose is in the form of DSM, while the other rows demonstrate the 3D Tile results obtained by incorporating the output of LoFT, the output of ONBuildingNet, and building labels as building features into our proposed pipeline. The leftmost column displays the overall reconstructed model from an image, while the right side provides an enlarged view of individual buildings. For a more dynamic and intuitive exploration of more 3D model results, please visit our website at https://shaoruizhe.github.io/building-stand-up.github.io/.

Figure 8 shows the reconstruction results of off-nadir remote sensing images using different methods. It can be observed that, compared to the DSM constructed by the Geopose, the 3D models built by our pipeline exhibit better visualization. This improvement can be attributed to our pipeline incorporating LoD1 building priors into the single-image 3D

reconstruction process and extracting textures for the 3D building models. In contrast to LoFT, our ONBuildingNet yields more accurate building shapes during reconstruction. Even for complex-shaped buildings, ONBuildingNet faithfully retains the shape and details, resulting in excellent visual quality for the 3D building models produced by our method.

Table 2 Height RMSE of the 3D results

Method	Height RMSE/meter ↓
LOFT	3.746
CMRCNN+offset	3.959
ViTAE+offset	3.416
Geopose	9.180
ONBuildingNet	3.302

Table 2 reports the Height RMSE for the 3D reconstruction results, which numerically validates that our method achieves state-of-the-art accuracy in this single image 3D reconstruction task. Although our method does not directly optimize for height estimation accuracy like DSM-aimed methods (such as Geopose) do, but rather focuses on the accuracy of building feature extraction, our approach still significantly outperforms Geopose in terms of height accuracy. This demonstrates that our pipeline, by incorporating LoD1 building structure priors, brings a strong advantage in 3D building reconstruction accuracy.

We organized experiments to validate the 3D modeling time efficiency and computational efficiency of our method. The experimental find out that our model can perform 3D construction on a portable laptop equipped with an NVIDIA GeForce RTX 3070 Ti Laptop GPU and an Intel Core i7-12700H CPU. The time efficiency on this device is as follows: on average, it takes 17.09 seconds to reconstruct 3D building models for each 1024x1024 off-nadir image and save the results in 3D Tile format files.

5.4 Ablation Study

We also conduct a series of experiments to investigate the function of each component in the proposed method. The detailed comparisons are given in the following.

Influence of the two proposed component. In the proposed ONBuildingNet, this paper introduces two components aimed at enhancing the performance of building polygon extraction: the cross-field auxiliary task and the multiscale mask head. In the following section, we design ablation experiments to analyze the contributions of these two components to the accuracy of polygon extraction.

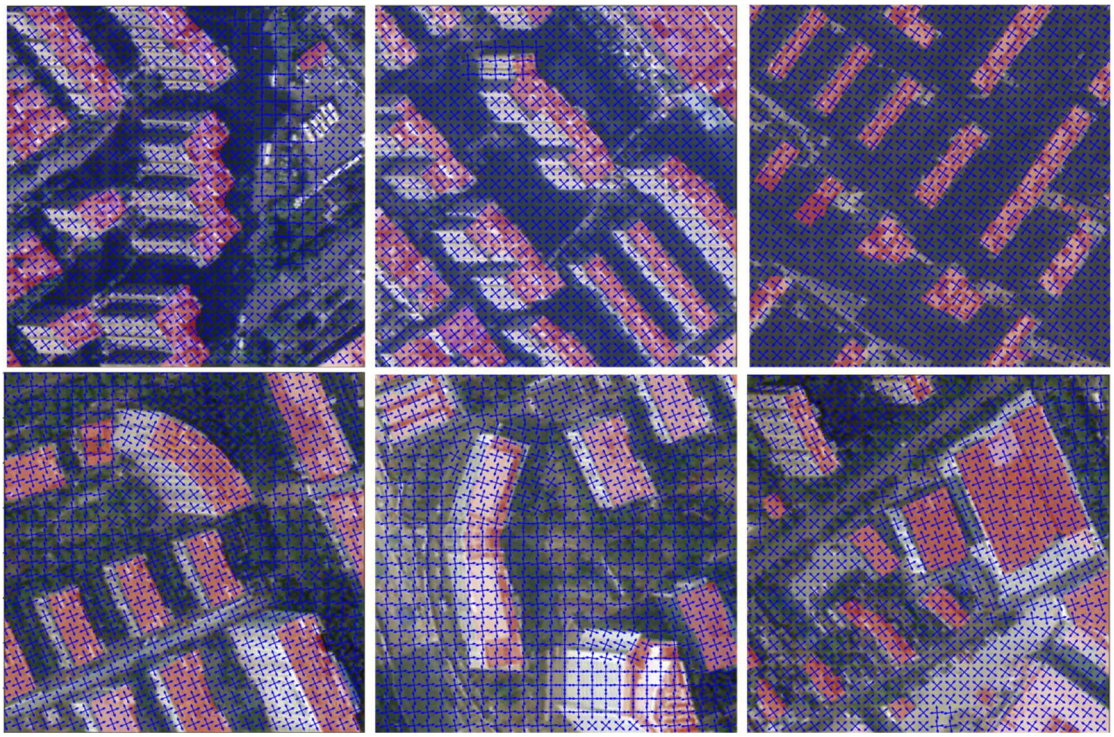

Figure 9 Visualization of the extracted cross-field. The blue crosses are the cross-field extracted by our ONBuildingNet; the red masks represent the extracted roofs.

To validate the effectiveness of the cross-field auxiliary task, we visualize the cross-field and roof mask, as shown in Figure 9. It can be observed that the extracted cross-field exhibits high consistency with the edges of the building roofs. Moreover, the extraction of the roof mask also achieves more accurate shapes with the assistance of the cross-field auxiliary task.

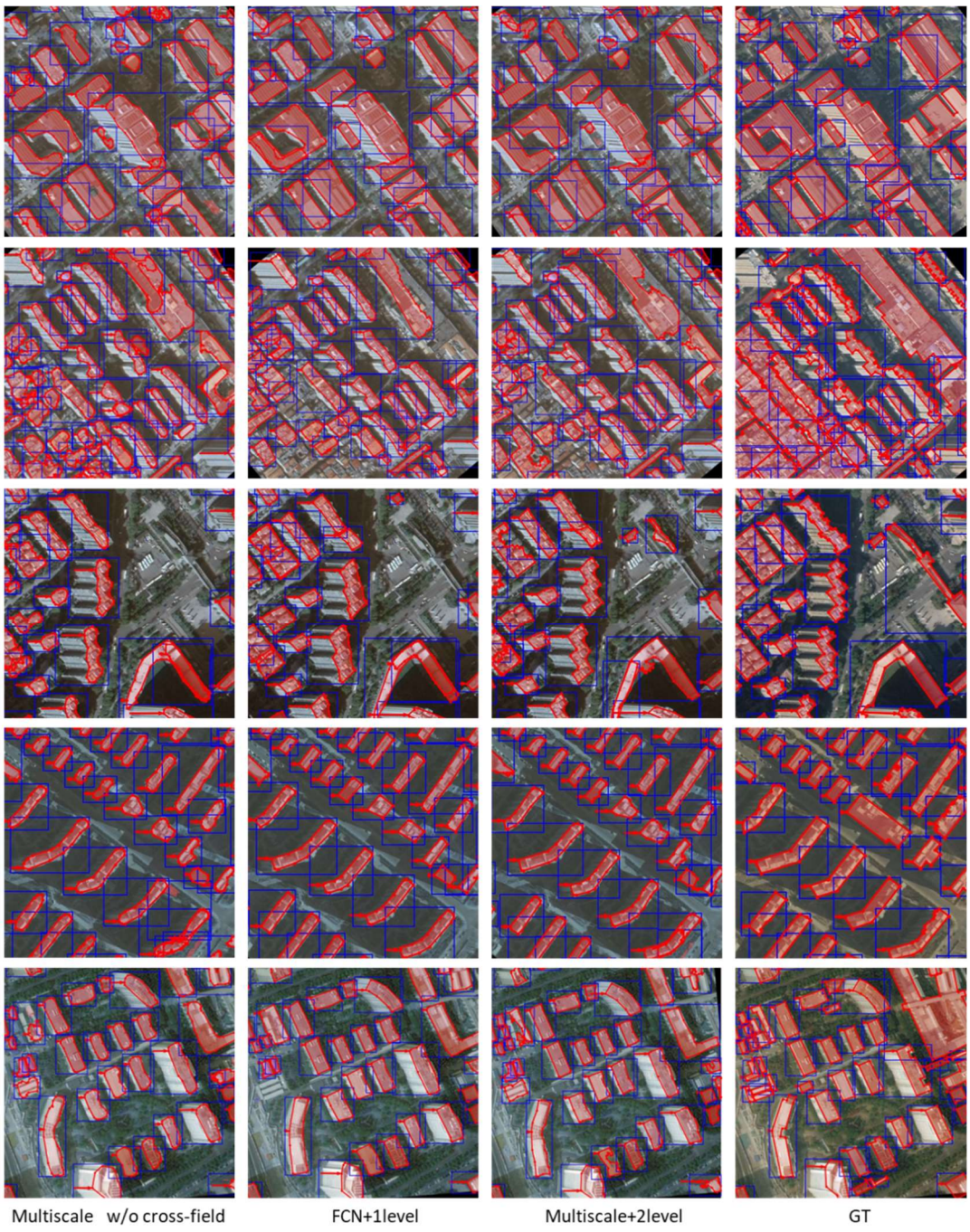

Figure 10 Ablation study results for the two proposed components. The first column displays the polygon extraction result from the network using only the multiscale mask head, without cross-field auxiliary task. The second column shows the result of the network with cross-field auxiliary task and general FCN mask head, rather than our multiscale mask head. The third column illustrates the outcome of the full version of the proposed ONBuildingNet. The last column is the ground truth.

From the first column of Figure 10, it can be observed that, compared to the building extraction results using the FCN mask head (see Figure 7 column 1), the multiscale mask head yields richer details in the extracted buildings, particularly for complex-shaped structures. However, at this stage, there is still a lack of sharp corners in the extracted results. The second column of Figure 10 is an evident that after incorporating the cross-field auxiliary task, the model can accurately extract building polygons with clear corners, especially for simple

rectangular structures. However, for complex-shaped buildings, there may be instances of insufficient resolution in the extraction. Finally, in the third column of the figure, with both of our proposed components enabled, the model precisely extracts distinctly angular building polygons, even for complex-shaped structures.

Table 3 Numerically validates of ablation study results for the proposed components. In the "cross-field" column, ×, 1-level, and 2-level represents without cross-field task, using 1-level cross-field discrepancy function, and using 1-level cross-field discrepancy function respectively.

Mask Head	Cross-field	AP_50 ↑	IoU ↑	MTAE/° ↓
FCN	×	0.532	0.484	12.11
multiscale	×	0.568	0.516	13.12
FCN	1-level	0.551	0.506	6.94
multiscale	1-level	0.566	0.511	10.17
FCN	2-level	0.586	0.508	9.45
multiscale	2-level	0.564	0.515	8.37

Figure 10 only displays representative results from the component ablation experiments due to space limitations. More comprehensive experimental results are reported in Table 3. These experimental findings demonstrate that our proposed multiscale task head, cross-field auxiliary task, and 2-level cross-field discrepancy function all contribute to improving the accuracy of polygon extraction. Additionally, we observed that the combination of FCN with 1-level cross-field discrepancy function outperforms other combinations except for the final combination of multiscale and 2-level, achieving a suboptimal result.

The impact of the weight of each loss in multitask learning. As shown in Figure 3, the multitask neural network ONBuildingNet involves multiple objectives to optimize and balance, and the final loss is computed using Equation (9). To discuss the impact of the weights $\lambda_1, \lambda_2, \lambda_3$,and λ_4 for each loss term in Equation (9) on the model's extraction performance, we conducted ablation experiments for them, and the results are presented in Table 4.

Table 4 ablation study results for the weights of the losses.

Line	Weight of cross-field align loss λ_1	Weight of cross-field smooth loss λ_2	Weight of mask loss λ_3	Weight of mask grad align loss λ_4	IoU ↑	MTAE/° ↓
1	3	0.5	2	0.1	0.497	10.69
2	3	0.1	2	0.1	0.514	9.27
3	3	0.2	2	0.05	0.507	13.76
4	3	0.2	2	0.2	0.506	8.06
5	1.5	0.1	2	0.1	0.514	8.92
6	6	0.4	2	0.1	0.502	15.20
7	3	0.2	2	0.1	0.515	8.37

The first two rows of Table 4 indicate that excessively high or low weights for the smooth loss are detrimental to accurately extracting building shapes. Rows 3, 4, and 7 demonstrate that the mask grad align loss, in conjunction with the mask loss, must strike a balance between pixel accuracy and shape accuracy during mask extraction to achieve optimal extraction results. Rows 5, 6, and 7 highlight the impact of balancing the cross-field task and mask task on the final outcome. Ultimately, we found that the weight combination 3, 0.2, 2, and 0.1 yields the

best results for building polygon extraction. Therefore, this paper adopts this group of weight configuration.

6 Conclusion

This paper proposes SingleRecon, which automatically reconstructs LoD1 3D building models from single non-orthorectified images. The 3D result exhibits high accuracy and excellent visual effects. To achieve this, we introduce a multi-task neural network, called ONBuildingNet, which extracts building features such as building polygons and height. ONBuildingNet incorporates two innovative components: a cross-field auxiliary task and a multiscale mask head, effectively enhancing the shape accuracy of building polygons. Experimental results demonstrate the superiority of our proposed ONBuildingNet, and the effectiveness of our single-image 3D reconstruction process.

References:

1. Domingo, D., J. van Vliet, and A.M. Hersperger, *Long-term changes in 3D urban form in four Spanish cities.* Landscape and Urban Planning, 2023. **230**.
2. Wu, C., et al., *Construction of spatial information model of 3D real estate: case study of the Nanjing gulou central business district.* Survey Review, 2022. **54**(386): p. 391-403.
3. Zi, W., et al., *UrbanSegNet: An urban meshes semantic segmentation network using diffusion perceptron and vertex spatial attention.* International Journal of Applied Earth Observation and Geoinformation, 2024. **129**: p. 103841.
4. Cheng, M.-L., et al., *Near-real-time gradually expanding 3D land surface reconstruction in disaster areas by sequential drone imagery.* Automation in Construction, 2022. **135**: p. 104105.
5. Liang, F., et al., *A novel skyline context descriptor for rapid localization of terrestrial laser scans to airborne laser scanning point clouds.* ISPRS Journal of Photogrammetry and Remote Sensing, 2020. **165**: p. 120-132.
6. Bizjak, M., et al., *Novel Half-Spaces Based 3D Building Reconstruction Using Airborne LiDAR Data.* Remote Sensing, 2023. **15**(5).
7. Chen, X., et al., *Camera and LiDAR Fusion for Urban Scene Reconstruction and Novel View Synthesis via Voxel-Based Neural Radiance Fields.* Remote Sensing, 2023. **15**(18).
8. Marí, R., G. Facciolo, and T. Ehret. *Multi-Date Earth Observation NeRF: The Detail Is in the Shadows*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.
9. Yu, D., et al., *Automatic 3D building reconstruction from multi-view aerial images with deep learning.* ISPRS Journal of Photogrammetry and Remote Sensing, 2021. **171**: p. 155-170.
10. Yu, D., et al., *Advanced approach for automatic reconstruction of 3d buildings from aerial images.* The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2020. **43**: p. 541-546.
11. Yang, B., et al., *A novel approach of efficient 3D reconstruction for real scene using unmanned aerial vehicle oblique photogrammetry with five cameras.* Computers and Electrical Engineering, 2022. **99**.
12. Mahmud, J., et al. *Boundary-aware 3D building reconstruction from a single overhead image*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

13. Chen, S., et al., *HTC-DC Net: Monocular Height Estimation From Single Remote Sensing Images.* IEEE Transactions on Geoscience and Remote Sensing, 2023. **61**: p. 1-18.
14. Buyukdemircioglu, M., S. Kocaman, and M. Kada, *Deep Learning for 3d Building Reconstruction: A Review.* The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2022. **XLIII-B2-2022**: p. 359-366.
15. Li, W., et al., *3D Building Reconstruction from Monocular Remote Sensing Images with Multi-level Supervisions.* arXiv preprint arXiv:2404.04823, 2024.
16. Li, Z., et al., *Reconstruction of 3D Information of Buildings from Single-View Images Based on Shadow Information.* ISPRS International Journal of Geo-Information, 2024. **13**(3): p. 62.
17. Aicardi, I., et al., *UAV photogrammetry with oblique images: First analysis on data acquisition and processing.* The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2016. **41**: p. 835-842.
18. SIG3D. *Manual for the Modeling of 3D Objects. Part 2: Modeling of Buildings–LoD1, LoD2, LoD3*. 2014 [cited 2024 5.14]; Available from: https://files.sig3d.org/file/ag-qualitaet/201311_SIG3D_Modeling_Guide_for_3D_Objects_Part_2.pdf.
19. Weir, N., et al. *Spacenet mvoi: A multi-view overhead imagery dataset*. in *Proceedings of the ieee/cvf international conference on computer vision*. 2019.
20. Wang, J., et al., *Learning to extract building footprints from off-nadir aerial images*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022. **45**(1): p. 1294-1301.
21. Li, W., et al., *Weakly-supervised 3D Building Reconstruction from Monocular Remote Sensing Images.* IEEE Transactions on Geoscience and Remote Sensing, 2024.
22. Sun, C., et al., *SemiSANet: A Semi-Supervised High-Resolution Remote Sensing Image Change Detection Model Using Siamese Networks with Graph Attention.* Remote Sensing, 2022. **14**: p. 2801.
23. Sun, C., et al., *SemiBuildingChange: A Semi-Supervised High-Resolution Remote Sensing Image Building Change Detection Method With a Pseudo Bitemporal Data Generator.* IEEE Transactions on Geoscience and Remote Sensing, 2023. **61**: p. 1-19.
24. Yang, L., et al., *EasySeg: an Error-Aware Domain Adaptation Framework for Remote Sensing Imagery Semantic Segmentation via Interactive Learning and Active Learning.* IEEE Transactions on Geoscience and Remote Sensing, 2024: p. 1-1.
25. Song, J., et al., *RSMT: A Remote Sensing Image-to-Map Translation Model via Adversarial Deep Transfer Learning*. Remote Sensing, 2022. **14**: p. 919.
26. Yingxiao, X., et al., *NBR-Net: A Non-rigid Bi-directional Registration Network for Multi-temporal Remote Sensing Images*. IEEE Transactions on Geoscience and Remote Sensing, 2022. **60**: p. 1-1.
27. Shao, R., et al., *SUNet: Change Detection for Heterogeneous Remote Sensing Images from Satellite and UAV Using a Dual-Channel Fully Convolution Network.* Remote Sensing, 2021. **13**(18): p. 3750.
28. Li, Z., et al., *RemainNet: Explore Road Extraction from Remote Sensing Image Using Mask Image Modeling.* Remote Sensing, 2023. **15**(17).
29. Girard, N., et al. *Polygonal building extraction by frame field learning*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.
30. *Boundary-aware 3D Building Reconstruction from a Single Overhead Image.* 2020.
31. Li, W., *Joint Semantic-Geometric Learning for Polygonal Building Segmentation.* 2021.

32. Zorzi, S., et al. *Polyworld: Polygonal building extraction with graph neural networks in satellite images*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
33. Chen, J., Y. Qian, and Y. Furukawa. *Heat: Holistic edge attention transformer for structured reconstruction*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.
34. Xu, B., et al., *HiSup: Accurate polygonal mapping of buildings in satellite imagery with hierarchical supervision.* ISPRS Journal of Photogrammetry and Remote Sensing, 2023. **198**: p. 284-296.
35. Zhang, H., G. Ma, and Y. Zhang, *Intelligent-BCD: A Novel Knowledge-Transfer Building Change Detection Framework for High-Resolution Remote Sensing Imagery.* IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022. **15**: p. 5065-5075.
36. Li, W., et al. *3D building reconstruction from monocular remote sensing images*. in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.
37. Yan, Y., et al., *GEOP-Net: Shape Reconstruction of Buildings From LiDAR Point Clouds.* IEEE Geoscience and Remote Sensing Letters, 2023. **20**: p. 1-5.
38. Zhao, L., et al., *A review of 3D reconstruction from high-resolution urban satellite images.* International Journal of Remote Sensing, 2023. **44**(2): p. 713-748.
39. Alidoost, F., H. Arefi, and F. Tombari, *2D image-to-3D model: Knowledge-based 3D building reconstruction (3DBR) using single aerial images and convolutional neural networks (CNNs).* Remote Sensing, 2019. **11**(19): p. 2219.
40. Christie, G., et al. *Learning geocentric object pose in oblique monocular images*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.
41. Mao, Y., et al., *Elevation Estimation-Driven Building 3D Reconstruction from Single-View Remote Sensing Imagery.* IEEE Transactions on Geoscience and Remote Sensing, 2023.
42. *ShadowNeuS: Neural SDF Reconstruction by Shadow Ray Supervision.* CVPR, 2023.
43. Mildenhall, B., et al., *Nerf: Representing scenes as neural radiance fields for view synthesis.* Communications of the ACM, 2021. **65**(1): p. 99-106.
44. Xiangli, Y., et al. *Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering*. in *European conference on computer vision*. 2022. Springer.
45. Kerbl, B., et al., *3D Gaussian Splatting for Real-Time Radiance Field Rendering.* ACM Transactions on Graphics, 2023. **42**(4): p. 1-14.
46. Wei, Y., G. Vosselman, and M.Y. Yang. *BuilDiff: 3D Building Shape Generation using Single-Image Conditional Point Cloud Diffusion Models*. in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.
47. Rematas, K., et al. *Urban radiance fields*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
48. Zou, Z.-X., et al., *Triplane Meets Gaussian Splatting: Fast and Generalizable Single-View 3D Reconstruction with Transformers.* arXiv preprint arXiv:2312.09147, 2023.
49. Zhuang, X., et al., *Synthesis and generation for 3D architecture volume with generative modeling.* International Journal of Architectural Computing, 2023. **21**(2): p. 297-314.
50. He, K., et al. *Mask r-cnn*. in *Proceedings of the IEEE international conference on computer vision*. 2017.

51. Ronneberger, O., P. Fischer, and T. Brox. *U-net: Convolutional networks for biomedical image segmentation*. in *International Conference on Medical image computing and computer-assisted intervention*. 2015. Springer.
52. Wang, D., et al., *An empirical study of remote sensing pretraining*. IEEE Transactions on Geoscience and Remote Sensing, 2022.
53. Cai, Z. and N. Vasconcelos. *Cascade r-cnn: Delving into high quality object detection*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.