BSKNet: 3D Building Reconstruction from Single Off-Nadir Remote Sensing Image with Semi-Weak Supervisions

1 Abstract

3D building reconstruction using monocular remote sensing imagery has emerged as a critical research frontier in geospatial AI, driven by its superior cost-effectiveness in terms of both in data acquisition and processing time. However, the reliance of existing deep learning methodologies on a large number of manually annotated labels poses a significant barrier to their practical application. Although recent studies have contributed to reducing label dependency, a considerable amount of annotation effort is still required to establish baseline reconstruction credibility. To address this challenge, we propose a BuildingSKeletonNetwork (BSKNet) for 3D building reconstruction under semi-weak supervisions combining limited full 3D annotations with abundant building-footprint-only labels. BSKNet innovatively designs and employs a skeletal representation of buildings, which **implicitly encapsulates knowledge from two dimensions: building structural knowledge and corner local information**, while providing ample data for 3D reconstruction. We designed BSKNet to learn the capability of building skeleton extraction from these two knowledge dimensions. Under weak supervision, the knowledge acquired from these two dimensions is synergistically enhanced via **mutual complementation and cross-validation.** Additionally, we propose a self-supervised training method based on consistency constraints and pseudo-labels to further improve the effect of weakly supervised training. The experimental results demonstrate that the proposed BSKNet achieves excellent reconstruction performance by utilizing only 3% fully annotated data combined with weakly supervised samples. This performance represents a significant improvement compared to current state-of-the-art methods. Please visit our project page at https://shaoruizhe.github.io to gain a more intuitive understanding of our method.

2 Introduction

3D urban building data provides an intuitive and comprehensive representation of urban morphology while serving as a foundational dataset for critical applications such as urban planning[1], disaster response[2], and skyline analysis[3]. Consequently, the construction of 3D urban building models has long attracted substantial scholarly attention. Researchers have proposed diverse approaches to 3D reconstruction from multiple perspectives, including LiDAR-based methods[4, 5], multi-view imagery methods[6, 7, 8, 9], and monocular image-based methods[10, 11, 12, 13, 14, 15]. Among these, single-view approaches have gained prominence in time-sensitive or data-constrained scenarios due to their efficiency and cost-effectiveness[15, 16]. For instance, in disaster relief operations, single-view methods enable immediate reconstruction from a

single remotely sensed image, delivering critical data support during initial response phases.

Within the realm of single-view 3D building reconstruction methods, the utilization of **off-nadir remote sensing imagery** represents a significant paradigm[17, 18, 13, 19]. Off-nadir imagery refers to obliquely captured remote sensing data acquired through sensors angled away from the vertical (nadir) axis[20]. Off-nadir based 3D building reconstruction approach holds dual advantages: first, off-nadir images substantially outnumber near-nadir counterparts in typical remote sensing datasets; second, parallax effects in off-nadir imagery produce measurable offsets between building rooftops and base footprints, providing crucial height estimation cues. These characteristics have driven recent research interest in off-nadir-based 3D reconstruction.

Despite achieving notable reconstruction accuracy, current off-nadir-based methods face persistent challenges in data annotation. Model training necessitates extensive manual annotations encompassing bounding boxes, rooftops, footprints, and offset values—a labor-intensive process incompatible with time-sensitive reconstruction tasks. To mitigate this bottleneck, scholars have explored leveraging crowd-sourced building footprint data from platforms like OpenStreetMap (OSM), Google Maps, and Amap. As illustrated in Figure 1, while such annotations lack the completeness required by supervised neural networks, they enable weakly supervised training paradigms that dramatically reduce labeling costs. Subsequent research on weakly supervised frameworks has yielded promising results, yet critical limitations persist. Existing methods still require non-trivial amounts of fully annotated data for effective training, and their accuracy under weak supervision remains suboptimal, suggesting substantial room for improvement in annotation efficiency and model performance.

To address the aforementioned challenges, we propose BuildingSKeletonNetwork (BSKNet), an end-to-end framework for 3D building reconstruction. As illustrated in Figure 2, conventional methods follow a two-stage instance segmentation paradigm: first detecting building bounding boxes from images, then extracting Region-of-Interest (RoI) features to predict footprints and elevation offsets. However, under **weak supervision with only footprint annotations**, this two-stage architecture fails to obtain reliable bounding box supervision, thereby invalidating subsequent footprint and offset estimation derived from RoI features. To resolve this fundamental limitation, we abandon the traditional **bounding box-offset representation** and instead introduce a **skeletal representation** for buildings. Our BSKNet network pioneers end-to-end skeletal building prediction, achieving effective supervision under weak supervision limited to footprint annotations.

The BSKNet integrates a **structure decoder** that encodes structure priors of building skeletons and a **corner decoder** that learns discriminative local visual features of building corners. These two components engage in **mutual complementation and cross-validation** during weakly supervised training, mutually reinforcing **building structural knowledge** and **corner local information** to progressively enhance skeletal extraction capability. Furthermore, we introduce a contrastive learning-based self-supervised strategy to regularize

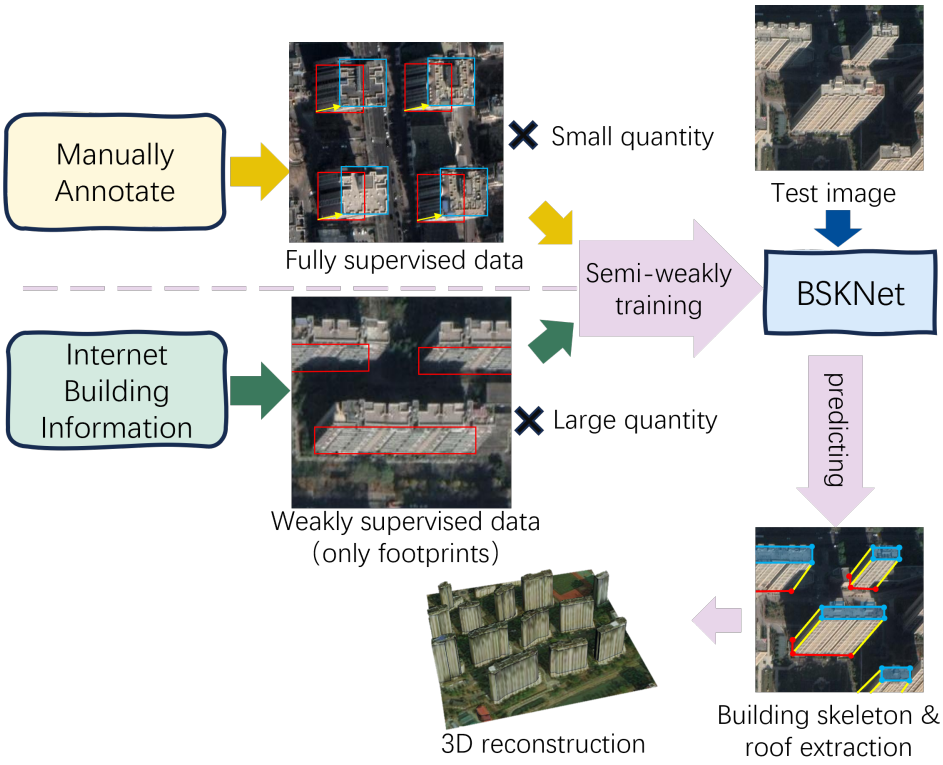

Figure 1: The technical workflow for 3D reconstruction in practical applications in this study: a large amount of weakly supervised data is collected through network mapping, while a small amount of fully supervised data is obtained through manual annotation. These data are then used to perform semi-weak training on the model, resulting in a network for extracting building skeletons and roofs from remote sensing images that can be utilized for 3D building reconstruction

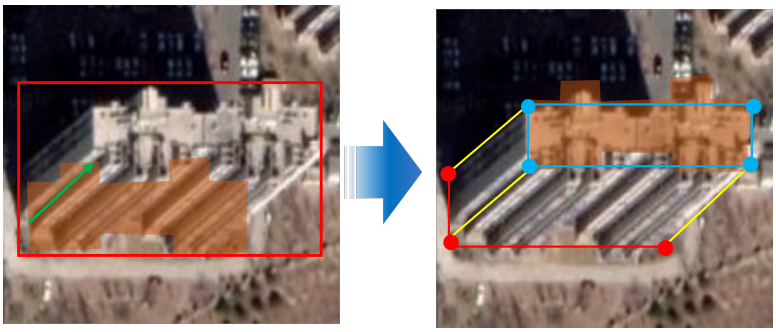

Figure 2: Two representations of buildings: the traditional Bounding Box-Footprint-Offset representation versus to the skeleton representation designed in this study.

skeletal geometry and a pseudo-label-based roof mask optimization method, collectively improving the extraction accuracy of the skeletal building and the roof.

Our primary contributions are summarized as follows:

1. We propose a BuildingSKeletonNetwork (BSKNet) for joint building 3D reconstruction, which innovatively designs a skeleton for building representation. BSKNet constructs knowledge from two dimensions: building structural knowledge and corner local information, which can be synergistically enhanced through mutual complementation and cross-validation under semi-weak supervision of building footprints.

2. We develop a weakly-supervised training framework for 3D building reconstruction, incorporating two key components: self-supervised consistency constraints and pseudo mask. This significantly enhances the network's capability to learn building extraction and 3D reconstruction from weakly-annotated data.

3. We conduct extensive experiments demonstrating that our method outperforms current state-of-the-art (SOTA) approaches under building footprint semi-weak supervision. Comprehensive ablation studies validate the effectiveness of each proposed component.

3 Related works

3.1 Monocular 3D building reconstruction

Monocular 3D building reconstruction has gained significant research attention due to its cost-effectiveness and the ease of data acquisition. Existing approaches generally fall into two categories: 1) Pixel-wise height estimation

methods [21, 10, 22] that generate 3D building representations in Digital Surface Model (DSM) format through monocular height estimation 2) Building-wise height estimation methods detect buildings and estimate its height for 3D reconstruction, where building shadows and the offset of buildings in off-nadir images emerge as two predominant height indicators. Shadow-based methods have been explored in [14, 23]. Since off-nadir observations are predominant in remote sensing images[20] and do not rely on specific solar illumination angles, approaches based on off-nadir data demonstrate broader applicability [24]. Recent years have witnessed growing interest in off-nadir-based reconstruction techniques, exemplified by LoFT [17], MLS-BRN [13, 24] and SingleRecon [19].

However, both paradigms face a persistent challenge: their heavy reliance on labor-intensive manual annotations, which remains a critical barrier to their practical deployment.

3.2 Weakly-Supervised and Semi-Supervised Learning

Both weakly- and semi-supervised learning researches aim to address annotation scarcity challenges by developing label-efficient training strategies. Weakly-supervised learning trains models under imperfect supervision characterized by incomplete, inexact, or noisy annotations[25]. This paradigm encompasses diverse problem formulations, including object detection with image-level supervision[26] and semantic segmentation using scribble annotations[27, 28], among others, prompting numerous methodological innovations[29, 30]. Semi-supervised learning operates by strategically training neural networks with a limited set of labeled samples with extensive unlabeled data[31]. Semi-supervised learning is typically implemented through two predominant mechanisms: pseudo-label generation[32, 33] and consistency constraints[34, 35, 36].

The present work addresses a semi-weakly supervised scenario prevalent in off-nadir 3D building reconstruction, where the supervision comprises two components: incomplete supervision from building footprint annotations and limited strong supervision with full 3D labels[17, 24]. Addressing this problem, recent works [18, 13] have initiated substantial research efforts with promising outcomes. However, these approaches suffer from limited learning efficacy from weakly-annotated data, consequently requiring over 30% fully supervised samples[13] to achieve practical usability—a requirement that significantly hinders real-world scalability.

3.3 Keypoint Detection Network

This paper introduces a structural skeleton representation for buildings and propose a skeleton extraction network based on advancements in keypoint detection research. This section reviews the research in keypoint detection, which serves as the foundational knowledge for the proposed network.

Contemporary keypoint detection methodologies primarily address the extraction of discriminative anatomical points (e.g., human joints[37, 38] or hand gestures [39, 40] through three dominant paradigms: bottom-up, top-down,

and end-to-end methods[41]. Bottom-up[37, 42, 43] approaches initially extract keypoint through localized features, then group them to associate points with individual instances. Conversely, Top-down[44, 45, 46, 47] methodologies initially identify candidate instances through object-level semantic analysis prior to performing instance-wise keypoint localization. End-to-end approaches[38, 41] simultaneously focus on local keypoint features and object-level semantic and structural characteristics of targets, directly extracting multiple keypoints belonging to distinct objects through bidirectional feature fusion and interaction between these two aspects.

Our framework adopts the end-to-end architecture due to its inherent capacity for bidirectional feature interaction. We aim to leverage this capability to enable the model to learn and train effectively, even when the supervisory information is incomplete, by allowing mutual supplementation and supervision between object's local and global features.

4 Method

4.1 Building SKeleton Network (BSKNet)

We propose a building skeleton network (**BSKNet**) to simultaneously acquire knowledge from two dimensions: building structural knowledge and corner local information during the learning process. While these two knowledge types function in distinct dimensions with relative autonomy, the task of building skeleton extraction serves as their cross-dimensional "product", enabling synergistic interaction between the knowledge types. Under footprint weak supervision, this design enables mutual verification between structural and corner knowledge, thereby constantly reinforcing the both components.

To achieve this, we observed and summarized the morphological characteristics of buildings in off-nadir remote sensing images and define a building skeleton structure as illustrated in Figure 3 (a). The skeleton comprises seven keypoints that encapsulate two critical aspects of building representation: In terms of **building structural features**, the edges of the roof and footprint remain parallel, while three connecting edges between the roof and footprint align with the offset. In terms of **corner local image features**, all keypoints correspond to prominent corners in the building image, as demonstrated in Figure 3 (b). Thus, we constructed a building skeleton representation that implicitly encapsulates both building structural knowledge and the corner local information.

Building on this skeleton design, we develop an end-to-end network for joint skeleton extraction and roof mask prediction, as shown in Figure 4. Inspired by PETR [43], the skeleton extraction branch employs two query embeddings: structure query embeddings encoding building structural knowledge and corner query embeddings capturing corner local information. Through cross-attention interactions between these queries and image features, the network adaptively fuses building structural knowledge with local evidence to localize building corners.

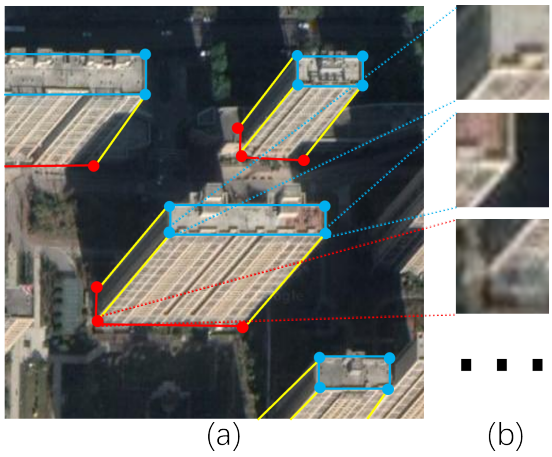

Figure 3: Building structural features and corner local image features of the building skeleton. (a): The skeleton consists of seven key points: the leftmost, rightmost, and bottommost points of the footprint, and the leftmost, rightmost, topmost, and bottommost points of the roof, with directional references defined relative to the offset vector. (b): Magnified visualization of building corner features at specific keypoint locations.

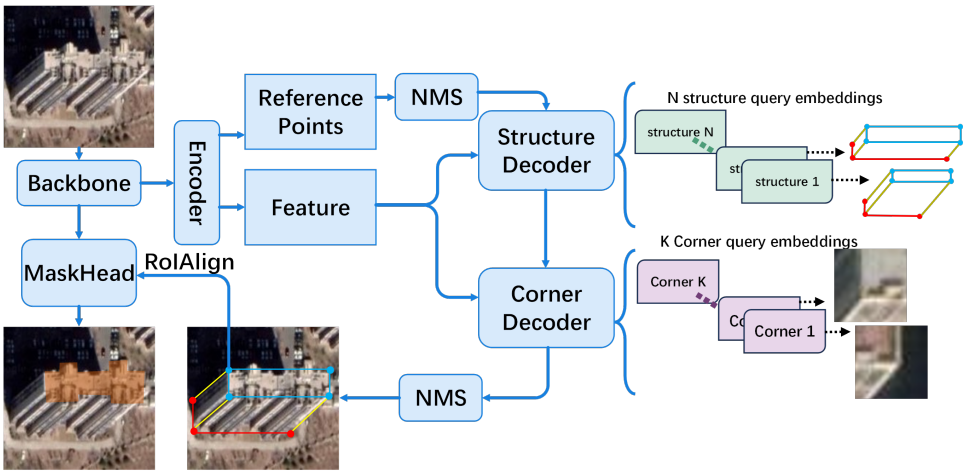

Figure 4: Structural diagram of proposed BSKNet

To reconstruct 3D building models, a mask head is integrated to predict roof regions within the bound of extracted roof keypoints. By sharing backbone features between the skeleton detection and mask prediction branches, the network synergizes building discriminability capabilities and roof recognition accuracy.

Addressing the unique challenges of remote sensing imagery—small targets and spatially distributed instances rather than single dominant objects—we incorporate Non-Maximum Suppression (NMS) during training. This ensures balanced attention across dispersed building instances.

Through the aforementioned model design, we have achieved the proposition outlined at the beginning of this section. Specifically, BSKNet has been empowered to construct knowledge across two dimensions, thereby enhancing building skeleton extraction capability under footprint weak supervision through mutual complementation and cross-validation. Next, we will introduce the loss function used for training under full supervision, while the methods for weakly-supervised training will be discussed in the following section.

When training with fully supervised data, the loss functions used include keypoints loss and mask loss:

$$l_{total\_fully} = \lambda_{kpt}\ l_{kpt} + \lambda_{mask} l_{mask} \tag{1}$$

where λ_{f_kpt} and λ_{mask} denote the weights for these two components. The keypoints loss is calculated using a combination of the smooth L1 loss and the OKS loss[43], while the mask loss is calculated using cross entropy:

$$l_{f\_kpt} = L_{kpt}\left(\widehat{p}^{kpt}, g^{kpt}\right) = L_{l1}\left(\widehat{p}^{kpt}, g^{kpt}\right) + L_{OKS}\left(\widehat{p}^{kpt}, g^{kpt}\right) \tag{2}$$

$$l_{mask} = cross\_entropy\left(\widehat{m}^{roof}, m_{gt}^{roof}\right) \tag{3}$$

4.2 Weakly-Supervised Training

As mentioned earlier, under weak supervision, we propose BSKNet to leverage the input of footprint weak supervision to enable mutual reinforcement between these two knowledge dimensions: building structural knowledge and corner local information, thereby progressively enhancing building skeleton extraction capability. However, two critical limitations persist: (1) Incomplete supervision may lead to erroneous skeleton predictions due to the lack of roof annotation. (2) Current frameworks lack training mechanisms for roof mask prediction under weak supervision.

To address these challenges, we introduce two novel weakly-supervised strategies illustrated in Figure 5:

We propose a contrastive learning approach to train the network under footprint weak supervision through enforcing prediction consistence across geometrically augmented inputs (Figure 5, blue arrows). We pass the augmented and original images through the network for building skeleton prediction. Then, we will assign the obtained building skeleton predictions to the footprint ground

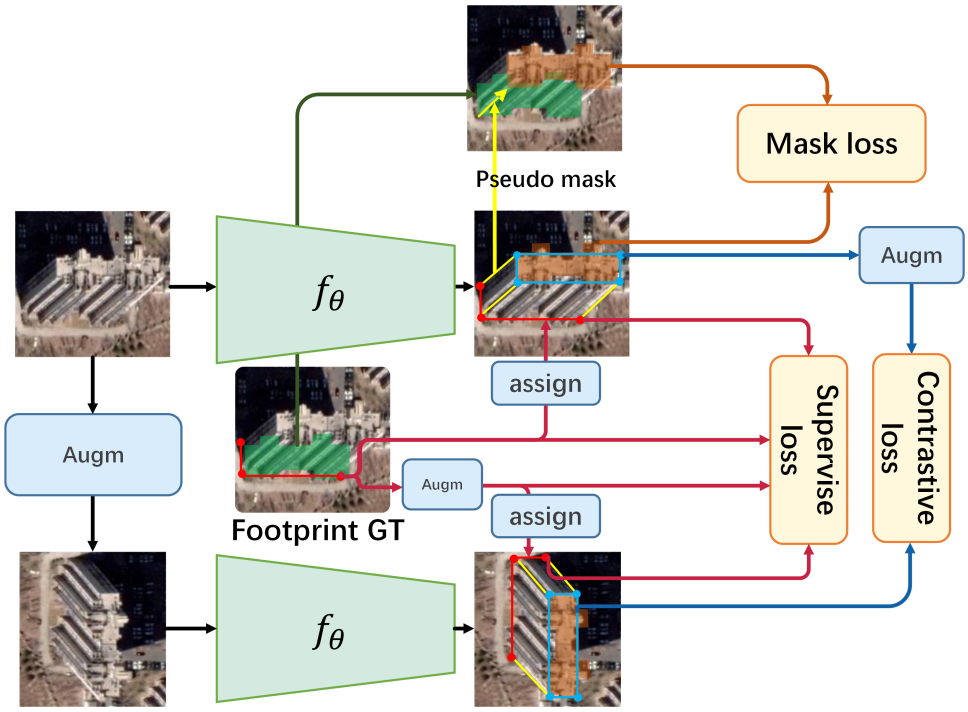

Figure 5: Training method when using footprint weakly-supervised data. The blue arrow represents the calculation of the contrastive learning loss for roof keypoints. The red arrow indicates the calculation of the supervised loss for footprint keypoints. The brown arrow shows the mask loss calculated by the pseudo roof mask method.

truth. For the skeleton predictions assigned to the same footprint ground truth, we calculate the consistency loss between their roof keypoints:

$$l_{c\_roof} = L_{kpt}\left(R_\theta\left(\widehat{p}^{roof}\right), \widehat{p}_R^{roof}\right) \tag{4}$$

where L_{kpt} is consistent with that in full supervision as shown in Equation (2). R_θ represents the rotation of the keypoints. $\widehat{p}^{roof}$ and $\widehat{p}_R^{roof}$ denote roof keypoint predictions assigned to the same footprint ground truth before/after augmentation:

$$\widehat{p}^{roof} = f_p(I), \widehat{p}_R^{roof} = f_p\left(A\left(R_\theta(I)\right)\right) \tag{5}$$

herein, A represents image augmentation operations, which include exposure changes, etc.

Since the footprint has ground truth under weak supervision, footprint keypoints utilize direct annotation supervision (Figure 5, red arrows):

$$l_{s\_footprint} = L_{kpt}\left(\widehat{p}^{foot}, g^{foot}\right) \tag{6}$$

To leverage weakly supervised data for training the model's ability to extract roof masks, we propose a method for generating pseudo roof labels for roof extraction training. (Figure 5, brown arrows) We first derive pseudo-offsets from predicted skeletons, then generate roof masks by translating footprint annotations accordingly. Finally, these pseudo roof masks are used to supervise the model's roof prediction. The pseudo mask loss calculation is as follows:

$$l_{pesudo\_mask} = cross\_entropy\left(\widehat{m}^{roof}, T\left(m_{gt}^{footprint}, offset\right)\right) \tag{7}$$

where T represents the translation of the mask, and $offset$ is the building offset calculated based on the three pairs of corresponding keypoints of the roof and footprint in the building skeleton prediction.

$$offset = \frac{1}{3}\sum_{i=1}^{3}\left(p_i^{roof} - p_i^{footprint}\right) \tag{8}$$

Finally, the loss for weakly supervised training is:

$$l_{total\_weakly} = \lambda_{c\_roof}\ l_{c\_roof} + \lambda_{s\_footprint} l_{s\_footprint} + \lambda_{pesudo\_mask}\ l_{pesudo\_mask} \tag{9}$$

the three λs are the weights of the three parts of the loss respectively.

5 Experiments

5.1 Dataset

This paper conduct experimentation on two benchmark datasets:

(1) **BONAI Dataset [17]**: Contains 3,300 off-nadir satellite images (1024×1024 pixels at 1-meter resolution) covering six geographically representative Chinese

cities: Shanghai, Beijing, Harbin, Jinan, Chengdu, and Xi'an. The dataset is partitioned into 3,000 training and 300 test images.

(2) **HK Dataset [13]**: Comprises 500 training and 119 test satellite images (1024×1024 pixels at 1-meter resolution) captured over Hong Kong.

Under the semi-weakly supervised setting, both datasets are randomly divided into fully- and footprint weakly-annotated subsets according to predetermined ratios. The **fully-supervised subset** provides complete annotations including: building bounding boxes, footprint masks roof-to-footprint offsets, off-nadir angles. The **weakly-supervised subset** retains only footprint masks and off-nadir angles, discarding other annotations. Utilizing these two datasets, we organized two sets of comparative experiments: the first set employs only the BONAI dataset, while the second set utilizes a mixed dataset comprising both the BONAI and HK datasets (denote as BH in the subsequent discussion).

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 set the maximum number of training epochs to 300 and the initial learning rate to 0.02, with a decay rate of 0.1 applied every 100 epochs. During testing in the 3D reconstruction pipeline, we used a confidence threshold of 0.5 for building instances. To optimize GPU memory utilization during training, we process all imagery as 512×512-pixel patches in our experiments.

When evaluating experimental results, we measure the accuracy of building footprint extraction by calculating the F1-score of the footprints at an Intersection over Union (IoU) threshold of 0.5. Additionally, offset estimation performance and 3D reconstruction accuracy is assessed the accuracy of computing the height mean absolute error (denote by MAE(H) in the following).

5.3 Experiment Results

For comparison, we chose three state-of-the-art instance segmentation methods as baselines for our PETR building: MLS-BRN[13], LOFT[17], ViTAE[48]+offset head, and Cascade Mask-RCNN [49]+offset head (referred to as CMRCNN+offset in the subsequent discussion). Among these, the first methodology is specifically designed for semi-weakly supervised learning scenarios, whereas the latter three constitute fully-supervised approaches. For the fully-supervised implementations, model training exclusively utilizes the annotated subset of the dataset.

We first compare the performance of different building feature extraction methods. Then, we evaluate the 3D reconstruction results. The 3D reconstruction results obtained by incorporating different building feature extraction methods into the 3D reconstruction pipeline described in [19]. Experiments were conducted on each of the two aforementioned experimental datasets, and the results are reported in TABLE I and TABLE II, respectively.

Figure 6 illustrates the building extraction results of our method compared to the state-of-the-art semi-weakly-supervised baseline. For conciseness, we

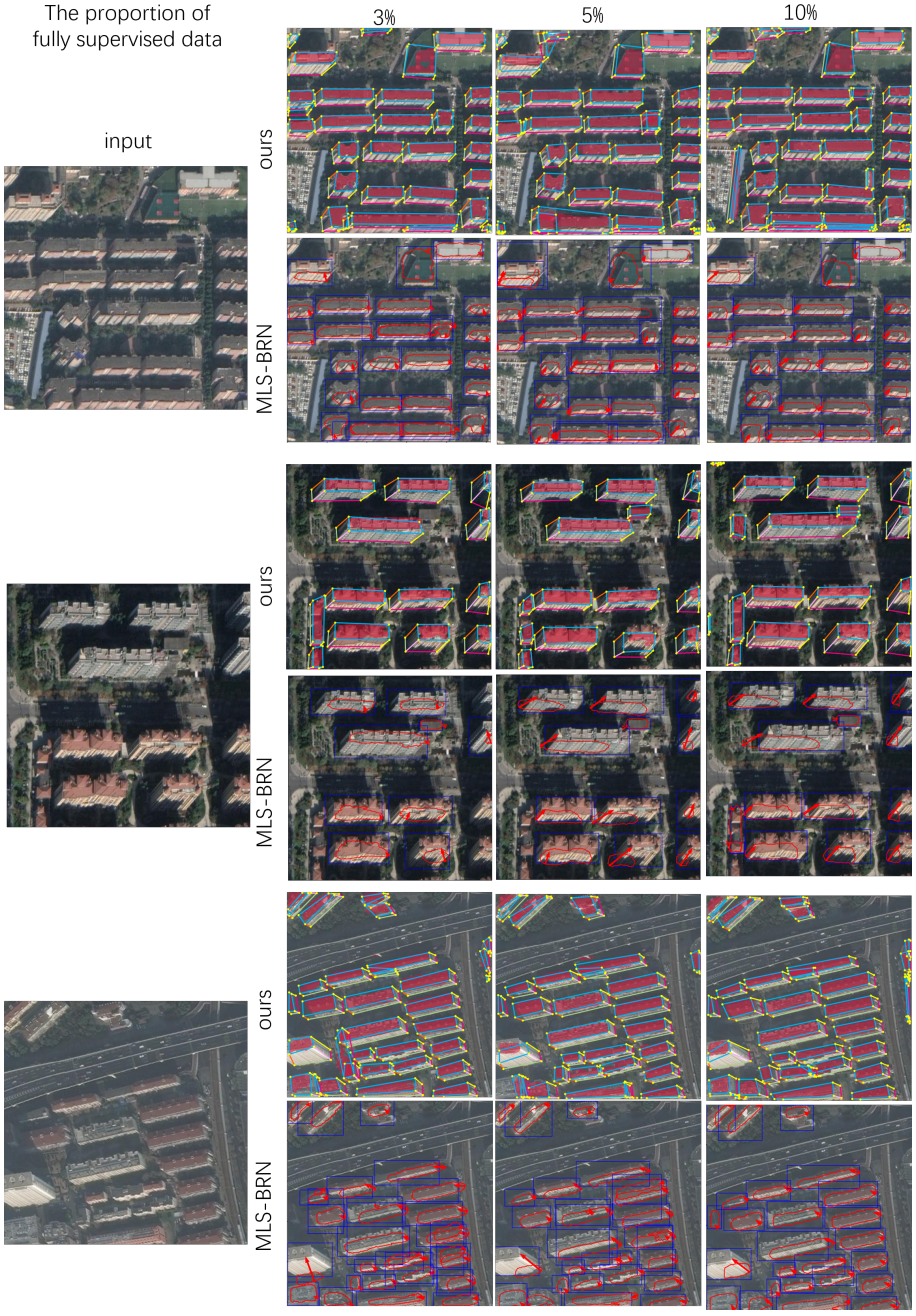

Figure 6: Comparative Visualization of Building Feature Extraction. Our method extracts building skeletons composed of seven points and building rooftops from the images, while the MLS-BRN method extracts building footprints (represented by red polygons) and building offsets (indicated by red arrows).

Table 1: the experiment results on BONAI

Fully supervised proportion	3 %(90 images)		5 %(150 images)		10 %(300 images)	
	F1↑	MAE(H)/m↓	F1↑	MAE(H)/m↓	F1↑	MAE(H)/m↓
CMRCNN[49]+offset	0.060	3.58	0.109	3.90	0.114	4.01
ViTAE[48]+offset	0.048	8.41	0.032	9013	0.397	7.85
LOFT-FOA[17]	0.143	3.86	0.229	4.13	0.254	3.77
MLS-BRN[13]	0.308	4.85	0.475	5.16	0.577	5.77
BSKNet	0.479	2.75	0.556	2.78	0.564	2.73

Table 2: the experiment results on BH

Fully supervised proportion	3 %(90 images)		5 %(150 images)		10 %(300 images)	
	F1↑	MAE(H)/m↓	F1↑	MAE(H)/m↓	F1↑	MAE(H)/m↓
CMRCNN+offset						
ViTAE[50]+offset						
LOFT-FOA	0.143	3.86	0.229	4.13	0.254	3.77
MLS-BRN	0.308	4.85	0.475		0.577	5.77
BSKNet			0.540	2.72	0.565	2.63

showcase results against MLS-BRN[13], the best-performing weakly-supervised approach in our benchmark studies. The visualization validates the critical advantages of BSKNet: BSKNet can still accurately and clearly extract building skeletons from images under semi-weakly supervised conditions. In contrast, MLS-BRN suffers from estimation bias in offsets due to its inability to learn offset estimation capabilities from footprint weakly supervised data, leading to inaccuracies in footprint position estimation. Notably, the building structural understanding from skeleton extraction enhances backbone building feature discriminability, thereby improving the accuracy of roof extraction.

Figure 7 illustrates the effectiveness of our method for 3D reconstruction under semi-weak supervision. The results prove that our proposed BSKNet can accurately extract building heights and footprints while achieving visually compelling 3D models, even with limited manual annotation. **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.

Components in Network Training: To adapt BSKNet to the characteristics of building targets in remote sensing images and to enhance the learning capability of the network during weakly supervised training, we incorporated Non-Maximum Suppression (NMS) into the network and designed a weakly supervised training process. We conducted ablation experiments to test the effectiveness of these components. This set of ablation experiments utilized the

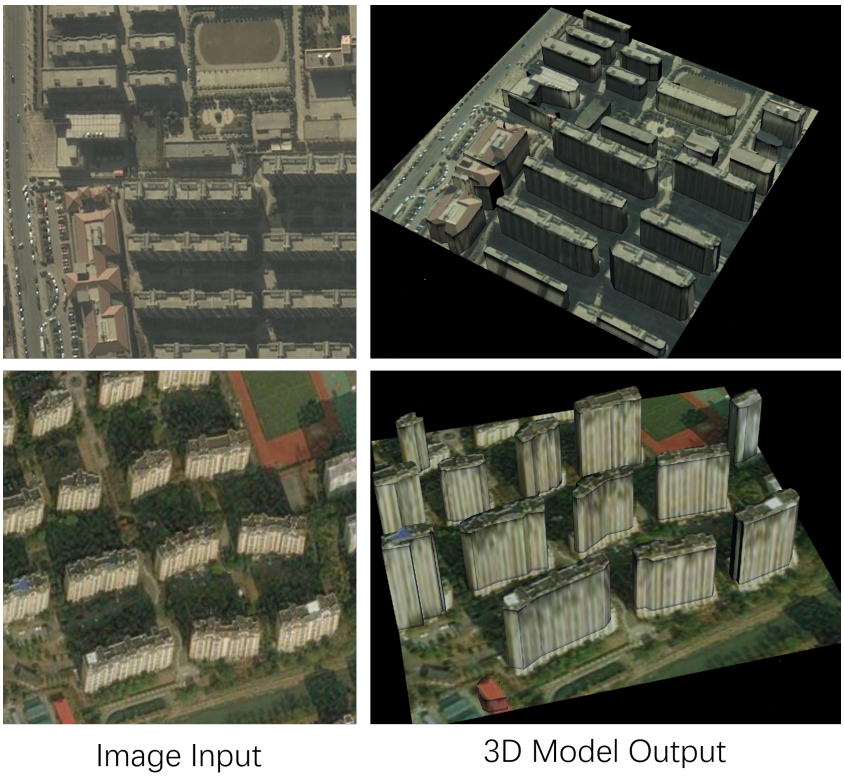

Figure 7: The 3D reconstruction results of buildings in the BONAI test set using BSKNet trained with semi-weakly supervised data (5% fully supervised)

semi-weakly supervised BONAI dataset with a fully supervised ratio of 5%.

Table 3: Ablation experiments targeting components

	F1↑	Precision↑	Recall↑	MAE(H)/m↓
w/o NMS				
w/o consistence loss	0.511	0.448	0.616	3.12
w/o pseudo mask loss				
ours	0.556	0.492	0.640	2.778

Loss Function Weights: In the context of weakly supervised training, we proposed two loss functions(Equation (9)). To assess the impact of the weights of these loss functions on model training, we performed ablation experiments on the weights. This set of ablation experiments also utilized the semi-weakly supervised BONAI dataset with a fully supervised ratio of 5%.

Table 4: Ablation experiments targeting components

Line	λ_{c_roof}	λ_{pesudo_mask}	F1↑	MAE(H)/m↓
1	10	20		
2	20	20	0.556	2.778
3	50	20	0.525	2.851
4	50	10		
5	50	50		
6	50	100		

The Supervision Ratio Boundary of our Method: To reveals our method's efficiency under extreme annotation scarcity scenarios., we conducted experiments varying the ratio of fully supervised data in training data.

6 Conclusion

In this paper, we propose a semi-weakly supervised framework for 3D building reconstruction. Experiments demonstrate that our method requires only a small amount of fully supervised 3D building annotations and a large number of weakly supervised building footprint annotations to achieve single-image 3D reconstruction of buildings, surpassing state-of-the-art (SOTA) methods. Ablation studies validate the advantages of our proposed building skeleton extraction strategy in semi-supervised learning, as well as the efficacy of the semi-supervised training strategy. We believe our approach holds significant potential for real-world applications where footprint weakly supervision data is widely available. Future work will focus on further improving the quality and accuracy of 3D building reconstruction.

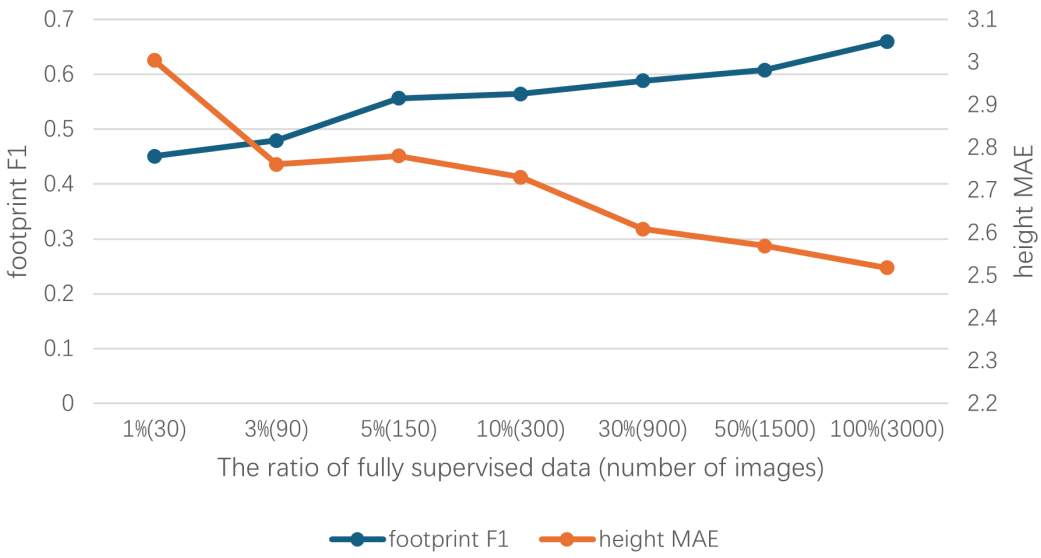

Figure 8: Relationship between the performance and the proportion of fully supervised training data

7 references

References

[1] Darío Domingo, Jasper van Vliet, and Anna M. Hersperger. Long-term changes in 3d urban form in four spanish cities. *Landscape and Urban Planning*, 230, 2023.

[2] Min-Lung Cheng, Masashi Matsuoka, Wen Liu, and Fumio Yamazaki. Near-real-time gradually expanding 3d land surface reconstruction in disaster areas by sequential drone imagery. *Automation in Construction*, 135:104105, 2022.

[3] Fuxun Liang, Bisheng Yang, Zhen Dong, Ronggang Huang, Yufu Zang, and Yue Pan. 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*, 165:120–132, 2020.

[4] Marko Bizjak, Domen Mongus, Borut Žalik, and Niko Lukač. Novel half-spaces based 3d building reconstruction using airborne lidar data. *Remote Sensing*, 15(5), 2023.

[5] Xuanzhu Chen, Zhenbo Song, Jun Zhou, Dong Xie, and Jianfeng Lu. Camera and lidar fusion for urban scene reconstruction and novel view synthesis via voxel-based neural radiance fields. *Remote Sensing*, 15(18), 2023.

[6] Roger Marí, Gabriele Facciolo, and Thibaud 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*, pages 2034–2044.

[7] Dawen Yu, Shunping Ji, Jin Liu, and Shiqing Wei. Automatic 3d building reconstruction from multi-view aerial images with deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 171:155–170, 2021.

[8] D Yu, S Wei, J Liu, and S Ji. Advanced approach for automatic reconstruction of 3d buildings from aerial images. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:541–546, 2020.

[9] Boxiong Yang, Faizan Ali, Bo Zhou, Shelei Li, Ying Yu, Tingting Yang, Xiaofei Liu, Zhiyong Liang, and Kaicun Zhang. A novel approach of efficient 3d reconstruction for real scene using unmanned aerial vehicle oblique photogrammetry with five cameras. *Computers and Electrical Engineering*, 99, 2022.

[10] Jisan Mahmud, True Price, Akash Bapat, and Jan-Michael Frahm. Boundary-aware 3d building reconstruction from a single overhead image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 441–451.

[11] Sining Chen, Yilei Shi, Zhitong Xiong, and Xiao Xiang Zhu. Htc-dc net: Monocular height estimation from single remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–18, 2023.

[12] M. Buyukdemircioglu, 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*, XLIII-B2-2022:359–366, 2022.

[13] Weijia Li, Haote Yang, Zhenghao Hu, Juepeng Zheng, Gui-Song Xia, and Conghui He. 3d building reconstruction from monocular remote sensing images with multi-level supervisions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27728–27737.

[14] Zhixin Li, Song Ji, Dazhao Fan, Zhen Yan, Fengyi Wang, and Ren Wang. Reconstruction of 3d information of buildings from single-view images based on shadow information. *ISPRS International Journal of Geo-Information*, 13(3):62, 2024.

[15] Vivek Verma, Rakesh Kumar, and Stephen Hsu. 3d building detection and modeling from aerial lidar data. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2213–2220. IEEE.

[16] Liuyun Duan and Florent Lafarge. Towards large-scale city reconstruction from satellites. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 89–104. Springer.

[17] Jinwang Wang, Lingxuan Meng, Weijia Li, Wen Yang, Lei Yu, and Gui-Song Xia. Learning to extract building footprints from off-nadir aerial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1294–1301, 2022.

[18] Weijia Li, Zhenghao Hu, Lingxuan Meng, Jinwang Wang, Juepeng Zheng, Runmin Dong, Conghui He, Gui-Song Xia, Haohuan Fu, and Dahua Lin. Weakly-supervised 3d building reconstruction from monocular remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[19] Ruizhe Shao, JiangJiang Wu, Jun Li, Shuang Peng, Hao Chen, and Chun Du. Singlerecon: Reconstructing building 3d models of lod1 from a single off-nadir remote sensing image. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[20] Nicholas Weir, David Lindenbaum, Alexei Bastidas, Adam Van Etten, Sean McPherson, Jacob Shermeyer, Varun Kumar, and Hanlin Tang. Spacenet mvoi: A multi-view overhead imagery dataset. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 992–1001.

[21] Gordon Christie, Rodrigo Rene Rai Munoz Abujder, Kevin Foster, Shea Hagstrom, Gregory D Hager, and Myron Z Brown. Learning geocentric object pose in oblique monocular images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14512–14520.

[22] Yongqiang Mao, Kaiqiang Chen, Liangjin Zhao, Wei Chen, Deke Tang, Wenjie Liu, Zhirui Wang, Wenhui Diao, Xian Sun, and Kun Fu. Elevation estimation-driven building 3d reconstruction from single-view remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[23] Shadowneus: Neural sdf reconstruction by shadow ray supervision. *CVPR*, 2023.

[24] Weijia Li, Lingxuan Meng, Jinwang Wang, Conghui He, Gui-Song Xia, and Dahua Lin. 3d building reconstruction from monocular remote sensing images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12548–12557.

[25] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018.

[26] Bo Li, Yongqiang Yao, Jingru Tan, Xin Lu, Fengwei Yu, Ye Luo, and Jianwei Lu. Improving long-tailed object detection with image-level supervision by multi-task collaborative learning. *arXiv preprint arXiv:2210.05568*, 2022.

[27] Hao Chen, Shuang Peng, Chun Du, Jun Li, and Songbing Wu. Sw-gan: Road extraction from remote sensing imagery using semi-weakly supervised adversarial learning. *Remote Sensing*, 14(17):4145, 2022.

[28] Mathias Micheelsen Lowes, Jakob L Christensen, Bjørn Schreblowski Hansen, Morten Rieger Hannemose, Anders Bjorholm Dahl, and Vedrana Dahl. Interactive scribble segmentation. In *Proceedings of the Northern Lights Deep Learning Workshop*, volume 4.

[29] Wentian Cai, Yijiang Li, Yandan Chen, Jing Lin, Zihao Huang, Ping Gao, Thippa Reddy Gadekallu, Wei Wang, and Ying Gao. Enhancing weakly supervised semantic segmentation with multi-label contrastive learning and llm features guidance. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[30] Xiangqi Chen, Chengzhuan Yang, Jiashuaizi Mo, Yunliang Jiang, and Zhonglong Zheng. End-to-end point supervised object detection with low-level instance features. *Applied Soft Computing*, 156:111513, 2024.

[31] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *IEEE transactions on knowledge and data engineering*, 35(9):8934–8954, 2022.

[32] Chengzhe Sun, Hao Chen, Chun Du, and Ning Jing. 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*, 61:1–19, 2023.

[33] Youngtaek Oh, Dong-Jin Kim, and In So Kweon. Daso: Distribution-aware semantics-oriented pseudo-label for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9786–9796.

[34] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

[35] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

[36] Mingkai Zheng, Shan You, Lang Huang, Fei Wang, Chen Qian, and Chang Xu. Simmatch: Semi-supervised learning with similarity matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14471–14481.

[37] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11977–11986.

[38] Huan Liu, Qiang Chen, Zichang Tan, Jiang-Jiang Liu, Jian Wang, Xiangbo Su, Xiaolong Li, Kun Yao, Junyu Han, and Errui Ding. Group pose: A simple baseline for end-to-end multi-person pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15029–15038.

[39] Prachetas Padhi and Mousumi Das. Hand gesture recognition using densenet201-mediapipe hybrid modelling. In *2022 International Conference on Automation, Computing and Renewable Systems (ICACRS)*, pages 995–999. IEEE.

[40] Alexander Kapitanov, Karina Kvanchiani, Alexander Nagaev, Roman Kraynov, and Andrei Makhliarchuk. Hagrid–hand gesture recognition image dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4572–4581.

[41] Dahu Shi, Xing Wei, Liangqi Li, Ye Ren, and Wenming Tan. End-to-end multi-person pose estimation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11069–11078.

[42] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306.

[43] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *Proceedings of the European conference on computer vision (ECCV)*, pages 269–286.

[44] Weian Mao, Yongtao Ge, Chunhua Shen, Zhi Tian, Xinlong Wang, Zhibin Wang, and Anton van den Hengel. Poseur: Direct human pose regression with transformers. In *European conference on computer vision*, pages 72–88. Springer.

[45] Ke Li, Shijie Wang, Xiang Zhang, Yifan Xu, Weijian Xu, and Zhuowen Tu. Pose recognition with cascade transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1944–1953.

[46] Weian Mao, Zhi Tian, Xinlong Wang, and Chunhua Shen. Fcpose: Fully convolutional multi-person pose estimation with dynamic instance-aware convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9034–9043.

[47] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703.

[48] Di Wang, Jing Zhang, Bo Du, Gui-Song Xia, and Dacheng Tao. An empirical study of remote sensing pretraining. *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

[49] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162.