FINE-GRAINED BUILDING ROOF INSTANCE SEGMENTATION BASED ON DOMAIN ADAPTED PRETRAINING AND COMPOSITE DUAL-BACKBONE

Guozhang Liu†, Baochai Peng†, Ting Liu, Pan Zhang, Mengke Yuan,
Chaoran Lu, Ningning Cao, Sen Zhang, Simin Huang, Tao Wang*

PIESAT Information Technology Co, Ltd., Beijing, China

ABSTRACT

The diversity of building architecture styles of global cities situated on various landforms, the degraded optical imagery affected by clouds and shadows, and the significant inter-class imbalance of roof types pose challenges for designing a robust and accurate building roof instance segmentor. To address these issues, we propose an effective framework to fulfill semantic interpretation of individual buildings with high-resolution optical satellite imagery. Specifically, the leveraged domain adapted pretraining strategy and composite dual-backbone greatly facilitates the discriminative feature learning. Moreover, new data augmentation pipeline, stochastic weight averaging (SWA) training and instance segmentation based model ensemble in testing are utilized to acquire additional performance boost. Experiment results show that our approach ranks in the first place of the 2023 IEEE GRSS Data Fusion Contest (DFC) Track 1 test phase (mAP_{50}:50.6%). Note-worthily, we have also explored the potential of multimodal data fusion with both optical satellite imagery and SAR data.

Index Terms— Roof instance segmentation, Self-supervised pretraining, Multimodal data fusion

1. INTRODUCTION

Automated and in-depth building interpretation is meaningful in remote sensing scenario, as the investigation results are of great significance in urban planning, smart city, and emergency management applications [1]. Unfortunately, the growing data volume and rapidly evolving deep learning techniques can still hardly meet the requirements of practical building roof detection and fine-grained roof type classification due to the lack of finely-annotated and well-organized dataset, the variable instance sizes, the diverse architectural styles and the great inter-class quantity discrepancy, etc.

The DFC 2023 [2] has constructed a large-scale, multimodal and elaborately labeled dataset in promoting the performance of building detection and roof type classification methods. The dataset contains a total of 12 predefined roof

---
† Equal contributions
* Corresponding author, wangtao@piesat.cn

Fig. 1: Distribution of instance pixels among predefined roof categories of DFC 2023. Gray, orange and blue represent large, medium and small instances respectively.

categories and about 190,000 high-quality mask annotations. The statistics of track1 training dataset (Fig. 1) reflects practical difficulties that there are significant proportion disparities of different classes and high-density small targets. The instances of the largest category "Flat roof" are more than 200 times of the least category "Revolved roof". Small objects account for 71% of the total number of objects. Huang et al. [1] validate the representative single-stage (SOLOv2), two-stage (Mask-RCNN, Cascade Mask RCNN), and query-based (QueryInst) methods which can not achieve desirable performance in similar UBC [1] dataset. In summary, the three challenges in fine-grained roof instance segmentation are: **1) the long-tail distribution of diverse roof styles**; **2) the detection of crowded small objects**; **3) the ambiguous visual features among different categories**.

This paper proposes a simple but robust instance segmentation framework based on Cascade Mask R-CNN [3] with domain-adapted pretraining and modernized dual-backbone to address these problems. We first use the domain adaptive pretraining model to initialize the parameters to improve the stability. Secondly, we adopt a composite dual-branch back-

Fig. 2: This figure displays our inference results (the second row) for satellite imagery input (the first row).

bone structure to construct a more robust and discriminative feature extractor, which alleviates the improper segmentation for small-size instances and misclassification of minority categories. The dual-branch backbone can also facilitate multi-modal data fusion. Moreover, dedicated data augmentation pipeline with modified copy-paste, SWA[4] training strategy, and an instance segmentation model aggregation in inference jointly improve the precision of our fine-grained roof instance segmentor. Our visual instance segmentation results are displayed in Fig. 2, in which small and rare (Revolved) objects can both be detected accurately. Comprehensive experiments and ablation study demonstrate the superiority of proposed method, which has achieved a mAP_{50} of 50.6% and won the first place on track 1 test phase of DFC 2023.

2. METHODS

This section will detail the proposed fine-grained roof instance segmentation framework. The overall network structure diagram is shown in Fig. 3 armed with ConvNeXt V2 [5] based composite dual-backbone and domain adapted pretraining, modified copy-paste data augmentation.

2.1. Composite dual-backbone feature extractor

For the purpose of constructing a robust feature extractor, we adopt a dual-branch structure backbone containing two densely connected sub-backbones, which is validated in CBNet [6]. Intuitively, it is a flexible structure for both single-modality and muti-modality input. Feature maps from all higher stages of the auxiliary sub-backbone are nearest interpolated and added to lower-level stages of the lead sub-backbone, as illustrated in equation (1). This densely connected structure deeply fuses the high-level semantic information and low-level semantic information of two sub-backbones to boost our feature extractor. Furthermore, there are some competitive alternativs to our sub-backbone, such

as ConvNeXt V2 [5], Swin Transformer [7].

$$F_{lead}^{j} = G_{lead}^{j}(\sum_{i=j}^{L} N(F_{aux}^{i}) + F_{lead}^{j-1}) \tag{1}$$

Where L represents the total stage number, F_{aux}^{i}, F_{lead}^{j} represent the feature map of i_{th}, j_{th} stage in the auxiliary sub-backbone and lead sub-backbone, $G_{lead}^{j}(x)$ represents model block of j_{th} stage in lead the sub-backbone, N represents nearest interpolation function.

2.2. Domain adapted pretraining

Model weights initialization strategy plays a key role in the entire optimization. Typically, Imagenet22k pretrained weights can not only accelerate the convergence, but also improve the performance. In the training process, we attempt to employ a better initialization through domain adaptation pretraining. Fully convolutional masked autoencoder (FC-MAE) [5] is a self-supervised pretraining method for conv-based models, which is beneficial for the capability of domain adaption on specific dataset. Therefore, we pretrain the ConvNeXt V2 weights on the RGB modality dataset and then initialize our two sub-backbones with pretraining weights.

2.3. Modified copy-paste

Simple copy-paste [8] is an effective way of instance-level data augmentation. To further enrich the dataset, we propose a modified version which first crops instances from images, and then apply random resize, rotation, and flip to these pixel-level instances before pasting them onto the chosen images. These processes can hardly damage the semantic information in the nadir-viewing images. Synthetic images can be found in Fig. 4. To prevent overall distribution shift caused by synthetic images, we separate the training process into two phases: 1) training with modified copy-paste data for the first 90% total epochs; 2) fine-tuning without the modified copy-paste data for the last 10% total epochs. In addition our data augmentation pipeline contains some regular data augmentation strategies like random rotate, random resize, random crop, etc.

2.4. Model ensemble

To further enhance the performance of our model, we employ a model ensemble strategy to fuse several different models. For object detection, weighted bounding boxes fusion (WBF) [9] is a popular approach for model ensemble. We modify WBF as Weightd Segmentation Fusion(WSF), which is suitable for instance segmentation. First, we adopt the WBF strategy to obtain fused bounding boxes, then fuse masks from different models to get the final results.

Fig. 3: The overview training workflow and network structure diagram of our proposed roof instance segmentor.

Fig. 4: Synthetic images produced by data augmentation.

2.5. Additional training strategy

In order to address the issues of long-tail distribution of roof categories and distribution disparity between training and validation datasets, we try to build a more robust model through various training techniques.

Tactics for long-tail distribution: We adopt a balanced sample strategy to increase the probability of samples containing tail classes of being chosen. Additionally, we use seesaw loss [10] to reduce the gradients from negative samples on tail classes.

Stochastic Weight Average: We employ the SWA technique with cyclical learning rates to train for the next 12 epochs after all training epochs. Subsequently, we average these 12 weights to obtain our final model weights.

3. IMPLEMENTATION DETAILS

Fig. 5: Training process of only RGB modality input.

Single-modality input. In our experiments, training process of our best result for single RGB modality contains four steps described by Fig. 5. 1) We employ the FCMAE method to pretrain a ConvNeXt V2 model for 100 epochs, with Imagenet22k pretrained weights as initialization. We combine the the RGB parts of the training data and validation data as our pretraining dataset. The model is trained under AdamW optimizer with a base learning rate of 10^{-4} and decay to 10^{-6} during the whole process. The main data augmentation techniques including random resizing, random cropping, and random flipping, the image mask ratio is 0.6 and L2 loss is used for supervision. 2) After pretraining, we train and finetune our fine-grained instance segmetor for 45 epochs and 5 epochs. 3) During finetuning, simple data augmentation strategies that do not involve modified copy-paste are employed. 4) After completing all training and finetuning steps, another 12 epochs SWA cyclical training procedure are adopted.

Muti-modality input. Several experiments are conducted to explore the potential of multi-modal inputs consisting of RGB and SAR data. However, the results of the multi-modal approach are not as good as those obtained with the single RGB modality,

4. ABLATION STUDY

Table 1 shows the ablation study. Our best result mAP_{50} 50.6% is acquired by fusing several strong detectors trained under different hyperparameters and backbones with WSF. From our experiments, we find that SAR data does not enhance model performance as shown in Table 1*, accuracy drops significantly compare to single RGB modality input.

5. CONCLUSION

After comprehensive experiments, we build a robust model for building roof instance segmentation with domain adapted

Table 1: The details of our ablation study, CMR=Cascade Mask Rcnn, SCP=Simple Copy-Paste, MCP=Modified Copy-Paste, db=DUal-Backbone, DAP=Domain Adapted Pretraining, CNV2=ConvNeXt V2, WSF= Weighted Segmentation Fusion.

	Method	Development phase (mAP50)	Test phase (mAP50)
	CMR + db-swin-base	0.443	/
*	CMR + db-swin-base **+ SAR-input**	0.371	/
	CMR + db-swin-base**+ seesaw**	0.461	/
	CMR + db-swin-base + seesaw **+ SCP**	0.469	/
	CMR **+ db-CNV2-base-DAP**+ seesaw + SCP	0.476	/
	CMR + db-CNV2-base-DAP + seesaw + SCP **+ SWA**	0.485	/
	CMR + db-CNV2-base-DAP + seesaw **+ MCP** + SWA	0.496	0.426
	CMR + db-CNV2-base-DAP + seesaw **+ MCP+finetune** + SWA	/	0.441
	CMR + **db-CNV2-large-DAP** + seesaw + MCP+finetune + SWA	/	0.448
	WSF based model ensemble	0.5510	0.5060

pretraining and dual backbone by utilizing optical satellite images as input. The misalignment and heterogeneity between optical and SAR data bring difficulties for multimodual data fusion and performance improve. More exploration is needed to investigate whether SAR image data can actually contribute to this scenario and how to make full use of them in the future.

6. ACKNOWLEDGEMENT

7. REFERENCES

[1] Xingliang Huang, Libo Ren, Chenglong Liu, Yixuan Wang, Hongfeng Yu, Michael Schmitt, Ronny Hänsch, Xian Sun, Hai Huang, and Helmut Mayer, "Urban building classification (ubc)-a dataset for individual building detection and classification from satellite imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1413–1421.

[2] C. Persello et al., "2023 ieee grss data fusion contest: Large-scale fine-grained building classification for semantic urban reconstruction [technical committees]," *IEEE Geoscience and Remote Sensing Magazine*, vol. 11, no. 1, pp. 94–97, 2023.

[3] 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*, 2018, pp. 6154–6162.

[4] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sünderhauf, "Swa object detection," *arXiv preprint arXiv:2012.12645*, 2020.

[5] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie, "Convnext v2: Co-designing and scaling convnets with masked autoencoders," *arXiv preprint arXiv:2301.00808*, 2023.

[6] Tingting Liang, Xiaojie Chu, Yudong Liu, Yongtao Wang, Zhi Tang, Wei Chu, Jingdong Chen, and Haibin Ling, "Cbnet: A composite backbone network architecture for object detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 6893–6906, 2022.

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.

[8] Yuxin Zhang, Kai Chen, Xinggang Wang, Hongsheng Li, and Dahua Lin, "Simple copy-paste is a strong data augmentation method for instance segmentation," *arXiv preprint arXiv:2006.09845*, 2020.

[9] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva, "Weighted boxes fusion: Ensembling boxes from different object detection models," *Image and Vision Computing*, pp. 1–6, 2021.

[10] Jiaqi Wang, Wenwei Zhang, Yuhang Zang, Yuhang Cao andJiangmiao Pang, Tao Gong, Kai Chen, Ziwei Liu, Chen Change Loy, and Dahua Lin, "Seesaw loss for long-tailed instance segmentation," *arXiv preprint arXiv:2008.10032*, 2021.