Large-Scale Fine-Grained Building Classification and Height Estimation for Semantic Urban Reconstruction: Outcome of the 2023 IEEE GRSS Data Fusion Contest

Guozhang Liu, Baochai Peng, Ting Liu, Pan Zhang, Mengke Yuan, Chaoran Lu, Ningning Cao, Sen Zhang, Simin Huang, Tao Wang, Xiaoqiang Lu, *Graduate Student Member, IEEE*, Licheng Jiao, *Fellow, IEEE*, Qiong Liu, Lingling Li, *Senior Member, IEEE*, Fang Liu, Xu Liu, *Senior Member, IEEE*, Yuting Yang, *Graduate Student Member, IEEE*, Kaiqiang Chen, *Member, IEEE*, Zhiyuan Yan, Deke Tang, Hai Huang, *Member, IEEE*, Michael Schmitt, *Senior Member, IEEE*, Xian Sun, *Senior Member, IEEE*, Gemine Vivone, *Senior Member, IEEE*, Claudio Persello, *Senior Member, IEEE*, and Ronny Hänsch, *Senior Member, IEEE*

***Abstract*—This article presents the scientific outcomes of the 2023 Data Fusion Contest (DFC23) organized by the Image Analysis and Data Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society. The contest consists of two tracks investigating the fusion of optical and synthetic aperture radar data for: 1) fine-grained roof type classification and 2) height estimation. During the development phase, 1000 people registered for the contest, while at the end 55 and 35 teams competed during the test phase in the two tracks, respectively. This article presents the methods and results obtained by the first and second-ranked teams of each track. In Track 1, both winning teams leveraged pretraining, modern network architectures, model ensembles, and measures to cope with the imbalanced class distribution. The solutions to Track 2 are more diverse and are characterized by modern multitask learning approaches. The data of this contest is openly available to the community for further research, development, and refinement of machine learning methods.**

***Index Terms*—Convolutional neural networks, data fusion, deep learning, fine-grain building classification, transformers, monocular height estimation (MHE).**

I. Introduction

Buildings play a vital role in urban areas, yet the focus of research in the extraction and 3-D reconstruction from remote sensing data often neglects detailed information about the specific roof types of buildings. This oversight restricts the potential for in-depth analysis, particularly in the context of urban planning applications. Classifying building roof types at a fine-grained level using satellite imagery poses significant challenges due to the presence of ambiguous visual features within the images. Furthermore, the lack of datasets specifically designed for fine-grained building roof type classification exacerbates the difficulty of this task.

The 2023 IEEE GRSS Data Fusion Contest (DFC23), organized by the Image Analysis and Data Fusion Technical Committee (IADF TC) of the IEEE Geoscience and Remote Sensing Society (GRSS), the Aerospace Information Research Institute under the Chinese Academy of Sciences, the Universität der Bundeswehr München, and GEOVIS Earth Technology Company, Ltd. aims to push current research on building extraction, classification, and 3-D reconstruction towards urban reconstruction with fine-grained semantic information of roof types and height information.

Manuscript received 5 February 2024; revised 18 April 2024; accepted 7 May 2024. Date of publication 20 May 2024; date of current version 17 June 2024. *(Corresponding author: Claudio Persello.)*

Guozhang Liu, Baochai Peng, Ting Liu, Pan Zhang, Mengke Yuan, Chaoran Lu, Ningning Cao, Sen Zhang, Simin Huang, and Tao Wang are with the AI Research Institute, Piesat Information Technology Company, Ltd., Beijing 100094, China (e-mail: liuguozhang@piesat.cn; pengbaochai@piesat.cn; tliu068@gmail.com; whu_zp@163.com; mengkeyuan@foxmail.com; luchaoran@piesat.cn; shenlan0927@gmail.com; zhangsen.sanmu@gmail.com; herlik_h@hotmail.com; wangtao@piesat.cn).

Xiaoqiang Lu, Licheng Jiao, Qiong Liu, Lingling Li, Fang Liu, Xu Liu, and Yuting Yang are with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University, Xi'an, Shaanxi Province 710071, China (e-mail: luxiaoqiang5903@163.com; lchjiao@mail.xidian.edu.cn; 22171214880@stu.xidian.edu.cn; llli@xidian.edu.cn; f63liu@163.com; xuliu361@163.com; ytyang_1@stu.xidian.edu.cn).

Kaiqiang Chen, Zhiyuan Yan, and Xian Sun are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China (e-mail: chenkq@aircas.ac.cn; yanzy@aircas.ac.cn; sunxian@aircas.ac.cn).

Deke Tang is with the Geovis Technology Company Ltd., Beijing 101399, China (e-mail: tangdk@geovis.com.cn).

Hai Huang is with the Institute for Applied Computer Science, University of the Bundeswehr Munich, 85577 Neubiberg, Germany (e-mail: hai.huang@unibw.de).

Michael Schmitt is with the Department of Aerospace Engineering, University of the Bundeswehr Munich, 85577 Neubiberg, Germany (e-mail: michael.schmitt@unibw.de).

Gemine Vivone is with the Institute of Methodologies for Environmental Analysis, National Research Council, 85050 Tito, Italy (e-mail: gemine.vivone@imaa.cnr.it).

Claudio Persello is with the Department of Earth Observation Science, Faculty of Geo-information Science and Earth Observation (ITC), University of Twente, 7522 NH Enschede, The Netherlands (e-mail: c.persello@utwente.nl).

Ronny Hänsch is with the Department SAR Technology, German Aerospace Center (DLR), 82234 Weßling, Germany (e-mail: ronny.haensch@dlr.de).

Digital Object Identifier 10.1109/JSTARS.2024.3403201

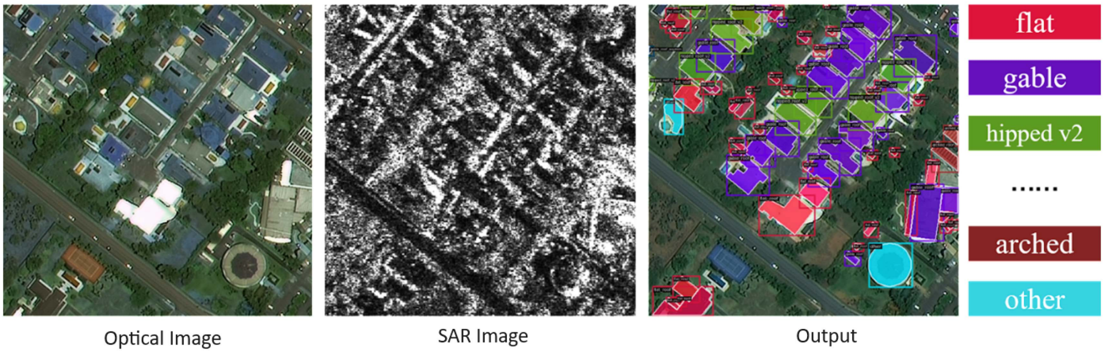

Fig. 1. Example image tile of the provided multimodal data (optical and SAR) for building detection and roof type classification.

To this aim, the DFC23 establishes a large-scale, fine-grained, and multimodal benchmark for the classification of building roof types [1]. It consists of two challenging competition tracks investigating the fusion of optical and synthetic aperture radar (SAR) data focusing on roof type classification and building height estimation, respectively. The data provided by the DFC23 includes several novel properties. They are as follows.

1) *Globally Distributed and Large-Scale:* Buildings are distributed across seventeen cities on six continents to cover a wide range of different building styles. This allows capturing the heterogeneity of cities in different continents with various landforms.
2) *Fine-grained Roof Type Categories:* The buildings are labeled according to a detailed (fine-grained) categorization of roof types. The DFC23 provides nearly 300k instances with twelve different types of building roofs, which renders building roof-type classification significantly more challenging.
3) *Multimodal Data:* To facilitate multimodal data fusion, optical imagery and SAR images are provided. The information captured by these different modalities can be jointly exploited, potentially resulting in the development of more accurate building extraction and classification models.

The contest is designed as a benchmark competition following previous editions [2], [3], [4], [5], [6], [7] and consists of two parallel tracks as follows.

1) *Track 1:* Building Detection and Roof Type Classification.
2) *Track 2:* Multitask Learning of Joint Building Extraction and Height Estimation.

Track 1: Building Detection and Roof Type Classification

This track focuses on the detection and classification of building roof types from high-resolution optical satellite imagery and SAR images. The SAR and optical modalities are expected to provide complementary information. The given dataset covers seventeen cities worldwide across six continents. The classification task consists of twelve fine-grained, predefined roof types. Fig. 1 shows an example.

Track 2: Multitask Learning of Joint Building Extraction and Height Estimation

This track defines the joint task of building extraction and height estimation. Both are fundamental tasks for building reconstruction. As in Track 1, the input data are multimodal optical and SAR satellite imagery. Building extraction and height estimation from single-view satellite imagery depend on semantic features extracted from the imagery. Multitask learning provides a potentially superior solution by jointly analyzing features from the different data sources and forming implicit constraints between multiple tasks compared to conventional single-mode methods. Satellite images are provided with reference data, i.e., building annotations and normalized Digital Surface Models (nDSMs). The participants are required to reconstruct building heights and extract building footprints. Fig. 2 shows an example.

II. Dataset

The images of the DFC23 dataset are acquired by the SuperView-1 (or "GaoJing" in Chinese), Gaofen-2 and Gaofen-3 satellites, with spatial resolutions of 0.5 m, 0.8 m, and 1 m, respectively. nDSMs provided for reference in Track 2 are produced from stereo images captured by Gaofen-7 and WorldView 1 and 2 with a ground sampling distance of roughly 2 m. Data was collected from seventeen cities on six continents to provide a large and representative dataset of high diversity regarding landforms, architecture, and building types. Roof type categories are organized according to twelve fine-grained roof type classes based on the geometry of the roof.

The data of this contest remains openly available to the community[1].

III. Contest Organization and Submissions

The contest consisted of two phases as follows.

Phase 1: Participants are provided training data and additional validation images (without corresponding reference data) to train and validate their algorithms. Participants can submit results for the validation set to the CodaLab competition websites

[1] [Online]. Available: https://ieee-dataport.org/competitions/2023-ieee-grss-data-fusion-contest-large-scale-fine-grained-building-classification

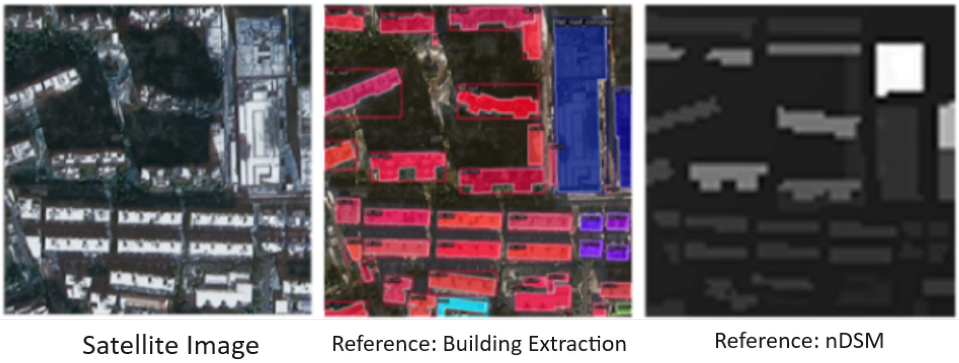

Fig. 2. Example of joint building extraction and height estimation.

TABLE I
REGISTRATION AND SUBMISSION STATISTICS OF THE TWO TRACKS

	Development phase		Test phase	
	Registrations	Submissions	Teams	Submissions
Track 1	680	4921	55	420
Track 2	320	1692	35	274

for Track 1[2] and Track 2[3] to get feedback on their performance. The performance of the best submission from each account will be displayed on the leaderboard. In parallel, participants submit a short description of the approach used to be eligible to enter Phase 2.

Phase 2: Participants receive the test dataset (without the corresponding reference data) and submit their results within seven days from the release of the test dataset. After evaluation of the results, four winners for each track are announced. We received 1000 registrations at the CodaLab competition website during the development phase (see Table I).

For Track 1, there were 680 unique registrations at the CodaLab competition website during the development phase, and 55 teams entered the test phase after screening the descriptions of their approaches submitted by the end of the development phase. 4921 submissions were received during the development phase, with active participation from all registered teams. During the test phase, the maximum number of submissions per team was limited to 5 per day, and 420 submissions were received.

For Track 2, there were 320 unique registrations at the CodaLab competition website during the development phase, and 35 teams entered the test phase after screening the descriptions of their approaches submitted by the end of the development phase. In total, 1692 submissions were received during the development phase, with active participation from all registered teams. During the test phase, the maximum number of submissions per team was limited to 5 per day, and 274 submissions were received.

Participants of the two tracks come from 32 countries. The specific number and proportion of participants in each country are shown in Fig. 3.

The first- to fourth-ranked teams were awarded as winners of the DFC2023 for each track and were invited to present their solutions during the 2023 IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2023). However, the fourth-ranked team of Track 1 decided to withdraw from the contest. In the following, we list the winning teams of the DFC2023 in Track 1 as follows.

1) *1st place: PIESAT-AI* team—Guozhang Liu, Baochai Peng, Ting Liu, Pan Zhang, Mengke Yuan, Chanran Lu, Ningning Cao, Sen Zhang, Simin Huang, Tao Wang from PIESAT Information Technology Company, Ltd., Beijing, China [8].
2) *2nd place: IPIU-XDU* team—Xiaoqiang Lu, Licheng Jiao, Qiong Liu, Lingling Li, Fang Liu, Xu Liu, Yuting Yang from Xidian University, China [9].
3) *3rd place: carryhjr* team—Jiarui Hu (Wuhan University), Zijun Huang (Guangdong University of Technology), Fei Shen (Nanjing University of Science and Technology), Dian He (Tsinghua University), Qingyu Xian (Tsinghua University) [10].

and in Track 2 as follows.

1) *1st place: PIESAT-AI* team—Chaoran Lu, Ningning Cao, Pan Zhang, Ting Liu, Baochai Peng, Guozhang Liu, Mengke Yuan, Sen Zhang, Simin Huang, Tao Wang, from PIESAT Information Technology Company, Ltd., Beijing, China [11].
2) *2nd place*: *IPIU-XDU* team—Xiaoqiang Lu, Licheng Jiao, Qiong Liu, Lingling Li, Fang Liu, Xu Liu, Yuting Yang from Xidian University, China [12].
3) *3rd place: ZheWang* team—Yuxuan Guo, Zhe Wang from Wuhan University, China [13].
4) *4th place: carryhjr* team—Jiarui Hu (Whuhan University), Zijun Huang (Guangdong University of Technology), Fei Shen (Nanjing University of Science and Technology), Dian He (Tsinghua University), Qingyu Xian (Tsinghua University) [14].

The two best-ranked teams in both tracks are invited to provide a deeper discussion on their respective approaches to win the DFC23 in this article.

IV. Track 1—First Place: Team PIESAT-AI

A. Method

Team PIESAT-AI introduced an effective instance segmentation framework designed to tackle the complex challenges of

[2] CodaLab Track 1: https://codalab.lisn.upsaclay.fr/competitions/8987
[3] CodaLab Track 2: https://codalab.lisn.upsaclay.fr/competitions/8988

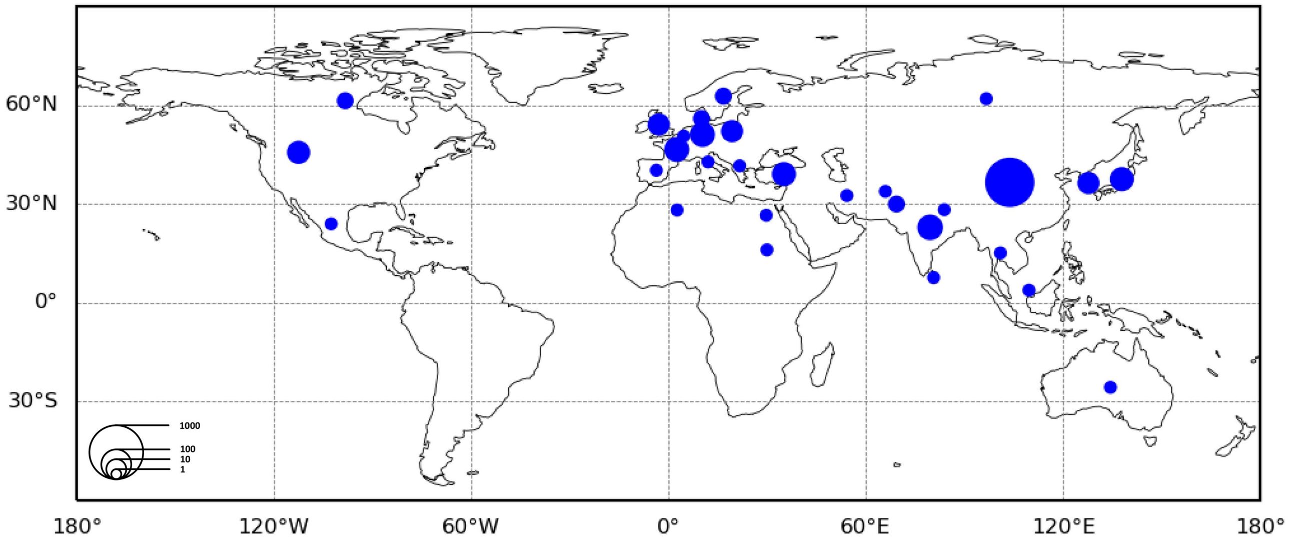

Fig. 3. Nationality statistics of participants of the two tracks.

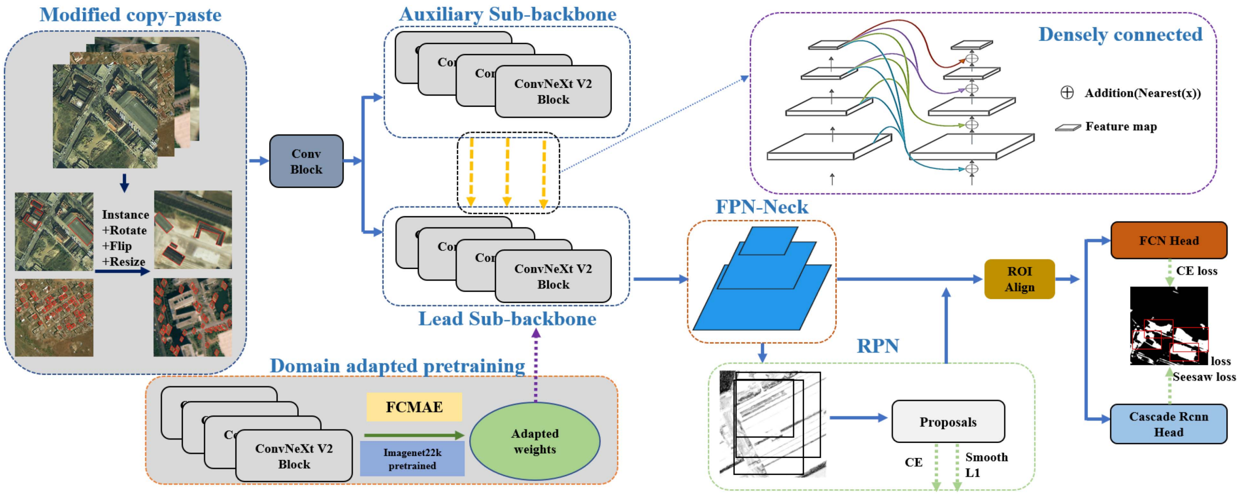

Fig. 4. Overview training workflow and network structure of the Track 1 winning approach.

fine-grained building roof classification. The approach builds upon the Cascade Mask R-CNN [15] architecture and is enhanced with domain-adapted pretraining and a modernized dual-backbone structure. These techniques collectively address a range of issues, such as the significant interclass imbalance of roof types, the diversity of building styles of global cities, and the distinction of imaging satellites. The overall structure is shown in Fig. 4. To improve training stability and to realize effective weight initialization, a domain-adaptive pretraining model is employed to optimize the parameter initialization. A composite dual-branch backbone structure is integrated as a more robust and discriminative feature extractor. The newly designed backbone mitigates problems associated with small-size instance segmentation and classification of minority categories. The dual-branch backbone is applicable for optical, as well as SAR inputs, which enhances the multimodal data fusion performance. A tailored data augmentation pipeline is adopted that incorporates modified copy-paste techniques, Stochastic Weight Average (SWA [16]) training strategy, and a novel instance segmentation model aggregation method during inference. These strategies enhance the precision of the fine-grained roof instance segmentation.

Domain adapted pretraining: The strategy for initializing model weights holds a pivotal role within the optimization framework. The utilization of pre-trained weights from ImageNet-22 k has two advantages: 1) convergence acceleration and 2) overall performance improvement. The Fully Convolutional Masked Autoencoder (FCMAE) [17] demonstrates a superior performance by improving the domain adaptation capacity for specific datasets with self-supervised pretraining tailored to convolution-based models. To this end, the pretraining weights of ConvNeXt V2 on the RGB modality dataset are leveraged and form the bedrock for initializing the dual subbackbones. This elaborated weight initialization, coupled with the FCMAE pretraining, boosts the detection performance and domain-specific adaptability.

Composite dual-backbone feature extractor: To obtain a robust feature extractor, a dual-branch backbone is used encompassing two interconnected subbackbones. This structure is validated in CBNet [18]. The configuration exhibits remarkable adaptability across both single-modality and multimodality inputs. The lead subbackbone and auxiliary subbackbone are densely connected, where high-level low-resolution feature maps undergo nearest interpolation before the fusion with low-level high-resolution ones. The fusion scheme is defined by

$$F^j_{\text{lead}} = G^j_{\text{lead}}\left(\sum_{i=j}^{L} N(F^i_{\text{aux}}) + F^{j-1}_{\text{lead}}\right) \tag{1}$$

where L represents the number of downsampling stages of the subbackbone, F^i_{aux}, F^j_{lead} represent the feature map of the ith and jth stages in the auxiliary subbackbone and lead subbackbone, respectively, $G^j_{\text{lead}}(x)$ represents a model block of the jth stage in lead subbackbone, and N represents the nearest neighbor interpolation operator.

The composite dual-backbone design promotes the extraction of global context and local details and achieves notable feature extraction capability.

Modified copy-paste: Copy-paste [19] is a simple yet powerful technique for instancing-level data augmentation. To enhance the diversity of the dataset, PIESAT-AI presents an improved modification of this approach. The design involves an initial extraction of instances from images, followed by the application of randomized resizing, rotation, and flipping to these pixel-level instances. Subsequently, these augmented instances are seamlessly integrated into the selected images. This synthesizing process preserves the authenticity of the nadir-viewing images. Due to the risk that synthetic images might introduce distributional shifts, PIESAT-AI implements a two-phased training pipeline. In the initial phase, the model is trained using the dataset enriched by the modified copy-paste technique, covering 90% of the total training epochs. The subsequent phase focuses on fine-tuning, during which the model refines its performance over the remaining 10% of epochs without the modified copy-paste data augmentation. In addition to the distinctive augmentation strategies, PIESAT-AI's data augmentation pipeline encompasses various routine techniques, including random rotation, resizing, cropping, etc.. This holistic augmentation approach synergizes diverse methodologies to ensure the model's adaptability to a wide range of real-world scenarios. *Additional training strategy:* To effectively tackle the challenges arising from the long-tailed distribution of roof categories and the distributional disparities between the training and validation datasets, the approach focuses on the construction of a more generalized model involving the strategic incorporation of diverse training techniques to strengthen the model's performance as follows.

1) *Tactics for long-tail distribution:* The statistics of the Track 1 training dataset (see Fig. 5) reflect the practical difficulty that there are significant proportion disparities for different classes. Thus, a balanced sampling strategy is utilized with the aim of increasing the probability of selecting minority-class samples. Furthermore, the see-saw loss [20] mechanism is employed to alleviate the impact of gradients stemming from less common negative class samples.
2) *SWA:* The SWA [16] technique with cyclical learning rates is employed during 12 further epochs after all training epochs are completed. The weights of the models of these 12 epochs are averaged to obtain the final model weights.
3) *Model ensemble:* To improve the generalization capability of the model, a model ensemble strategy is employed to fuse several different models. When it comes to object detection, the widely adopted methodology of weighted boxes fusion (WBF) [21] has emerged as a prevalent strategy for orchestrating model ensembles. PIESAT-AI introduces Weighted Segmentation Fusion (WSF), which is suitable for instance segmentation: First, the WBF strategy is adapted to obtain fused bounding boxes. Then, the masks from different models are fused to obatin the final results.
4) *Multimodality input:* Several experiments have been carried out to evaluate the potential of employing multimodal inputs, i.e., using both RGB and SAR data. However, the outcomes of the multimodal approach fail to match the high performance achieved through utilizing the RGB modality only.

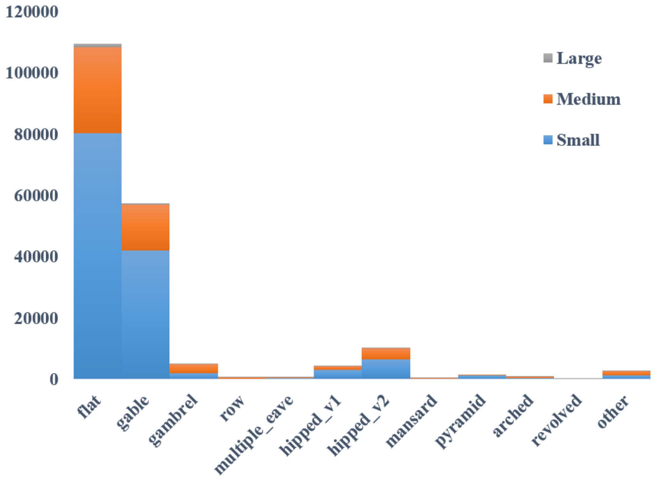

Fig. 5. Distribution of instance pixels among predefined roof categories of DFC 2023. Gray, orange, and blue represent large, medium and small instances, respectively.

B. Results

Table II contains the accuracy of different experiments. These ablation studies show that the best result with an mAP_{50} of 50.6% is achieved by fusing several strong detectors trained under different hyperparameters and backbones with WSF. Notably, SAR data does not enhance model performance, but accuracy drops significantly compared to the single RGB modality input. The visualization results depicted in Fig. 6 demonstrate the model's capability to accurately identify and delineate the boundaries of roofs, even for small-scale and poorly defined

TABLE II
DETAILS OF TEAM PIESAT-AI'S 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

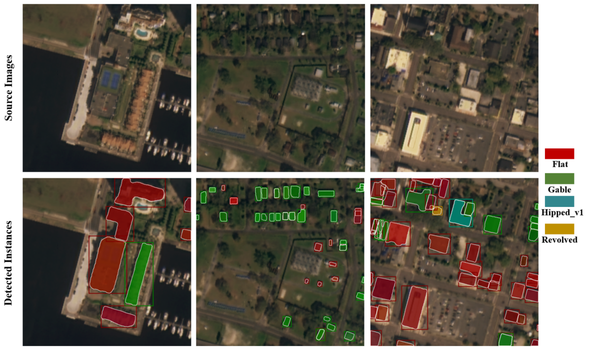

Fig. 6. Qualitative results of the first place of Track 1.

targets. Despite these promising results, some instances of misclassification persist and indicate the potential of further refinement and optimization of the model.

C. Discussion

An extensive series of experiments led to a robust and successful model tailored for fine-grained building roof instance segmentation. The approach integrates domain-adapted pretraining and a dual backbone framework. It utilizes optical satellite imagery as the primary input source. The inherent misalignment and heterogeneity between optical and SAR data presents great challenges to enhance performance in a multimodal data fusion setting.

V. TRACK 1—SECOND PLACE: TEAM IPIU-XDU

A. Method

To address the problem of a long-tailed class distribution, general data insufficiency, interclass similarity, and intraclass differences in the DFC23 dataset, team IPIU-XDU proposes a three-stage learning framework consisting of pretraining, supervised training, and semisupervised training shown in Fig. 7.

Pretraining: Several computer vision approaches make extensive use of models pretrained on ImageNet. However, there are considerable differences between natural and remote sensing images, such as less-defined shapes and a rather dispersed distribution of objects in remote sensing imagery. Team IPIU-XDU cropped every building box to create a fine-grained classification dataset used to fine-tune the ImageNet pretrained model into a building-specific model. BEiTv2-L [22] is used as a feature extractor. Following the process of fine-tuning, a vision transformer adapter [23] is used to enhance the representations of the model.

TABLE III
CLASS-INSTANCE AND CLASS-IMAGE DISTRIBUTIONS

Category	id	N_{ins}	r_{ins}	N_{img}	r_{img}
Flat	1	109460	0.5635	3461	0.9304
Gable	2	57426	0.2956	2965	0.7970
Gambrel	3	5087	0.0262	524	0.1409
Row	4	943	0.0049	477	0.1282
Multiple	5	738	0.0038	176	0.0473
Hipped_v1	6	4320	0.0222	1083	0.2911
Hipped_v2	7	10220	0.0526	1422	0.3823
Mansard	8	510	0.0026	339	0.0911
Pyramid	9	1520	0.0078	631	0.1696
Arched	10	833	0.0043	536	0.1441
Revolved	11	154	0.0008	138	0.0371
Other	12	2872	0.0148	1099	0.2954

Supervised training: The following three techniques aim to improve the supervised model to derive a better result for semisupervised training.

1) *Long-tailed data distribution:* Table III shows that the UBC [24] dataset is severely imbalanced. To address this issue, image-level resampling and object-level balanced copy-paste strategies are introduced. Specifically, the number N_{ins} of each class instance is used to compute the ratio r_{ins} of class instances and the ratio r_{img} of class images. Two thresholds $\delta_{\text{ins}} = 0.005$ and $\delta_{\text{img}} = 0.1$ are defined to determine whether to copy-paste and resample, respectively, i.e., depending on r_{ins} or r_{img} being less than δ_{ins} or δ_{img}.
2) *Cross-modal fusion:* The key obstacles in building detection and classification include occlusions, shadows, and diverse roof types of urban coverings. In terms of the benefits and drawbacks of various forms of data, RGB images provide excellent spatial resolution and textural information but are vulnerable to weather, whereas SAR images work in all weather conditions and highlight foreground items. As a result, using them both augments the input feature space and might potentially improve the resulting

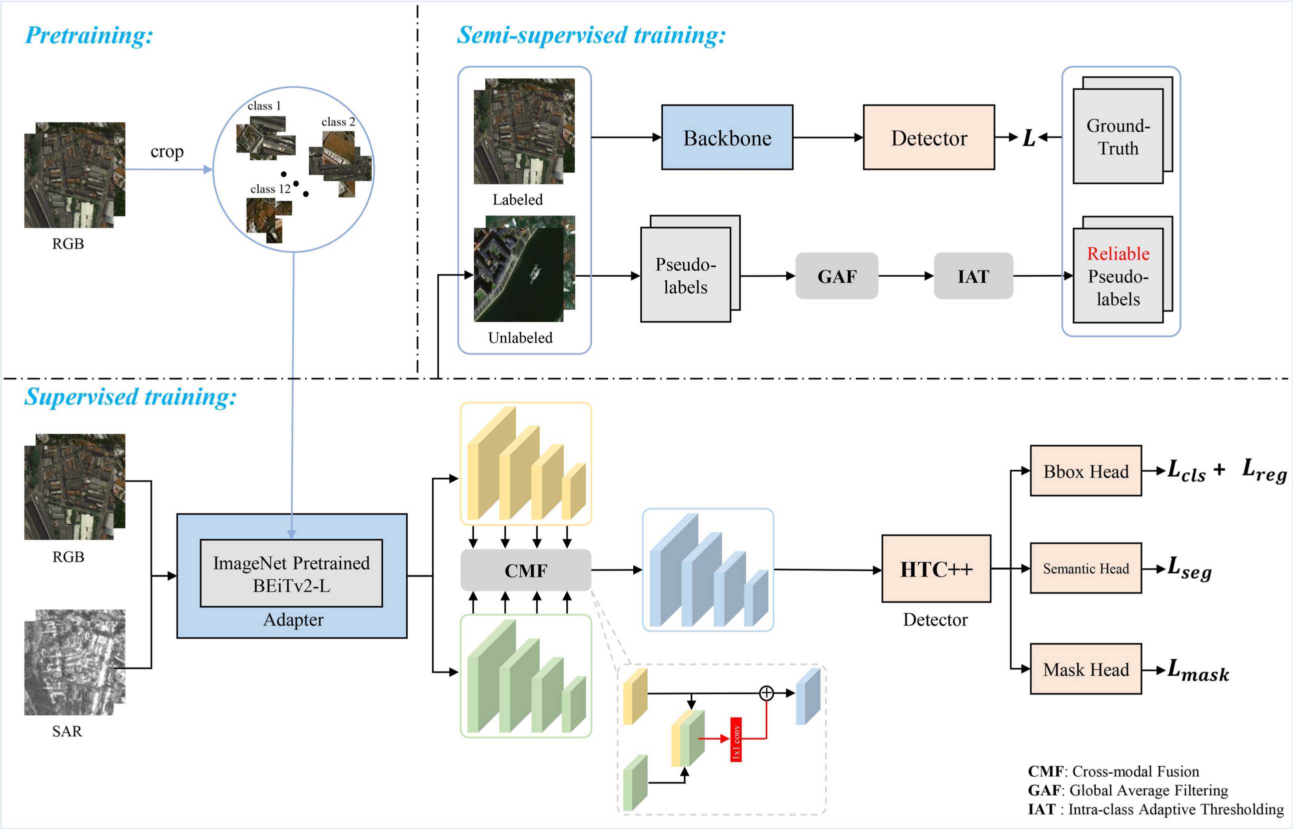

Fig. 7. Illustration of the three-stage learning framework of the second place for Track 1.

model. To that end, a simple but effective cross-modal fusion (CMF) module is created that independently concatenates hierarchical features extracted from RGB and SAR. Feature fusion and channel dimensionality reduction are achieved by using 1 × 1 convolutions and residual connections to improve the RGB feature representation.

3) *MHEM loss and training optimization:* The majority of currently used fine-grained visual classification techniques overfit challenging examples in the training set but do not learn to generalize to unseen examples in the test set. In order to effectively modulate the hard instances and encourage the model to avoid overfitting them, a moderate hard example modulation (MHEM) loss is incorporated [25]. Except for the bounding box classification head loss L_{cls}, the losses of the bounding box regression head L_{reg}, semantic head L_{seg}, and mask head L_{mask} are set according to [26].

Semisupervised training: Semisupervised learning (SSL) for visual dense prediction in the downstream tasks has been established to be effective [27], [28], [29]. Self-training, which is a key branch of SSL, can help a model that was only trained on a small amount of labeled data by expanding it to new situations and training samples. However, because of the complex unlabeled data and model bias, it suffers greatly from noise being present in the initial pseudolabels. To this end, two core denoising techniques are proposed to explore a safe boundary for using unlabeled data: 1) Image-level denoising by global average filtering (GAF) and 2) object-level denoising by intraclass adaptive thresholding (IAT) are performed in a global-to-instance manner.

1) *Global average filtering:* Given the ith unlabeled image x_i and a trained model F, the initial pseudolabels of x_i can be formulated as $y_i = F(x_i)$. The maximum total number of image objects is set to $N_{\text{dets}} = 2000$. The pseudolabel of the jth object in y_i consists of a corresponding confident score s_j and a 4-D coordinate representing the position. An image-level confidence S_{x_i} of x_i is computed as

$$S_{x_i} = \frac{1}{N_{\text{dets}}} \sum_{j=1}^{N_{\text{dets}}} s_j. \tag{2}$$

A threshold $\delta_{\text{GAF}} = 0.01$ is used to determine whether to drop x_i depending on S_{x_i} being less than δ_{GAF}.

2) *Intraclass adaptive thresholding (IAT):* For class c, a class-level confidence $S^c_{x_i}$ of x_i can be formulated as

$$S^c_{x_i} = \frac{1}{N^c_{\text{dets}}} \sum_{j=1}^{N^c_{\text{dets}}} s^c_j \tag{3}$$

where N^c_{dets} represents the number of objects belonging to class c. A majority class m_i of x_i can be derived by $m_i = \arg\max(\mathbf{S}^c_{x_i})$. For m_i, by comparison with a threshold

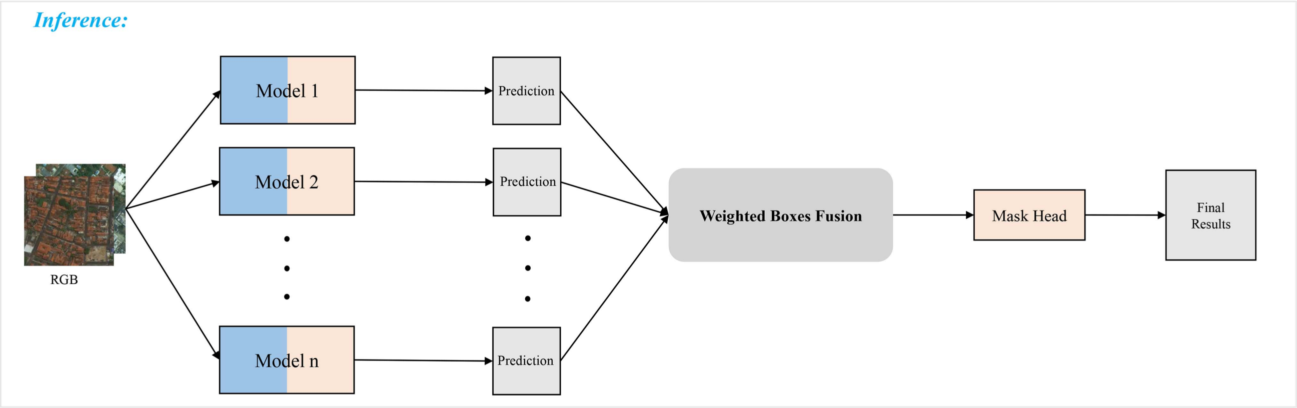

Fig. 8. Two-stage test-time augmentation results ensemble pipeline used by the second place of Track 1.

$\delta_{\text{IAT}} = 0.2$, the ratio r_i is computed as

$$r_i = \frac{1}{N_{\text{dets}}^{m_i}} \sum_{j=1}^{N_{\text{dets}}^{m_i}} \mathbb{1}\left(s_j^{m_i} \geq \delta_{\text{IAT}}\right) \tag{4}$$

where $\mathbb{1}(\cdot)$ is the indicator function. Finally, for any other class c $(c \neq m_i)$, $\mathbf{S}_{x_i}^c$ is sorted in descending order and only the top r_i objects are kept ensuring reliable pseudolabels with the same confidence-level. The corresponding thresholds can be selected for each class and each unlabeled image in an adaptive manner.

Two-stage results ensemble pipeline: As shown in Fig. 8 , for n predictions from n trained models, (WBF [21]) is used to ensemble the corresponding bounding boxes predictions, resulting in more accurate proposals. These are then passed to the best model's mask head to perform segmentation, deriving the final result. Fifteen models trained with different hyperparameters are assigned equal weights for integration in the test phase of the contest.

B. Results

All experiments are conducted on 8 NVIDIA V100. The pretraining setting is based on [22], except for batch size and learning rate using a linear scale. For supervised and semisupervised training, the model is optimized by AdamW with a base learning rate of 0.0002, and batch size of 8. The training image base size is set to 1024–1536, and the test multiscale size is increased by 0, 64, 128, 192, 256 based on the base size. Random flipping contains horizontal and vertical flipping, both with a probability of 0.25. The training assigner is set to three times the default setting. The test-time nonmaximum suppression uses a score threshold of 0.00001 and a maximum object number of 2000.

To investigate each component of the proposed method, extensive experiments are performed on the UBC dataset. Table IV shows the results which led to the first place ($\text{AP}_{50} = 0.559$) in the development phase and second place ($\text{AP}_{50} = 0.495$) in the test phase. Table V shows the effectiveness of the three supervised training strategies.

TABLE IV
Ablation Study on Various Components of Track 1's Second Place

Baseline	Pre	Sup	Semi	TTA	AP_{50}(val)	AP_{50}(test)
✓					0.438	–
✓	✓				0.447	–
✓	✓	✓			0.508	–
✓	✓	✓		✓	0.530	–
✓	✓	✓	✓		0.534	–
✓	✓	✓	✓	✓	**0.559**	**0.495**

Pre: Pretraining. Sup: Supervised training. Semi: Semisupervised training. TTA: Test-time augmentation.

TABLE V
Ablation Study on the Three Supervised Training Strategies of Track 1's Second Place

Baseline + Pre	BCP	CMF	MHEM	AP_{50}(val)
✓				0.447
✓	✓			0.489
✓	✓	✓		0.501
✓	✓	✓	✓	**0.508**

Pre: Pretraining. BCP: Balanced copy-paste. CMF: Cross-modal fusion. MHEM: Moderate hard example modulation loss.

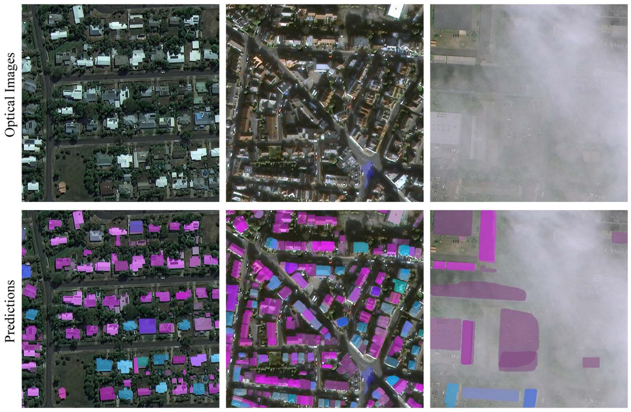

Fig. 9. Qualitative results of the second place of Track 1.

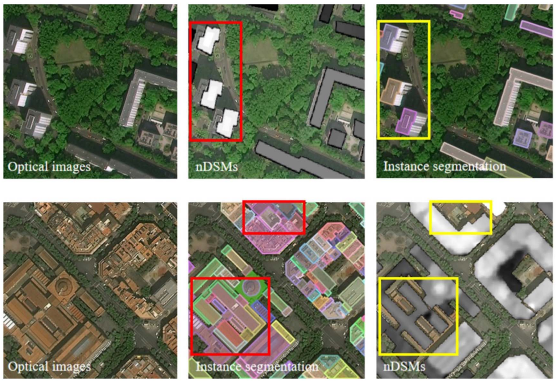

Fig. 10. Misaligned Example in DFC2023 Dataset.

Fig. 9 visually demonstrates that the method of Track 1's second place can accurately extract and identify various building instances. Moreover, thanks to the cross-modal fusion module, the method is capable of mitigating the negative effects caused by cloud occlusion.

C. Discussion

Team IPIU-XDU addressed several challenges that hinder a good performance of building classification detectors, including long-tailed class distributions, data insufficiency, interclass similarity, and intraclass differences. Specifically, IPIU-XDU proposed a three-stage training framework consisting of pretraining on the cropped buildings, supervised training with a strong vision transformer adapter with added cross-modal fusion module, and semisupervised training with two image- and object-level denoising techniques. In addition, a two-stage test-time ensemble pipeline is designed to further boost the performance and generalization of the model. Extensive experiments on the DFC23 data demonstrate the effectiveness of the method.

VI. Track 2—First Place: Team PIESAT-AI

A. Method

As depicted in Fig. 10, there are several spatial misalignments between building footprints and stereo-reconstructed nDSM height labels in the DFC2023 datasets. To address this issue, first, a Height-hierarchy Guided Dual-decoder Network (HGDNet) is proposed to estimate building height. In contrast to methods like PopNet [30] and SCENet [31], which rely on high-quality aligned semantic labels for height estimation, HGDNet performs the joint task of height estimation and hierarchical classification by utilizing hierarchical information about heights. Under the guidance of a synthesized discrete height-hierarchical nDSM, an auxiliary height-hierarchical building extraction branch enhances the height estimation branch with implicit constraints, yielding an accuracy improvement of more than 6% on the DFC2023 Track 2 dataset. Second, to achieve a precise building instance segmentation, an additional two-stage cascade architecture is adopted to extract individual building contours. The improved feature extraction and the fusion of outcomes from distinct structural models significantly contribute to the performance increase.

Height Estimation: The architecture of HGDNet is shown in Fig. 11. A ConvNeXt V2 base is adopted as the encoder module. The dual decoders consist of the height estimation branch and the height-hierarchical segmentation branch. They are as follows.

1) *Height Estimation Branch:* Building heights exhibit an uneven distribution, primarily concentrated within the range of 0–50 m and even as low as in the range of 0–10 m. To address this issue, a logarithmic transformation is applied to the nDSMs labels to approximate a Normal distribution as

$$DSM_{\text{renorm}} = \frac{\ln(\text{nDSMs})}{\max(\ln(\text{nDSMs}))}. \tag{5}$$

UPerNet [32] is employed as the foundational decoding layer. The up-down pathway and lateral connections work cooperatively to enhance multiscale features extracted from the backbone. These augmented features are subsequently upsampled to a predetermined scale and then fused. This fusion process generates the final feature for accurate height estimation. Additionally, to accommodate the height regression task, an additional one-channel convolutional layer is appended to the height branch, followed by a sigmoid layer at the end of the network.

2) *Height-hierarchical Segmentation Branch:* An additional height-hierarchical segmentation branch is seamlessly integrated into the network architecture to enhance the precision of height estimation. This branch leverages the same UPerNet decoder employed in the height estimation branch. However, the fused features are fed to a single convolutional layer comprising n channels, where n corresponds to the number of height hierarchies. The proposed methodology avoids the direct employment of instance segmentation annotations featuring only a single class. Instead, the nDSM is partitioned into a discrete hierarchy. This partitioning is achieved through the distribution analysis of nDSMs and the application of clustering algorithms. Subsequently, nDSMs are classified into n classes, thereby establishing height-hierarchy labels. These distinct labels are used to steer the process of building height estimation.

3) *The Weighted Loss Function:* The height-hierarchical segmentation branch uses a cross-entropy loss while the height branch uses a smoothed $L1$ loss. Given the disparity in magnitudes between the different loss functions, weighting coefficients α, β are introduced to balance the influence of the two branches. The final form of the loss function is

$$Loss = \alpha \cdot L_{CE} + \beta \cdot L_{\text{smooth}_{L1}}. \tag{6}$$

Building Extraction: To extract accurate contours of buildings, various two-stage cascade detection frameworks and backbones are employed as follows and is shown in Fig. 11.

1) *Cascade Framework:* The approach integrates several efficient cascade networks, including Cascade Mask R-CNN [33] and HTC++[34]. These networks perform target

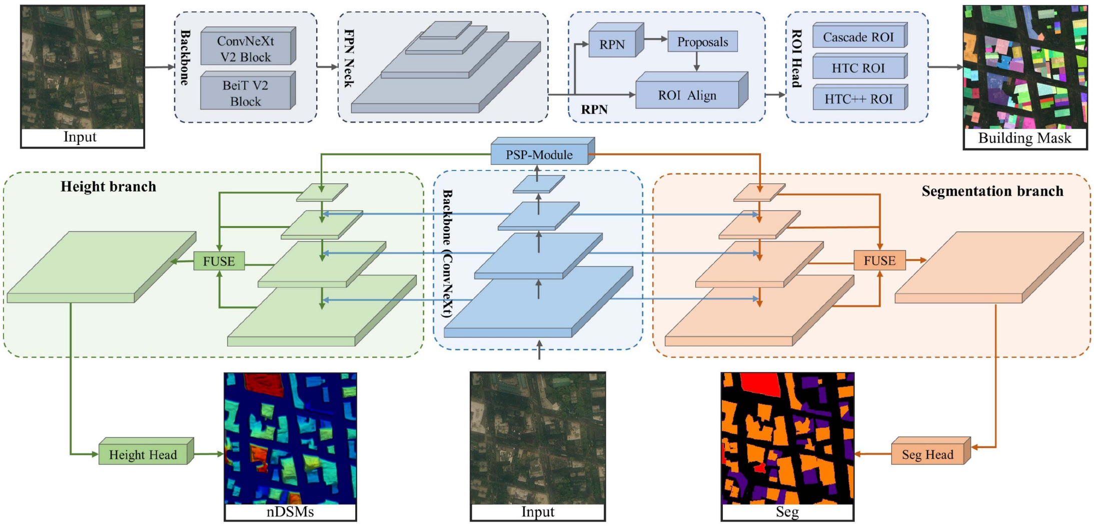

Fig. 11. Overview of the building extraction (top) and height estimation (bottom) framework proposed for the winning solution in Track 2.

extraction in two stages. The initial stage emphasizes locating the targets to generate proposals and then classifying, segmenting, and refining the proposals.

2) *Backbone:* After conducting comparative experiments, ConvNeXt V2-Base [17] and BEiT V2 [22] networks are selected to extract semantic features. These backbone networks extract profound features across distinct scales from the input data, ensuring robust extraction for buildings of various styles.

3) *Loss Function:* To enhance the fidelity of bounding box prediction, the model employs the smoothed $L1$ and GIOU loss functions. For building mask learning, the cross-entropy loss function is employed.

4) *Model Ensemble:* A modified version of WBF ensembles all results from different models. First, WBF is employed to fuse bounding boxes. Subsequently, the segmentation masks are combined. Finally, the fused masks are cropped using the merged boxes. This strategy facilitates the integration of diverse models and is applicable for multiscale scenarios.

B. Results

The performance of HGDNet is compared with other methods. Table VI shows the specific quantitative comparison. Notably, while "ConvNextv2+Uper" and the proposed "HGDNet:ConvNeXt V2" differ only in the presence of a height-hierarchical semantic branch, the discernible 6% performance gap underscores the clear effectiveness of the height-hierarchical semantic branch.

The accuracy metrics of each model are shown in Table VII. The optimal performance is attained through a selection of seven models, not necessarily the top seven highest-scoring ones, but rather a set encompassing diverse architectural configurations.

TABLE VI
COMPARISON OF TEAM PIESAT-AI'S METHOD AND OTHER METHODS ON THE VALIDATION SET

Method	δ_1	δ_2	δ_3
Baseline[4]	0.7042	0.7613	0.7922
DSMNet [35]	0.7339	0.7866	0.8110
SCENet-50 [31]	0.7731	0.8214	0.8438
ConvNextv2+Uper	0.7313	0.7822	0.8084
HGDNet:Without seg branch	0.7313	0.7822	0.8084
HGDNet:ResNet-50	0.7824	0.8037	0.8502
HGDNet:ConvNeXt V1 Base	0.7930	0.8371	0.8544
HGDNet:ConvNeXt V2 Base	**0.7966**	**0.8383**	**0.8581**

TABLE VII
METRICS FOR SEVEN BUILDING EXTRACTION MODELS AND THEIR FUSION RESULT ON THE TEST SET

Method	id	AP_{50}	mAP
HTC++_BEiT V2	1	0.730	0.402
HTC++_ConvNeXt V2-Base	2	0.710	0.388
HTC_BEiT V2	3	0.722	0.400
Cascade_BEiT V2	4	0.751	0.428
Cascade_ConvNeXt V2-Base	5	0.737	0.422
HTC_ConvNeXt V2-Base	6	0.732	0.420
Cascade_ ConvNeXt V2-Base	7	0.721	0.410
WSF based model ensemble		**0.773**	**0.450**

Fig. 12 illustrates the high performance of the proposed method in building extraction and height estimation. For downstream tasks such as 3-D reconstruction of buildings, it is essential to further refine the segmentation of building edges and to ensure the smoothness of building heights.

C. Discussion

A straightforward yet highly efficient architecture termed HGDNet is introduced for accurate building height estimation.

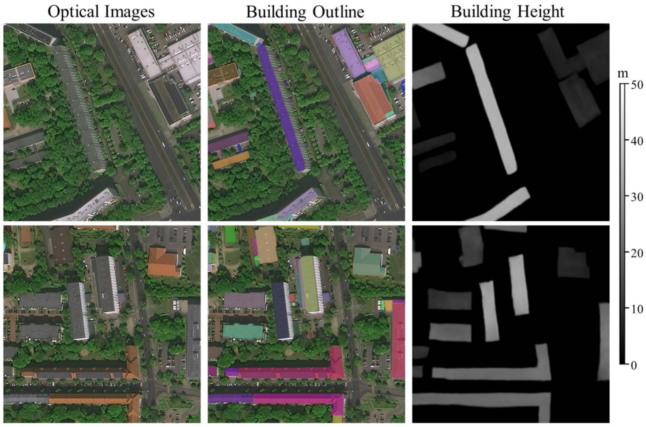

Fig. 12. Qualitative results of the first place of Track 2.

Using an advanced backbone, the architecture unifies semantic segmentation and height estimation tasks and leverages shared multiscale features and implicit constraints to enhance the overall performance. Furthermore, the fusion of outcomes from multiple two-stage instance segmentation models boosts the model's capability to discern buildings of various styles. The efficacy of the proposed approach is validated through extensive experimentation on the DFC2023 dataset.

VII. Track 2—Second Place: Team IPIU-XDU

A. Method

To address the problems of data limitation and foreground-background confusion of the DFC23 datasets, team IPIU-XDU proposes a novel trident cooperation network (TCNet) to perform end-to-end building extraction and height estimation using RGB and SAR data, as shown in Fig. 13.

Trident cooperation network: The proposed TCNet contains a CMF module and three task-specific heads [26], [36], [37]. To perform end-to-end building extraction and height estimation, the shared encoded features play an essential role in multitask learning. Therefore, a simple yet effective CMF module is proposed to derive more representative features. To perform each visual subtask, the following three task-specific decoders are introduced.

1) *Hybrid task cascade (HTC):* HTC [26] is an efficient cascade architecture for instance segmentation, which is mainly used in natural image interpretation. The key components include a cascade structure that detects instances at different scales and refines the results from the previous stages, a semantic segmentation branch that helps to improve the detection accuracy of small objects in crowded scenes, and a feature alignment module that aligns the features of the semantic segmentation and the bounding box regression tasks to generate spatial context to better utilizing the shared features. Based on this strong detection baseline, a robust training pipeline consisting of single-scale training is constructed leveraging random flipping and edge enhancement as used in [28]. The losses of the bounding box classification head L_{cls}, bounding box regression head L_{reg}, semantic head L_{seg}, and mask head L_{mask} are defined according to [26].
2) *All-MLP:* Unlike traditional segmentation decoder architectures that use convolutional layers, the All-MLP [36] decoder utilizes a series of pure multilayer perceptrons to decode the features from the encoder, which avoids the hand-crafted and computationally demanding components. This additional binary segmentation branch is introduced to highlight the foreground predictions and to suppress the wrong background predictions in the height estimation maps. The used loss L_{seg} is the standard cross-entropy loss.
3) *PixelFormer:* Monocular height estimation, similar to monocular depth estimation (MDE), aims to predict pixel-wise height given a single-view image. PixelFormer [37] is used as an efficient MDE network. It proposes a skip attention module to fuse the encoder and decoder features and a bin center predictor module that estimates bin centers adaptively per image using the global information from the coarsest-level feature maps. Since PixelFormer uses a typical encoder-decoder architecture, its original input features are replaced with the shared hierarchical features extracted from RGB and SAR data. Its optimization target is further weighted by the binary mask predicted by the All-MLP segmentation head to force the model to focus on the foreground building objects. The height estimation loss L_{hei} is the scale-invariant loss used in [37].

Finally, the total loss is the uniformly weighted sum of each subtask loss.

Self-training: Semisupervised learning (SSL) has been proven to be effective for downstream dense prediction visual tasks [27], [28], [29]. Self-training, as a main branch of SSL, can achieve better performance and robustness by expanding various scenarios and training samples to assist the model trained only on limited labeled data. Since the dataset of the DFC23 Track 2 is a subset of the dataset of the DFC23 Track 1, the difference set of them can be a suitable unlabeled dataset as it has similar scenarios and data distribution.

Given several trained models, the following two-stage trident cooperation results fusion pipeline is used to derive more reliable initial pseudolabels. From the pseudolabels of the detection task, predicted bounding boxes with confidence scores lower than $\delta_d = 0.8$ are removed while their initial pseudolabels are maintained for the binary segmentation and height estimation tasks. Finally, by combining both labeled data and unlabeled data with their corresponding processed pseudolabels, the proposed TCNet is retrained to boost the performance and generalization capabilities.

Two-stage trident cooperation results fusion pipeline: As shown in Fig. 14, for n detection predictions from n trained models, the proposed pipeline uses weighted boxes fusion to only ensemble their bounding boxes predictions, resulting in more accurate proposals, which are then passed to the best model's mask head to perform segmentation, deriving final detection results. For n binary segmentation masks, a hard voting strategy with a mask threshold of 0.5 is used to obtain majority predictions. These are further employed to remove wrong background predictions of the height estimation.

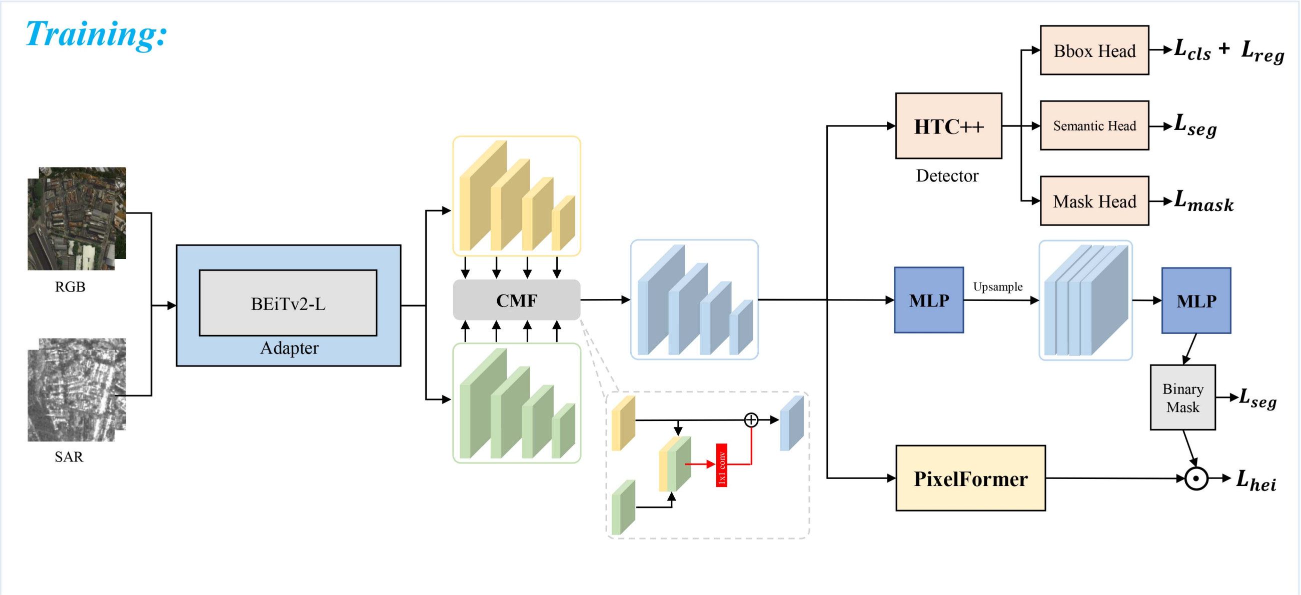

Fig. 13. Illustration of the proposed trident cooperation network (TCNet). The backbone is pretrained on the DFC23 Track-1 data for initialization. A cross-modal fusion (CMF) module provides more informative shared features for downstream visual tasks. Finally, the trident task-specific heads are leveraged, consisting of a detector [26], a lightweight all-MLP decoder [36], and a pixel query regression head [37]. The binary mask predicted by the all-MLP is additionally used to alleviate fore-background confusion existing in the estimated map.

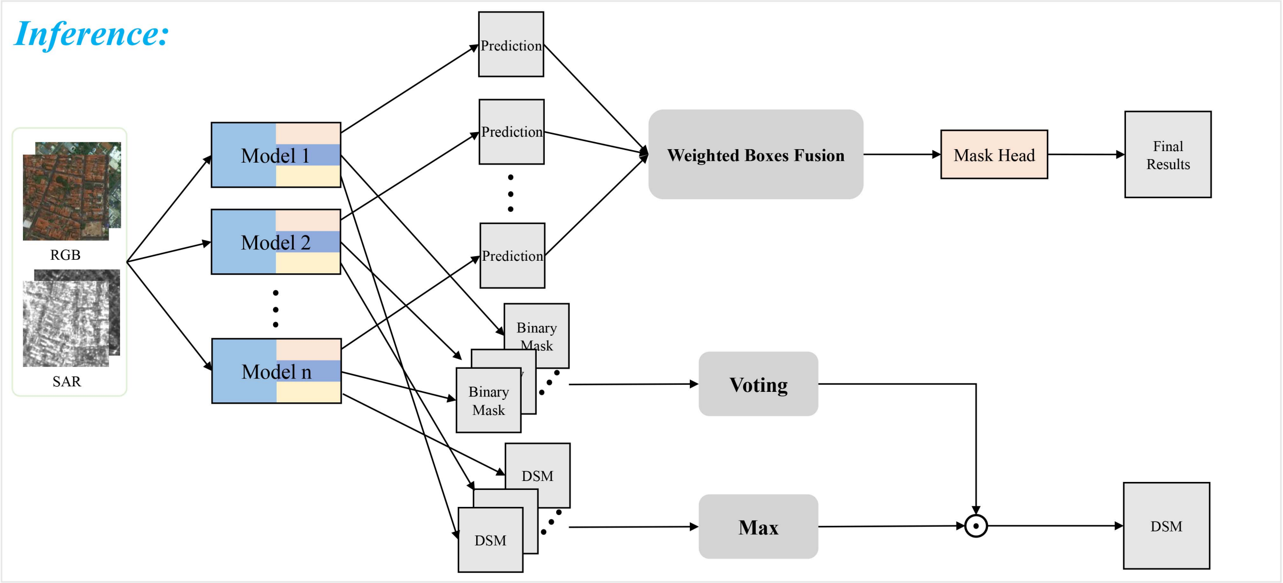

Fig. 14. Illustration of the two-stage trident cooperation results fusion framework proposed by the second place of Track 2 consisting of task-specific ensemble strategies, including weighted boxes fusion [21], hard voting, and pixel-wise maximum for merging the detection results, the binary masks, and the DSM results, respectively. The merged binary masks are employed to element-wise multiply with the height nDSM maps to filter the confusing background.

B. Results

Table VIII shows the effectiveness of each component of Team IPIU-XDU's method, indicating that the proposed TCNet can benefit from shared information of multitask learning.

Fig. 15 visually demonstrates that the method of Track 2's second place can accurately extract various building instances and estimate their corresponding height. Moreover, thanks to the cross-modal fusion module, the method is capable of mitigating the negative effects caused by cloud occlusion.

C. Discussion

Team IPIU-XDU proposed a novel trident cooperation network, TCNet, to perform end-to-end building extraction and height estimation. TCNet has a strong vision transformer adapter backbone that added a cross-modal fusion module, which provides more informative shared hierarchical features for the following trident task-specific decoders. To boost the performance and robustness of TCNet, team IPIU-XDU introduced a simple yet effective self-training. In addition, a two-stage trident cooperation results fusion framework including three task-specific

TABLE VIII
ABLATION STUDY ON VARIOUS COMPONENTS

Baseline	Sup	Semi	TCNet	TTA	AP_{50}(val)	AP_{50}(test)	δ_1(val)	δ_1(test)
✓					0.7230	–	0.7561	–
✓	✓				0.7460	–	0.7708	–
✓	✓	✓			0.7510	–	0.7722	–
✓	✓	✓	✓		0.7570	–	0.7746	–
✓	✓	✓	✓	✓	0.7790	0.7830	0.7766	0.7855

Sup: Supervised training. Semi: Semisupervised training. TTA: Test-time augmentation.

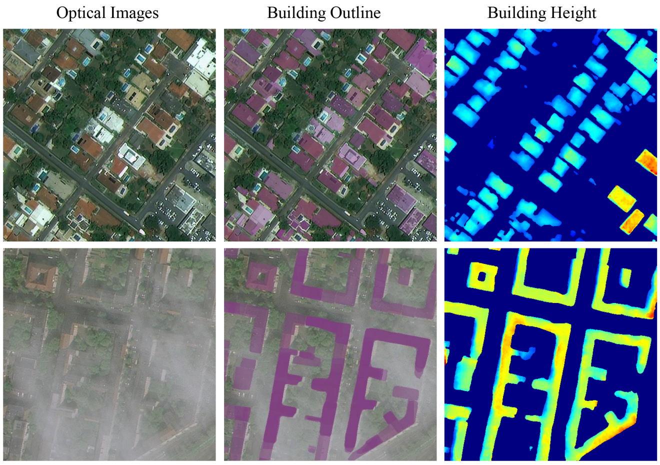

Fig. 15. Qualitative results of the second place of Track 2.

ensemble strategies is designed to further improve the proposed TCNet. Extensive experiments on DFC23 demonstrate the effectiveness of the proposed method.

VIII. Conclusion

The capability to classify fine-grained building types according to different roof typologies promises to offer valuable insights into the urban fabric that are useful for urban planning and management. Furthermore, the capability to estimate building heights provides an additional layer of spatial information, enabling a 3-D understanding of urban structures. Elevation data, when integrated with other geospatial information, contributes to the creation of realistic 3-D city models and digital twins, fostering a virtual representation that mirrors the physical world.

This article summarized the outcomes of the DFC23, describing the methodological choices made by the teams that ranked first and second in the contest. The strategies adopted by the winning teams provide valuable insights, including advanced deep learning methods, as well as best practices to boost the performance of instance segmentation and monocular height estimation techniques. For Track 1, team PIESAT-AI adopted an instance segmentation framework based upon the Cascade Mask R-CNN [15] architecture enhanced with domain-adapted pretraining and a modernized dual-backbone structure. Team IPIU-XDU resorted to a three-stage learning framework containing pretraining, supervised training, and semisupervised training. They adopted the BEiTv2-L [22] network for feature extraction, followed by fine-tuning with a vision transformer adapter [23] used to enhance the representations of the model. Both teams paid attention to the imbalance of class distributions adopting resampling and copy-paste data balancing/augmenting strategies. Interestingly, both teams improved the model performance adopting an ensemble strategy to combine multiple models.

For Track 2, team PIESAT-AI proposed a HGDNet to estimate building heights. HGDNet performs a joint task of height estimation and hierarchical classification by utilizing hierarchical height information. Team IPIU-XDU proposed a novel trident cooperation network, TCNet, to perform end-to-end building extraction and height estimation. In addition, a fusion framework including three task-specific ensemble strategies is designed to further improve the results.

All in all, DFC23 has fostered the development of advanced image analysis and data fusion techniques to tackle complex tasks such as building instance segmentation with fine-grained roof type classification and height estimation. The methods adopted by the two winning teams reveal many relevant contributions in terms of deep learning architectures, learning strategies, class balancing techniques, ensemble modelling, and several tricks and tweaks to boost the model performance.

The datasets remain open to the scientific community to further advance the development of image analysis and data fusion methods.

Acknowledgment

The IADF TC chairs would like to thank the Aerospace Information Research Institute under the Chinese Academy of Sciences, the Universität der Bundeswehr München, and GEOVIS Earth Technology Company, Ltd. for providing the data and the IEEE GRSS for continuously supporting the annual Data Fusion Contest through funding and resources.

References

[1] C. Persello et al., "2023 IEEE GRSS data fusion contest: Large-scale fine-grained building classification for semantic urban reconsstruction[technical committees]," *IEEE Geosci. Remote Sens. Mag.*, vol. 11, no. 1, pp. 94–97, 2023.

[2] N. Yokoya et al., "Open data for global multimodal land use classification: Outcome of the 2017 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1363–1377, May 2018.

[3] Y. Xu et al., "Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1709–1724, Jun. 2019.

[4] S. Kunwar et al., "Large-scale semantic 3D reconstruction: Outcome of the 2019 IEEE GRSS data fusion contest—Part A," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 922–935, 2020.

[5] Y. Lian et al., "Large-scale semantic 3D reconstruction: Outcome of the 2019 IEEE GRSS data fusion contest—Part B," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1158–1170, 2020.

[6] C. Robinson et al., "Global land-cover mapping with weak supervision: Outcome of the 2020 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3185–3199, 2021.

[7] Y. Ma et al., "The outcome of the 2021 IEEE GRSS data fusion contest - track dse: Detection of settlements without electricity," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 12375–12385, 2021.

[8] G. Liu et al., "Fine-grained building roof instance segmentation based on domain adapted pretraining and composite dual-backbone," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2023, pp. 670–673.

[9] X. Lu et al., "A strong vision transformer adapter with adaptive thresholding for fine-grained building classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2023, pp. 674–677.

[10] J. Hu, Z. Huang, F. Shen, D. He, and Q. Xian, "A bag of tricks for fine-grained roof extraction," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2023, pp. 678–680.

[11] C. Lu et al., "HGDNet: A height-hierarchy guided dualdecoder network for single view building extraction and height estimation," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2023, pp. 758–761.

[12] X. Lu et al., "Trident cooperation network for building extraction and height estimation," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2023, pp. 762–765.

[13] Y. Guo and Z. Wang, "Height estimation based on semantic segmentation," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2023, pp. 766–769.

[14] J. Hu, Z. Huang, F. Shen, D. He, and Q. Xian, "A robust method for roof extraction and height estimation," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2023pp. 770–771.

[15] Z. Cai and N. Vasconcelos, "Cascade r-CNN: Delving into high quality object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6154–6162.

[16] H. Zhang, Y. Wang, F. Dayoub, and N. Sünderhauf, "SWA object detection," 2020, *arXiv:2012.12645*.

[17] S. Woo et al., "Convnext v2: Co-designing and scaling convnets with masked autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 16133–16142.

[18] T. Liang et al., "Cbnet: A composite backbone network architecture for object detection," *IEEE Trans. Image Process.*, vol. 31, pp. 6893–6906, 2022.

[19] G. Ghiasi et al., "Simple copy-paste is a strong data augmentation method for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2918–2928.

[20] J. Wang et al., "Seesaw loss for long-tailed instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9695–9704.

[21] R. Solovyev, W. Wang, and T. Gabruseva, "Weighted boxes fusion: Ensembling boxes from different object detection models," *Image Vis. Comput.*, vol. 107, 2021, Art. no. 104117.

[22] Z. Peng, L. Dong, H. Bao, Q. Ye, and F. Wei, "Beit v2: Masked image modeling with vector-quantized visual tokenizers," 2022, *arXiv:2208.06366*.

[23] Z. Chen et al., "Vision transformer adapter for dense predictions," in *Proc. Int. Conf. Learn. Representations*, 2023.

[24] X. Huang et al., "Urban building classification (UBC)-a dataset for individual building detection and classification from satellite imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2022, pp. 1413–1421.

[25] Y. Liang, L. Zhu, X. Wang, and Y. Yang, "Penalizing the hard example but not too much: A strong baseline for fine-grained visual classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 5, pp. 7048–7059, May 2022.

[26] K. Chen et al., "Hybrid task cascade for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4974–4983.

[27] X. Lu, G. Cao, and T. Gou, "Semi-supervised landcover classification with adaptive pixel-rebalancing self-training," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 4611–4614.

[28] X. Lu et al., "Simple and efficient: A semisupervised learning framework for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5543516.

[29] X. Lu et al., "Weak-to-strong consistency learning for semisupervised image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5510715.

[30] Z. Zheng, Y. Zhong, and J. Wang, "Pop-Net: Encoder-dual decoder for semantic segmentation and single-view height estimation," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 4963–4966.

[31] S. Xing, Q. Dong, and Z. Hu, "Sce-Net: Self-and cross-enhancement network for single-view height estimation and semantic segmentation," in *Proc. Remote Sens.*, 2022, vol. 14, Art. no. 2252.

[32] T. Xiao et al., "Unified perceptual parsing for scene understanding," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 418–434.

[33] Z. Cai and N. Vasconcelos, "Cascade r-CNN: High quality object detection and instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1483–1498, May 2021.

[34] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[35] M. Elhousni, Z. Zhang, and X. Huang, "Height prediction and refinement from aerial images with semantic and geometric guidance," *IEEE Access*, vol. 9, pp. 145638–145647, 2021.

[36] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 12077–12090, 2021.

[37] A. Agarwal and C. Arora, "Attention attention everywhere: Monocular depth prediction with skip attention," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 5861–5870.