HigherNet-DST: Higher-Resolution Network With Dynamic Scale Training for Rooftop Delineation

Hongjie He, Lingfei Ma, *Member, IEEE*, and Jonathan Li, *Fellow, IEEE*

Abstract— **High-definition (HD) maps of building rooftops or footprints are important for urban application and disaster management. Rapid creation of such HD maps through rooftop delineation at the city scale using high-resolution satellite and aerial images with deep-learning methods has become feasible and has drawn much attention. However, the scale variance issue in rooftop delineation limited the overall performance. Existing methods exhibit considerably poor performance in the rooftop delineation of small buildings. In this article, we propose a new method, namely the higher-resolution network with dynamic scale training (HigherNet-DST) to overcome the scale variance problem in rooftop delineation. Specifically, dynamic scale training (DST) is applied in the model training phase to reduce the negative impact of scale variance. Then, a scale-aware backbone, namely the Higher-Resolution Network, is adopted to enhance the feature representation. Finally, the high-resolution supervision targets are used to further boost the delineation performance. Our method was tested on four publicly accessible building datasets and the results demonstrated that our method achieved the highest performance in rooftop delineation among the existing methods. Extensive experiments showed the superior performance of our method with an average precision (AP) of 68.5% on the AICrowd Building Dataset and an intersection of union (IoU) of 82.6% on the Inria Building Dataset, respectively, which surpassed many state-of-the-art (SOTA) methods. On the WHU Building Dataset and the Waterloo Building Dataset (WBD), our method also achieved the highest performance among the benchmarked methods, showing the high performance of our method for building boundary delineation.**

Index Terms— **Dynamic scale training (DST), high-resolution supervision targets, rooftop delineation, scale variance, scale-aware higher resolution network.**

I. Introduction

HIGH-DEFINITION (HD) maps of building footprints or rooftops are basic data for urban applications and disaster management [1], [2], [3]. High-resolution remotely sensed images, in particular, aerial imagery, have been widely used in citywide building extraction [4], [5], [6], [7]. However, canopy occlusion [8], scale-variance, and intraclass variation [9] inhibit its practical use

In recent years, deep convolutional neural networks (DCNNs) dominated in computer vision tasks [10] and were successfully applied to remote sensing. The DCNN-based instance segmentation and semantic segmentation methods were widely used in rooftop delineation from aerial images. In addition, with a large volume of aerial images with building annotations, solving the scale-variance and intraclass variation problems became possible. Consequently, these methods achieved high accuracy, while occlusion issues remained under exploration. Moreover, blurred rooftop boundaries still existed in extraction results [9], [11].

The end-to-end DCNN-based building extraction has drawn much attention recently. These methods were first introduced to directly generate vectorized building maps from remote-sensing images without any postprocessing [12]. In recent research, the extraction results became more accurate with sharp and regularized rooftop boundaries by employing advanced techniques, such as the convolutional gate recurrent unit (ConvGRU) [7] and the graph neural network (GNN) [13], and by supervising model training with new targets, such as vertices [7], [12], [13], [14], frame field [15], attraction field maps (AFMs) [14], and permutation matrices [13]. In addition, by directly outputting vectorized rooftops, the problems caused by occlusion and blurred rooftop boundaries have been significantly reduced. Nonetheless, scale-variance problems still exist.

Scale variance, in the context of object detection, recognition, and segmentation, pertains to the significant differences in sizes among samples within a dataset. For example, in the MicroSoft Common Objects in COntext (MS COCO) dataset, medium-sized and large objects are detected in 71% and 83% of the images, respectively, whereas small objects are observed in only 52% of the dataset [16]. This imbalanced distribution leads to biased optimization toward different scales [16]. Therefore, when datasets exhibit substantial scale variance, the performance of models on different scales may differ significantly. Specifically, scale variance issues often result in notably poor model performance on small objects. Effectively addressing scale variance problems hinges on enhancing the detection or segmentation performance of small objects to

Manuscript received 23 July 2023; revised 11 December 2023 and 10 January 2024; accepted 31 January 2024. Date of publication 5 February 2024; date of current version 15 February 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 42101451; in part by the Emerging Interdisciplinary Project of the Central University of Finance and Economics, China; and in part by the China Scholarship Council for their support via a Doctoral Scholarship under Grant 201906180088. *(Corresponding authors: Lingfei Ma; Jonathan Li.)*

Hongjie He is with the Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: h69he@uwaterloo.ca).

Lingfei Ma is with the School of Statistics and Mathematics, Central University of Finance and Economics, Beijing 102206, China (e-mail: l53ma@cufe.edu.cn).

Jonathan Li is with the Department of Geography and Environmental Management and the Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: junli@uwaterloo.ca).

Digital Object Identifier 10.1109/TGRS.2024.3362601

improve overall model performance. In the context of rooftop delineation, researchers have employed model architecture refinement and data augmentation techniques to tackle these challenges. However, the data augmentation methods they used were static and did not benefit from training results. Consequently, improvements in performance for one fixed scale often led to a decrease in performance for other scales. Refinements in model architecture have concentrated on enhancing feature extraction and representation across multiple scales. Feature extraction and representation are fundamental aspects influencing the performance of deep-learning models, as elaborated in Section II-C. The associated techniques in this field have advanced rapidly. Improved feature extraction and representation directly translate to enhanced deep-learning model performance. The approach presented by Liu et al. [9] incorporated both new model architecture and data augmentation. However, it is evident that it still struggles to effectively address the challenges posed by scale variance. Hence, the current state-of-the-art (SOTA) techniques in rooftop delineation do not adequately prioritize the challenge of scale variance. Moreover, the existing methods that do address this concern are not notably effective in mitigating its impact. Thus, further advancements are crucial to developing more robust solutions for addressing scale variance issues in rooftop analysis.

In this article, we present a new method to solve the scale-variance problem in an end-to-end manner for automated delineation of rooftops in aerial imagery. Following the approach introduced by Liu et al. [9], our method integrates data augmentation and multiscale feature fusion. Diverging from conventional static data augmentation techniques, our method adopts a dynamic scale training (DST) strategy [16]. This DST strategy dynamically adjusts the data augmentation process based on real-time feedback during the model training, setting it apart from previously mentioned static augmentation strategies. Regarding multiscale feature fusion, we leverage a scale-aware higher-resolution network called HigherHRNet [17], an extension of the high-resolution network (HRNet v2) [18], specifically tailored to address scale variation challenges. Furthermore, our approach incorporates high-resolution supervision targets, enhancing the detection accuracy of smaller objects. The contributions of this article are as follows.

1) We introduce a new powerful end-to-end rooftop delineation training model.
2) We mitigate scale-variance issues in rooftop delineation without additional computational resource overhead by employing the DST strategy.

The rest of this article is organized as follows. Section II provides a brief literature review on building extraction with a consideration of the scale variance issue. Section III details our method. Section IV describes the datasets used and presents experimental results obtained. Section V discusses the effectiveness of each part of our method and provides a comparison of the performance between DST and MS training and testing. Section VI summarizes and concludes the article with our findings.

II. Related Work

A. DCNN-Based Rooftop Delineation

To the best of our knowledge, the earliest use of DCNN in building extraction can be traced back to the work of Shu [8] and Mnih [19]. In their work, DCNNs were used to extract features, with fully connected layers for feature flattening, pixel-level image classification, and building extraction. However, this kind of method exhibited low efficiency with limited input size.

With the proposal of the fully convolutional networks (FCNs) [20] and the U-Net [21], pixel-wise image classification, also known as semantic segmentation, has grown rapidly with a large number of new methods invented yearly. As for rooftop delineation from aerial images, from ConvNet [22] to the capsule feature pyramid network (CapsFPN) [23] and the coarse-to-fine boundary refinement network (CBR-Net) [24], different deep-learning techniques have been introduced to this task. These techniques include, but are not limited to, the attention scheme [25], the capsule network [23], and the MS feature extraction [24]. The SOTA methods can extract accurate building masks, but postprocessing is still required to generate vectorized rooftop polygons.

To generate vectorized rooftops from aerial images, an intuitive way is to regularize the polygons converted from building masks extracted by the DCNN-based methods. The regularization can be conducted separately. For example, Zhao et al. [26] employed the mask region-based convolutional neural networks (R-CNNs) first, for instance, segmentation and instance mask generation. Instance masks were then converted to polygons using the Douglas-Peucker algorithm and the minimum description length (MDL) optimization with generated hypothesis hypotheses. Regularization methods, such as the active contour model (ACM) [27], also known as snake, can be embedded into the DCNN architectures and generate polygons in an end-to-end manner. In this direction, Marcos et al. [28] proposed deep structured active contours (DSACs), which combined deep learning and the ACM for image segmentation. Gur et al. [29] proposed an end-to-end trainable ACM via differentiable rendering. Similarly, Hatamizadeh et al. [30] proposed the trainable deep active contour (TDAC) model to directly delineate building polygons from aerial images. Cheng et al. [31] combined the active ray network with deep leaning and proposed the deep active ray network (DARNet).

Concurrently, another family of algorithms has been developed to generate regular rooftop polygons. The most representative method is the PolyMapper [12], which applied the convolutional long short-term memory (ConvLSTM) to predict the sequence of vertices of building boundaries from CNN features. Zhao et al. [7] refined the PolyMapper by replacing the ConvLSTM with the ConvGRU and decorating the original backbone with the global context block (GCB) and the boundary refinement block (BRB). Recently, Girard et al. [15] proposed frame field learning for building delineation by introducing the frame field targets for optimizing models. Zorzi et al. [13] proposed the PolyWorld by employing the GNN and a sophisticatedly designed loss function. Xu et al. [14] proposed the hierarchical supervisions (HiSups) learning scheme with hierarchical building representations, including the low-level convex and concave building vertices, the mid-level AFMs for line segments and the high-level regional masks of buildings. These three methods have demonstrated high performance in building extraction and represent the SOTA approaches. As HiSup is the most recently developed method and demonstrates the highest performance, we took HiSup as a starting point and refined it to solve scale-variance issues, especially addressing low performance on small objects.

B. Scale Variation in Computer Vision Tasks

In natural photography and remote sensing, balanced distribution with regard to object scale cannot be guaranteed. This leads to significant performance vary in common image processing tasks among different scales, which is named scale variation [16]. In addition, scale variation also limits the overall performance. Compared to large-scale (area > 96^2 pixels) and middle-scale (32^2 < area < 96^2 pixels) objects [32], small-scale (with area < 32^2 pixels) objects contributed less to the total loss [16]. This results in less supervision during training and lower performance at smaller-scale objects. Therefore, small-scale objects should be focused on when alleviating scale variance problems in deep learning. In literature, data preparation and model optimization are the focus when dealing with scale variance.

Data preparation adjusts data distribution before model training or optimizing. Methods such as resampling [16] and image pyramid [9] are intuitive. However, as tested in Chen et al. [16], resampling hurts model performance at other scales. The image pyramid is a more robust technique, but arbitrarily selected scales may not be suitable for overcoming scale variance. Other image pyramid-type methods, such as the scale normalization for image pyramids (SNIPs) [33] and the SNIP with efficient resampling (SNIPER) [34], increase inference burden. In contrast, the collage style data augmentation, as adopted in [16], [35], and [36], has been effective in handling scale variance and has shown high performance.

Feature pyramid and dilation-based methods are optimization-based [16]. In feature pyramid style methods, different scales of feature maps are learned and aggregated. The feature pyramid network (FPN) [37] is the most representative method in this category. The HRNet [18] aggregates feature maps from four different scales in each stage of each branch (or scale). HRNet has shown high performance in feature representation. By refining the HRNet, the HigherHRNet was proposed in Cheng et al. [17], which showed better performance in feature representation, especially for small objects. The dilation-based methods, such as the deformable convolution networks (DCNs) [38] and the trident networks (TridentNets) [39], can generate scale-sensitive feature representations with high resolution but suffer from storage issues.

C. Scale Variation in Rooftop Delineation

The scale variation, complex architectures, and diverse appearances are the main obstacles that make rooftop delineation challenging [9], [23], [40], [41], [42] Even with auxiliary data as input, the issue of scale variance cannot be properly overcome [39]. Following the taxonomy of methods for dealing with scale variance in natural images, the methods used for rooftop delineation can also be classified into model optimization and data preparation.

In literature, model optimization methods are more commonly used. For example, Liu et al. [9] proposed an MS U-shaped CNN building instance extraction framework with edge constraint (EMU-CNN). The EMU-CNN consists of an MS fusion U-shaped network (MFUN), a region proposal network (RPN), and an edge-constrained multitask network (ECMN). The MFUN module collects feature information from input images with three different spatial resolutions and fuses the features for a U-shaped deconvolution network. The method showed good rooftop delineation performance on both large-scale and small-scale buildings. A similar method with sole input was proposed by Zhu et al. [40]. In their work, a multiple attending path neural network (MAP-Net) was proposed, in which the spatial location-preserved MS features were learned by a multiparallel path taking a sole image as input. The learned MS features enabled the method to be able to extract exact building edges and recognize small buildings. In addition, Guo et al. [41] proposed the deep-supervision convolutional neural network (DS-Net) for rooftop delineation also with MS feature learning. Three stages including encoder, decoder, and deep supervision, make up the DS-Net. The experiments showed the high performance of the DS-Net in depicting the boundary of a small building. Furthermore, in recent research, Liu et al. [42] proposed an end-to-end MS geoscience network (MS-GeoNet). Various embedding modules and loss functions were explored and applied in the network for better performance in rooftop delineation. Specifically, with the CoordConv module, the method performed well on small building extraction. In addition, Wu et al. [43] proposed a topography-aware loss (TAL) for better performance on rooftop delineation in semantic segmentation-based methods. Combining MS feature learning by the HRNet, TAL not only showed better performance on regular-size buildings, but also on small-size buildings reporting its high performance in dealing with scale variation issues. Overall, MS feature learning is the basis of methods in the model optimization category.

As for data preparation-based methods, we can only find one research [9]. As we mentioned earlier, images with three different spatial resolutions were taken as input in the EMU-CNN bringing MS features and resulting in better performance in rooftop delineation especially for small buildings.

In summary, inspired by Liu et al. [9], we employed both data preparation and model optimization methods in our method to deal with scale variation issues. Specifically, we employed the DST and a scale-aware backbone-HigherHRNet. To further improve the performance of our method, instead of using downsampled targets for model supervision [14], we calculated loss using upsampled final outputs and targets with original resolution.

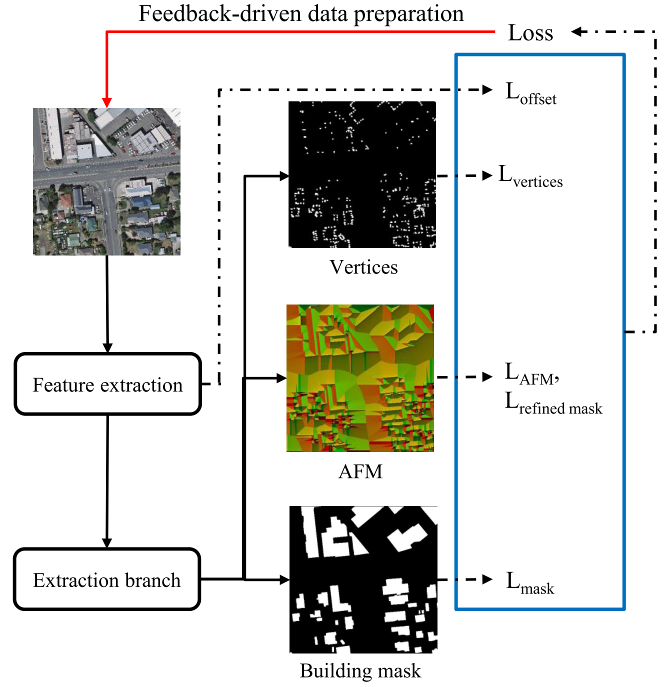


Fig. 1. Overview of the HigherNet-DST.

III. Methodology

A. Overview

In this article, we introduced a novel approach that aimed at addressing scale-variance challenges in rooftop delineation. Our method is built upon the SOTA method, HiSup [14] (Section III-B), ensuring the robustness and efficiency of our approach. Furthermore, we integrated the DST for data preparation in model training (Section III-C) and leveraged the scale-aware HigherNet for feature extraction (Section III-D). Notably, to enhance the delineation of small objects, we utilized high-resolution supervision targets instead of smaller ones, as employed in HiSup [14], for model optimization (Section III-E). The overview of our method is illustrated in Fig. 1.

B. HiSup Learning for Rooftop Delineation

To mitigate the performance gap between mask prediction and polygon extraction caused by mask reversibility, hierarchical supervision learning was proposed in Xu et al. [14]. Specifically, after feature extraction by the backbone, four branches were attached for mask prediction, AFM prediction (used for line segmentation) [44], vertex location prediction, and offset prediction [14]. For detailed information regarding the HiSup, we direct readers to Xu et al. [14]. In their experiments, HiSup showed the highest performance on the AICrowd Building Dataset [45] and a competitive performance on the Inria Building Dataset [46] against other methods, achieving the SOTA performance in learning-based rooftop delineation. Therefore, we took it as the basis of our method.

As previously outlined, our proposed method incorporates hierarchical supervision learning for the prediction of vertices, attractive field maps, masks, refined masks, and vertex offsets [14]. As depicted in Fig. 1, the proposed higher-resolution network with DST (HigherNet-DST) model integrates a composite loss function comprising L_{offset}, L_{vertices}, L_{AFM}, $L_{\text{refined mask}}$, and L_{mask}. The cross-entropy loss function is applied to both mask prediction and vertex location prediction. Additionally, the L1 loss function is utilized for the line segment prediction and the refinement of vertex locations with offsets. The specific loss functions employed for model training are detailed as follows:

$$\begin{aligned}\mathrm{Loss} &= \mathrm{w}_1\mathrm{L}_{\text{offset}} + \mathrm{w}_2\mathrm{L}_{\text{vertices}} + \mathrm{w}_3\mathrm{L}_{\text{AFM}} \\ &\quad + \mathrm{w}_4\mathrm{L}_{\text{mask}} + \mathrm{w}_5\mathrm{L}_{\text{refined mask}}\end{aligned} \tag{1}$$

$$\mathrm{L}_{\text{offset}} = \begin{cases} \dfrac{t}{w}|\mathrm{sigmoid}(\mathrm{logits}) - 0.5 - \mathrm{targets}|, \\ \qquad\qquad \text{if } \mathrm{v} \neq \mathrm{None} \\ |\mathrm{sigmoid}(\mathrm{logits}) - 0.5 - \mathrm{targets}|, \\ \qquad\qquad \text{if } \mathrm{v} = \mathrm{None} \end{cases} \tag{2}$$

$$\mathrm{L}_{\text{vertices}} = -\frac{1}{N}\sum_{i=1}^{N}\left(v_i \log\left(\mathrm{softmax}\left(\mathrm{pv}_i\right)\right)\right) \tag{3}$$

$$\mathrm{L}_{\text{AFM}} = \frac{1}{N}\sum_{i=1}^{N}|\mathrm{AFM}_i - \mathrm{A\hat{F}M}_i| \tag{4}$$

$$\mathrm{L}_{\text{mask}} = -\frac{1}{N}\sum_{i=1}^{N}\left(m_i \log\left(\mathrm{pm}_i\right) + (1 - m_i)\log\left(1 - \mathrm{pm}_i\right)\right) \tag{5}$$

$$\begin{aligned}\mathrm{L}_{\text{refined mask}} = -\frac{1}{N}\sum_{i=1}^{N}&\left(m_i \log\left(\mathrm{prm}_i\right)\right. \\ &\left.+(1 - m_i)\log\left(1 - \mathrm{prm}_i\right)\right)\end{aligned} \tag{6}$$

where the variables w_1, w_2, w_3, w_4, and w_5 correspond to the weights assigned to $\mathrm{L}_{\text{offset}}$, $\mathrm{L}_{\text{vertices}}$, L_{AFM}, L_{mask}, and $\mathrm{L}_{\text{refined mask}}$, respectively. The terms "sigmoid" and "softmax" denote the sigmoid and softmax functions, respectively. The "t/w" value represents a weighting factor obtained by dividing the tensor "t," derived from specific vertices in the ground-truth mask, by the tensor "w," which computes the mean of these vertices. This computation aims to modulate the influence of different vertex elements on $\mathrm{L}_{\text{vertices}}$, allowing targeted emphasis on different parts of the mask. This adjustment facilitates enhanced model learning by focusing on specific mask features. "logits" and "targets" denote the predicted offsets and ground-truth offsets, respectively. Similarly, "v," "pv," "m," "pm," and "prm" represent ground-truth vertices, predicted vertices, ground-truth masks, predicted masks, and predicted refined masks, respectively. "N" symbolizes the number of samples within each batch.

C. DST in Rooftop Delineation

The DST (Stitcher) overcomes scale-variance by collaging images and supervision targets which is guided by dynamic feedback [16]. Specifically, the feedback is the proportion of loss contribution of small objects against that of all objects. For instance, if L_small/L<= τ, in the next iteration, k images are randomly selected from the next batch of data to create a new image. In the inequality, "small" and "L" represent

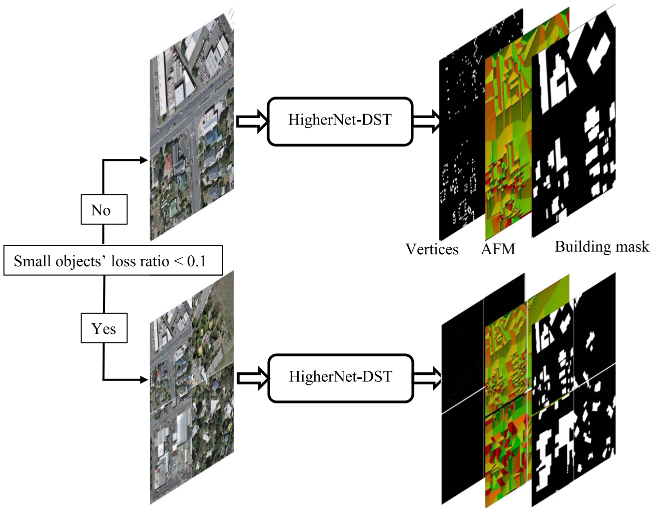


Fig. 2. Principle of DST in HigherNet-DST.

losses calculated on small objects and all objects in each batch. In addition, τ and k are two hyperparameters representing the threshold for "Stitcher" and the number of images used for creating the collage, respectively. The collected images and supervision targets are downsampled and stitched together, as shown in Fig. 2. If the ratio is larger than τ, the model is trained with the usual pipeline in the next iteration.

D. Scale-Aware HigherHRNet (HigherNet)

HRNet has shown excellent performance in feature extraction and representation using multilevel features with repeated information exchange in each stage [18]. However, in the final stage of the HRNet, the highest resolution of features is 1/4 of the input. Information loss and scale variance suppress the performance of the HRNet. Cheng et al. [17] proposed the HigherHRNet by adding a scale-aware module on top of the HRNet. The scale-aware module is mainly composed of a deconvolutional module and four Residual blocks (or "Basic Blocks") [47]. To save computational resources, we downsampled and concatenated features with different spatial resolutions to 128 × 128 pixels, which is similar to the output size of HRNet in the HiSup. The architecture of our scale-aware HigherHRNet is provided in Fig. 3. The "feature" in Fig. 3 is used in the extraction branch as shown in Fig. 4. The "output" is the predicted vertex offset which will be used in the final polygon generation process as described in [14].

E. High-Resolution Supervision Targets

As discussed in Section II-B, using higher-resolution input [17] can, to some extent, overcome scale-variance issues. However, this brings an overwhelming computational burden. In the HiSup [14], the models were trained with lower resolution targets (1/4 of input resolution), which significantly reduced the memory cost in model training and deployment. However, lower-resolution supervision targets also cause information loss and poor performance, especially for small objects. To balance the memory cost and performance, we applied high-resolution supervision targets that maintain the same resolution as the input. There are two reasons why we adopted the high-resolution supervision targets instead of low-resolution ones as in Xu et al. [14].

1) High-resolution supervision targets have more detailed information compared to low-resolution ones [48]. Such detailed information is helpful for small object detection [17].
2) When supervised with upsampled targets, the model is trained to consider the upsampling process including both advantages and potential errors and can learn to take advantage or deal with these accordingly. Fig. 4 shows the rooftop delineation part of our method.

In our method, we applied the mask-and-vertices attraction [14] which was used in HiSup. Predicted vertices and masks were taken as input to initialize polygons. The local nonmaximum suppression (NMS) was applied to sparse vertices. Refined vertices with the aid of predicted offset vectors were used to simplify initialized polygons by removing redundant vertices and low-confidence vertices from initialized polygons. Adjacent edges in each polygon were further merged if they were almost paralleled. We direct readers to Xu et al. [14] for more information.

F. Implementation Details

In the training phase, the weights assigned to L_{offset}, L_{vertices}, L_{AFM}, L_{mask}, and $L_{\text{refined mask}}$, denoted as w_1, w_2, w_3, w_4, and w_5, respectively, are set to 0.25, 8.0, 0.1, 1.0, and 1.0. For the hyperparameters setting, the initial learning rate, the weight decay, the max epoch, and the batch size were set as 1e-4, 1e-4, 100, and 16, respectively. After the first 25 epochs, the learning rate was divided by 10. In all experiments including DST, we set τ and k to 0.1 and 4 following Chen et al. [16]. For the ablation study and the comparative study with different backbones, we set the batch size as 7 due to memory limitation with a large backbone. To ease the computational burden, we applied automatic mixed precision in this work. It is denoted as "*" in the rest of this article. We used PyTorch 1.7 and trained our network on a Nvidia[1] RTX 3090 GPU. To test our method on the AICrowd Building Dataset, we used two Nvidia[1] RTX 3090 GPUs with the same parameters set.

IV. Results and Analysis

A. Datasets and Evaluation Metrics

1) Building Datasets Preparation: To extensively evaluate the performance of our network, and test its robustness, we selected four widely used public building datasets for rooftop delineation. These datasets are the AICrowd Building Dataset [45], the Inria Building Dataset [46], the WHU Building Dataset [49], and the Waterloo Building Dataset (WBD) [1]. Each dataset comprises red, green, and blue bands but differs in spatial resolution and covers distinct geographic locations.

The AICrowd Building Dataset was first used in the AICrowd (previously CrowdAI) mapping challenge [45]. The

[1] Registered trademark.

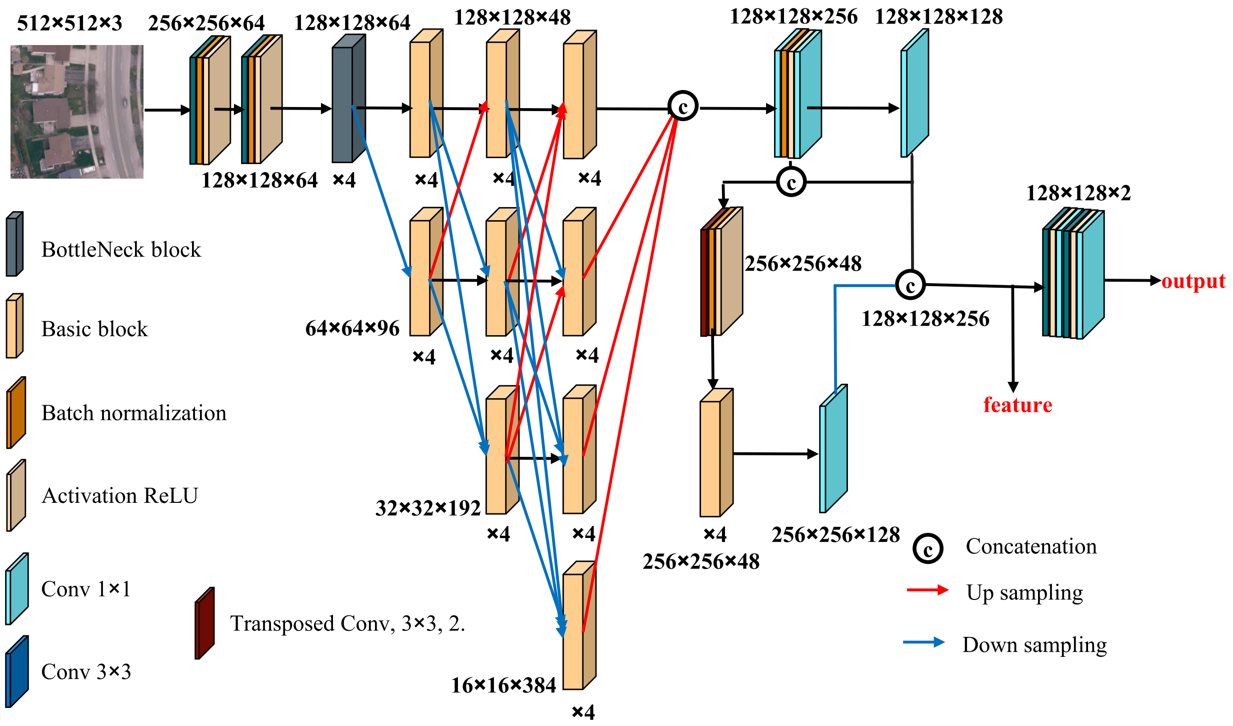


Fig. 3. Architecture of the scale-aware HigherHRNet.

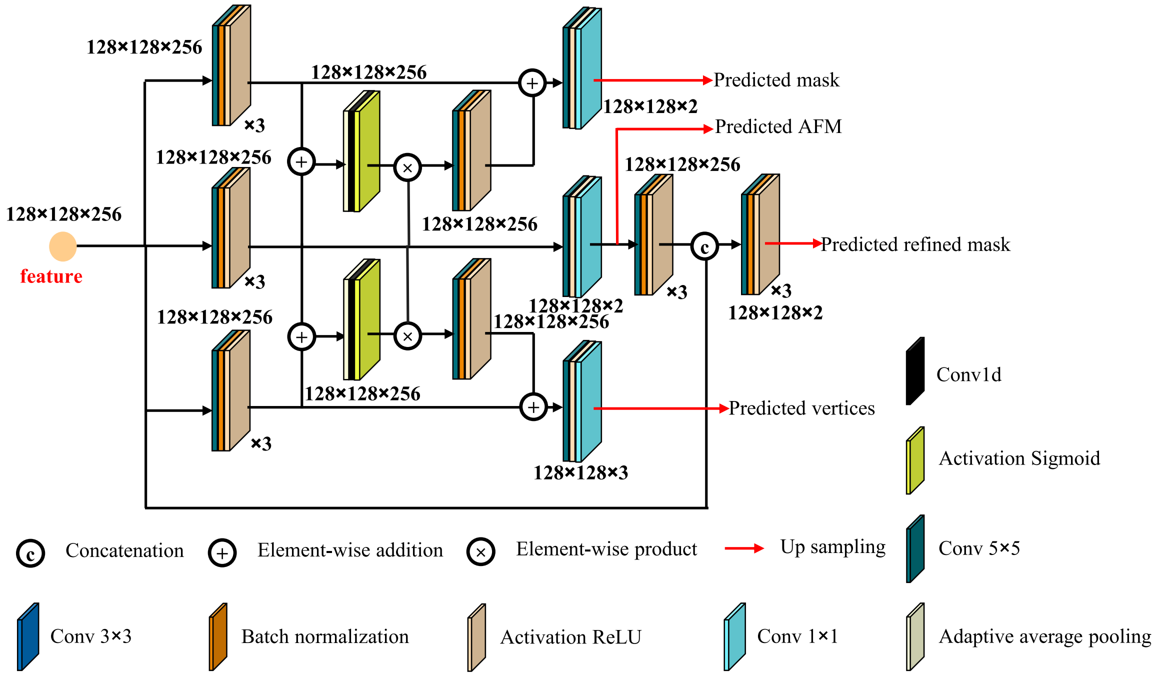


Fig. 4. Architecture of the object extraction branch (modified from Xu et al. [14]).

satellite images have a spatial resolution of 0.30 m/pixel. Annotation files were provided in MS COCO format [32]. All images were cropped to 300 × 300 pixels. Because of the missing testing dataset, we followed previous work [12], [13], [14], [15] and used the training dataset and validation dataset for model training and testing, respectively. The training dataset is composed of 280 741 images, and the validation dataset contains 60 317 images [45].

The Inria Building Dataset was also widely used in rooftop delineation [14], [15], [50]. This building dataset was proposed in 2017 to alleviate the generalization problems in building outline delineation by splitting ten cities from the USA and Austria into training and testing datasets [46]. A total of 360 orthorectified aerial images were evenly separated into training and testing datasets, but the testing datasets were not publicly released. The dataset was composed of 180 images with an image size of 5000 × 5000 pixels and a pixel size of 0.30 m (405 km^2 in total). Only semantic labeling in binary mask for each image was provided in the original dataset. A simple polygonization was necessary to convert binary masks to the MS COCO format annotations and to enable model training for end-to-end manner rooftop delineation. Following previous work, we collected the first five images of each location for model evaluation.

The WHU Building Dataset [49]) has been widely used in semantic segmentation for building outline delineation. The aerial images in the dataset covered 450 km^2 of Christchurch, New Zealand, with 220 000 independent buildings. The original images were collected with a spatial resolution of 0.075 m/pixel and then downsampled to 0.30 m/pixel. Finally, the cropped small patches, with an image size of 512 × 512 pixels, were split into 4736, 1036, and 2416 tiles for training, validation, and test subsets, respectively. Similar to processing the Inria Building Dataset, a polygonization process was also necessary to generate the MS COCO format annotation files for end-to-end rooftop delineation. Due to the WHU Building Dataset having a relatively small data volume compared to other datasets used in this work, we selected it for the ablation study.

The WBD was released in 2022, which covers 205.83 km^2 area of Kitchener-Waterloo area in Ontario, Canada [1]. The building dataset was developed for semantic segmentation methods for rooftop delineation. This dataset consists of 242 aerial images of a size of 8350 × 8350 pixels and a spatial resolution of 0.12 m/pixel. Manually labeled binary masks for building rooftops were provided for all images. Both images and binary masks were cropped to small patches of size 512 × 512 pixels. Then patches with geometric distortion were removed. Finally, 42 147, 6887, and 18 945 pairs of images and masks were assigned into training, validation, and test subsets, respectively. The polygonization process was also required here to convert binary masks to MS COCO annotation files.

2) Evaluation Metrics: The object-level evaluation metrics proposed in Lin et al. [32] are widely used in instance segmentation and object detection in computer vision and remote-sensing applications. In literature, average precision (AP), average recall (AR), AP$_{50}$, and AP$_{75}$ were used to evaluate different methods on the AICrowd Building Dataset. As we also focused on scale variance in this work, in addition to these metrics, we also used AP-Small (AP$_s$), AP-Medium (AP$_m$), AP-Large (AP$_L$), AR-Small (AR$_s$), AR-Medium (AR$_m$), and AR-Large (AR$_L$). Small, medium, and large sizes denote 32 × 32 pixels, between 32 × 32 pixels and 96 × 96 pixels, and larger than 96 × 96 pixels, respectively. In our work, we empirically adopted the MS COCO criterion to define small, medium, and large sizes. Small buildings under this criterion exhibited poorer performance compared to medium and large building objects in existing rooftop delineation research. These methods include the Mask R-CNN-based method [45], the path aggregation network (PANet) [51], the PolyMapper [12], the PolyWorld [13], and the HiSup [14]. Additionally, considering the proportion of small buildings over the total number of buildings is indispensable, improving the accuracy of delineating small building objects is expected to enhance the overall accuracy. Therefore, we employed the MS COCO criterion to define small objects in our work. Because MS COCO metrics are widely used, we omitted the detailed introduction here and direct readers to Lin et al. [32] and He et al. [52] for more information.

Following Xu et al. [14], we also adopted the restricted metric AP^{boundary} [53]. AP^{boundary} is AP calculated based on boundary intersection of union (IoU) [14] instead of mask IoU in Lin et al. [32]. The boundary IoU can be calculated as

$$\text{BoundaryIoU}\left(C, \hat{C}\right) = \frac{|(C_d \cap C) \cap \left(\hat{C} \cap \hat{C}_d\right)|}{|(C_d \cap C) \cap \left(\hat{C} \cap \hat{C}_d\right)|} \tag{7}$$

where C and $\hat{C}$ are ground-truth building masks and predicted building masks, respectively. C_d and $\hat{C}_d$ represent pixels within distance d from building boundaries. In this work, we set d to 0.02. For the comparative study on the Inria Building Dataset, we also calculated IoU and overall accuracy.

In addition to COCO metrics and AP^{boundary}, to evaluate the predicted building structures, we also employed PoLiS and C-IoU metrics [14]. The PoLiS and the C-IoU metrics offer distinct approaches to assess performance. The former one is defined to describe the difference between two polygons. The average distance between each vertex in one polygon and its closet vertex in another polygon defines the PoLiS metric. C-IoU takes both segmentation accuracy and polygonization complexity into account, which is the normal IoU weighted by a coefficient defined using the number of vertices in two different polygons. PoLiS and C-IoU are defined as

$$\text{PoLiS}\left(P, \hat{P}\right) = \frac{1}{2q}\sum_{a_j \in P}\min_{b \in \partial \hat{P}}\|a_j - b\| + \frac{1}{2r}\sum_{b_k \in \hat{P}}\min_{a \in \partial P}\|b_k - a\| \tag{8}$$

$$\begin{aligned}\text{RD(Relative Difference)}&(N_P, N_{\hat{P}})\\ &= \frac{|N_P - N_{\hat{P}}|}{N_P + N_{\hat{P}}}\end{aligned} \tag{9}$$

$$\text{C-IoU}\left(P, \hat{P}\right) = \text{IoU}\left(P_m, \hat{P}_m\right) \cdot (1 - \text{RD}(N_P, N_{\hat{P}})) \tag{10}$$

where P and $\hat{P}$ denote the predicted polygon and the ground-truth polygon, respectively, with a and b as vertices in two polygons. ∂P and $\partial \hat{P}$ represent the boundaries of P and $\hat{P}$, respectively. The variables q and r correspond to the respective numbers of vertices in P and $\hat{P}$, respectively. Specifically, $1 \leq i \leq q$ and $1 \leq k \leq r$. For C-IoU, the IoU(.) calculates the IoU between two polygon masks. The RD(.) is short for the relative difference between the total number of vertices from two polygons.

B. Results on the AICrowd Building Dataset

As introduced previously, the AICrowd Building Dataset was released in 2018 and widely used in recent years for rooftop delineation. In this work, for comparison, we visualized extraction results generated by the PolyWorld [13], the HiSup [14], and our method ordered from top to bottom in Fig. 5. As shown, the performance on extracting medium and large building objects has limited differences. However, on small objects, such as objects in the top right of the first column and objects in the bottom middle of the second

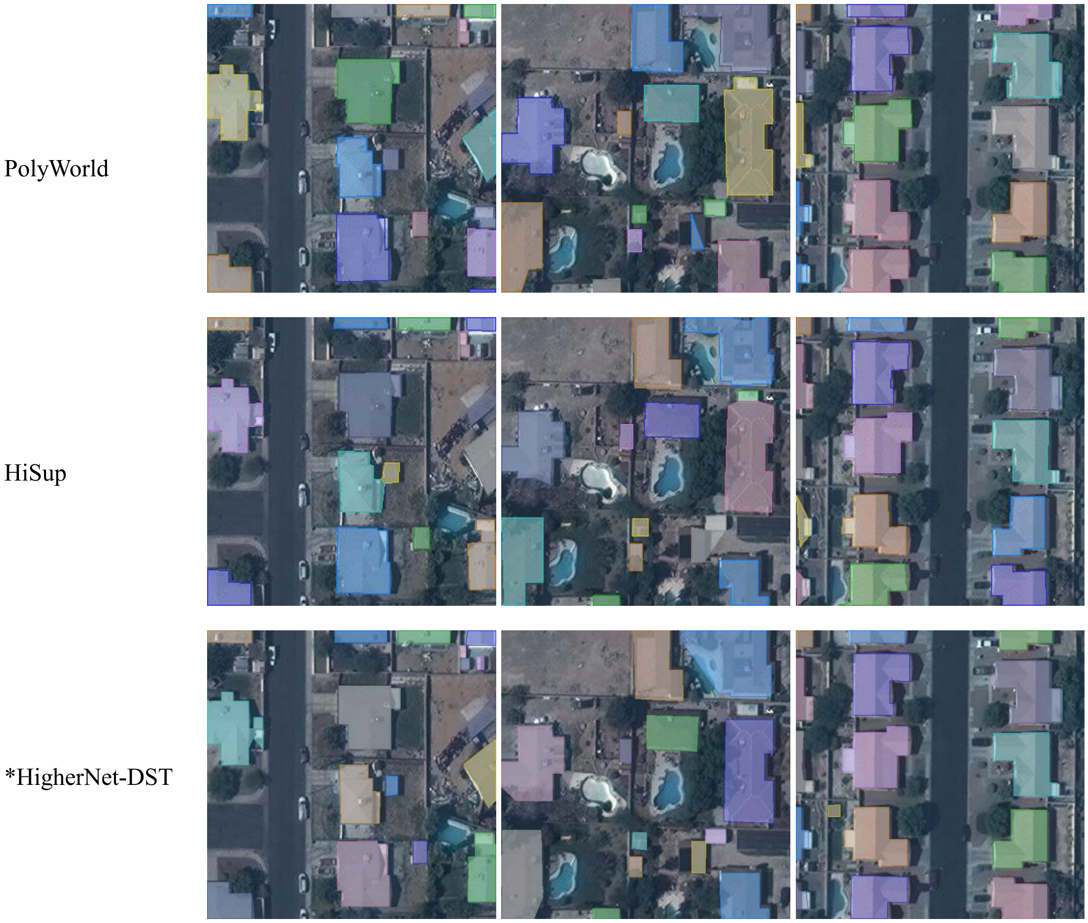


Fig. 5. Rooftop delineation results on the AICrowd Building Dataset obtained using PolyWorld, HiSup, and *HigherNet-DST.

TABLE I

EVALUATION RESULTS ON THE AICROWD BUILDING DATASET (IN % EXCEPT FOR POLIS AND TIME)

Methods	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_L	AR	AR_s	AR_m	AR_L	$AP^{boundary}$	PoLiS	IoU	C-IoU	Time(s)
Mask RCNN[45, 54]	41.9	67.5	48.8	12.4	58.1	51.9	47.6	18.1	65.2	63.3	15.4	3.454	61.3	50.1	-
Path Aggregation Network (PANet)[51]	50.7	73.9	62.6	19.8	68.5	65.8	54.4	21.8	73.5	75.0	-	-	-	-	-
PolyMapper[12]	55.7	86.0	65.1	30.7	68.5	58.4	62.1	39.4	75.6	75.4	22.6	2.215	77.6	67.5	-
Frame Field Learning (FFL)[15]	67.0	92.1	75.6	-	-	-	73.2	-	-	-	34.4	1.945	84.3	73.8	-
Li et al[55]	73.8	92.0	81.9	-	-	-	72.6	-	-	-	-	-	-	-	-
PolyWorld[13]	63.3	88.6	70.5	37.2	83.6	87.7	75.4	52.5	88.7	95.2	50.0	0.962	91.2	88.4	-
HiSup[14]	**79.4**	**92.7**	**85.3**	**55.4**	**92.0**	**96.5**	**81.5**	**60.1**	**94.1**	**97.8**	**66.5**	**0.726**	**94.3**	**89.6**	**4660**
*HigherNet-DST	68.5	88.4	77.5	41.9	82.6	88.8	71.3	46.6	85.6	91.7	48.0	1.293	89.9	84.5	4741

column, our method surpassed the PolyWorld and the HiSup. In the last column, concerning the performance of objects located on the left side, our method outperformed the HiSup but was inferior to the PolyWorld.

Table I[2] provides quantitative evaluation results of our method and other SOTA methods on the AICrowd Building Dataset. Our method showed a competitive performance compared to other SOTA methods but was inferior to that of Li et al. [55] and the HiSup. Specifically, our method achieved 68.5% of AP, which was competitive compared to other methods but lower than 73.8% and 79.4% of AP reported by Li et al. [55] and the HiSup, respectively. We believe that this result was caused by dataset interpolation which may have resulted in uncertainty and was harmful to the performance. Regarding inference time, our method requires more time for rooftop delineation in the AICrowd Building Dataset, approximately 81 s for processing 60 317 images.

C. Results on the Inria Building Dataset

In recent work [14], [15], [50], the Inria Building Dataset was also used to test the performance of new methods in rooftop delineation and specifically to test generalizability. We presented qualitative results in Fig. 6. In the first row,

[2] Missing metrics from the original publications are denoted with "-" in Table I. Evaluation results with references are collected from Li et al. [12], Zorzi et al. [13], and Xu et al. [14].

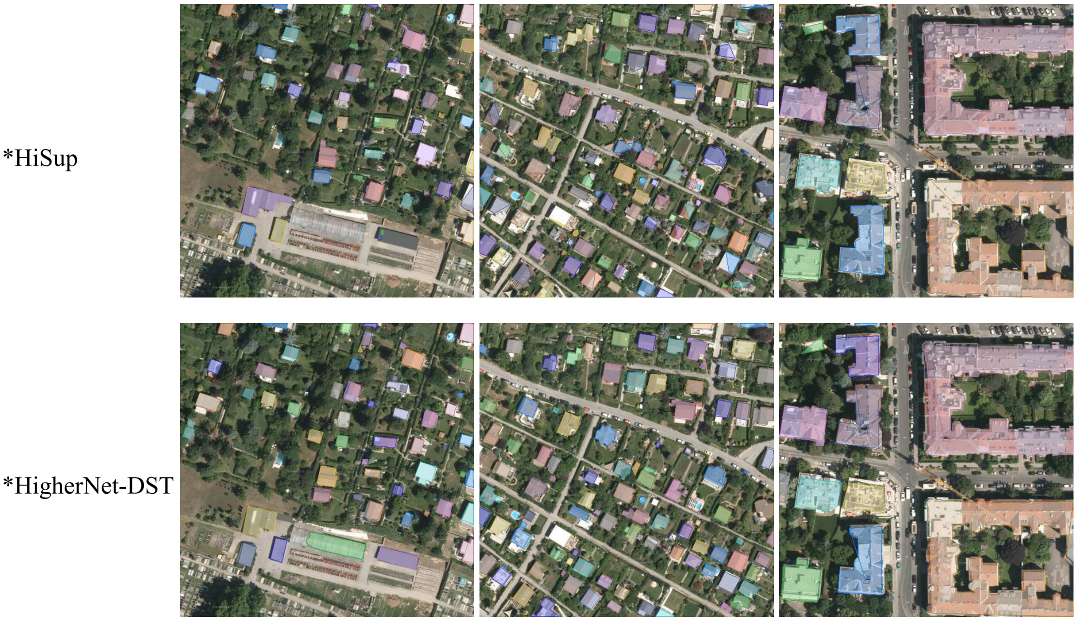


Fig. 6. Rooftop delineation results on the Inria Building Dataset obtained using the *HiSup and *HigherNet-DST.

TABLE II

EVALUATION RESULTS ON THE INRIA BUILDING DATASET—OBJECT LEVEL (IN % EXCEPT FOR POLIS AND TIME)

Methods	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_L	AR	AR_s	AR_m	AR_L	$AP^{boundary}$	PoLiS	IoU	C-IoU	Time(s)
*HiSup[14]	29.0	50.0	29.8	17.8	41.6	**49.8**	34.0	21.5	45.7	59.2	24.2	3.057	**69.6**	**52.5**	**329**
*HigherNet-DST	**38.4**	**64.2**	**40.8**	**26.8**	**52.2**	40.2	**46.1**	**34.0**	**58.1**	**59.6**	**34.0**	**2.888**	67.4	49.9	382

we tabulated extraction results generated by the HiSup model released by Xu et al. [14]. In the second row, we supplied extraction results generated by our HigherNet-DST. As shown in the first two columns, our model detected more building objects with accurate boundaries than the HiSup. As shown, our method excelled in extracting small building objects, outperforming the other benchmarked methods. In the last column, we showed the extraction performance on large objects. As shown, both methods showed high performance, but unexpected lines appeared, which may be caused by the wrong order of junctions in generating final polygons. We can claim the better performance in the qualitative results in Fig. 6 by checking the yard detection. The yard of the middle bottom building was detected by our method but missed by the HiSup, which further proved the superior performance of our method when extracting small objects.

To quantitatively evaluate two models, we applied IoU and Accuracy (pixel/overall accuracy) following Xu et al. [14] and object-level metrics as used in Section IV-A. As shown in Tables II and III[3], our model achieved the highest values on both pixel-level metrics and object-level metrics except for AP_L. In addition, our method also possessed a lower PoLiS value. This demonstrated the high performance of our models in image segmentation and boundary delineation. Specifically, our model obtained an AP of 38.4%, which is 9.4% higher than HiSup. The value of AR_L was increased by more than 20% with our model. And the values of AP_{50}, AP_{75}, AP_m, AR, AR_s, and AR_m were also increased more than 10% with our model. Similarly, concerning inference time, our method requires more time for rooftop delineation on the Inria Building Dataset. Specifically, an additional 53 s are needed for processing 25 large images sized at 5000 × 5000 pixels each. The experiment confirmed the success of our proposal.

TABLE III

EVALUATION RESULTS ON THE INRIA BUILDING DATASET—PIXEL LEVEL (IN %)

Methods	IoU	Accuracy
FFL[15]	74.8	96.0
*HiSup[14]	80.7	97.0
*HigherNet-DST[14]	**82.6**	**97.4**

D. Results on the WHU Building Dataset

The WHU Building dataset was released with binary building masks [49]. To the best of our knowledge, the dataset has never been used in polygon delineation in literature. Therefore, in this work, we first converted binary building masks to polygon annotations in the MS COCO format. We conducted experiments on the WHU Building Dataset primarily to assess the transferability of our method. Therefore, we exclusively tested HiSup [14] alongside our proposed method on this dataset.

[3] We collected evaluation results of the FFL from Xu et al. [14] and assessed the HiSup using the released model.

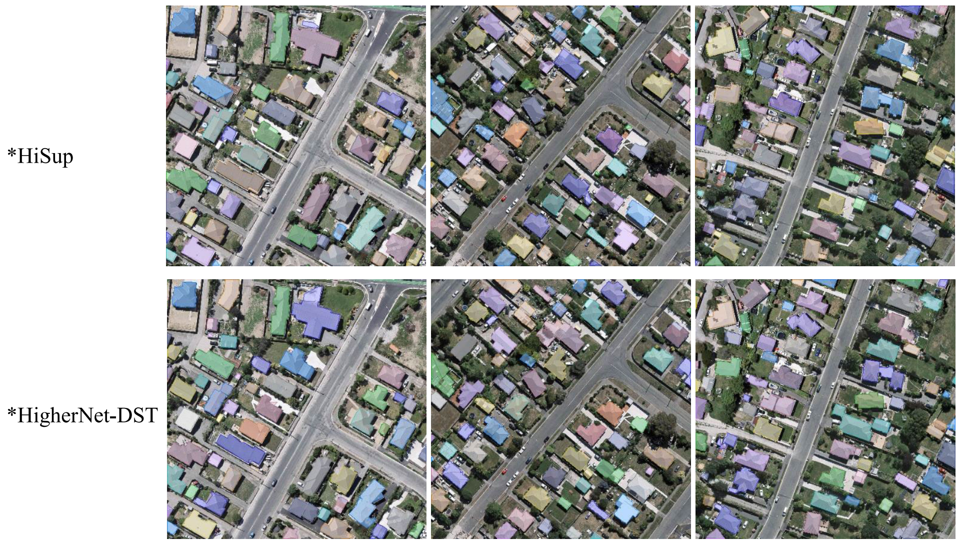


Fig. 7. Rooftop delineation results on the WHU Building Dataset obtained using the *HiSup and *HigherNet-DST.

TABLE IV

EVALUATION RESULTS ON THE WHU BUILDING DATASET—OBJECT LEVEL (IN % EXCEPT FOR POLIS AND TIME)

Methods	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_L	AR	AR_s	AR_m	AR_L	$AP^{boundary}$	PoLiS	IoU	C-IoU	Time(s)
*HiSup[14]	58.3	80.4	65.8	43.7	76.0	**75.3**	63.2	47.8	**81.0**	**84.0**	56.7	**1.302**	**82.9**	61.5	**314**
*HigherNet-DST	**60.1**	**82.8**	**69.0**	**46.2**	**77.1**	74.9	**63.6**	49.3	80.2	80.7	**58.5**	1.463	82.1	**71.9**	446

Fig. 7 provides the extraction results from this experiment. As shown, our improvement can be found in small objects. Small objects in the output of the HiSup, especially in the first and the last columns, had incomplete polygons, which were fixed when using our method.

As shown in Table IV, our method achieved the best performance when compared to the HiSup. Specifically, our method obtained an AP of 60.1% compared to the 58.3% of the HiSup. Our method obtained an AR of 63.6% compared to the 63.2% of the HiSup. In addition, our method delivered a higher value of C-IoU and decent values of PoLiS and IoU. For small objects, our method obtained an AP_s of 46.2% surpassing the performance of the PolyMapper (41.6%) and the HiSup (43.7%). Although the AR value of our method is lower than the PolyMapper, it is higher than that of the HiSup. We have also documented the total time taken for rooftop delineation on the WHU Building Dataset's test dataset. For 2416 tiles, our method requires 132 s. The performance on large objects dropped, but AP and AR increased. Therefore, with this experiment, we proved the success of our proposal in terms of dealing with poor performance caused by small objects.

E. Results on the WBD

To test the robustness of our method with regard to different spatial resolutions, we downsampled the WBD to 0.30 m/pixel and tested the performance of our method. In this experiment, aiming at evaluating the transferability of our method, we solely tested HiSup and our proposed method on the WBD.

We first showed the visualization results in Fig. 8. In Fig. 8 (a), we provided results predicted by the *HiSup on the 0.12 m/pixel WBD, followed by extraction results generated by our method on the same dataset. In Fig. 8(b), we tabulated results predicted by the HiSup and our method on the 0.30 m/pixel WBD. As shown in Fig. 8(a), two methods can delineate building polygons with similar performance, but our method had high sensitivity when distinguishing building rooftops from building walls. In addition, our method can extract small objects with higher performance as shown in the third example. As shown in Fig. 8(b), building polygons become completer and more accurate going from top to bottom. In addition, as shown in the second and last columns, compared to the HiSup, the advantage of our method is apparent when segmenting buildings that are very close. It can also be attributed to the high performance when delineating small objects.

Table V provides the quantitative evaluation results. As shown in Table V, our method has a similar performance with HiSup on the 0.12 m/pixel dataset, and higher performance when delineating small and medium objects. On the 0.30 m/pixel dataset, our method achieved an AP of 51.5% and an AR of 55.4%. The values of AP_{50} and AP_{75} increased more than 10%. In both spatial resolutions, our method achieved higher C-IoU values but lower values of PoLiS and IoU. By outperforming the previous SOTA, these results demonstrated the effectiveness of our proposed method. Taking into account the extended inference time alongside the achieved performance improvements, as demonstrated in Table V, the experiment reaffirmed the success of the proposed approach.

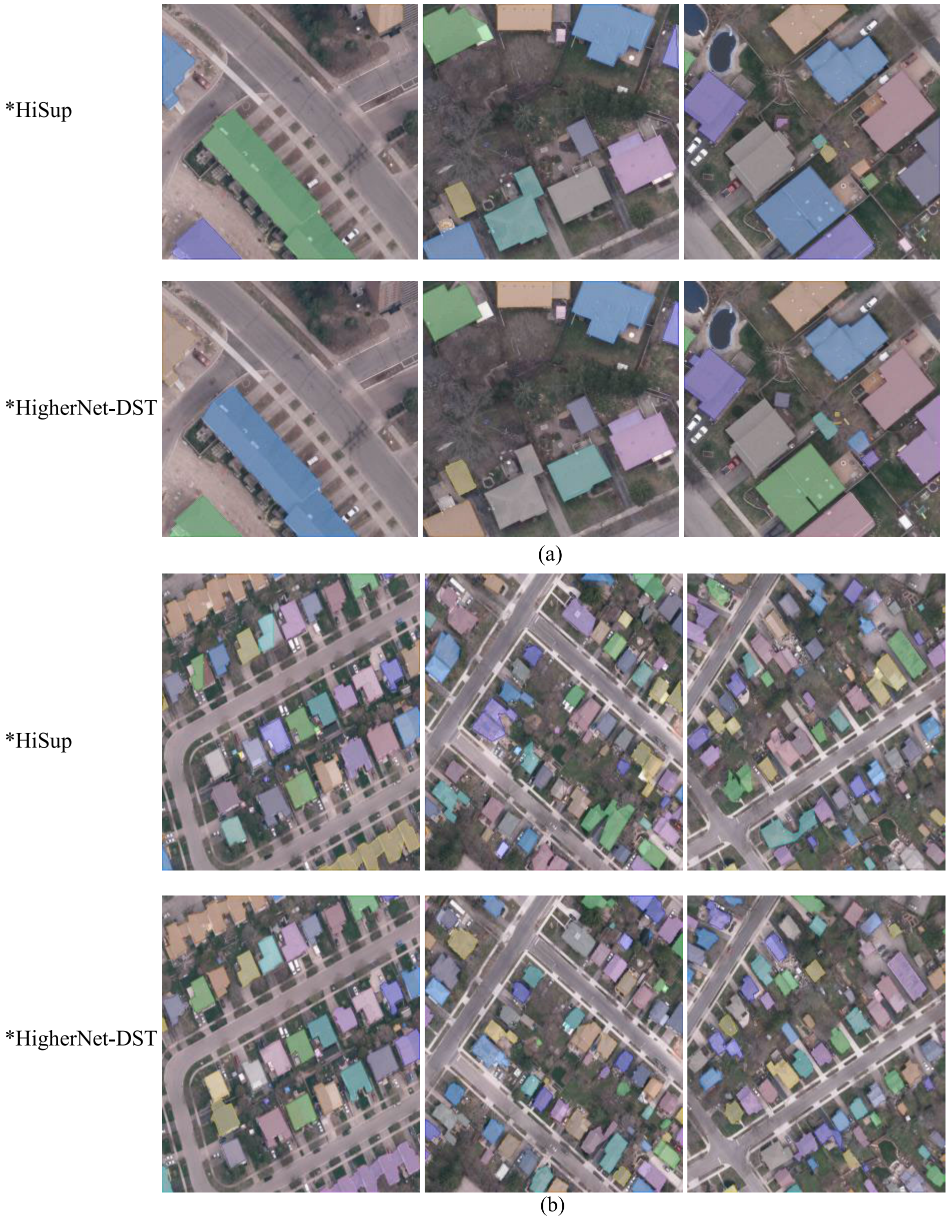


Fig. 8. Rooftop delineation results on the WBD obtained using the *HiSup and *HigherNet-DST. (a) Rooftop delineation results on the 0.12 m/pixel WBD. (b) Rooftop delineation results on the 0.30 m/pixel WBD.

TABLE V

Evaluation Results on the WBD—Object Level (in % Except for PoLiS and Time)

Pixel Size	Methods	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_L	AR	AR_s	AR_m	AR_L	$AP^{boundary}$	PoLiS	IoU	C-IoU	Time(s)
0.12 m/pixel	*HiSup	**66.9**	**82.7**	**74.1**	30.5	75.5	**84.0**	70.8	36.5	78.3	**88.9**	**52.5**	**2.409**	**85.6**	66.5	**1522**
	*HigherNet-DST	66.5	82.5	74.0	**31.5**	**76.2**	82.8	**70.8**	**37.4**	**78.7**	87.4	51.0	2.556	85.4	**70.6**	1584
0.30 m/pixel	*HiSup	42.8	62.3	47.7	30.7	56.1	**70.7**	48.0	33.3	61.0	**78.2**	40.2	**1.890**	**77.6**	60.3	**455**
	*HigherNet-DST	**51.5**	**73.2**	**59.6**	**37.3**	**66.8**	64.4	**55.4**	39.5	**70.2**	70.3	**49.0**	2.098	74.9	**65.0**	543

V. Discussion

A. Ablation Study

In this section, we explored the effectiveness of each part of our method with respect to the baseline. Specifically, on the WHU Building Dataset, we took the HiSup as the baseline and tested the performance of the automatic mixed precision, the DST, the higher-resolution network, and the high-spatial-resolution supervision targets. In addition, we showed the performance of our method trained with full precision. In Table VI, we showed their performance on the WHU Building Dataset and denoted them as "+amp," "+DST," "+HigherNet," "*HigherNet-DST," and "HigherNet-DST," respectively. As shown in Table VI, by applying the DST, replacing high-resolution network (HRNet v2) with Higher Resolution Network, using higher-spatial-resolution

TABLE VI

ABLATION STUDY CONDUCTED ON THE WHU BUILDING DATASET (IN % EXCEPT FOR POLIS AND TIME)

Methods	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_L	AR	AR_s	AR_m	AR_L	$AP^{boundary}$	PoLiS	IoU	C-IoU	Time(s)
HiSup	59.1	80.6	67.2	43.8	77.0	77.5	63.4	47.7	81.5	83.2	57.5	1.282	**83.6**	61.9	**315**
+amp	58.3	80.4	65.8	43.7	76.0	75.3	63.2	47.8	81.0	**84.0**	56.7	1.302	82.9	61.5	317
+DST	59.4	80.6	67.1	44.4	77.4	**77.6**	63.7	48.3	81.6	83.3	57.6	**1.251**	82.6	61.5	437
+HigherNet	59.6	80.7	67.4	44.9	77.4	76.1	64.0	48.7	**81.8**	82.8	58.0	1.253	83.2	60.4	446
*HigherNet-DST	60.1	82.8	69.0	46.2	77.1	74.9	63.6	49.3	80.2	80.7	58.5	1.463	82.1	**71.9**	446
HigherNet-DST	**61.4**	**83.8**	**70.9**	**47.3**	**78.1**	75.2	**64.8**	**50.6**	81.3	81.1	**59.6**	1.404	81.9	70.8	469

TABLE VII

PERFORMANCE OF MS TRAINING AND TESTING ON THE WHU BUILDING DATASET (IN % EXCEPT FOR POLIS AND TIME)

Methods		AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_L	AR	AR_s	AR_m	AR_L	$AP^{boundary}$	PoLiS	IoU	C-IoU	Time(s)
HiSup		59.1	80.6	**67.2**	43.8	77.0	77.5	63.4	47.7	81.5	83.2	57.5	1.282	**83.6**	**61.9**	**315**
+amp		58.3	80.4	65.8	43.7	76.0	75.3	63.2	47.8	81.0	84.0	56.7	1.302	82.9	61.5	317
+MS training	1024	41.0	63.9	46.2	29.8	55.4	56.5	46.3	33.5	61.0	64.4	38.9	1.831	68.8	48.7	501
	512	58.8	**80.7**	67.1	43.6	76.7	**77.7**	63.1	47.5	81.0	83.7	57.0	1.277	82.8	61.6	399
	256	28.6	56.5	26.0	9.9	48.9	66.9	32.2	12.3	54.5	74.7	26.3	2.603	69.7	48.7	324
	Combination	35.3	45.6	40.6	31.0	57.0	43.3	61.3	42.1	83.4	**88.1**	34.1	1.253	77.3	56.8	1224
+DST	1024	38.8	60.9	43.2	29.4	52.1	46.7	44.1	32.7	57.4	54.5	37.0	1.858	66.6	47.1	410
	512	**59.4**	80.6	67.1	**44.4**	**77.4**	77.6	**63.7**	**48.3**	81.6	83.3	**57.6**	1.251	82.6	61.5	437
	256	30.0	58.4	28.4	10.6	51.1	67.3	33.2	12.9	56.1	73.3	27.9	2.570	70.7	50.5	362
	Combination	37.2	47.8	42.5	31.8	60.0	46.7	61.8	42.8	**83.6**	87.4	36.0	**1.231**	77.8	56.7	1209

supervision targets and adding extra semantic segmentation branch on the backbone, the delineation performance increased gradually. In addition, by applying the DST, the PoLiS value decreased; by employing the higher-resolution network, the IoU value increased; by applying the high-resolution supervision targets, the C-IoU value increased; by adding the extra branch, all metrics became better; and the full precision training of our method achieved a lower value of PoLiS. Each modification brought an increase in at least one of three metrics. The increased performance confirmed the effectiveness of each modification.

To further explore the performance increase, we added an extra semantic segmentation branch taking the output feature from the backbone (as shown in Fig. 4). With the extra semantic segmentation branch, our method can be improved further. However, the increase is marginal compared to the increase in computational burden. Specifically, by adding the branch to "*ours," the AP value increased from 60.1% to 60.2%, and the AR value increased from 63.6% to 63.8%. Therefore, we did not include the branch in our method.

Regarding the inference time, as indicated in Table VI, the increased inference time primarily stems from the adoption of the DST. This potentially arises from the redesignation of the data preparation pipeline, which may require further refinement.

B. MS Training/Testing

To further show the superior performance of the DST in rooftop delineation, we compared it with MS training and testing, which were commonly used to deal with scale-variance issues. The baseline architecture in this section is the HiSup with auto-mixed precision. Following Liu et al. [9], three scales, including 256 × 256 pixels, 512 × 512 pixels, and 1024 × 1024 pixels, were used in MS training and testing. Specifically, in the training phase, input images were resized to three scales followed by the combination of convolution layers and batch normalization layers. The features generated by three scales were concatenated as the input of the first stage in the backbone (as shown in Fig. 9). For MS testing, input images were resized to three scales (2×, 1×, and 0.5×) before flowing into the deep network. In Table VII, we denoted the HiSup, the HiSup with auto mixed precision, the HiSup with auto mixed precision and MS training, and the HiSup with auto mixed precision and the DST as "HiSup," "+amp," "+multiscale training," and "+DST." MS testing with different spatial resolution input were noted by the side length of the input. We also evaluated the performance of the output combination from different scales input and noted it as a "combination."

As shown in Table VII, MS training with the 512 × 512 pixels size input indeed improved the whole performance while the improvement is less than that brought by employing DST. For example, by applying MS training with the 512 × 512 pixels size input, the AP value was increased from 58.3% to 58.8%. However, by applying the DST with the 512 × 512 pixels size input, the AP value was increased from 58.3% to 59.4%. In addition, employing MS training consumes more computational resources than applying DST. Therefore, MS training is not effective compared to DST in rooftop delineation, evidenced by the difference in inference time consumed, as shown in Table VII. As for MS testing, Table VII shows using both models with 512 × 512 pixels size input gives the best performance, which means it has a negative impact on rooftop delineation.

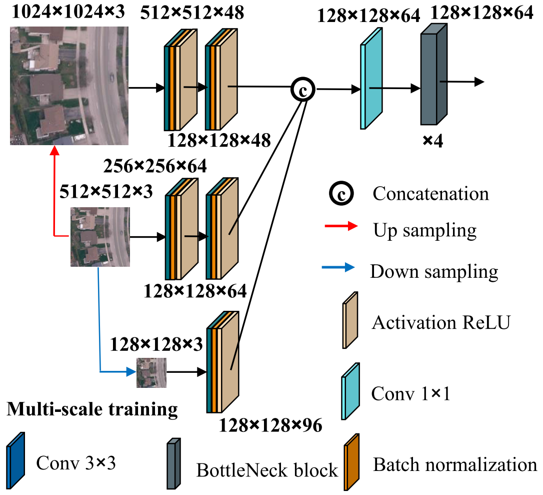


Fig. 9. Feature extraction part of MS training.

C. Robustness Analysis

Our method exhibits the highest performance on the AICrowd Building Dataset, followed by strong performance on the 0.12 m/pixel WBD, the WHU Building Dataset, and the 0.30 m/pixel WBD, except the Inria Building Dataset.

Regarding transferability, our method demonstrates notable adaptability with an AP value exceeding 50% on all datasets, except for the Inria Building Dataset. The exceptional performance on the AICrowd Building Dataset can be attributed to its homogenous nature. Conversely, the lower performance on the Inria Building Dataset could be linked to its high diversity, covering images from five cities across Austria and the USA.

The consistent and satisfactory performance observed on both the WHU Building Dataset and the WBD further confirms the transferability and effectiveness of our method. Notably, our findings highlight that higher spatial resolutions yield enhanced performance, as evidenced by the results on the WBD.

VI. Conclusion

In this article, we proposed a new deep-learning network, namely the HigherNet-DST for rooftop delineation. By applying the DST, adopting the scale-aware Higher-Resolution Network, and using higher-resolution supervision targets based on the HiSup, our method can relieve the scale-variance issue and improve the performance of building boundaries delineation. By conducting an extensive comparative study, our method showed competitive performance on the AICrowd Building Dataset and better performance on the Inria Building Dataset, the WHU University Building Dataset, and the WBD compared to other SOTA methods. The ablation study further showed the effectiveness of each module of our method. Precisely, experiments on the AICrowd Building Dataset showed the competitive performance of our method with an AP of 68.5%. On the Inria Building Dataset, with an IoU of 82.6% and an accuracy of 97.4%, our method achieved the best pixel classification performance among all benchmarked methods. In terms of building boundaries delineation, our method has 9.4%–27.2% higher values on all object level metrics, except for AR_L, compared to the HiSup, the previous SOTA method. On the WHU Building Dataset, our method achieved 60.1% of AP, 82.8% of AP_{50}, 69.0% of AP_{75}, 46.2% of AP_s, 77.1% of AP_m, 63.6% of AR, and 58.5% of AP^{boundary}, which were higher than the HiSup, while other metrics were also competitive. On the 0.30 m/pixel WBD, our methods surpassed the HiSup by 6.2%–11.9% on all metrics except for AP_L and AR_L. On the original WBD, our method showed competitive performance compared to the HiSup while better performance on small- and medium-size objects. Our experiments showed the effectiveness of our new network in dealing with scale-variance issues, especially excelling at the small building's regime, which is a long-standing problem in building boundary delineation.

For future research, two directions are promising in building boundary delineation. They are: 1) adapting new backbones to building boundaries delineation for better feature representation and better performance and 2) relieving the labeling cost to reduce the total cost for building boundaries delineation. Newly developed networks, such as vision transformer networks, have shown high performance in computer vision tasks in recent research. How to adapt these advanced networks to building boundary delineation is promising to be explored. The performance of DCNN models heavily lies in training samples. For building boundaries delineation, building polygons are costly to label. How to relieve the labeling process while keeping the high performance is also important to make more efforts.

Acknowledgment

The authors would like to acknowledge Kyle Gao and Lanying Wang from the Geospatial Intelligence and Mapping (GIM) Group, University of Waterloo, Waterloo, ON, Canada, for their contributions and assistance in editing this article.

References

[1] H. He et al., "Waterloo building dataset: A city-scale vector building dataset for mapping building footprints using aerial orthoimagery," *Geomatica*, vol. 75, no. 3, pp. 99–115, 2022.

[2] H. He et al., "Super-resolving and composing building dataset using a momentum spatial-channel attention residual feature aggregation network," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 111, Jul. 2022, Art. no. 102826.

[3] P. A. Pelizari et al., "Automated building characterization for seismic risk assessment using street-level imagery and deep learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 180, pp. 370–386, Oct. 2021.

[4] J. Xiao, M. Gerke, and G. Vosselman, "Building extraction from oblique airborne imagery based on robust façade detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 68, pp. 56–68, Mar. 2012.

[5] B. Johnson and Z. Xie, "Classifying a high resolution image of an urban area using super-object information," *ISPRS J. Photogramm. Remote Sens.*, vol. 83, pp. 40–49, Sep. 2013.

[6] J. Huang, X. Zhang, Q. Xin, Y. Sun, and P. Zhang, "Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network," *ISPRS J. Photogramm. Remote Sens.*, vol. 151, pp. 91–105, May 2019.

[7] W. Zhao, C. Persello, and A. Stein, "Building outline delineation: From aerial images to polygons with an improved end-to-end learning framework," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 119–131, May 2021.

[8] Y. Shu, "Deep convolutional neural networks for object extraction from high spatial resolution remotely sensed imagery," Ph.D. dissertation, Dept. Geography Environ. Manag., Univ. Waterloo, Waterloo, ON, Canada, 2014.

[9] Y. Liu, D. Chen, A. Ma, Y. Zhong, F. Fang, and K. Xu, "Multiscale U-shaped CNN building instance extraction framework with edge constraint for high-spatial-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6106–6120, Jul. 2020.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.

[11] Y. Shi, Q. Li, and X. X. Zhu, "Building segmentation through a gated graph convolutional neural network with deep structured feature embedding," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 184–197, Jan. 2020.

[12] Z. Li, J. D. Wegner, and A. Lucchi, "Topological map extraction from overhead images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1715–1724.

[13] S. Zorzi, S. Bazrafkan, S. Habenschuss, and F. Fraundorfer, "PolyWorld: Polygonal building extraction with graph neural networks in satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1848–1857.

[14] B. Xu, J. Xu, N. Xue, and G.-S. Xia, "HiSup: Accurate polygonal mapping of buildings in satellite imagery with hierarchical supervision," *ISPRS J. Photogramm. Remote Sens.*, vol. 198, pp. 284–296, Apr. 2023.

[15] N. Girard, D. Smirnov, J. Solomon, and Y. Tarabalka, "Polygonal building extraction by frame field learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5891–5900.

[16] Y. Chen et al., "Dynamic scale training for object detection," 2020, *arXiv:2004.12432*.

[17] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5386–5395.

[18] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5693–5703.

[19] V. Mnih, *Machine Learning for Aerial Image Labeling*. Toronto, ON, Canada: Univ. Toronto, 2013.

[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2015, pp. 234–241.

[22] J. Yuan, "Learning building extraction in aerial scenes with convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2793–2798, Nov. 2018.

[23] Y. Yu et al., "Capsule feature pyramid network for building footprint extraction from high-resolution aerial imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 895–899, May 2021.

[24] H. Guo, B. Du, L. Zhang, and X. Su, "A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 183, pp. 240–252, Jan. 2022.

[25] Y. Cai et al., "A comparative study of deep learning approaches to rooftop detection in aerial images," *Can. J. Remote Sens.*, vol. 47, no. 3, pp. 413–431, 2021.

[26] K. Zhao, J. Kang, J. Jung, and G. Sohn, "Building extraction from satellite images using mask R-CNN with building boundary regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 247–251.

[27] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 321–331, Jan. 1988.

[28] L. Zhang et al., "Learning deep structured active contours end-to-end," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8877–8885.

[29] S. Gur, T. Shaharabany, and L. Wolf, "End to end trainable active contours via differentiable rendering," 2019, *arXiv:1912.00367*.

[30] A. Hatamizadeh, D. Sengupta, and D. Terzopoulos, "End-to-end trainable deep active contour models for automated image segmentation: Delineating buildings in aerial imagery," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 730–746.

[31] D. Cheng, R. Liao, S. Fidler, and R. Urtasun, "DARNet: Deep active ray network for building segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7431–7439.

[32] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[33] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection-SNIP," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3578–3587.

[34] B. Singh, M. Najibi, and L. S. Davis, "SNIPER: Efficient multiscale training," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 9333–9343.

[35] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[36] D. Zhou, X. Zhou, H. Zhang, S. Yi, and W. Ouyang, "Cheaper pretraining lunch: An efficient paradigm for object detection," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 258–274.

[37] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[38] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.

[39] Y. Li, Y. Chen, N. Wang, and Z.-X. Zhang, "Scale-aware trident networks for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6054–6063.

[40] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6169–6181, Jul. 2020.

[41] H. Guo, X. Su, S. Tang, B. Du, and L. Zhang, "Scale-robust deep-supervision network for mapping building footprints from high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10091–10100, 2021.

[42] T. Liu et al., "Multi-scale attention integrated hierarchical networks for high-resolution building footprint extraction," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 109, May 2022, Art. no. 102768.

[43] Y. Wu, L. Xu, Y. Chen, A. Wong, and D. A. Clausi, "TAL: Topography-aware multi-resolution fusion learning for enhanced building footprint extraction," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6506305.

[44] N. Xue, S. Bai, F. Wang, G.-S. Xia, T. Wu, and L. Zhang, "Learning attraction field representation for robust line segment detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1595–1603.

[45] S. P. Mohanty. (2018). *CrowdAI Mapping Challenge 2018: Baseline With Mask-RCNN*. [Online]. Available: https://github.com/crowdai/crowdai-mapping-challenge-mask-rcnn

[46] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The Inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 3226–3229.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[48] Z. Ma, M. Xia, L. Weng, and H. Lin, "Local feature search network for building and water segmentation of remote sensing image," *Sustainability*, vol. 15, no. 4, p. 3034, Feb. 2023.

[49] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2018.

[50] S. Zorzi, K. Bittner, and F. Fraundorfer, "Machine-learned regularization and polygonization of building segmentation masks," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 3098–3105.

[51] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.

[52] H. He et al., "Mask R-CNN based automated identification and extraction of oil well sites," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, Aug. 2022, Art. no. 102875.

[53] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, and A. Kirillov, "Boundary IoU: Improving object-centric image segmentation evaluation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15334–15342.

[54] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[55] W. Li, W. Zhao, H. Zhong, C. He, and D. Lin, "Joint semantic-geometric learning for polygonal building segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1958–1965.

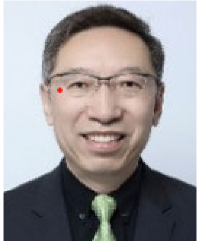
Hongjie He received the B.Sc. degree in geomatics from the China University of Petroleum, Qingdao, China, in 2016, the M.Sc. degree in cartography and geographic information systems from Lanzhou University, Lanzhou, China, in 2019, and the Ph.D. degree in geography specializing in applied earth observations from the Geospatial Sensing and Data Intelligence Group, University of Waterloo, Waterloo, ON, Canada, in October 2023.

He has authored papers in *International Journal of Applied Earth Observation and Geoinformation*, *Canadian Journal of Remote Sensing*, and *Geomatica*, and flagship conferences, including IGARSS and ISPRS. His research interests include AI-based algorithms and software tools for information extraction from Earth observation images.

Lingfei Ma (Member, IEEE) received the Ph.D. degree in geomatics engineering from the University of Waterloo, Waterloo, ON, Canada, in 2020.

He is currently an Associate Professor with the Central University of Finance and Economics, Beijing, China. He has authored more than 50 papers in refereed journals and conferences, including IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, *ISPRS Journal of Photogrammetry and Remote Sensing*, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, and IEEE-CVPRW. His research interests include autonomous driving, mobile laser scanning, intelligent processing of point clouds, 3-D scene modeling, and machine learning.

Dr. Ma was a recipient of the 2020 National Best Ph.D. Thesis Award granted by the Canadian Remote Sensing Society. He served as the Guest Editor for the *International Journal of Applied Earth Observation and Geoinformation*.



Jonathan Li (Fellow, IEEE) received the Ph.D. degree in geomatics engineering from the University of Cape Town, Cape Town, South Africa, in 2000.

He is currently a Professor of geomatics and systems design engineering with the University of Waterloo, Waterloo, ON, Canada. He has supervised nearly 200 master's/Ph.D. Students as well as Post-Doctoral Fellows/Visiting Scholars to completion. He has authored or coauthored almost 600 publications, more than 150 of which were published in top remote-sensing journals, including *Remote Sensing of Environment*, *ISPRS Journal of Photogrammetry and Remote Sensing*, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, and *International Journal of Applied Earth Observation and Geoinformation*. He has also published papers in flagship conferences in computer vision and AI, including CVPR, AAAI, and IJCAI. His main research interests include AI-based information extraction from Earth observation images and LiDAR point clouds, pointgrammetry, remote sensing, GeoAI and 3-D vision for digital twin cities, and autonomous driving.

Dr. Li is a fellow of the Canadian Academy of Engineering, the Royal Society of Canada (Academy of Science), and the Engineering Institute of Canada. He is the Editor-in-Chief of *International Journal of Applied Earth Observation and Geoinformation* and an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS.