Optimization of Rooftop Delineation from Aerial Imagery with Deep Learning

by

Hongjie He

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Geography

Waterloo, Ontario, Canada, 2023

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Dr. Jinfei Wang Dept. of Geography & Environment, University of Western Ontario
Supervisor(s):	Dr. Jonathan Li Dept. of Geography & Environmental Management, University of Waterloo
Internal Member:	Dr. Grant Gunn Dept. of Geography & Environmental Management, University of Waterloo
Internal Member:	Dr. Michael A. Chapman Dept. of Civil Engineering (Adjunct GEM), Toronto Metropolitan University
Internal-External Member:	Dr. Linlin Xu Dept. of Systems Design Engineering, University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This doctoral thesis is accomplished in the manuscript option, following the graduation guidelines administered by the joint Waterloo-Laurier Graduate Program in Geography. It consists of three required manuscripts. The first paper was published in the International Journal of Applied Earth Observation and Geoinformation; the second paper is under revision for the IEEE Transactions on Geoscience and Remote Sensing. The third paper was submitted to the IEEE Transactions on Geoscience and Remote Sensing. These three papers are presented in Chapters 3, 4, and 5, respectively, with a minor revision for a consistent format. The introduction and literature review sections of the three papers comprise Chapters 1 and 2 of the thesis, respectively.

1. Hongjie He, Kyle Gao, Weikai Tan, Lanying Wang, Nan Chen, Lingfei Ma, and Jonathan Li. Super-resolving and composing building dataset using a momentum spatial-channel attention residual feature aggregation network. International Journal of Applied Earth Observation and Geoinformation, 111:102826, 2022.

2. Hongjie He, Lingfei Ma, and Jonathan Li. HigherNet-DST: Higher resolution network with dynamic scale training for rooftop delineation. IEEE Transactions on Geoscience and Remote Sensing, under revision.

3. Hongjie He, Lingfei Ma, and Jonathan Li. Box2Boundary: Delineation of rooftops with box supervision. IEEE Transactions on Geoscience and Remote Sensing, submitted.

I am the first author for all those three manuscripts mentioned above, and my supervisor, Prof. Dr. Jonathan Li, is the corresponding author. These papers are predominantly the result of my intellectual effort, including designing the ideas and methods, conducting experiments, and writing the manuscripts. Other co-authors contributed to these papers through document preparation, review, and paper editing.

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 leaning methods has become feasible and draw much attention. In the context of rooftop delineation, the end-to-end Deep Convolutional Neural Networks (DCNNs) have demonstrated remarkable performance in accurately delineating rooftops from aerial imagery. However, several challenges still exist in this task, which are addressed in this thesis. These challenges include: (1) the generalization issues of models when test data differ from training data, (2) the scale-variance issues in rooftop delineation, and (3) the high cost of annotating accurate rooftop boundaries.

To address the challenges mentioned above, this thesis proposes three novel deep learning-based methods. Firstly, a super-resolution network named Momentum and Spatial-Channel Attention Residual Feature Aggregation Network (MSCA-RFANet) is proposed to tackle the generalization issue. The proposed super-resolution network shows better performance compared to its baseline and other state-of-the-art methods. In addition, data composition with MSCA-RFANet shows high performance on dealing with the generalization issues. Secondly, an end-to-end rooftop delineation network named Higher Resolution Network with Dynamic Scale Training (HigherNet-DST) is developed to mitigate the scale-variance issue. The experimental results on publicly available building datasets demonstrate that HigherNet-DST achieves competitive performance in rooftop delineation, particularly excelling in accurately delineating small buildings. Lastly, a weakly supervised deep learning network named Box2Boundary is developed to reduce the annotation cost. The experimental results show that Box2Boundary with post processing is effective in dealing with the cost annotation issues with decent performance. Consequently, the research with these three sub-topics and the three resulting papers are thought to hold potential implications for various practical applications.

Acknowledgements

I would like to express my sincere gratitude to the following individuals and organizations for their support, funding, and contributions to this doctoral thesis.

First and foremost, I am deeply grateful to my thesis advisor Prof. Dr. Jonathan Li for his guidance, encouragement, and patience throughout my doctoral study journey. His insights, expertise and passion have been invaluable in shaping my research direction and improving the quality of my work.

Then, I would like to thank my thesis committee members: Prof. Dr. Michael A. Chapman, Department of Civil Engineering, Toronto Metropolitan University; Prof. Dr. Grant Gunn, Department of Geography and Environmental Management, University of Waterloo; Prof. Dr. Linlin Xu, Department of Systems Design Engineering, University of Waterloo; and Prof. Dr. Jinfei Wang for their valuable comments and suggestions. Their help significantly improves the quality of the thesis.

I am grateful to all my colleagues in the Geospatial Intelligence and Mapping (GIM) Lab for creating a collaborative and warm work environment. They are always ready to help each other. During the COVID-19 pandemic, this inclusive environment is particularly invaluable. I would like to give special thanks to Kyle Gao, Dr. Lingfei Ma, Yuwei Cai, Dr. Ming Liu, Dr. Ke Yang, Weikai Tan, Hongzhang Xu, Lanying Wang, Sarah Narges Fatholahi, Dening Lu, Dr. Nan Chen, Liangzhi Li, Qiutong Yu, Dr. Kun Zhao, Dr. Junbo Wang, Bingxu Hu, Zijian Jiang, Dr. Yan Liu, Zhehan Zhang, Liyuan Qing, Hasti Andon Petrosians, Dr. Awase Syed, Xuanchen Liu, Jirui Hu, Jingyi Hu, Zihao Yang, Longxiang Xu, Yuxiang Fang, Wenxuan Zhu and Zhuoran Pan for their academic assistance and lifelong friendships.

I would also like to thank Alan Anthony, Graduate Program Administrator in the Department of Geography and Environmental Management, for answering trivial questions during my PhD career with extreme patience and willingness. Thanks to the Geospatial Centre of the University of Waterloo for providing valuable data for my research. Thanks to Jennifer Keir and Michael Tjendra from the Centre for Mapping, Analysis & Design (MAD) for their technical support during my PhD comprehensive examination.

I would extend my thanks to my roommates Ying Wu, Peiqi Wang, Jinyan Wu and Lutong Li, as well as my close friend Tianjiao Chen, who helped me adapt to the new study career and environment in Canada quickly and happily. Thanks for sharing the happy time with me.

Finally, I would show my deepest gratitude to my parents and my sisters for their emotional encouragement and endless love. I would like to thank my country and the

grant from the China Scholarship Council, so that I could have the opportunity to pursue my PhD degree and to have an unforgettable experience in Canada.

Thanks again for all individuals and groups listed above. Thank you all for making this journey possible.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

3D	Three-dimensional
ACGC	Adversarial Climbing Gated Convolution
ACM	Active Contour Model
AFM	Attraction Field Map
AI	Artificial Intelligence
AIRS	Aerial Imagery for Roof Segmentation
AP	Average Precision
AR	Average Recall
BANA	Background-Aware pooling and Noise-Aware loss
BBAM	Bounding Box Attribution Map
BDD	Berkeley DeepDrive dataset
BoxCaSeg	Box-supervised Class-Agnostic object Segmentation
BRB	Boundary Refinement Block
CAM	Class Activation Map
CapsFPN	Capsule Feature Pyramid Network
CBR-Net	the Coarse-to-fine Boundary Refinement Network
CFA	Cross-scale Feature Aggregation
CNN	Convolutional Neural Network
ConvLSTM	Convolutional Long-Short Term Memory
ConvNet	Convolutional Neural Network
ConvGRU	Convolutional Gated Recurrent Unit
CRF	Condional Random Field
DARNet	Deep Active Ray Network
DCN	Deformable Convolution Network
DDRNet	Deep Dual-Resolution Network
DL	Deep Learning
DP	Douglas-Peucker

DRM	Dual Relationship Module
DRCN	Deeply Recursive Convolutional Network
DRRN	Deep Recursive Residual Network
DSAC	Deep Structured Active Contour
DSM	Digital Surface Model
DS-Net	Deep-Supervision convolutional neural Network
DST	Dynamic Scale Training
ECHO	Extraction and Classification of HOmogeneous
ECMN	Edge-Constrained Multitask Network
EDSR	Enhanced Deep Super-Resolution network
EMU-CNN	Multiscale U-shaped Convolutional Neural Network building instance extraction framework with Edge constraint
EPS	Explicit Pseudo-pixel Supervision
ERI	Edge Regularity Indices
SLI	Shadow Line Indices
ESA	Enhanced Spatial Attention
ESPCN	Efficient Sub-Pixel Convolutional Neural Network
FCN	Fully Convolutional Networks
FPN	Feature Pyramid Network
FOCA	First-Order Channel Attention
FSG	Fuzzy Stacked Generalization
GC	Global Context
GELU	Gaussian Error Linear Unit
GIS	Geographic Information System
GSD	Ground Sampling Distance
GCB	Global Context Block
GFLOPs	Giga Floating Point Operations
GPU	Graphic Processing Unit
GNN	Graph Neural Network
DCNN	Deep Convolutional Neural Network
HD	High Definition
HiSup	Hierarchical Supervised learning
HigherNet	Higher resolution Network
HigherNet-DST	Higher resolution Network with Dynamic Scale Training
HR	High Resolution
HRNet	High Resolution Network
HSR	High Spatial Resolution

IAD	Instance-Aware Decoder
IoU	Intersection over Union
ISPRS	International Society for Photogrammetry and Remote Sensing
ISODATA	Iterative Self-Organizing Data Analysis Techniques
LapSRN	Laplacian Pyramid Super-Resolution Network
LiDAR	Light Detection and Ranging
LSC	Long Skip Connection
LSTM	Long-Short Term Memory
LR	Low Resolution
LN	Layer Normalization
M2S	Multi to Single Module
MAE	Mean Absolute Error
MAP-Net	Multiple Attending Path neural Network
Mask R-CNN	Mask Region Based Convolutional Neural Network
MBI	Morphological Building Index
MDL	Maximum Description Length
MDSR	Multi-scale Deep Super-Resolution system
MFR	Multiscale Feature Retrieval
PGC	Pseudo-mask Generation and Correction
MFUN	Multiscale Fusion U-shaped Network
MIL	Multiple Instance Learning
mIoU	mean Intersection over Union
MRF	Markov Random Fields
MSCA-RFANet	Momentum Spatial Channel Attention RFANet
MS COCO	MicroSoft Common Objects in COntext
MS-GeoNet	Multi-Scale Geoscience Network
MSG-SR-Net	MultiScale Generation and Superpixel Refinement
MSI	Morphological Shadow Index
MSPA	Morphological Spatial PAttern
nDSM	normalized Digital Surface Model
NMS	Non-Maximum Suppression
OA	Overall Accuracy
OSM	OpenStreetMap
PANet	Pixelwise Affinity Network
PVT	Pyramid Vision Transformer
PSNR	Peak Signal-to-Noise Ratio
RCAN	Residual Channel Attention Network

RED-Net	Residual Encoder-Decoder Network
ResNet	Residual Network
RF	Random Forest
RFANet	Residual Feature Aggregation Network
RGB	Red, Green, and Blue
RL-NL	Region-Level Non-Local
RMSE	Root Mean Square Error
RPN	Region Proposal Network
SAN	Second-order Attention Network
SAR	Synthetic Aperture Radar
SAWSN	Structure-Aware Weakly Supervised Network
SCA	Spatial Channel Attention
SEAM	Self-supervised Equivariant Attention Mechanism
SGD	Stochastic Gradient Descent
SNIP	Scale Normalized for Image Pyramid
SOCA	Second-Order Channel Attention
SPMF	Superpixel Pooling and Multi-scale Feature
SRCNN	Super-Resolution Convolutional Neural Network
SRDenseNet	Dense Network for Super-Resolution
SRGAN	Super-Resolution Generative Adversarial Network
SSIM	Structural Similarity Index Measure
SSRG	Share-Source Residual Group
Swin	Shifted windows
SWOOP	Southwestern Ontario Orthophotography Project
SISR	Single-Image Super-Resolution
SVM	Support Vector Machine
TAL	Topography-Aware Loss
TDAC	Trainable Deep Active Contour
TM-SPOT	Thematic Mapper-Satellite Pour l'Observation de la Terre
TridenNet	Trident Network
URSS	Uncertainty Reduction and Self-Supervision
VDSR	Very Deep Super-Resolution
VHSR	Very High Spatial Resolution
VOC	Visual Object Class
WHU	Wuhan University
WSF-NET	Weakly Supervised Feature-fusion NETwork
WSOD	Weakly-supervised Salient Object Detection

Chapter 1

Introduction

1.1 Background and Motivations

Buildings, as one of the key elements in urban areas, have been used as an indicator for the urban change detection. Since urbanization developed rapidly in recent years, automatic building mapping has drawn much attention [20, 157]. Specifically, building rooftop or footprint maps play key roles in several municipal applications, such as urban planning and management, urban cadastral management, geo-databased updating, population estimation, infrastructure development and smart city construction [165, 122, 51]. The population density maps generated by population estimation with rooftops or building footprints can further be used in epidemic or pandemic control, as illustrated in malaria control studies [41]. In addition, building maps are also important to homeowners, local authorities, and insurance companies in natural hazard management and damage estimation [134, 144]. For example, accurate building maps are required to estimate damage and assess the risk brought by earthquakes, storms and other geological disasters [128, 160]. In such emergent events, building maps should be obtained effectively and efficiently at any cost once these disasters occur [160]. For these applications, the remote sensing imagery-based, especially the aerial imagery-based, methods provide promising solutions.

Detailed building information is required for the aforementioned applications. According to the Nyquist–Shannon sampling theorem, the sampling frequency, i.e., the inverse of the ground sampling distance (GSD), must be higher or equal to twice the highest spatial frequency of the signal [38]. Take rooftop delineation as an example, to extract 0.3 m wide eaves, a spatial resolution of 0.15 m or higher one is required. In the literature, submetre aerial images are commonly used in rooftop delineation. As for methods used for this

task, they can be categorized into (1) manual annotation, (2) hand-crafted feature-based methods, and (3) deep learning (DL)-based methods [111]. Manual annotation is notorious for being time-consuming and labor-intensive. In hand-crafted feature-based methods, spectral information, geometric information and/or height information are collected based on expert knowledge to delineate rooftops. Hypotheses about the geometric features of the building are employed to simplify the delineation process. Obviously, expert knowledge is required, and these methods can only be used in certain limited situations. In contrast, deep learning-based methods can learn features directly from data, achieving higher levels of accuracy and speed. While these methods require significant computational resources, advancements in computing technology are mitigating this issue. Consequently, the Deep Convolutional Neural Networks (DCNN) dominated in computer vision tasks [76] have been successfully applied in remote sensing and rooftop delineation.

Despite recent advances, the generalization ability of models to data outside the training set remains a challenge. Applying certain methods such as regularization in model training can alleviate the problem to some extent. Another way is to construct a dataset with a wide variety of characteristics and distributions [78]. In addition, the models may not perform as expected in datasets which have different spatial resolutions with the training data. For rooftop delineation, given the wide application of high spatial resolution imagery [11] and the abundance of public building datasets released in the past decade [125, 109, 103, 66, 148, 124, 51], composing these datasets to overcome generalization errors can be a more promising solution. Since these datasets have varying spatial resolutions, ranging from 5 cm (the ISPRS Vaihingen and Potsdam Datasets [125]) to 1 m (the Massachusetts Building Dataset [109]), as well as different spectral bands, it is necessary to perform spatial resolution enhancement and band selection in order to process and compose these datasets. For the band selection, one can preserve the most discriminative and commonly used combinations of Red, Green, and Blue (RGB) bands[22]. For the spatial resolution integration, powerful methods are required to process different datasets used in rooftop delineation.

Due to the availability of a large volume of aerial imagery with building annotations, DCNN-based rooftop delineation methods have achieved high accuracy by partially addressing the scale-variance issue, the intra-class variation issue, and the inter-class similarity [134]. However, the occlusion issue and blurred rooftop boundary still exist in the extraction results [98, 133]. The end-to-end DCNN for rooftop delineation has drawn much attention recently. These methods were initially introduced to directly generate vectorized building maps from remote sensing imagery without any or with simple post-processing [90]. In addition, by directly outputting vectorized results with building corners, the

occlusion and blurred boundaries are relieved significantly. Nonetheless, scale[1]-variance problems still exist in practice. Specifically, rooftop delineation gives poor performance on small buildings compared to large and medium sized buildings. Therefore, advanced building delineation methods are required to improve the overall performance by addressing scale-variance issues on small buildings.

The quality and data volume required in rooftop delineation bring high cost in data annotation. As a result, costly annotation inhibits the use of DL-based rooftop delineation methods in practice [178]. Deep learning methods, including weakly supervised learning, semi-supervised learning, and self-supervised learning, can be used to liberate this constraint from traditional full supervised DL, using weak annotations, limited annotations, and image representation learning, respectively [120]. For rooftop delineation, weakly supervised learning is widely used given its flexibility, although current research mainly employs weakly supervised semantic segmentation instead of instance segmentation. However, weakly supervised instance segmentation methods are more suitable for rooftop delineation as individual rooftop can be extracted. In computer vision, weakly supervised instance segmentation methods have been well explored. Therefore, advanced weakly supervised instance segmentation methods are achievable, which are required for weakly supervised rooftop delineation.

In summary, this thesis focuses on three challenges in rooftop delineation for the rapid creation of High-Definition (HD) maps.

(1) Generalization issues regarding differences in training and test data in terms of spatial resolution and image characteristics. DL-based rooftop delineation methods have poor performance when test data differ from training data. Data composition is a promising way to overcome the issue. In the same geo-location, spatial resolution and spectral resolution are the most different characteristics. Compared to spectral differences, spatial resolution differences require better methods.

(2) Scale-variance issues in rooftop delineation when building sizes vary extremely. Current DCNN methods are typically robust against scale variance issues. However, poor performance on small objects cannot be fully addressed. In cases with many small buildings, scale-variance has a severe negative impact on objectives and total performance. Therefore, scale-variance issues must be addressed.

(3) Costly annotation hinders the widespread use of DL-based methods for rooftop delineation. The high performance of DL-based methods heavily relies on the high quality

[1]The term "scale-variance" is commonly employed to characterize the issue addressed in this thesis and is used in preference to other alternatives. In this context, 'scale' refers to the size of objects, distinct from its geographical meaning.

and large volume of well annotated training data. But the annotation work is labor-intensive and expensive. How to fully exploit the potential of deep learning and reduce the annotation cost should be the key focus in rooftop delineation.

1.2 Objectives

To deploy DL-based methods for rooftop delineation, the challenges mentioned above need to be carefully addressed. Specifically, in this thesis, several DL-based methods are proposed to tackle generalization issues, scale variance issues and costly annotation issues in rooftop delineation, respectively (as shown in Figure 1.1).

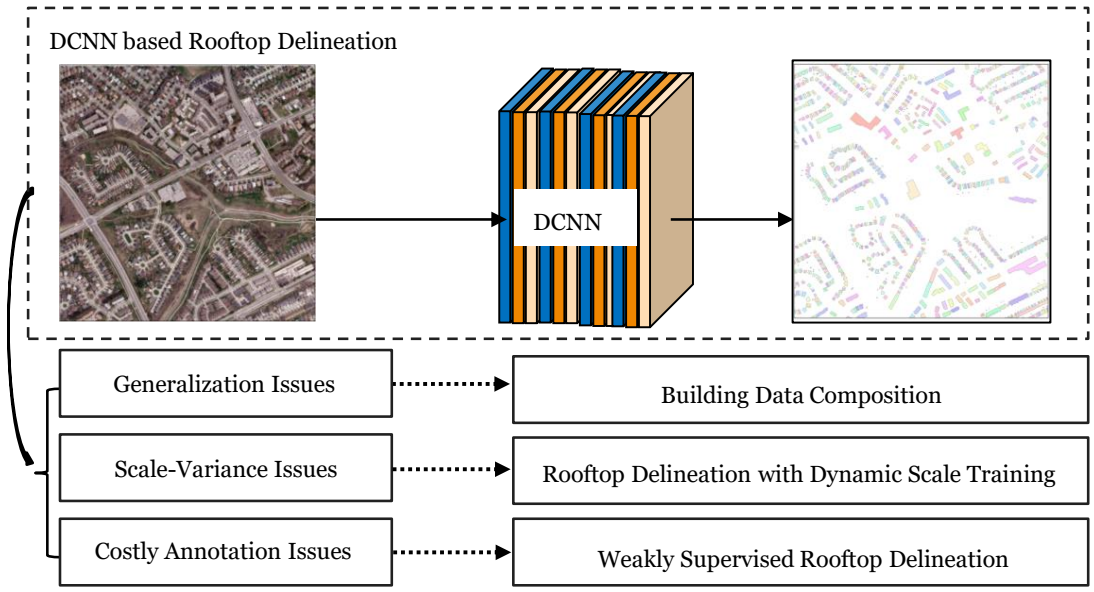

Figure 1.1: Logic flow of the thesis

(1) In order to address the generalization issues, it is necessary to compose diverse datasets with varying characteristics for model training. To achieve this, the integration of publicly available building datasets with varying spatial resolutions necessitates the utilization of advanced deep learning-based super-resolution techniques.

(2) In order to address the scale-variance issues in rooftop delineation and improve the extraction of small buildings, it is essential to develop a new method that can effectively mitigate the scale-variance issue and enhance the performance of rooftop delineation.

(3) To overcome the costly annotation issues, it is crucial to employ weak supervision targets for rooftop delineation while striving for high performance. Therefore, it is imperative to reassess weakly supervised rooftop delineation techniques and introduce novel methods that offer improved performance.

1.3 Thesis Structure

This thesis consists of six chapters. Figure 1.2 shows the structure of the thesis.

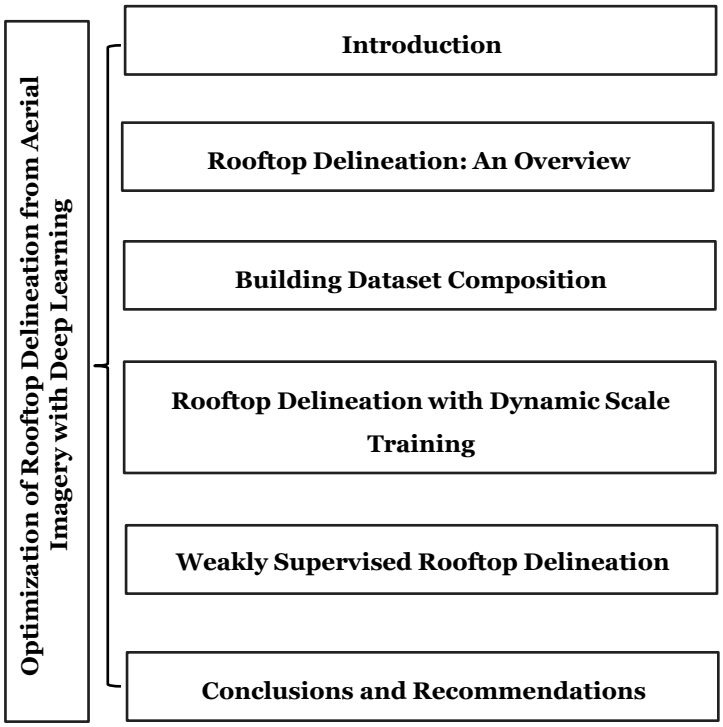

Figure 1.2: Structure of the thesis

Chapter 1 introduces the background and the motivation, as well as the objectives, of the thesis. Three challenges in the deployment of DL-based rooftop delineation in practice are placed as the focus.

Chapter 2 briefly reviews the publicly available building datasets, the applications of deep learning in rooftop delineation, the methods used for dealing with generalization

issues, scale variance issues and costly annotation issues in literature.

Chapter 3 introduces the data composition with the Momentum Spatial Channel Attention Residual Feature Aggregation Network (MSCA-RFANet). With extensive experiments, the impact of super-resolution on rooftop delineation is well explored, and the new method shows the state-of-the-art performance in super-resolution compared to baselines. Better performance in super-resolution results in more positive impact on rooftop delineation.

Chapter 4 describes the method used for dealing with the scale variance issue in end-to-end manner rooftop delineation with the Higher Resolution Network with Dynamic Scale Training (HigherNet-DST) method. The extensive experiments demonstrate its superior performance compared to other state-of-the-art methods.

Chapter 5 details box-supervised rooftop delineation with the Box2Boundary method. Extensive experiments show the superiority of the Box2Boundary surpassing all other weakly supervised methods with competitive performance compared to fully supervised rooftop delineation methods.

Chapter 6 concludes the thesis and indicates the potential future research directions.

Chapter 2

Rooftop Delineation: An Overview

2.1 Publicly Available Building Datasets

With the advancement of imaging technologies, there is increasing accessibility to high spatial resolution satellite and aerial images. Consequently, several publicly available building datasets have been released in recent years.

(1) The Vaihingen and Potsdam datasets, released by the International Society for Photogrammetry and Remote Sensing (ISPRS) [125], are two relatively small building datasets. These datasets consist of six classes, including impervious surfaces, buildings, low vegetation, trees, cars, and clutter. The Vaihingen dataset comprises 33 images, offering a spatial resolution of 0.09 m/pixel. It encompasses spectral bands of Red, Green, Blue, and Near Infrared, with an image size of approximately 2,500×2,500 pixels. On the other hand, the Potsdam dataset consists of 38 images with a spatial resolution of 0.05 m/pixel, matching the spectral resolution of the Vaihingen dataset. The image size for the Potsdam dataset is approximately 6,000×6,000 pixels. Furthermore, both datasets include corresponding Digital Surface Model (DSM) data alongside the image data. These datasets are notable for having the highest spatial resolution among existing datasets, despite covering only a 5 km^2 area.

(2) The Massachusetts building dataset [109] consists of 151 aerial images with a spatial resolution of 1 m/pixel. These images, featuring RGB bands and a size of 1,500×1,500 pixels, are classified into building and non-building categories. This dataset covers approximately 340 km^2 in the Boston area. It is split into training (137 images), validation (4 images), and testing (10 images) sets. According to the author, the dataset demonstrates high accuracy, with an average omission of building classification of less than 5

(3) The Inria building dataset [103] also has building and non-building classes. It contains aerial images covering 810 km^2 of 10 cities in the USA and Austria. The training set and the testing set evenly take 360 images with a spatial resolution of 0.3 m/pixel and RGB bands. The dataset aims at exploring the generalizability ability of convolutional neural networks, so adjacent images are split into training and testing sets.

(4) The WHU (Wuhan University) building dataset [66] encompasses both an aerial image dataset and a satellite image dataset for building extraction. The aerial image dataset consists of 8,189 tiles, each with a spatial resolution of 0.3 m/pixel. These tiles are captured in RGB bands and possess dimensions of 512×512 pixels. This dataset covers an extensive area of 450 km^2 in Christchurch, New Zealand. The satellite image dataset is divided into two distinct sets. The first set includes 204 images from 6 cities worldwide, with spatial resolutions ranging from 0.3 m/pixel to 2.5 m/pixel. Each image in this set has a size of 512×512 pixels. The second satellite image dataset comprises 17,388 tiles from 6 adjacent images. These tiles have a spatial resolution of 0.45 m/pixel and dimensions of 512×512 pixels. The coverage area for this dataset extends over 860 km^2 in East Asia. It is worth noting that these two satellite image datasets employ different sensors, resulting in varying spectral resolutions.

(5) The SpaceNet building dataset [148] was released via two SpaceNet challenges for building detection. For this dataset, five cities from various regions were selected as areas of interest: Las Vegas, USA; Paris, France; Rio de Janeiro, Brazil; Shanghai, China; and Khartoum, Sudan. WorldView-2 and WorldView-3 images were used, which were cropped into smaller patches with dimensions of 650×650 pixels. These image patches collectively cover a total area of 5,555 km^2 across different continents.

(6) The AIRS (Aerial Imagery for Roof Segmentation) dataset [21] was constructed using the same aerial images as the WHU building dataset, where the original spatial resolution of images (0.075 m/pixel) is preserved.

(7) The Semcity Toulouse dataset [124] is made based on WorldView-2 imagery for building instance segmentation, which covers 50 km^2 area of Toulouse, France. The satellite images in the dataset are classified into 8 classes, namely impervious surface, building, pervious surface, high vegetation, car, water, sport venue and void. These images have a spatial resolution of 0.5 m/pixel for the panchromatic band and 2 m/pixel for the other bands. Each image is split into 16 tiles. Eventually, the panchromatic band has a size of 3,504×3,452 pixels and other bands have a size of 876×863 pixels.

(8) The Waterloo Building Dataset, released in 2022, covers an area of 205.83 km^2 in the Kitchener-Waterloo region of Ontario, Canada [51]. This dataset was specifically developed for semantic segmentation, focusing on rooftop delineation. It comprises 242 aerial images,

each with a size of 8,350×8,350 pixels and a spatial resolution of 0.12 m/pixel. Manually labelled binary masks for rooftop are 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, 6,887, and 18,945 pairs of images and masks were assigned into training, validation and test subsets, respectively.

Table 2.1: Publicly available building datasets

Dataset	Location	Spectral Bands	Classes	Coverage (km^2)	Pixel Size (m/pixel)
ISPRS Vaihingen/ Postdam	Vaihingen and Potsdam, Germany	NIR, R, G, B, DSM	6 land-cover classes	1.4/3.4	0.05/0.09
Massachusetts	Massachusetts, USA	R, G, B	Building and non-building	340	1
WHU (aerial)	Christchurch, New Zealand	R, G, B	Building and non-building	457	0.3
Inria	10 regions in the USA and Austria	R, G, B	Building and non-building	810	0.3
SpaceNet	4 cities around the world	WorldView-3 8 bands	Building, road and background	5555	0.3/0.5
AIRS	Christchurch, New Zealand	R, G, B	Building and non-building	457	0.075
ISPRS Semcity Toulouse	Toulouse, France	WorldView-2 8 bands	8 land-cover classes	50	0.5
Waterloo	Kitchener-Waterloo, Canada	R, G, B	Building and non-building	205.83	0.12

In addition to the previously discussed datasets (as summarized in Table 2.1), there are several other building datasets available. These include datasets created for the DeepGlobe Building Extraction Challenge [33], the Open Cities AI Challenge [50], and the Crowd-AI Mapping Challenge [110]. The DeepGlobe building dataset is based on the SpaceNet dataset and focuses on annotating building footprints instead of rooftops. The dataset for the Open Cities AI Challenge consists of building footprints from 10 cities in Africa

but is known for having inconsistent annotation accuracy. The dataset for the Crowd-AI Mapping Challenge comprises over 40,000 tiles of RGB images, each with a size of 300×300 pixels. However, the buildings are homogeneous in the dataset, making them easier to be segmented from the background when compared with other datasets [124].

2.2 Rooftop Delineation

In this section, existing rooftop delineation methods are comprehensively reviewed. The initial approach, manual editing, involves visual interpretation and the subsequent generation of polygons. In the realm of hand-crafted feature-based methods, features are meticulously crafted through expert knowledge to encapsulate pertinent information crucial for distinguishing various classes or categories within the image. These methods were the dominant paradigm before the advent of DCNN-based methods in rooftop delineation. Thus, this section will focus on the review of hand-crafted feature-based methods and DCNN-based methods.

2.2.1 Hand-crafted Feature-based Rooftop Delineation

In hand-crafted feature-based methods, spectral information, geometric information, and/or height information are collected based on expert knowledge to facilitate rooftop delineation.

The pioneering approaches employed geometric details, including edge, line, and corner information [62, 64], as well as shadow information [94, 108], which are fundamental for building extraction. Additionally, some methods integrated spectral information for building mapping [131, 81]. These methods are classified into three categories: classification strategies-based methods, active contour-based methods, and graph-based methods[113].

Classification Strategies-Based Methods: In the domain of spectral feature methods, two notable approaches stand out. Zhang[185] employed Iterative Self-Organizing Data Analysis Techniques (ISODATA) clustering on merged Thematic Mapper-Satellite Pour l'Observation de la Terre (TM-SPOT) data for preliminary building extraction, followed by gray value co-occurrence matrix filtering to eliminate non-building entities. Subsequent refinement involved additional ISODATA clustering and manual editing. The complexities of the method hindered its efficiency. Lee et al. [81] combined supervised (extraction and classification of homogeneous, ECHO) and unsupervised (ISODATA) methods for building extraction from IKONOS-2 images, yielding a modest 64.4% pixel accuracy due to supervised classification limitations.

Regarding geometric feature-based methods, Inglada [63] integrated geometric invariants and Fourier-Mellin descriptors, achieving a robust pixel accuracy of 92.93% using Support Vector Machines (SVM). Successful application, however, required precise centering of the target objects. Widely adopted, morphological features were employed for non-building entity filtration by Aytekin et al. [5] and Lefèvre et al. [85]. Huang and Zhang [60, 61] introduced morphological building index (MBI) and morphological shadow index (MSI), while morphological spatial pattern (MSPA) improved the classification of preliminary MBI results. Nevertheless, multi-stage methodologies introduced uncertainty and lower accuracy [181].

Concurrently, other strategies encompass diverse features and classification techniques. Integrating spectral and spatial information through a fuzzy pixel-based classifier using pan-sharpened multispectral IKONOS-2 images facilitated urban land cover classification. Enhanced by object-based classification, incorporating shape, spectral, and contextual traits, this approach demonstrated high performance, yet encountered challenges in effectively distinguishing between buildings and other impervious surfaces[131]. Conversely, the Fuzzy Stacked Generalization(FSG) method achieved a pixel accuracy of 84% in detecting building regions from a single QuickBird image, relying on a two-layer hierarchical ensemble learning model. The assumption of the method is the statistical stability of the training and testing data, which is hard to control in practice[130]. Another avenue involved an adaptive fuzzy-genetic algorithm for building detection in IKONOS-2 images, with efficacy dependent on parameter tuning[139]. Meanwhile, a novel Conditional Random Field (CRF) formulation aimed to extract building rooftops, yet pre-segmentation, non-building filtration, and CRF-based segmentation introduced uncertainty, limiting feasibility and accuracy[86]. Novel indices, Edge Regularity Indices (ERI) and Shadow Line Indices (SLI), proposed by Chen et al.[22], demonstrated value in building footprint extraction, validated by Adaboost, Random Forest (RF), and SVM classifiers.

Active Contour-Based Methods: Active contours, also known as snakes [69], have found widespread application in image processing for boundary localization. These contours seek to locate object boundaries by minimizing an energy function comprising external and internal energy terms. The total energy is the sum of these components: the external energy (image energy) pertains to image characteristics that guide contours towards object boundaries, while the internal energy (shape energy) contributes to boundary smoothness [1]. Cao and Yang [12] proposed a Chan-Vese model for building extraction, primarily targeting man-made regions rather than individual buildings. Extending this, Karantzalos and Paragios[68] incorporated prior shape knowledge into active contours to effectively extract buildings from high and very high spatial resolution (HSR/VHSR) images. Ahmadi et al. [2] initialized their snake models using sampled data from buildings

and background, achieving commendable performance in VHSR aerial image analysis. Notably, precise knowledge of building and background classes is essential for optimal results.

Graph-Based Methods: Markov Random Fields (MRF) have been employed to cluster line segments, generating preliminary shapes that are then input into active contour models to generate final results[75]. Katartzis and Sahli [70] employed MRFs to define dependencies among building hypotheses, validated through stochastic optimization, achieving accurate estimates but within certain hypotheses. Izadi and Saeedi [65] adopted graph-based searching to identify building rooftop hypotheses based on lines and intersections, which is effective only for flat roofs. Cui et al. [30] translated building corners and edges into graph vertices and edges. The accuracy of building extraction depended on the initial segmentation quality identifying building locations. Ok [113] presented a two-level graph partitioning framework, refining shadow detection, yet exhibiting unstable performance. Despite its efficiency compared to prior work, this method struggles to differentiate low-contrast buildings from the background.These methods, however, rely on specific hypotheses and are limited to extracting buildings with certain shapes or characteristics.

Numerous studies have utilized data-fusion techniques to enhance building extraction by integrating diverse sources like Light Detection And Ranging (LiDAR) data, Synthetic Aperture Radar (SAR) data, hyper-spectral data, existing Geographic Information System (GIS) building layers, and predefined models. These supplementary sources validate results and offer extra information such as building heights[113]. For example, LiDAR data especially for height information, are widely employed in building extraction[4, 43, 82, 138, 141], often combined with spectral and geometric data to boost accuracy for building extraction. However, the availability of free LiDAR datasets is limited.

In summary, hand-crafted methods are developed based on certain hypotheses about the geometric features of buildings or developed for certain environments. Consequently, those methods can only be used in certain scenarios. In addition, techniques used in those methods, such as classification methods and multi-stage processing framework, also limit the accuracy of these methods. Therefore, to develop a CNN-based building footprints extraction method has been drawing much attention because of its high performance.

2.2.2 DCNN-based Rooftop Delineation

Due to its performance in computer vision tasks, deep learning has been widely used in remote sensing applications [197]. To the best of my knowledge, the earliest applications of using DCNN in rooftop delineation can be traced back to the studies conducted by Mnih

[109] and Shu [134]. In their research, the Convolutional Neural Network (CNN) was used to extract features, with fully connected layers for features flattening, for pixel-level image classification and rooftop delineation. However, this kind of method has low efficiency with limited input sizes.

With the proposal of Fully Convolutional Networks (FCN) [101] and the U-Net [123], pixel-wise image classification, also known as semantic segmentation, has grown rapidly with a large number of new methods invented yearly. In the context of rooftop delineation from remote sensing imagery, various advanced deep learning techniques have been applied, ranging from the ConvNet [175] to the Capsule Feature Pyramid Network (CapsFPN) [174], and the Coarse-to-fine Boundary Refinement Network (CBR-Net) [45]. While these state-of-the-art methods can extract accurate building masks, additional post-processing is still necessary to generate vectorized building polygons, which are essential for creating building maps.

In order to generate vectorized rooftop boundaries 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. [187] employed the Mask Region based Convolutional Neural Networks (R-CNN) first for instance segmentation and instance masks generation. Instance masks were then converted to polygons using the Douglas-Peucker algorithm and the Minimum Description Length (MDL) optimization with generated hypotheses [187]. Regularization methods, such as the ACM [69], also known as snake, which can be embedded into DCNN architectures and generate polygons in an end-to-end manner. In this direction, Marcos et al. [106] proposed the Deep Structured Active Contours (DSAC), which combined deep learning and the ACM for image segmentation. Gur et al. [48] proposed an end-to-end trainable ACM via differentiable rendering. Similarly, Hatamizadeh et al. [49] proposed the Trainable Deep Active Contour (TDAC) model to directly delineate building polygons from aerial images. Cheng et al. [26] 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 of which is the PolyMapper [90], which utilizes the Convolutional Long-Short Term Memory (ConvLSTM) to predict the sequence of vertices of building boundaries from CNN features. Zhao et al. [189] improved upon the PolyMapper by replacing ConvLSTM with Convolutional Gated Recurrent Unit (ConvGRU) and decorating the original backbone with Global Context Block (GCB) and Boundary Refinement Block (BRB). Recently, Girard et al. [44] proposed the frame field learning for rooftop delineation by introducing the frame field targets to optimize models. Zorzi et al. [199] proposed the PolyWorld by employing the Graph Neural Network

(GNN) and a sophisticatedly designed loss function. Xu et al. [166] proposed the Hierarchical Supervisions (HiSup) learning scheme with hierarchical building representations, including the low-level convex and concave building vertices, the mid-level Attraction Field Maps (AFM) for line segments and the high-level regional masks of buildings. These three methods have demonstrated high performance in building extraction and represent the state-of-the-art approaches.

2.3 Generalization Issues

Data composition is an intuitive method for tackling generalization issues. It involves constructing large datasets using different sub-datasets, which enriches the training data with different characteristics. However, challenges may arise during the integration process for rooftop delineation, particularly when dealing with varying spatial resolutions. In such cases, super-resolution techniques can be employed as an effective solution.

2.3.1 Super-Resolution

Super-resolution methods are commonly categorized into two groups: the joint image super-resolution [107] and the Single Image Super-Resolution (SISR) [170]. The former is usually applied to hyperspectral imagery. It utilizes spectral information from Low Resolution (LR) hyperspectral imagery and spatial information from multispectral imagery [183]. The SISR methods directly process a LR image and output a high resolution (HR) image, which are flexible and easy to use. The pre-trained SISR models can be easily applied to new images. Therefore, the SISR methods are more suitable for composing building datasets.

The first DCNN-based SISR method, proposed by Dong et al. [35], surpassed conventional SISR methods and spurred rapid development in this field. To improve the accuracy of SISR, networks became increasingly deeper. Notable advancements in this direction include the Very Deep Super-Resolution (VDSR) [71], the Deeply Recursive Convolutional Network (DRCN) [72], the Residual Encoder-Decoder Networks (RED-Net) [105], and the Deep Recursive Residual Network (DRRN) [142]. With the development of new DL techniques such as the transposed convolution and the dense block, the Laplacian Pyramid Super-Resolution Network (LapSRN) [77], the Dense Network for Super-Resolution (SRDenseNet) [146], the Super-Resolution Generative Adversarial Network (SRGAN) [80], the Enhanced Deep Super-Resolution network (EDSR) and the Multi-scale Deep Super-Resolution system (MDSR) [91] were proposed.

In recent years, the attention mechanisms have become widely used in DCNNs, and recent DCNN-based SISR methods have adopted this innovation. Examples of such methods include the Residual Channel Attention Network [184], the Second-order Attention Network [32], and the Residual Feature Aggregation Network [95]. The RCAN and the SAN apply the channel attention, whereas the RFANet uses the spatial attention. The Efficient Sub-Pixel Convolutional Neural Network (ESPCN) [132], one of the classic SISR networks, is used as an up-sampling module in those three methods.

2.3.2 Data Composition

In computer vision, there are two types of data mixing: single-domain data mixing and cross-domain data mixing [78]. In single-domain data mixing, datasets with a similar purpose are mixed, such as combining various driving datasets. On the other hand, Lambert et al. [78] performed cross-domain data mixing by merging datasets from multiple domains for semantic segmentation. They presented the MSeg dataset, which included the MicroSoft® Common Objects in COntext (MS COCO) dataset, the ADE20K dataset, the Mapillary dataset, the India Driving Dataset (IDD), the Berkeley DeepDrive dataset (BDD), the Cityscapes dataset, and the SUN RGB-D dataset. Their experiments showed that training a model on the MSeg dataset resulted in greater robustness compared to training on a single dataset or mixing datasets from a single domain. For rooftop delineation, the Inria Building Dataset [103] was released to address generalization issues. Therefore, the split was carefully made to ensure that no adjacent images existed in the training or testing dataset. However, the dataset is limited to only two countries. In rooftop delineation, it is important to consider generalization issues on a global scale, covering both the Northern and Southern hemispheres and most continents.

2.4 Scale-Variance Issues

2.4.1 Scale-Variance Issues in Computer Vision

In georeferenced and non-georeferenced remote sensed images, a balanced distribution with regard to object scale cannot be guaranteed. This leads to significant variability in performance in common image processing tasks among different scales, which is known as scale variation[23]. In addition, scale variation also limits the overall performance. Compared to large scale and middle scale objects, small scale objects contribute less to the total loss

[93, 23]. This results in less supervision during training and lower performance at smaller scale objects. Therefore, small scale objects should be focused 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 [23] and image pyramid [98] are intuitive. However, as tested in Chen et al. [23], resampling hurts model performance at other scales. The image pyramid is a robust technique, but arbitrarily selected scales may not be suitable for overcoming scale variance. Other image pyramid type methods, such as the SNIP [135] and the SNIPER [136], increase inference burden. In contrast, the collage style data augmentation, as adopted in Bochkovskiy et al. [8], Zhou et al. [193] and Chen et al. [23], has been effective in handling scale variance and has shown high performance.

Feature pyramid and dilation-based methods are model optimization-based methods [23]. In feature pyramid style methods, different scales of feature maps are learned and aggregated. The Feature Pyramid Network (FPN) [92] is the most representative method in this category. The High-Resolution Network (HRNet) [140] aggregates feature maps from four different scales in each stage of each branch (or scale). The HRNet has shown high performance in feature representation. By refining HRNet, the HigherHRNet was proposed in Cheng et al. [25], which showed better performance in feature representation, especially for small objects. The dilation-based methods, such as the Deformable Convolution Networks (DCN) [31] and the Trident Networks (TridentNet) [89], can generate scale-sensitive feature representations with high resolution but suffer from storage issues. Therefore, using the HigherHRNet as the backbone is more useful compared to other optimization-based methods in dealing with scale-variance issues.

2.4.2 Scale-Variance Issues in Rooftop Delineation

The main obstacles that make rooftop delineation challenging include scale variation, complex architectures, and diverse appearances [98, 174, 196, 46, 97]. Despite the use of auxiliary data as input, the issue of scale variance cannot be effectively overcome [89]. 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 commonly used. For example, Liu et al. [98] proposed a multiscale U-shaped CNN building instance extraction framework with edge constraint (EMU-CNN). The EMU-CNN consists of a multiscale 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 resolution 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. [196]. In their work, a multiple attending path neural network (MAP-Net) was proposed, in which the spatial location-preserved multi-scale features were learned by a multi-parallel path taking a sole image as input. The learned multi-scale features enabled the method to be able to extract exact building edges and recognize small building. In addition, Zhu et al. [196] proposed the deep-supervision convolutional neural network (DS-Net) for rooftop delineation also with multi-scale 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 small building. Furthermore, in recent research, Liu et al. [97] proposed an end-to-end Multi-Scale 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. [163] proposed a topography-aware loss (TAL) for better performance on rooftop delineation in semantic segmentation-based methods. Combining multi-scale feature learning by the HRNet, TAL not only showed better performance on regular size building but also on small size building reporting its high performance on dealing with scale variation issues. Overall, multi-scale feature learning is the basis of methods in the model optimization category.

Regarding data preparation-based methods, there is only one study [98]. Images with three different spatial resolution were taken as input in the EMU-CNN bringing multi-scale features and resulting in better performance in rooftop delineation especially for small buildings.

In summary, inspired by Liu et al. [98], combining model optimization and data preparation seems to be effective in dealing with scale variance issues in rooftop delineation.

2.5 Costly Annotation Issues

2.5.1 Weakly Supervised Leaning in Rooftop Delineation

In weakly supervised learning, weak supervision signals, such as image tags and bounding boxes, are used to reduce annotation costs [178]. Weakly supervised semantic segmentation [3, 14, 126, 195], weakly supervised object detection [184, 171, 27, 149] and weakly

supervised instance segmentation [194, 198, 172, 178] are well explored in computer vision and remote sensing field in recent years.

For rooftop delineation, weakly supervised semantic segmentation is widely used. In weakly supervised semantic segmentation, pseudo labels generation is the focus of the research. With the pseudo labels, rooftop delineation can be conducted using fully supervised methods. Among different types of weak supervision signals, image-level annotation (image tag) is the most widely used [9, 40, 169, 180]. Based on a ratio of building pixels to all pixels within the patch, image patches can be tagged as building or non-building. These image-level annotations are used as supervision targets to generate pseudo masks for semantic segmentation model training. Optimizing or completing the pseudo masks has become a significant research focus in weakly supervised semantic segmentation for rooftop delineation. Decent results were reported in these studies. However, instance segmentation may be more suitable for rooftop delineation as it can generate individual rooftop boundaries.

To the best of my knowledge, there are few studies on building instance segmentation using weakly supervised learning. In contrast, in computer vision, weakly supervised instance segmentation has shown promising performance. Similar to weakly supervised semantic segmentation, the generation of pseudo masks is also the key focus in weakly supervised instance segmentation. To achieve high accuracy, bounding boxes are commonly used to generate pseudo masks in weakly supervised instance segmentation [178].

2.5.2 Box Supervised Instance Segmentation

Box supervised instance segmentation can be categorized into two types: multi-stage methods and methods with unified frameworks. For multi-stage methods, pseudo masks generation and instance segmentation are implemented separately. For example, in the Box-supervised Class-Agnostic object Segmentation (BoxCaSeg) [155], a multi-task learning model is first used to generate pseudo masks followed by pseudo masks refinement and instance segmentation. Specifically, the multi-task learning models take fine annotated salient images and box-supervised images as input, with the fine annotated salient images providing precise object localization guidance for the box-supervised images. A novel merged and dropped strategy is applied to refine the masks generated by the multi-task learning model. Instance segmentation is then performed using Mask R-CNN, which takes the input images and proxy masks as training data. Unlike the class agnostic mask generator in BoxCaSeg, the Bounding Box Attribution Map (BBAM) approach utilizes high-level information from the behavior of trained object detectors to identify the smallest area

within an image where the object detector yields nearly identical results to those obtained from the entire image [84]. The BBAM is comprised of these regions, which determine the target object's bounding box and effectively act as a substitute for ground-truth data in weakly supervised instance segmentation. After refinement, the BBAM can serve as the pseudo masks for either semantic segmentation or instance segmentation. Although this kind of method can achieve high performance, the multiple steps involved in the training pipeline and the need to tune numerous hyperparameters make them quite complex.

Weakly supervised instance segmentation with a unified framework focuses on the pixel affinity. Hsu et al. [58] proposed a Multiple Instance Learning (MIL) [56] loss based on Mask R-CNN, which replaces the original segmentation loss. This MIL-based method primarily employs the bounding box tightness prior, where all columns or rows that contain at least one object pixel are considered positive bags, and the remaining columns or rows are considered negative bags. In Mask R-CNN, the bags generated for each region proposal enable the generation of classification probabilities by max-pooling pixel classification probabilities. The bag classification loss can then be backpropagated to optimize the pixel classifier and enhance the accuracy of instance segmentation. Based on an anchor free instance segmentation method, BoxInst was proposed as a powerful weakly supervised instance segmentation with only box annotation [145]. The innovation of the BoxInst is the redesignation of the loss function. They added a surrogate term and a pairwise loss to the original instance segmentation loss. The first one aims at minimizing the discrepancy between the projection of the predicted masks and the ground truth bounding boxes. The pairwise loss employs a loss function that utilizes the prior knowledge that adjacent pixels with similar color values are more probable to belong to the same object category label. DiscoBox is a self-ensembling framework with guidance from a structured teacher network and box supervision. The teacher network models the pairwise pixels relationship both within and across the bounding boxes. Optimizing the teacher network would refine the object masks and generate dense correspondences between objects sharing the same class label. The former one will be taken as pseudo label for task network training and the latter one will help the dense contrastive learning by providing correspondence pairs either for negative pairs or positive pairs [79]. These approaches establish a pairwise affinity connection among neighboring pixel pairs, either partially or across all pairs. However, this oversimplifies the assumption that pixels or colour pairs sharing similarity should have the same label. Consequently, these methods are vulnerable to objects with similar appearances or complex backgrounds, leading to subpar performance in instance segmentation.

The current state-of-the-art method in box supervised instance segmentation, Box2Mask, is a level-set based approach, evolving objects' boundary in an end-to-end manner [88].

Both input images and their deep features are employed for boundary evolution. In addition, a local consistency module, which relies on a kernel that measures the affinity between pixels, is proposed and utilized to extract information about the local context and spatial relationships. Therefore, Box2Mask can be a good start point to develop a box-supervised rooftop delineation method.

2.5.3 Chapter Summary

This chapter presents a summary of publicly available building datasets and DCNN-based methods for rooftop delineation. Then the related work to deal with generalization issues, the scale-variance issues and costly annotation issues are reviewed. For generalization issues, data composition with image processing based on SISR methods is identified as a promising solution. For scale-variance issues, a powerful end-to-end rooftop delineation method with dynamic scale training and higher resolution network is recognized as a promising method. For costly annotation issues, box-supervised instance segmentation is found to be the most suitable method for weakly supervised rooftop delineation. In the following three chapters, these methods and solutions will be detailed.

Chapter 3

Building Datasets Composition[2]

3.1 Introduction

DCNN-based SISR methods have undergone significant development alongside deep learning techniques, exhibiting remarkable advancements in sophistication and performance. Notable progress has been made from the introduction of the Super-Resolution CNN (SR-CNN) [35] to more recent methods like RFANet [95]. In the field of SISR, state-of-the-art techniques such as the RCAN [184], the SAN [32], and the RFANet have emerged, with each achieving exceptional performance on different datasets. Super-resolution techniques have proven to be a promising approach for data composition, and these state-of-the-art methods serve as a solid foundation. In this chapter, two main objectives are focused on. The first objective is to examine super-resolution and dataset composition for the improvement of rooftop delineation. Specifically, a comparative study is conducted to explore the performance of the same DL-based rooftop delineation model on thfe original datasets, the super-resolved datasets, and the composited dataset. The second objective is to benchmark the newly developed MSCA-RFANet, which incorporates the advantages of the RFANet, the residual channel attention mechanism, and share-source skip connection. A comparative study is conducted with four other DL-based methods as well as bicubic interpolation. The objectives of this chapter include:

(1) exploring and examining the effects of super-resolution and data composition on rooftop delineation,

[2]The content of the chapter has been published on International Journal of Applied Earth Observation and Geoinformation with the paper entitled "Super-resolving and composing building dataset using a momentum spatial-channel attention residual feature aggregation network".

(2) and presenting a new SISR method, namely the MSCA-RFANet, and benchmarking it against the state-of-art SISR methods.

3.2 Datasets and Methods

3.2.1 Datasets

Data used for training SISR methods

The aerial images selected for training and testing the SISR methods are from two sources: the Southwestern Ontario Orthophotography Project 2010 (SWOOP 2010)[3] and aerial images [4] covering the Kitchener-Waterloo area. Specifically, aerial images from the Brant, Bruce, Chatham-Kent, Dufferin, Elgin, and Kitchener-Waterloo areas in Ontario are included in the dataset, with counts of 1,127, 4,910, 2,582, 1,478, 750, and 274, respectively. Specifically, 1,127, 4,910, 2,582, 1,478, 750 and 274 aerial images from Brant, Bruce, Chatham-Kent, Dufferin, Elgin, and Kitchener-Waterloo area in Ontario, Canada (as shown in Figure 3.1[5] and Table 3.1), respectively, are selected from the SWOOP 2010 Dataset and the Regional Municipality of Waterloo. Aerial images from the SWOOP dataset have a spatial resolution of 0.2 m/pixel with RGB bands. Each image has dimensions of 5,000 × 5,000 pixels and covers 1 km^2 area. Aerial images covering the Kitchener-Waterloo area have a spatial resolution of 0.12 m/pixel and s size of 8,350×8,350 pixels. Images from the SWOOP 2010 and the Kitchener-Waterloo area are resized and cropped into small patches with a size of 256×256 pixels as High Resolution (HR) (0.25 m) images and processed further to a size of 64×64 pixels as Low Resolution (LR) (1 m) images. In the rest of the chapter, this dataset is noted as the SWOOP 2010 Dataset. Consequently, a total of 1,708,032, 284,672, and 854,272 pairs of patches are prepared for the training, validation, and testing of the SISR networks, respectively.

Datasets for rooftop delineation

In this section, three datasets for rooftop delineation and data composition are selected: the Massachusetts Building Dataset, the WHU Building Dataset, and the Waterloo Building

[3]Produced by the Ontario Ministry of Natural Resources under License with the Ontario Ministry of Natural Resources © Queen's Printer for Ontario, 2010-2011.

[4]Those aerial images were used for constructing the Waterloo building dataset.

[5]Administrative areas shapefiles and SWOOP extent area shapefile are downloaded from http://www.diva-gis.org/gdata, and https://rb.gy/ti5y4, respectively.

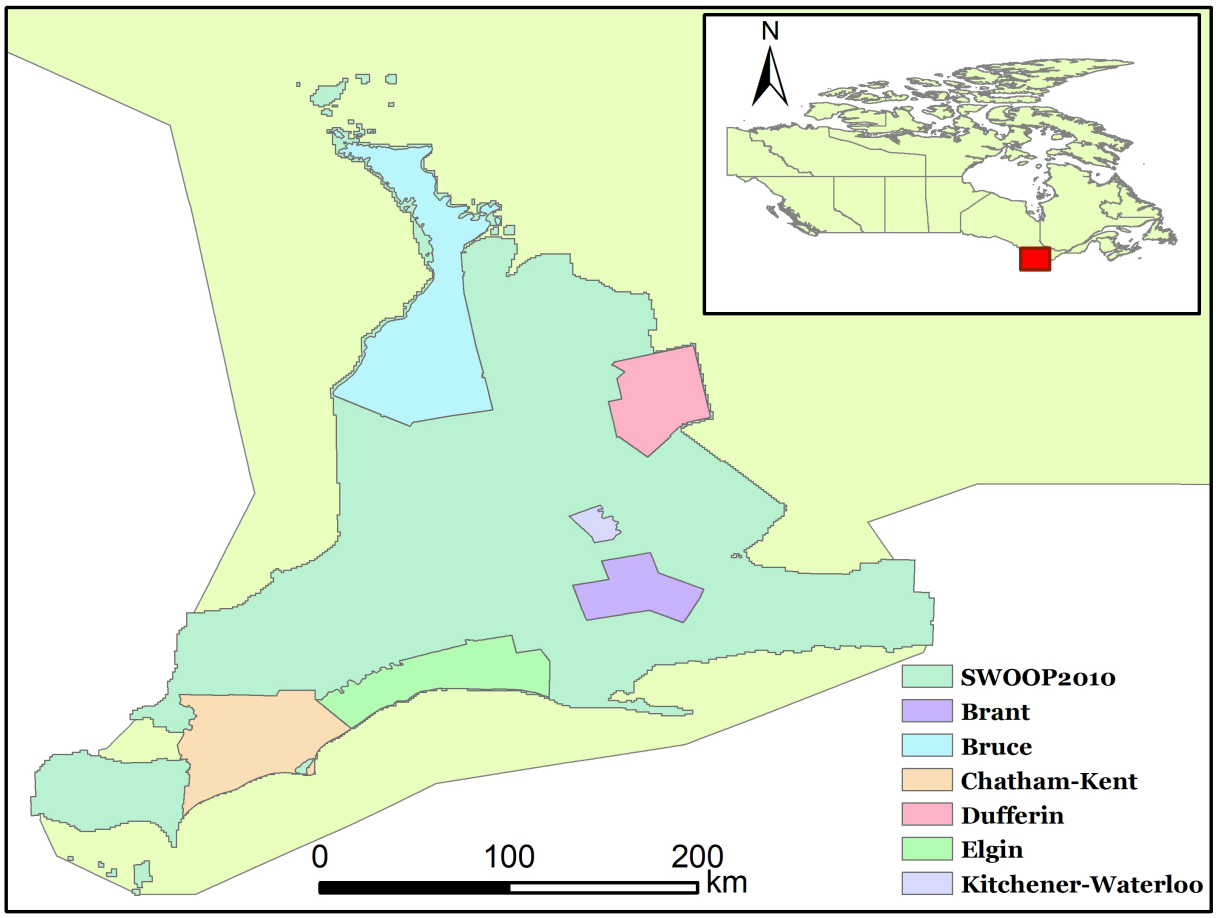

Figure 3.1: Geographical distribution map of SWOOP data

Table 3.1: Images used for SISR

Area	No. Images	Image size (pixels)	Pixel size (m/pixel)
Brant	1,127	5,000×5,000	0.20
Bruce	4,910	5,000×5,000	0.20
Chatham-Kent	2,582	5,000×5,000	0.20
Dufferin	1,478	5,000×5,000	0.20
Elgin	750	5,000×5,000	0.20
Waterloo	274	8,350×8,350	0.12

Table 3.2: Datasets for rooftop delineation

Dataset		No. Images	Image size (pixels)	Pixel size (m/pixel)
Waterloo Building Dataset	Training	42,147	512×512	0.12
	Validation	6,887	512×512	0.12
	Test	20,768	512×512	0.12
WHU Building Dataset	Training	4,736	512×512	0.30
	Validation	1,036	512×512	0.30
	Test	2,416	512×512	0.30
Massachusetts Building Dataset	Training	137	1,500×1,500	1.00
	Validation	4	1,500×1,500	1.00
	Test	10	1,500×1,500	1.00

Dataset. As mentioned in Section 2.1, it should be noted that most public building datasets have a spatial resolution of 0.3 m/pixel. Therefore, the selected datasets will be unified to this spatial resolution. The Massachusetts Building Dataset is selected as a relatively low spatial resolution dataset, while the other two datasets are selected for the data composition purpose. In addition, these three datasets are also utilized to examine the generalizability of trained models. The details of the three building datasets are listed in Table 3.2.

3.2.2 Methods

This section describes the proposed SISR method MSCA-RFANet, the rooftop delineation method and the evaluation metrics.

MSCA-RFANet

The state-of-the-art SISR deep learning methods typically have three parts [95]: the head part, the trunk part (base modules) and the reconstruction part (as shown in Figure 3.2). They are responsible for shallow feature extraction, deep feature extraction and image reconstruction, respectively. The MSCA-RFANet builds upon the powerful RFANet, which is considered the most recent and effective method in the SISR field. It preserves the core architecture of RFANet, particularly its RFA module. For the head and reconstruction parts, a standard convolution layer and the ESPCN are utilized for shallow feature extraction and image reconstruction, respectively. Consequently, the modifications and innovations primarily focus on the trunk part of RFANet.

In the trunk part, inspired by recent work [19, 162, 188, 186], the Channel Attention (CA) block is added after the Enhanced Spatial Attention (ESA) block resulting in a spatial-channel attention block (SCA block). In this way, the network could focus on both informative regions and features. The modified RFA module is named RFA+ in the rest of the thesis. Each RFA+ module is skip connected to the previous one with a momentum term. Share-source skip connection would relieve the deep model training and benefit the information flow, which was also used in the shared source residual group of SAN [32]. With the RFA+ module, the overall architecture becomes deeper and larger. To accelerate model training, the momentum scheme [129] is adopted in the trunk part to connect RFA+ modules. The difference between skip connection with and without the momentum term is described as follows:

Normal skip connection (ResNet):

$$x_{n+1} = x_n + f(x_n, \theta_n) \tag{3.1}$$

Skip connection with the momentum term (Momentum ResNet):

$$\begin{cases} v_{n+1} = \gamma v_n + (1-\gamma) f(x_n, \theta_n) \\ x_{n+1} = x_n + v_{n+1} \end{cases} \tag{3.2}$$

where x_n represents the convolutional layer generated feature. $f(x_n, \theta_n)$ stands for the convolution block in Residual Network (ResNet)[54], in which θ_n represent the learnable parameters in each block. γ is a constant value between 0 and 1. v_n is the momentum at layer n. The initial momentum can be 0 or pre-defined function. As described by Sander et al. [129], the Momentum ResNet can achieve comparable accuracy to ResNet in image classification while requiring a smaller memory footprint. Additionally, it offers advantages in transfer learning [129]. To mitigate the issue of gradient vanishing caused by the deep architecture, a batch normalization layer is incorporated after each skip connection.

HRNet v2

In rooftop delineation, both semantic and instance segmentation methods have been widely used [11, 124]. In this section, rather than focusing on comparing various sophisticated rooftop delineation methods, the HRNet v2 [140], a powerful network proposed recently is adopted for rooftop delineation. This network is designed to maintain high-resolution representations by preserving four levels of features with different spatial resolutions, which are subsequently concatenated in four stages. With the exception of the first stage, these stages

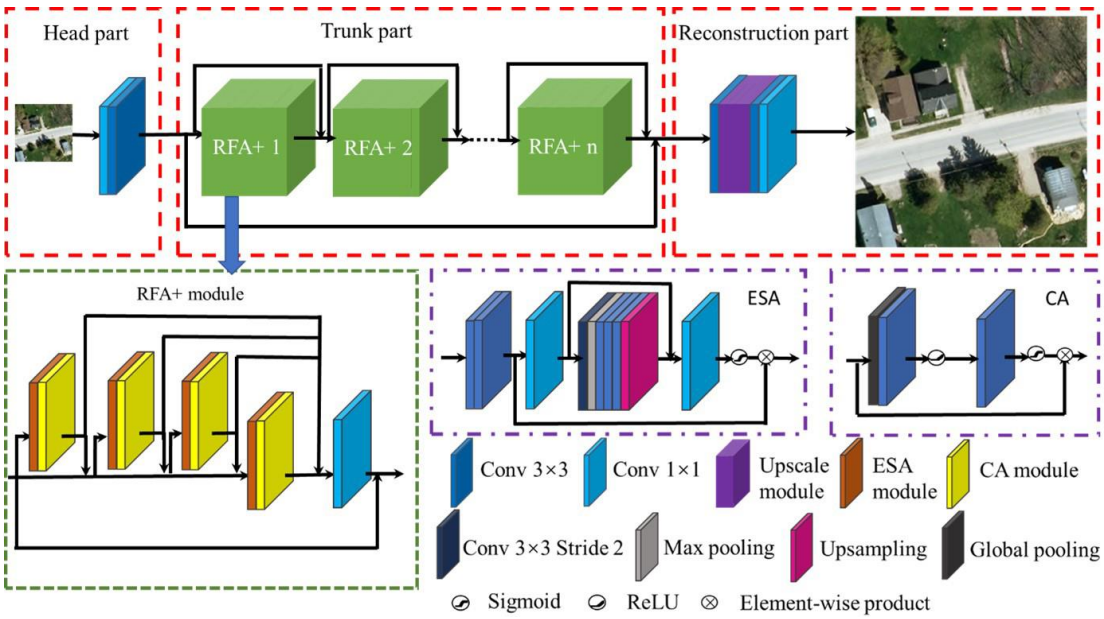

Figure 3.2: Architecture of the proposed MSCA-RFANet (modified from [95])

consist of repeated modularized multi-resolution blocks that incorporate multi-resolution group convolution and multi-resolution convolution. Detailed information about the architecture can be found in Sun et al. [140].

3.2.3 Evaluation Metrics

To evaluate super-resolution performance, the Mean Square Error (MSE), the Root Mean Square Error (RMSE), the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM) are calculated respectively by:

$$\mathrm{MSE} = \frac{1}{N}\sum_{n=1}^{N}(\hat{g}_i - g_i)^2 \tag{3.3}$$

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} \tag{3.4}$$

$$\mathrm{PSNR} = 20\log_{10}\left(\frac{L}{\mathrm{RMSE}}\right) \tag{3.5}$$

$$\mathrm{SSIM} = \frac{(2\mu_{\hat{g}}\mu_g + C_1)(2\sigma_{\hat{g}}\sigma_g + C_2)}{(\mu_{\hat{g}}^2 + \mu_g^2 + C_1)(\sigma_{\hat{g}}^2 + \sigma_g^2 + C_2)} \tag{3.6}$$

where $\hat{g}$ and g refer to the super-resolved images and the corresponding ground truth high spatial resolution images,respectively. N is the total number of pixels in the images, and i indexes individual pixels which ranges from $i = 1$ to $i = N$. L in the PSNR calculation denotes the maximum possible pixel value based on the bit depth of the images. For example, if images are normalized into 0 to 1, L will be 1. For images with an unsigned int 8 bits depth, L is 255. $\mu_{\hat{g}}$ and μ_g are the mean values of all pixels in the super-resolved images and the ground truth high spatial resolution images, respectively. Similarly, $\sigma_{\hat{g}}$ and σ_g represent the unbiased standard deviations.

To evaluate the accuracy of segmentation results, several metrics are used, including Overall Accuracy (OA), Intersection over Union (IoU), mean IoU (mIoU), precision, recall, and F1 score. OA indicates the proportion of correctly classified pixels. mIoU represents the average IoU calculated between the negative and positive classes. F1 score is the harmonic mean of precision and recall, which provides a balanced measure of performance. Detailed formulas for calculating these metrics are provided in Appendix A1.

Table 3.3: The head part of the MSCA-RFANet

Layer types	No. Filters	Size	Strides	Output size
Input	-	-	-	h×w×3
Convolutional layer	3	1×1	1	h×w×3
Convolutional layer	64	3×3	1	h×w×64

Table 3.4: CA modules in the trunk part of the MSCA-RFANet

Layer types	No. Filters	Size	Strides	Output size
Residual_CA	-	-	-	h×w×64
Global pooling	-	-	-	1×1×64
Convolutional layer	4	3×3	1	1×1×4
ReLU	-	-	-	1×1×4
Convolutional layer	64	3×3	1	1×1×64
Sigmoid	-	-	-	1×1×64
Multiply: × = Residual_CA	-	-	-	h×w×64

Table 3.5: ESA modules in the trunk part of the MSCA-RFANet

Layer types	No. Filters	Size	Strides	Output size
Convolutional layer	64	3×3	1	h×w×64
ReLU	-	-	-	h×w×64
Convolutional layer	64	3×3	1	h×w×64
Residual_ESA1	-	-	-	h×w×64
Convolutional layer	16	1×1	1	h×w×16
Residual_ESA2	-	-	-	h×w×16
Convolutional layer	16	3×3	2	h/2×w/2×16
Maxpooling	-	8×8	2	h/4×w/4×16
Convolutional layer	16	3×3	1	h/4×w/4×16
Convolutional layer	16	3×3	1	h/4×w/4×16
Convolutional layer	16	3×3	1	h/4×w/4×16
Upsampling	-	4×4	-	h×w×16
Add: +=Residual_ESA2	-	-	-	h×w×16
Convolutional layer	64	1×1	1	h×w×64
Sigmoid	-	-	-	h×w×64
Multiply: × = Residual_ESA1	-	-	-	h×w×64

Table 3.6: Reconstruction part of the MSCA-RFANet

Layer types	No. Filters	Size	Strides	Output size
Convolutional Layer	64	3×3	1	h×w×64
Convolutional Layer	256	3×3	1	h×w×256
Depth_to_space	-	-	-	2h×2w×64
Covnolutional layer	256	3×3	1	2h×2w×256
Depth_to_space	-	-	-	4h×4w×64
Convolutional Layer	64	3×3	1	4h×4w×64
Convolutional Layer	3	3×3	1	4h×4w×3

3.2.4 Implementation Detail

Following the configuration of RFANet, the number of RFA+ modules is set as 30 in MSCA-RFANet. The configurations of the head part and the reconstruction part are detailed in Tables 3.3 and 3.6. The configurations of the ESA and CA modules are detailed in Tables 3.5 and 3.4. The ESA and CA blocks are connected to construct the RFA+ module as presented in bottom left of Figure 3.2. A total of 30 RFA+ modules are connected using the share-source skip connection approach [32] with a momentum term initialized to 0. The initial learning rate is set to 5e-5 and is halved every 2e5 iterations. In addition, the Adam optimizer is used with the Mean Absolute Error (MAE) as the loss function. In the training of both SISR models, the maximum number of epochs and the batch size are set as 20 epochs and 16. In the context of the remote sensing, all images are directly input to the network and evaluated in RGB color space rather than YCBCr color space which is commonly used in the computer vision field.

For the training of building extraction models, the Adam optimizer is utilized for its high performance, instead of the Stochastic gradient descent (SGD) which is used in the original paper [140]. The learning rate is set to a constant value of 1e-4. The Jaccard loss [7] is used as the loss function to address binary class imbalance. It is important to note that for fair comparison, all HRNet v2 models discussed in the following sections are trained using an equal number of iterations. Specifically, in the training of each HRNet v2, the batch size is set as 8 and the models are iterated 5,400 times per epoch for a total of 100 epochs.

In this chapter, all experiments are implemented on a single Nvidia® GeForce RTX 3090 GPU and CUDA 11.2.

3.3 Experimental Results and Analysis

3.3.1 Evaluation of Super-resolution Methods

To visually compare the performance of MSCA-RFANet with other super-resolution methods used in the experiments, three image patches are selected from the SWOOP 2010 Dataset. As shown in Figure 3.3, the images from the first row to the last row are: the low-resolution images with a pixel size of 1 m/pixel, the high-resolution images with a pixel size of 0.25 m/pixel, the bicubic interpolated images, the RCAN super-resolved images, the SAN super-resolved images, the RFANet super-resolved images and the MSCA-RFANet

Figure 3.3: Examples of super-resolution

super-resolved images, denoted as "LR images", "HR images", "BI images", "RCAN images", "SAN images", "RFANet images", and "MSCA-RFANet images" in the first column of Figure 3.3. As shown in the figure, bicubic interpolation can generate high-resolution images but features in the images are blurred. The CNN-based super-resolution methods can generate high-resolution buildings and roads, but they also blur trees in the first and last columns. From the figure, it is hard to distinguish the differences between the CNN-based super-resolution methods. Therefore, the quantitative evaluation is conducted in the next section.

In this section, the performance of MSCA-RFANet is evaluated by super-resolving the SWOOP 2010 Dataset and the down sampled WHU Building Dataset generated by bicubically interpolating all images in the original WHU Building Dataset to a spatial resolution of 1.2 m/pixel. In addition, the RCAN [184], the SAN [32] and the RFANet [95] are trained on the SWOOP 2010 Dataset as baselines, and their performance is tested on the SWOOP 2010 Dataset (Table 3.7[6]) and the WHU Building Dataset (Table 3.8). In these tables, "BI" refers to the bicubic interpolation. RCAN, SAN, RFANet and MSCA-RFANet represent four DCNN based SISR methods. In addition, the SCA-RFANet denotes the method which only applies the SCA block on top of RFANet. The performance of SCA-RFANet is provided here to explore the contribution of the SCA block and the share-source skip connection between RFA+ modules by comparing it to RFANet and MSCA-RFANet.

Table 3.7: Performance of SISR models (tested on the SWOOP 2010 Dataset)

Models	MSE	RMSE	PSNR (dB)	SSIM	No. Parameters	GFLOPs
BI	43.05	6.33	29.13	0.69	0	0
RCAN	38.04	5.91	30.41	0.73	16,406,409	135.14
SAN	37.47	5.87	30.51	0.74	15,936,553	179.36
RFANet	36.94	5.81	30.66	0.75	10,692,489	87.76
SCA-RFANet	36.89	5.81	30.68	0.75	11,245,449	87.83
MSCA-RFANet	36.64	5.79	30.72	0.75	11,245,449	87.84

As shown in Tables 3.7 and 3.8, all DL-based SISR methods significantly outperform bicubic interpolation in terms of all metrics. As shown in Table 3.7, on the SWOOP 2010 Dataset, MSCA-RFANet outperforms the other state-of-the-art methods. Specifically, MSCA-RFANet achieves a PSNR value of 30.72 dB, exceeding those of RCAN, SAN,

[6]In Tables 3.7 and 3.8, 1 GFLOPs represents 1 billion Floating Point Operations.

Table 3.8: Performance of SISR models (tested on the WHU Building Dataset)

Models	MSE	RMSE	PSNR (dB)	SSIM	No. Parameters	GFLOPs
BI	76.54	8.73	19.39	0.44	0	0
RCAN	69.10	8.29	20.36	0.50	16,406,409	540.54
SAN	-	-	-	-	-	-
RFANet	69.42	8.31	20.35	0.50	10,692,489	351.06
SCA-RFANet	68.88	8.27	20.42	0.50	11,245,449	351.31
MSCA-RFANet	68.97	8.28	20.38	0.50	11,245,449	351.37

RFANet by 0.31 dB, 0.21 dB and 0.06 dB, respectively. On the WHU Building Dataset (Table 3.8), MSCA-RFANet achieves a PSNR value of 20.01 dB, which is higher than those of RCAN and RFANet by 0.02 dB and 0.03 dB, respectively. Because RFANet, SCA-RFANet and MSCA-RFANet have similar performance, the SSIM score is the same up to two decimal places in Tables 3.7 and 3.8. The evaluation scores of the SAN model are omitted because with the limited computational resource the SAN could not process the down sampled WHU Building Dataset while other methods could. In Tables 3.7 and 3.8, the RMSE values are not equal to the squared MSE values. This is because MSE and RMSE are calculated as averaged values across all images, rather than computing RMSE as the square root of MSE.

By examining the performance of SCA-RFANet in Tables 3.7 and 3.8, the positive contribution of using both spatial attention (ESA) and CA or SCA block in SISR can be observed. For instance, the PSNR value of SCA-RFANet on the WHU Building Dataset has increased from 20.35 dB to 20.42 dB. The contribution of the share-source skip connection between RFA+ modules needs further investigation because the share-source skip connection shows a positive effect on the spatial resolution enhancement of the SWOOP 2010 Dataset but a negative effect on that of the WHU Building Dataset. Overall, the MSCA-RFANet achieves superior performance in the conducted experiments.

3.3.2 Impact of SISR on Rooftop Delineation

Semantic models trained on single building dataset

For ease of comparison, the evaluation metrics of the extraction results from the two experiments are arranged according to the training dataset in Tables 3.9 and 3.10. It is

worth noting that for the images in the Massachusetts Building Dataset, SISR methods are used to process the images and bicubic interpolation is used to process the ground truth masks. The interpolated ground truth masks accurately depict the locations and shapes of the buildings, as shown in Figure 3.4. For the Waterloo Building Dataset, the bicubic interpolation is used to down sample both the images and the ground truth masks as shown in Figure 3.5.

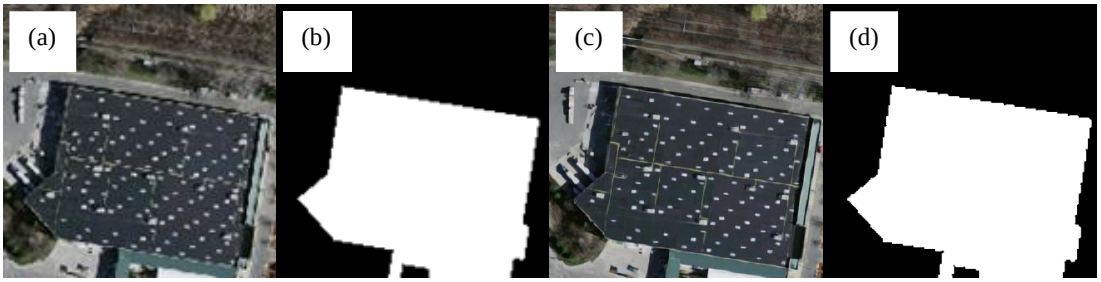

Figure 3.4: Examples of the super-resolved Massachusetts Building Dataset. (a-b) An original image and the matched original mask (1 m/pixel); (c-d) The matched super-resolved image and the interpolated mask (0.3 m/pixel).

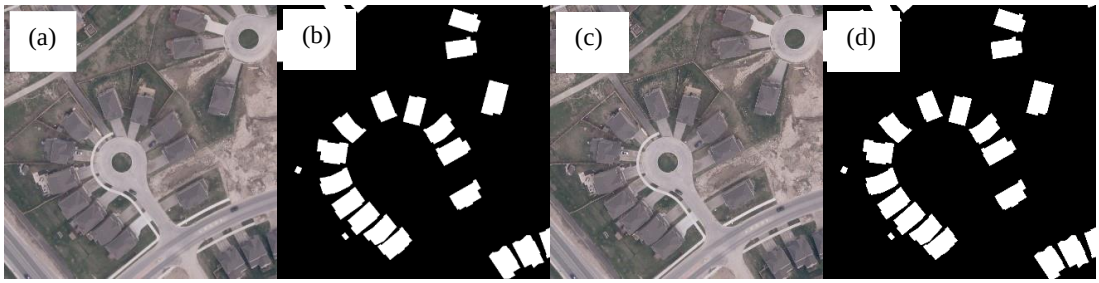

Figure 3.5: Examples of the processed Waterloo Building Dataset. (a-b) An original image and the matched original mask (0.12 m/pixel); (c-d) The matched interpolated image and the interpolated mask (0.3 m/pixel).

As shown in Table 3.9, two trends can be observed based on OA, mIoU, and F1 score. Firstly, the evaluation metrics tend to be higher when the test set is more similar to the training set in terms of data distribution and spatial resolution. Secondly, the quality of extraction results generally improves after super-resolution. However, there are cases where a significant discrepancy between the test set and training set in spatial resolution

and data statistics can affect the results. This can be observed in the models trained on the original Massachusetts Building Dataset (with the spatial resolution of 1 m/pixel). For example, the model trained on the original Waterloo Building Dataset obtains its highest mIoU of 87.12% on the same dataset but achieves its lowest mIoU of 43.31% on the original Massachusetts Building Dataset. This model achieves its second highest mIoU of 69.31% on the bicubically interpolated Waterloo Building Dataset. In addition, the model exhibits a higher mIoU when applied to the super-resolved Massachusetts Building Dataset compared to both the bicubically interpolated and original Massachusetts Building Dataset. Precisely, the mIoU value increases from 43.31% to 45.46% and 48.00% after performing interpolation and super-resolution on the Massachusetts Building Dataset.

Same trends could also be found in the extraction results on the test sets using models trained on the interpolated Waterloo Building Dataset (Table 3.9), the WHU Building Dataset (Table 3.9), the bicubically interpolated Massachusetts Building Dataset (Table 3.10), the super-resolved Massachusetts Building Dataset (Table 3.10) and the original Massachusetts Building Dataset (Table 3.10). One interesting thing is that the HRNet v2 model trained on the interpolated Waterloo Building Dataset performs poorly on the WHU Building Dataset. This could be attributed to the presence of different building types in the WHU dataset and minor interpolation errors that affect the model's performance.

The proposed MSCA-RFANet is expected to outperform RFANet in all scenarios. Super-resolving the training set using MSCA-RFANet produce significantly better results when tested on test sets that differ from the training set in terms of resolution (the original Massachusetts Building Dataset) or building distribution (the WHU Building Dataset). However, as indicated by the results in Tables 3.9[7] and 3.10, it is evident that super-resolving the test set with RFANet yields slightly improved results. This can be attributed to the fact that training on MSCA-RFANet enhances the model's ability to handle significant distribution shifts between the training and test datasets. However, using RFANet on the test set effectively mitigates minor distribution shifts that might exist between the test set and the training set. This effect is also observable in training sets D and E, as shown in Table 3.11 in the subsequent section.

Semantic models trained on composed building dataset

The effect of SISR on dataset composition for rooftop delineation is evaluated in this section. The HRNet v2 model is trained on the combination of the Waterloo Building

[7]In Tables 3.9,3.10, 3.11 and 3.12, the Waterloo Building Dataset, the WHU Building Dataset and the Massachusetts Building Dataset are noted as "Waterloo"," WHU" and "Massachusetts". Bicubic interpolation, super-resolution using RFANet and MSCA-RFANet are noted as BI, RFA, and new, respectively.

Table 3.9: Performance of rooftop delineation results using models trained on the Waterloo Building Dataset and the WHU Building Dataset (in %)

Training data (pixel Size (m/pixel))	Test Data (pixel size (m/pixel))	OA	IoU	mIoU	Precision	Recall	F1 score
Waterloo (0.12)	Waterloo (0.12)	97.78	76.63	87.12	92.48	81.72	86.77
	Waterloo (BI: 0.3)	94.67	44.18	69.31	79.07	50.03	61.29
	WHU (0.3)	89.54	18.62	53.95	58.04	21.52	31.39
	Massachusetts (BI: 0.3)	73.08	19.75	45.46	32.45	33.55	32.99
	Massachusetts (RFA: 0.3)	74.31	23.95	48.00	36.58	40.97	38.65
	Massachusetts (new:0.3)	74.18	23.63	47.78	36.24	40.44	38.22
	Massachusetts (1)	80.71	6.14	43.31	40.66	6.75	11.58
Waterloo (BI: 0.3)	Waterloo (0.12)	79.21	25.49	51.55	27.25	79.79	40.62
	Waterloo (BI: 0.3)	83.99	30.55	56.66	32.50	83.57	46.80
	WHU (0.3)	50.26	15.76	30.46	16.26	83.61	27.23
	Massachusetts (BI: 0.3)	73.49	21.01	46.25	33.80	35.70	34.73
	Massachusetts (RFA: 0.3)	75.48	29.37	51.04	40.53	51.62	45.41
	Massachusetts (new:0.3)	75.09	28.65	50.48	39.75	50.65	44.54
	Massachusetts (1)	76.97	11.44	43.85	28.95	15.90	20.52
WHU (0.3)	Waterloo (0.12)	88.63	15.30	51.85	31.29	23.05	26.55
	Waterloo (BI: 0.3)	87.37	18.93	52.96	29.21	34.99	31.84
	WHU (0.3)	95.22	67.74	81.21	73.09	90.26	80.77
	Massachusetts (BI: 0.3)	75.68	6.56	40.91	21.40	8.65	12.32
	Massachusetts (RFA: 0.3)	75.75	21.66	47.83	37.44	33.94	35.61
	Massachusetts (new:0.3)	75.81	21.83	47.94	37.63	34.21	35.83
	Massachusetts (1)	80.15	17.98	48.61	44.20	23.26	30.48

Table 3.10: Performance of rooftop delineation results using models trained on the Massachusetts Building Dataset (in %)

Training data (pixel size (m/pixel))	Test data (pixel size (m/pixel))	OA	IoU	mIoU	Precision	Recall	F1 score
Massachusetts (BI: 0.3)	Waterloo (0.12)	65.13	10.74	37.17	12.21	47.04	19.39
	Waterloo (BI: 0.3)	75.39	12.96	43.71	15.58	43.45	22.94
	WHU (0.3)	77.37	23.52	49.60	27.38	62.53	38.08
	Massachusetts (BI: 0.3)	81.48	44.88	61.54	52.14	76.32	61.95
	Massachusetts (RFA: 0.3)	80.98	45.20	61.32	51.19	79.45	62.26
	Massachusetts (new:0.3)	79.89	43.95	60.04	49.43	79.86	61.06
	Massachusetts (1)	52.03	24.76	33.90	25.95	84.38	39.69
Massachusetts (RFA: 0.3)	Waterloo (0.12)	68.82	9.61	38.68	11.48	37.21	17.54
	Waterloo (BI: 0.3)	76.83	11.21	43.67	14.21	34.70	20.16
	WHU (0.3)	77.61	19.45	47.89	24.49	48.59	32.57
	Massachusetts (BI: 0.3)	81.75	42.83	60.85	52.91	69.21	59.97
	Massachusetts (RFA: 0.3)	84.57	49.49	65.66	58.34	76.54	66.21
	Massachusetts (new:0.3)	83.81	48.34	64.63	56.67	76.67	65.17
	Massachusetts (1)	53.51	23.29	34.58	25.20	75.47	37.79
Massachusetts (new: 0.3)	Waterloo (0.12)	64.06	10.17	36.35	11.57	45.66	18.46
	Waterloo (BI: 0.3)	74.23	10.25	41.85	12.67	34.90	18.59
	WHU (0.3)	82.48	15.18	48.55	24.76	28.18	26.36
	Massachusetts (BI: 0.3)	78.08	38.52	56.55	46.34	69.56	55.62
	Massachusetts (RFA: 0.3)	80.69	44.40	60.79	50.73	78.07	61.50
	Massachusetts (new:0.3)	79.60	43.11	59.49	48.97	78.29	60.25
	Massachusetts (1)	64.92	20.77	41.07	26.45	49.16	34.40
Massachusetts (1)	Waterloo (0.12)	80.29	4.81	42.45	7.79	11.18	9.19
	Waterloo (BI: 0.3)	82.22	5.09	43.57	8.48	11.32	9.69
	WHU (0.3)	87.68	21.09	54.18	42.34	29.59	34.84
	Massachusetts (BI: 0.3)	79.85	8.94	44.19	45.45	10.01	16.41
	Massachusetts (RFA: 0.3)	81.78	21.28	51.06	59.23	24.93	35.10
	Massachusetts (new:0.3)	81.55	20.86	50.73	57.69	24.62	34.51
	Massachusetts (1)	87.46	47.68	66.76	68.49	61.08	64.57

Dataset, the WHU Building Dataset and Massachusetts Dataset. In training set A, all three datasets are combined using the original datasets. In training set B, the Massachusetts Building Dataset is bicubically interpolated to 0.3 m/pixel. In training set C, the Waterloo Building Dataset and the Massachusetts Building Dataset are bicubically interpolated to 0.3 m/pixel. In training set D, the Waterloo Building Dataset and the Massachusetts Building Dataset are processed to 0.3 m/pixel using bicubic interpolation and RFANet. In training set E, the Waterloo Building Dataset and the Massachusetts Building Dataset are processed to 0.3 m/pixel using bicubic interpolation and MSCA-RFANet.

In addition to the previously mentioned trends, the performance improvement caused by dataset composition is noticeable. For example, OA on the original Waterloo Building Dataset test set increases from 88.63% of the model trained on the WHU Building Dataset (Table 3.9) to 94.36% of the model trained on the training set A (Table 3.11), although it is lower than the 97.78% achieved by the model trained on the original Waterloo Building Dataset where both training and test sets are split from the same dataset (Table 3.9). In other words, by simply composing datasets, the generalizability of the trained model is improved significantly. By super-resolving the Massachusetts Building Dataset using DCNN based methods in the composed dataset, this improvement becomes more obvious. Across all composed training sets, the evaluation scores increase for different test sets, except for the original Massachusetts Building Dataset, which exhibits a large spatial resolution difference. However, this discrepancy is mitigated through the use of SISR super-resolution (or bicubic super-resolution to a lesser degree) to process the test set. For instance, when training on any composed datasets and applying super resolution as a preprocessing step, the model achieves a high degree of generalizability. Super-resolving the test set using RFANet yields the best results. Nonetheless, super-resolving the training set using MSCA-RFANet enhances the model's generalizability and yields even better results.

3.3.3 Impact Visualization

In this section, the generalization errors, and the impact of super-resolution, and combining super-resolution and data composition on rooftop delineation are visualized in Figures 3.6,3.7, and 3.8, respectively.

Figure 3.6 tabulates, from the first to the last row, the samples of super-resolved images from the Massachusetts Building Dataset and the matched ground truth mask, extraction results generated by models trained on the Waterloo Building Dataset with spatial resolution of 0.12 m/pixel and 0.3 m/pixel, the WHU Building Dataset and the Massachusetts

Table 3.11: Performance of rooftop delineation results using models trained on the Massachusetts Building Dataset (in %)

Training Data (pixel size (m/pixel))	Test Data (pixel size (m/pixel))	OA	IoU	mIoU	Precision	Recall	F1 score
A	Waterloo (0.12)	94.36	58.21	76.04	63.14	88.16	73.58
	Waterloo (BI: 0.3)	92.72	43.11	67.70	55.84	65.40	60.25
	WHU (0.3)	94.76	65.06	79.63	71.62	87.67	78.84
	Massachusetts (BI: 0.3)	81.25	12.58	46.66	61.44	13.66	22.35
	Massachusetts (RFA: 0.3)	84.92	34.96	59.28	70.25	41.04	51.81
	Massachusetts (new:0.3)	84.76	33.98	58.72	70.18	39.72	50.73
	Massachusetts (1)	88.17	51.93	69.19	68.42	68.29	68.36
B	Waterloo (0.12)	79.97	21.48	50.14	24.82	61.46	35.36
	Waterloo (BI: 0.3)	87.11	27.55	57.00	34.37	58.13	43.20
	WHU (0.3)	93.46	59.62	76.20	65.61	86.73	74.71
	Massachusetts (BI: 0.3)	78.71	41.12	58.06	47.54	75.25	58.27
	Massachusetts (RFA: 0.3)	79.27	41.68	58.67	48.39	75.03	58.84
	Massachusetts (new:0.3)	78.30	40.70	57.60	46.93	75.41	57.85
	Massachusetts (1)	77.89	23.14	49.72	39.82	35.58	37.58
C	Waterloo (0.12)	81.65	24.65	52.56	27.99	67.35	39.55
	Waterloo (BI: 0.3)	89.19	37.56	63.00	42.26	77.15	54.61
	WHU (0.3)	93.93	61.47	77.37	67.67	87.03	76.14
	Massachusetts (BI: 0.3)	82.57	43.77	61.80	54.68	68.68	60.89
	Massachusetts (RFA: 0.3)	82.97	45.48	62.81	55.29	71.95	62.53
	Massachusetts (new:0.3)	82.71	45.09	62.47	54.74	71.90	62.16
	Massachusetts (1)	80.87	15.93	48.04	47.21	19.38	27.48
D	Waterloo (0.12)	92.02	40.79	66.17	54.61	61.71	57.95
	Waterloo (BI: 0.3)	96.04	62.12	78.94	76.22	77.05	76.63
	WHU (0.3)	95.45	68.14	81.55	75.54	87.43	81.05
	Massachusetts (BI: 0.3)	84.67	42.04	62.39	62.39	56.31	59.20
	Massachusetts (RFA: 0.3)	87.67	52.92	69.30	68.29	70.15	69.21
	Massachusetts (new:0.3)	87.32	52.09	68.69	67.27	69.78	68.50
	Massachusetts (1)	81.54	14.51	47.73	52.06	26.75	25.34
E	Waterloo (0.12)	92.78	41.41	66.90	59.96	57.25	58.57
	Waterloo (BI: 0.3)	94.83	56.27	75.36	66.25	78.88	72.01
	WHU (0.3)	94.16	61.80	77.68	69.43	84.91	76.39
	Massachusetts (BI: 0.3)	86.35	40.98	62.95	73.77	47.97	58.14
	Massachusetts (RFA: 0.3)	88.32	54.21	70.33	70.63	69.99	70.31
	Massachusetts (new:0.3)	88.25	54.00	70.18	70.41	69.85	70.13
	Massachusetts (1)	82.79	16.29	49.24	64.32	17.91	28.02

Building Dataset. As shown in the red boxes, among these models, the model trained on the bicubically interpolated Waterloo Building Dataset shows higher performance than that trained on the original Waterloo Building Dataset; the model trained on the WHU Building Dataset shows the poorest performance; and the model trained on the Massachusetts Building Dataset shows the highest performance. In the last row, the trained model shows high performance on the DCNN super-resolved images, with slightly higher performance coming from the MSCA-RFANet super-resolved image. Therefore, it can be concluded from Figure 3.6 with the previously mentioned findings: the more similar the test set is to the training set with respect to data distribution and spatial resolution, the higher the model scored on the evaluation metrics.

Figure 3.7 tabulates, from the first to the last row, the samples of super-resolved images from the Massachusetts Building Dataset and the matched ground truth mask, extraction results generated by models trained on the original, the bicubically interpolated, the RFANet super-resolved, and the MSCA-RFANet super-resolved Massachusetts Building Dataset respectively. As shown in the red boxes, among these models, models trained on the super-resolved datasets show better performance than those trained on the original and bicubically interpolated datasets. The performance difference between two models trained on super-resolved data is marginal. Similarly, in the last row, the model shows high performance on CNN super-resolved images with marginal difference. The results confirm the second finding that after super-resolution the quality of extraction results can be improved.

Figure 3.8 displays, from the first to the last row, the samples of super-resolved images from the Massachusetts Building Dataset and the matched ground truth mask, extraction results generated by models trained on the original Massachusetts Building Dataset, training sets A, B, C, D and E. As shown in the red boxes, the model trained on the set E shows best performance among all models. CNN super-resolved images as test data exhibit better performance compared to bicubically interpolated images. The visualization results align with the previous findings. The extraction results in the last two row confirm the positive impact of combining super-resolution and data composition on rooftop delineation.

3.3.4 Test on "Unknown" Data

To further test the impact of super-resolution and data composition on rooftop delineation, the trained models from previous experiments, as well as the model trained specifically on the Inria Building Dataset, are compared on the Inria Building Dataset. As its test set is not released, the training set is split into training and test sets with a ratio of 7:3. As

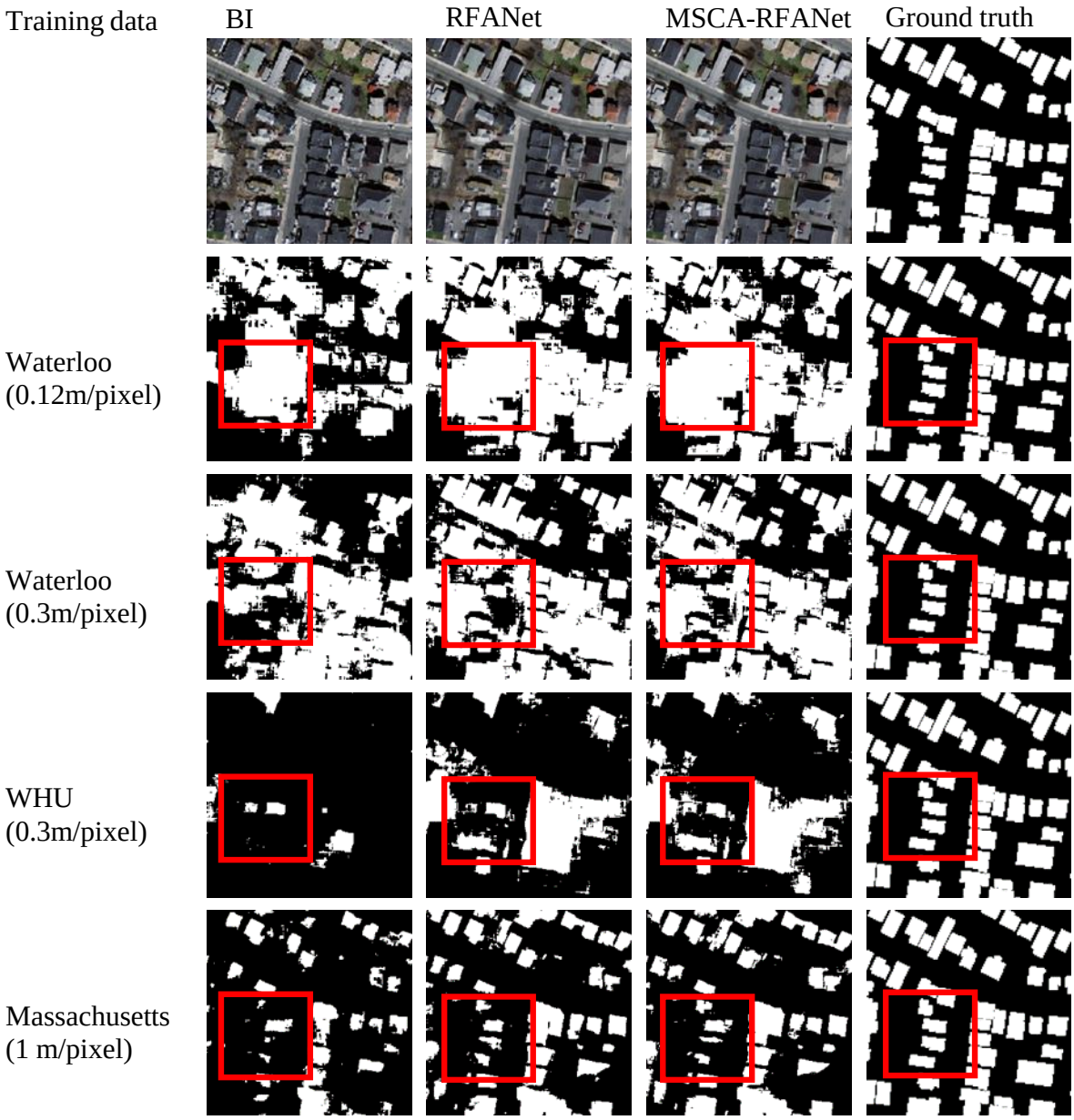

Figure 3.6: Visualization of generalization errors and extraction results using models trained on different building datasets

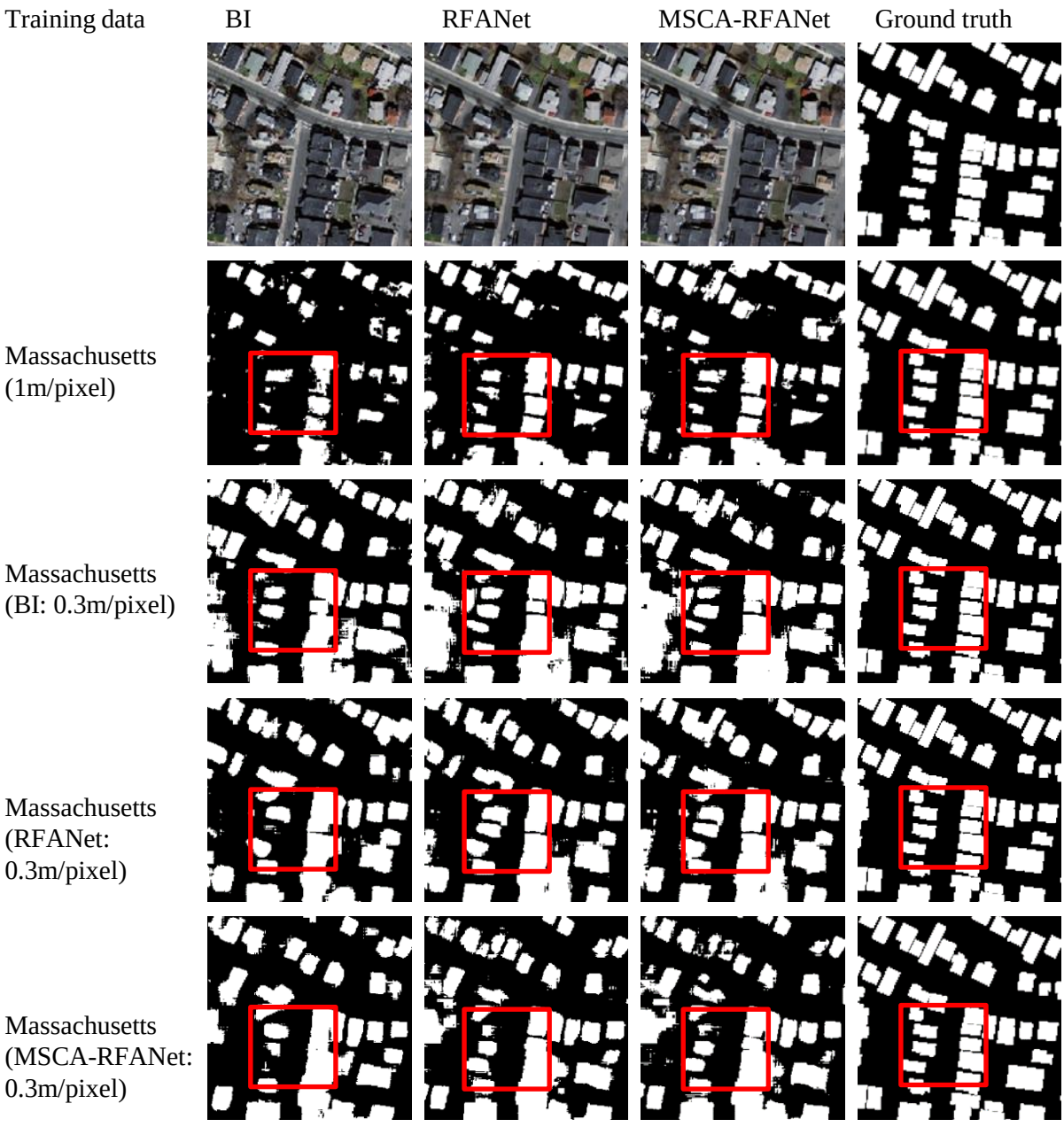

Figure 3.7: Visualization of the impact of super-resolution on rooftop delineation

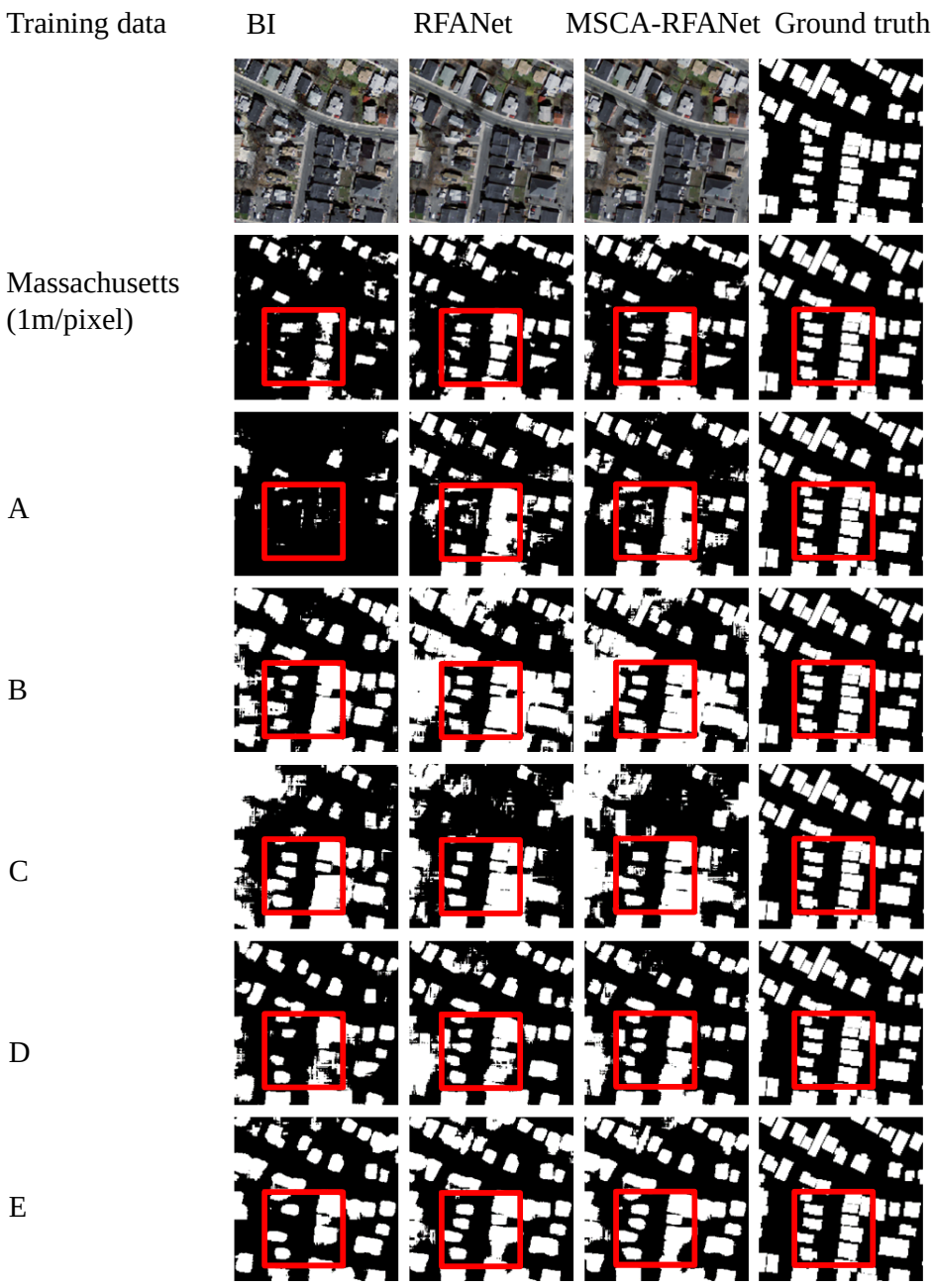

Figure 3.8: Visualization of the impact of data composition and super-resolution on rooftop delineation

shown in Table 3.12, the results confirm that the more similar the test set is to the training set with respect to data distribution and spatial resolution, the higher the model scores on the evaluation metrics. The performance of models trained on training sets C, D, and E further validates the second finding, indicating that the quality of extraction results generally improves after super-resolution. The higher performance of models trained on training sets A and B can be attributed to the larger volume of data used for training. Notably, the model trained on training set E achieves the best performance, highlighting the superior results achieved by combining super-resolution and data composition when constructing the training dataset for rooftop delineation.

Table 3.12: Test on "unknown" Inria Building Dataset (in %)

Training Data (pixel size (m/pixel))	OA	IoU	mIoU	Precision	Recall	F1 score
The Inria Building Dataset (0.3)	92.35	59.49	75.44	74.01	75.20	74.60
Waterloo (0.12)	86.18	15.90	50.85	63.53	17.49	27.43
Waterloo (0.3)	83.24	34.39	58.01	45.29	58.83	51.18
WHU (0.3)	82.06	26.70	53.76	40.65	43.75	42.15
Massachusetts (BI: 0.3)	83.92	30.27	56.49	46.19	46.76	46.47
Massachusetts (RFA: 0.3)	83.91	25.18	54.09	45.19	36.26	40.23
Massachusetts (new: 0.3)	83.14	25.04	53.59	42.70	37.70	40.04
Massachusetts (1)	86.19	12.35	49.13	70.31	13.03	21.99
A	88.11	36.24	61.74	64.52	45.26	53.20
B	86.13	33.38	59.24	54.15	46.53	50.05
C	86.41	32.99	59.21	55.57	44.81	49.61
D	87.60	33.14	59.96	62.93	41.19	49.79
E	88.78	39.93	63.90	66.53	49.97	57.07

3.4 Discussion

3.4.1 Accuracy Improvement

The key modules in RCAN and SAN play crucial roles in performance improvement. In RCAN, long skip connection (LSC), short skip connection and CA are key strategies, which

were explored in spatial resolution enhancement of Set 5 dataset [184]. LSC, which fuses features from the head and trunk parts via pixel-wise addition, contributed to a 0.32 dB increase in PSNR. Short skip connection, which fuses features from the input and output of each module, contributed to a 0.36 dB increase in PSNR. The CA block contributed to a 0.07 dB increase in PSNR. Short skip connection was inherited in SAN and RFANet; LSC was inherited in RFANet and upgraded to share-source residual group (SSRG) in SAN; CA was embedded in the SCA blocks of the MSCA-RFANet. Therefore, both of these key modules were studied in the MSCA-RFANet.

In SAN, the region-level non-local module (RL-NL), the SSRG and the Second-Order Channel Attention (SOCA) are key modules. These modules have been extensively studied for spatial resolution enhancement on the Set 5 dataset [32]. By considering feature inter-dependencies, SOCA outperformed the First Order Channel Attention (FOCA) and was adopted in SAN. However, the implementation of SOCA required the matrix calculation of a large-sized covariance matrix, which limited the size of input images and affected the performance of SISR [32]. Therefore, the SOCA module was not adopted in the MSCA-RFANet, even though it provided a 0.16 dB increase in PSNR according to Dai et al. [32]. The share-source skip connection in the SSRG, which represents the skip connection between each basic module (RFA+ module in the MSCA-RFANet), resulted in a 0.07 dB increase in PSNR, as reported by Dai et al. [32] and discussed in the previous section. This skip connection was subsequently incorporated into the MSCA-RFANet. The RL-NL modules in SAN evenly split the input features into the top left, top right, bottom left, and bottom right parts. Subsequently, non-local modules are applied to each part, enabling the computation of long-range dependencies in the images. Adding an RL-NL module before and after the trunk part of SAN increased the PSNR value by 0.04 dB and 0.06 dB, respectively. RL-NL modules have the potential to further enhance the performance of the MSCA-RFANet. Considering the superior performance of the Global Context (GC) module in recent work [13] compared to the non-local module, this section tests the former rather than the latter. Figure 3.9 shows the difference in architecture between the NL module and

Table 3.13: Effect of GC blocks on the performance of MSCA-RFANet

Datasets	Models	MSE	RMSE	PSNR (dB)	SSIM
SWOOP	MSCA-RFANet	36.64	5.79	30.72	0.75
	+GC block	36.70	5.79	30.70	0.75
WHU	MSCA-RFANet	68.97	8.28	20.38	0.50
	+GC block	71.01	8.40	20.01	0.47

the GC module. The detailed information about the GC module can be found in Cao et al. [13]. The effect of the GC module on the performance of MSCA-RFANet is provided in Table 3.13. The model with GC block before and after the trunk part of MSCA-RFANet is denoted as "+ GC block". As shown in Table 3.13, the super-resolution performance is decreased after adding GC blocks for both datasets. For example, the PSNR value of super-resolution performance on the WHU Building Dataset is significantly decreased from 20.38 dB to 20.01 dB after adding GC blocks to MSCA-RFANet. The experiment's result shows the detrimental effect of GC blocks on the performance of MSCA-RFANet. In the end, the combination of key modules from RCAN, SAN, and RFANet used in MSCA-RFANetis can be confirmed as the optimal combination. To further improve the SISR performance, powerful networks, such as the capsule network [127] and the transformer networks [36] should be considered.

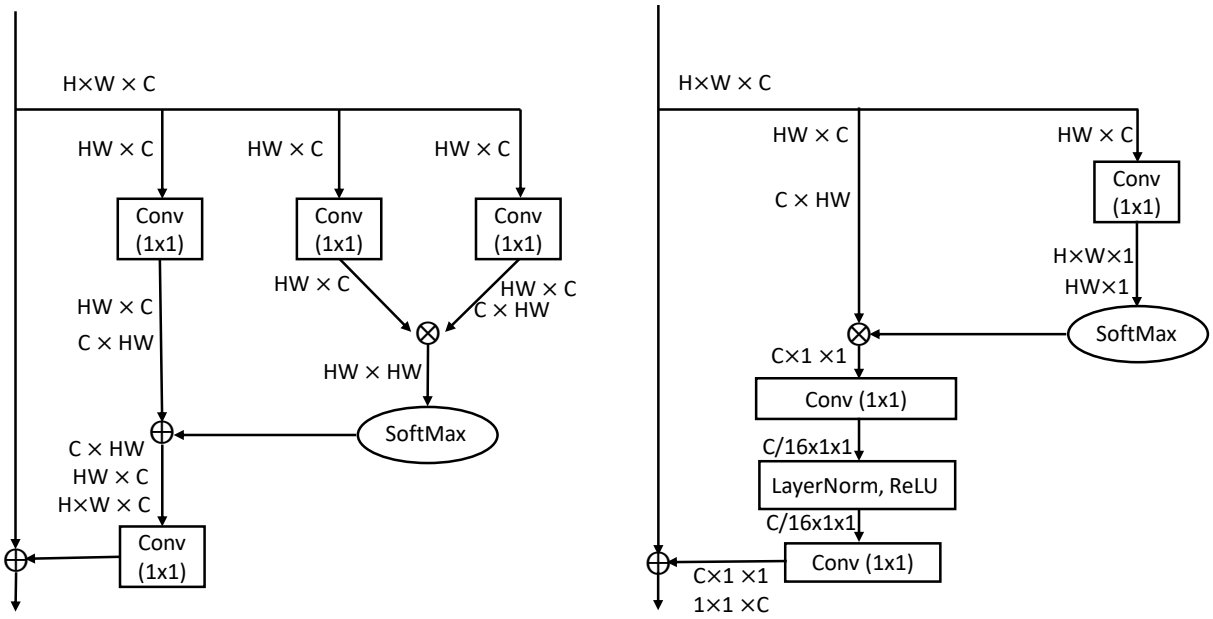

Figure 3.9: NL module(left) and GC module(right)

3.4.2 Computational Efficiency Improvement

In this section, RFANet is taken as an example to explore the performance of low-precision training and separable convolution methods on accelerating SISR methods. Low-precision

training employs the fact that current Graphic Processing Units (GPUs) (such as Nvidia® V100) perform low precision floating point operations much faster than full precision floating point operations [55]. Separable convolution defines a convolution group with fewer parameters compared to standard convolution, resulting in computational efficiency [28]. After thorough exploration, the most promising acceleration method is applied to RFANet to assess its super-resolution performance on the SWOOP 2010 Dataset

Table 3.14: Time consumed for model training in first epoch

Models	Time consumed (mins)
Original model	806
Separable convolution	3645
Mixed precision	525

As shown in Table 3.14, the low-precision training (Mixed precision) is identified as a viable acceleration method. The low speed of model training with the separable convolution is unexpected. Theoretically, reducing the number of trainable parameters would boost the speed of model training. The low training speed using the separable convolution is probably caused by a non-optimized network implementation in the deep learning framework [119], which could not make full use of GPU capacity.

Table 3.15: Performance of super-resolution with low-precision training

Models	MSE	RMSE	PSNR (dB)	SSIM
RFANet	36.94	5.81	30.66	0.75
+ low-precision training	39.47	6.06	30.03	0.72

Following the initial exploration, the low-precision training is employed in RFANet training to explore its impact on the super-resolution performance. Table 3.15 denotes RFANet with and without low-precision training as "+low-precision training" and "RFANet". As shown in Table 3.15, by applying low-precision training, the PSNR value of RFANet significantly drops from 30.66 dB to 30.03 dB. Although the PSNR value is still higher than that of bicubically interpolated images, it is unacceptable given its low accuracy compared to DL-based SISR methods in this chapter and low speed compared to bicubic interpolation method. In other words, the speed gain brought by low-precision training could not make up for the accuracy loss.

3.5 Chapter Summary

In this chapter, data composition is proposed to overcome the generalization errors and improve the accuracy in rooftop delineation. A new super-resolution method is proposed for super-resolving different datasets with different spatial resolution based on state-of-the-art methods, namely MSCA-RFANet. The impact of super-resolution and data composition on rooftop delineation is examined then. In the comparative study of different SISR methods, MSCA-RFANet shows higher performance on both the SWOOP 2010 Dataset and the WHU Building Dataset compared to bicubic interpolation, RCAN, SAN and RFANet. In the super-resolution impact examination, the experimental results show that using super-resolution to match spatial resolution across datasets resulted in higher performance of rooftop delineation. In addition, data composition achieves a positive impact on rooftop delineation, resulting in higher generalizability of trained models. By unifying the spatial resolution of different datasets, and training on the composed dataset, significant improvements are achieved in building extraction performance. For rooftop delineation, not only can MSCA-RFANet be used to compose the training set by unifying candidate training datasets to a single spatial resolution, but also as a preprocessing step during testing or deployment to up-sample input images to the spatial resolution used during training. Doing so would, according to the results, greatly alleviate the generalization error in the practical application of rooftop delineation models. When super-resolving the test set, despite being the better SISR network as demonstrated by the super-resolution metrics, using the MSCA-RFANet yields slightly inferior building extraction results compared to the RFANet it is based on, with around 0.1% to 1.1% OA difference. However, when super-resolving the training set (e.g., the Massachusetts Building Dataset from 1 m/pixel to 0.3 m/pixel), MSCA-RFANet produces better building extraction results compared to RFANet when applied to test sets that significantly differ from the training set in terms of spatial resolution (11.4% OA improvement on the 1 m/pixel Massachusetts Building Dataset) or building distribution (4.87% OA improvement on the WHU Building Dataset). In general, both methods outperform other SISR models, whether applied to the training or the test set. It probably is caused by how the two SISR models affected the distribution shift across the training and test sets, which needs to be investigated further.

Among the currently available building datasets, it is essential to consider building datasets with a variety of building types to overcome generalization errors. Additionally, the availability of data should also be taken into consideration. The results indicate that, in general, when training on a composite dataset that combines different training sets, the model exhibits increased robustness in out-of-distribution testing on an unknown dataset (the Inria Building Dataset).

Chapter 4

Rooftop Delineation with Dynamic Scale Training[8]

4.1 Introduction

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 post-processing [90]. In recent research, the extraction results became more accurate with sharp and regularized rooftop boundaries by employing advanced techniques, such as the ConvGRU [189] and the GNN [199], and by supervising model training with new targets, such as vertices [90, 189, 199, 166], frame field [44], attraction field maps [166], and permutation matrices [199]. In addition, by directly outputting vectorized rooftops with corners, the problems caused by occlusion and blurred rooftop boundaries have been significantly reduced. Nonetheless, scale-variance problems still exist.

In this Chapter, a new method is developed to overcome the scale-variance issue in an end-to-end manner for automated delineation of rooftops in aerial imagery. The new method incorporates a Dynamic Scale Training (DST) strategy and employs a scale-aware higher resolution network (HigherNet) as backbone, as well as the higher resolution supervision targets. The objectives of this chapter are as follows:

[8]The content of the chapter has been submitted to IEEE Transactions on Geoscience and Remote Sensing with a paper entitled "HigherNet-DST: Higher resolution network with dynamic scale training for rooftop delineation from aerial images".

(1) introducing a new powerful end-to-end rooftop delineation model,

(2) mitigating scale-variance issues in rooftop delineation without additional computational resource overhead by employing the DST strategy,

(3) and presenting the extensive experiments which are conducted on four publicly available datasets and show the competitive performance of the proposed method.

4.2 Datasets and Methods

Scale-variation, especially in datasets with a low proportion of small objects, inhibits the total performance of rooftop delineation. A new method is developed based on the HiSup [166] to alleviate scale variance issues. Specifically, the DST strategy is applied in model optimization; the original High-Resolution Network (HRNet) v2 is replaced with the scale aware HigherHRNet [25]; and the high-resolution supervision target is adopted instead of down sampled targets in Xu et al. [166] without introducing new trainable parameters.

4.2.1 Building Datasets Preparation

To extensively evaluate the performance of the proposed network, and test its robustness, four widely used public building datasets are selected in this chapter. These datasets are the AICrowd Building Dataset [110], the Inria Building Dataset [103], the WHU Building Dataset [66] and the Waterloo Building Dataset [51]. They consist of RGB bands but differ in terms of the spatial resolution and cover various geographic locations.

AICrowd Building Dataset was firstly used in the AICrowd (previously CrowdAI) mapping challenge [110]. The satellite images have a spatial resolution of 0.3 m/pixel. Annotation files were provided in MS COCO format [93]. All images were cropped to 300×300 pixels. Because of the missing of testing dataset, following previous work [90, 44, 199, 166], the training dataset and validation dataset are used for model training and testing, respectively. The training dataset is composed of 280,741 images, and the validation dataset contains 60,317 images [110].

In contrast to the MS COCO format annotations in the AICrowd Building Dataset, the Inria Building Dataset, the WHU Building Dataset, and the Waterloo Building Dataset only have binary masks as annotations in their original datasets. Therefore, a polygonization step is required to convert these annotations into the MS COCO format for model training. Additionally, since the Inria Building Dataset does not include test data, the first

five images of each city in the dataset are typically used for model evaluation [44, 199, 166]. The WHU Building Dataset is selected for ablation and comparative studies due to its smaller data volume compared to the other datasets.

4.2.2 Hierarchical Supervision Learning for Rooftop Delineation

To mitigate the performance gap between mask prediction and polygon extraction caused by mask reversibility, the Hierarchical Supervision (HiSup) learning was proposed in Xu et al. [166]. Specifically, after feature extraction by backbone, four branches were attached for mask prediction, attraction field map prediction (used for line segmentation) [167], vertex location prediction and offset prediction [166]. In their experiments, the HiSup showed the highest performance on the AICrowd Building Dataset [110] and a competitive performance on the Inria Building Dataset [103] against other methods, achieving the state-of-the-art performance in learning-based rooftop delineation. Therefore, the HiSup is taken as the basis for developing the new method in this chapter.

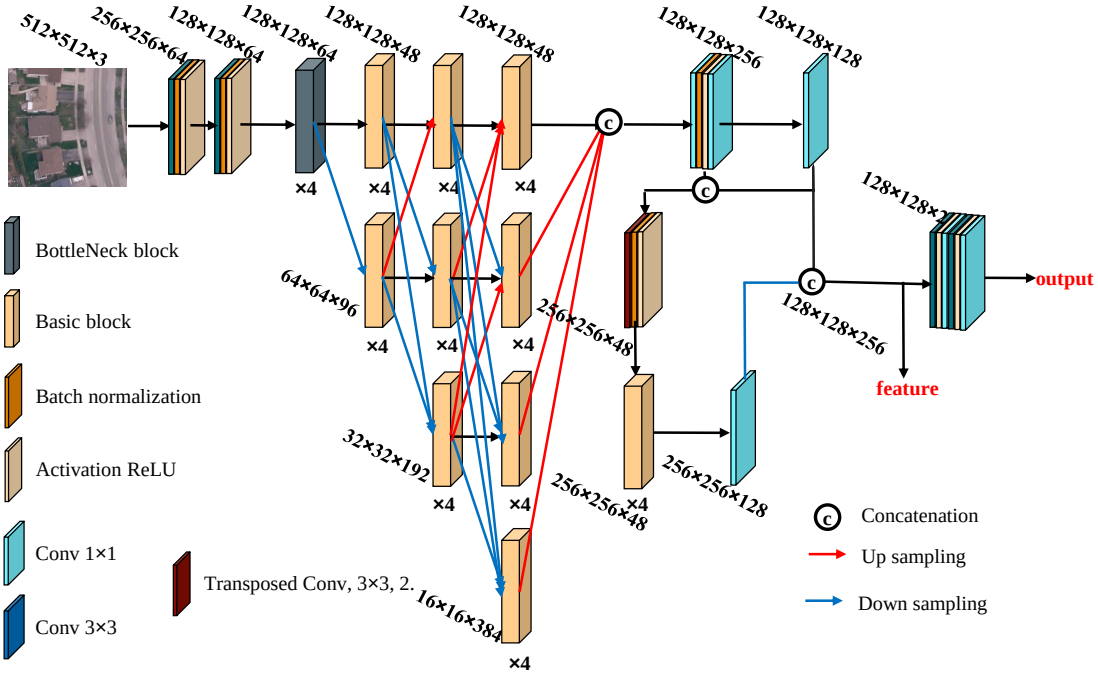

Figure 4.1: Architecture of the scale-aware HigherHRNet (modified from Cheng et al.[25])

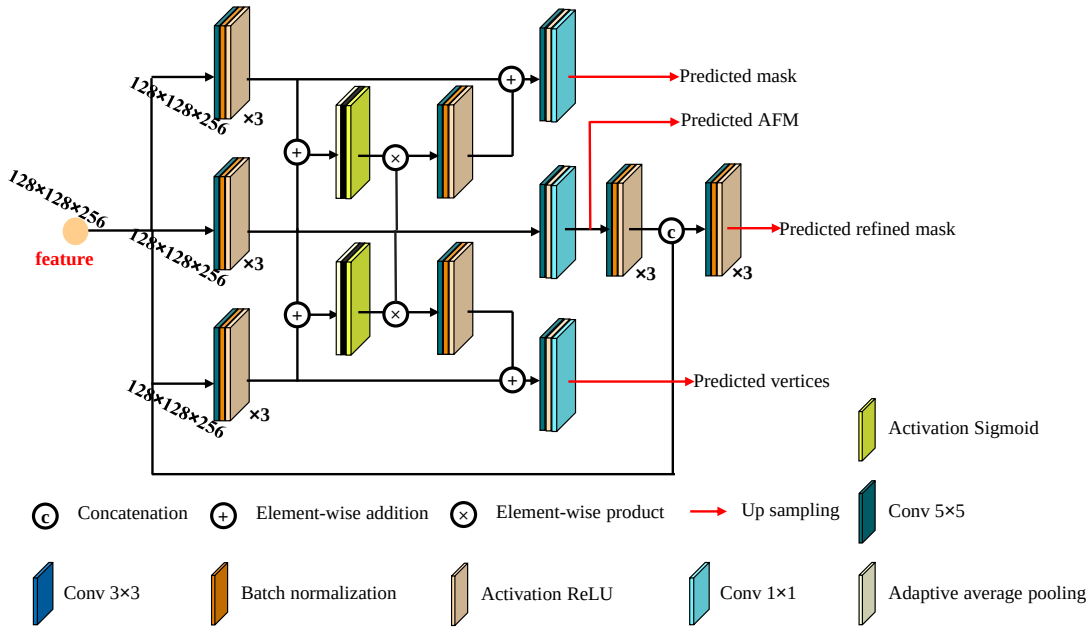

Figure 4.2: Architecture of the object extraction branch (modified from Xu et al.[166])

4.2.3 Scale-aware HigherHRNet (HigherNet)

HRNet v2 has shown excellent performance in feature extraction and representation using multi-level features with repeated information exchange in each stage [140]. However, in the final stage of HRNet v2, the highest resolution of features is 1/4 of the input. Information loss and scale variance suppresses the performance of HRNet v2. Cheng et al. [25] proposed the scale-aware HigherHRNet by adding a scale-aware module on top of the HRNet v2. The scale-aware module is mainly composed of a deconvolutional module and 4 residual blocks (or "Basic Blocks") [54]. To save computational resources, two features with different spatial resolution are down sampled to 128×128 pixels and concatenated, which is similar to the output size of HRNet in HiSup. The architecture of the scale aware HigherHRNet is shown in Figure 4.1. The "feature" in Figure 4.1 is used in the extraction branch as shown in Figure 4.2. The "output" in Figure 4.1 is the predicted vertex offset which will be used in the final polygon generation as described in Xu et al. [166].

4.2.4 Dynamic Scale Training

DST (Stitcher) overcomes scale-variance by collaging images and supervision targets which is guided by dynamic feedback [23]. Specifically, the feedback is the proportion of loss contribution of small objects against that of all objects. For instance, if $L_small/L <= \tau$, in next iteration, k images are randomly selected from next batch of data to create a new image. In the inequality, L_small and L represent losses calculated on small objects and all objects in each batch. In addition, τ and k are two hyper parameters representing the threshold for "Stitcher" and the number of images used for creating the collage, respectively. The collected images and supervision targets are down sampled and stitched together, as shown in Figure 4.3. If the ratio is larger than τ, the model is trained with the usual pipeline in the next iteration.

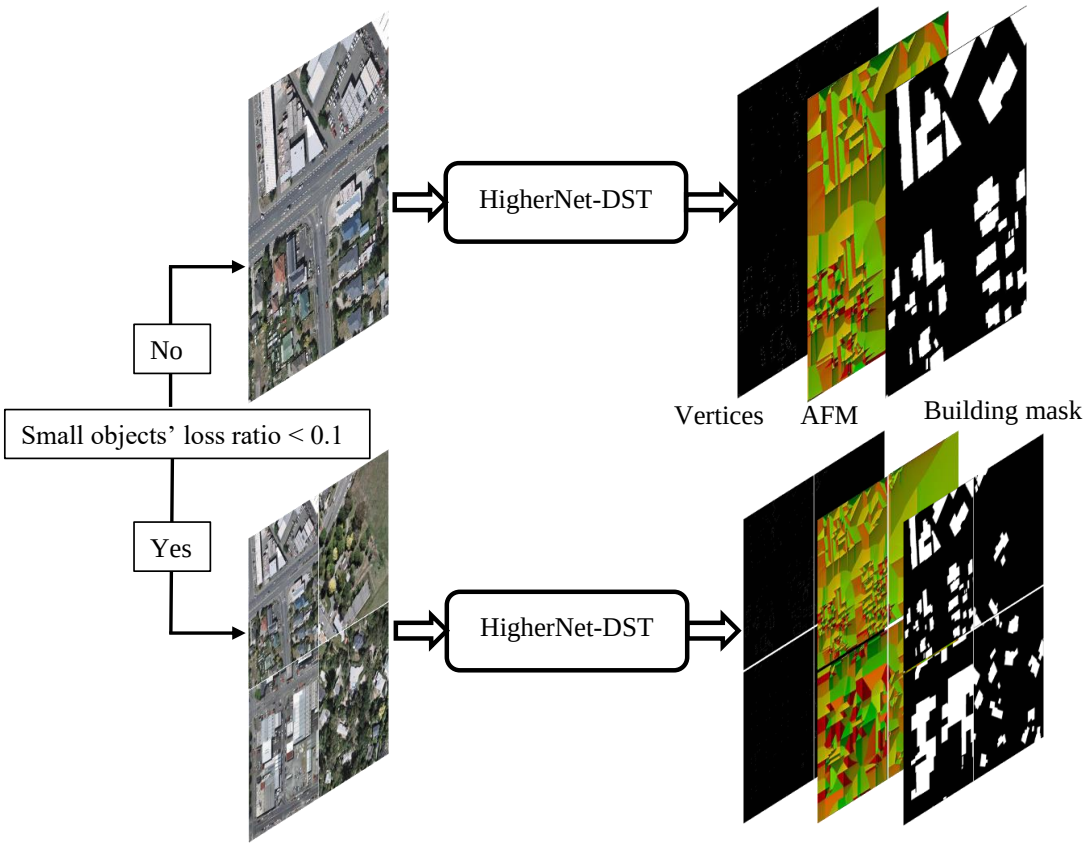

Figure 4.3: Principle of Dynamic Scale Training in HigherNet-DST

4.2.5 High-resolution Supervision Targets

As discussed in Chapter 2, using higher resolution input [23] can, to some extent, overcome scale variance. However, this approach brings an overwhelming computational burden. In the HiSup [166], 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, in HigherNet-DST, high resolution supervision targets are applied, which have the same resolution as the input. There are two reasons why the high-resolution supervision targets are adopted instead of low-resolution one as in Xu et al. [166]. (1) High-resolution supervision targets have more detailed information compared to low-resolution ones [102]. Such detailed information is helpful for small object detection [25]. (2) When supervised with up sampled targets, the model is trained to consider the up-sampling process including both advantages and potential errors and can learn to take advantages or deal with these accordingly. The boundary delineation part of the method is depicted in Figure 4.2.

In this method, the Mask-and-Vertices Attraction [166], which was used in HiSup, is employed for polygons construction. Predicted vertices and masks are taken as input to initialize polygons. Local Non-Maximum Suppression (NMS) is applied to sparse vertices. Refined vertices with the aid of predicted offset vectors is used to simplify initialized polygons by removing redundant vertices and low confidence vertices from initialized polygons. Adjacent edges in each polygon are further merged if they are almost paralleled.

4.2.6 Evaluation Metrics

The object levels evaluation metrics proposed in Lin et al. [93] are widely used in instance segmentation and object detection in computer vision and remote sensing applications. In the literature, Average Precision (AP), Average Recall (AR), AP_{50} and AP_{75} were used to evaluate different methods on the AICrowd Building Dataset. To deal with the scale variance in this chapter, in addition to these metrics, 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) are also applied. Small, Medium and Large size denote 32×32 pixels[9], between 32×32 pixels and 96×96 pixels, and larger than 96×96 pixels, respectively.

[9]The choice to use pixels instead of meters as the unit of measurement is rooted in the fact that deep learning models focus on input image pixel values and do not inherently account for the physical dimensions of objects. This allows us to align the measurement with the way these models process and analyze data.

In this chapter, the MS COCO criterion is utilized to define small, medium and large size. 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 [110], the PANet [96], the PolyMapper [90], the PolyWorld [199] and the HiSup [166]. 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.

Following Xu et al. [166], the restricted metric $AP^{boundary}$ [24] is also adopted. $AP^{boundary}$ is average precision calculated based on boundary IoU [166] instead of mask IoU in Lin et al. [93]. The boundary IoU is calculated as:

$$BoundaryIoU(C, \hat{C}) = \frac{|(C_d \cap C) \cap (\hat{C} \cap \hat{C}_d)|}{|(C_d \cap C) \cap (\hat{C} \cap \hat{C}_d)|} \tag{4.1}$$

where C and $\hat{C}$ are ground truth building mask and predicted building mask. C_d and $\hat{C}_d$ represent pixels within distance d from building boundaries. In this chapter, the parameter d is set to 0.02. For the comparative study on the Inria Building Dataset, IoU and OA are employed to evaluate the extraction results.

4.2.7 Implementation Details

In the training phase, the cross-entropy loss is used for mask prediction, and vertex location prediction. The L_1 loss is used for line segment prediction, and offset prediction. For hyperparameters setting, the initial learning rate, the weight decay, the max epoch and the batch size are set to 1e-4, 1e-4, 100 and 16 respectively. After the first 25 epochs, the learning rate is divided by 10. In all experiments including DST, the parameters of τ and k are set to 0.1 and 4, following Chen et al. [23]. For the ablation study and the comparative study with different backbones, the batch size is set to 7 due to the memory limitation with large backbones. Pytorch 1.7 and an Nvidia® RTX 3090 GPU are employed to train the networks. To train the HigherNet-DST on the AICrowd Building Dataset, 2 Nvidia® RTX 3090 GPUs with the same parameter settings are used.

4.3 Experimental Results and Analysis

4.3.1 Results on the AICrowd Building Dataset

As introduced at the beginning of Section 4.2.1, the AICrowd Building Dataset was released in 2018 for rooftop delineation. In this chapter, for comparison, the extraction results generated by PolyWorld [199], HiSup [166] and HigherNet-DST are visualized in Figure 4.4.

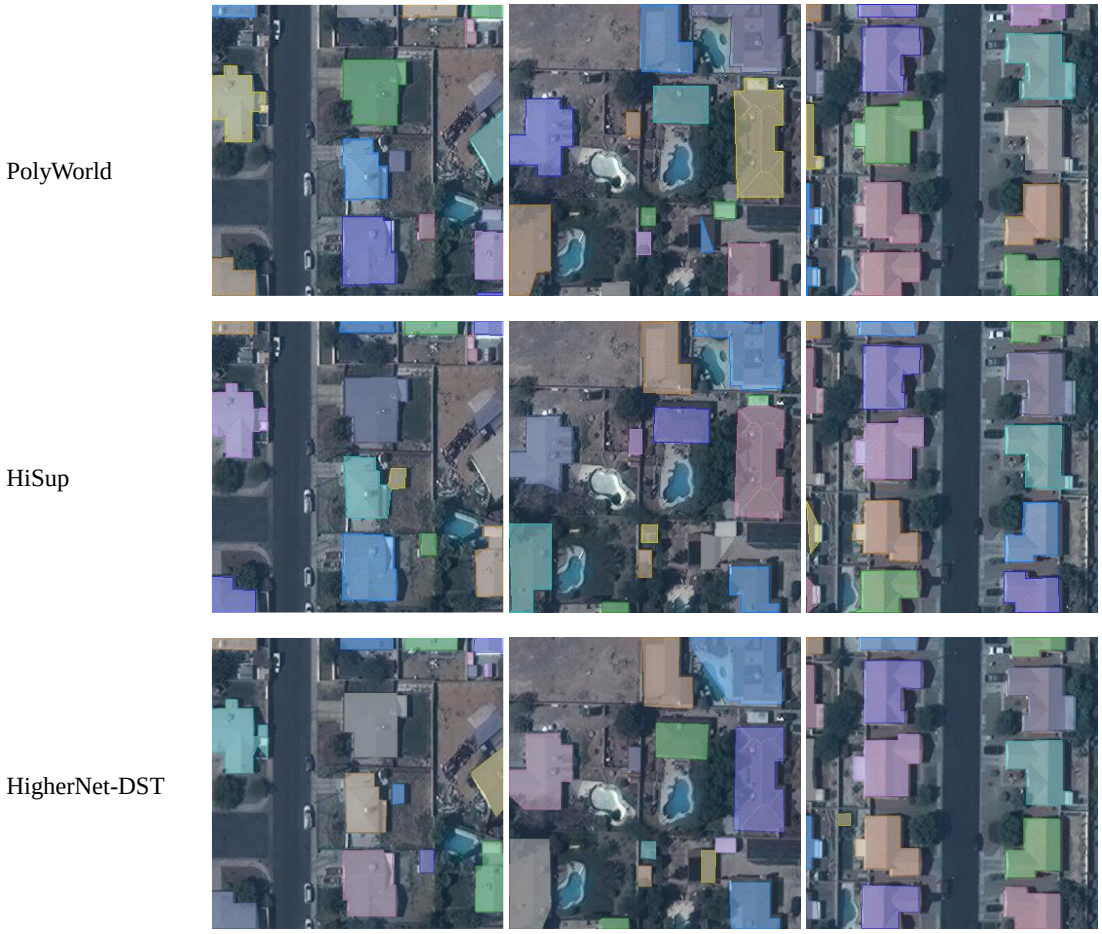

Figure 4.4: Building polygon delineation results on the AICrowd Building Dataset (from top to bottom: selected examples from PolyWorld, HiSup and HigherNet-DST, respectively)

As shown in the figure, the performance on extracting medium and large-sized building objects shows limited differences among the methods. However, on small objects, such as objects in the top right of the first column and the bottom middle of the second column, the proposed method outperforms PolyWorld and HiSup. In the last column, the performance on objects on the left side, the proposed method surpasses HiSup but is inferior to PolyWorld.

Table 4.1 provides quantitative evaluation results of HigherNet-DST and other state-of-the-art methods on the AICrowd Building Dataset. HigherNet-DST shows competitive performance compared to other state-of-the-art methods but is inferior to that of Li et al. [87] and HiSup [166]. Specifically, it achieves 68.5% of AP, which is competitive compared to other methods but lower than 73.8% and 79.4% of AP reported by Li et al. [87] and HiSup, respectively. This result is probably caused by dataset interpolation which may have resulted in uncertainty and is harmful to the performance. Further experiments are required to find the reason with more data and more computational resources.

Table 4.1: Evaluation results on the AICrowd Building Dataset

Methods	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_L	AR	AR_s	AR_m	AR_L	$AP^{boundary}$
Mask RCNN[53, 110]	41.9	67.5	48.8	12.4	58.1	51.9	47.6	18.1	65.2	63.3	15.4
PANet[96]	50.7	73.9	62.6	19.8	68.5	65.8	54.4	21.8	73.5	75.0	-
PolyMapper[90]	55.7	86.0	65.1	30.7	68.5	58.4	62.1	39.4	75.6	75.4	22.6
FFL[44]	67.0	92.1	75.6	-	-	-	73.2	-	-	-	34.4
Li et al[87]	73.8	92.0	81.9	-	-	-	72.6	-	-	-	-
PolyWorld[199]	63.3	88.6	70.5	37.2	83.6	87.7	75.4	52.5	88.7	95.2	50.0
HiSup[166]	79.4	92.7	85.3	55.4	92.0	96.5	81.5	60.1	94.1	97.8	66.5
*HigherNet-DST[10]	68.5	88.4	77.5	41.9	82.6	88.8	71.3	46.6	85.6	91.7	48.0

4.3.2 Results on the Inria Building Dataset

In recent work [44, 199, 166], the Inria Building Dataset was also used to test the performance of new methods in rooftop delineation and specifically to test generalizability. In this section, to visually evaluate rooftop delineation performance on the Inria Building Dataset, the qualitative results are tabulated in Figure 4.5. In the first row, extraction results generated by the HiSup model released by Xu et al. [166] are presented. In the

second row, extraction results generated by HigherNet-DST are presented. As shown in the first two columns of Figure 4.5, HigherNet-DST detects more building objects with accurate boundaries than HiSup. It excels in extracting small building objects, outperforming HiSup. In the last column, the extraction performance on large objects is depicted. Both methods show high performance, but unexpected lines appear, which may be caused by errors in the junctions ordering when generating final polygons. In addition, HigherNet-DST achieves better performance in the qualitative results in Figure 4.5 by checking the yard detection. The yard of the middle bottom building is detected by HigherNet-DST but missed by HiSup, which further proves the superior performance of the proposed method when extracting small objects.

Figure 4.5: Building polygon delineation results on the Inria Building Dataset (top: results from HiSup, bottom: results from HigherNet-DST)

To quantitatively evaluate two models, IoU and OA (pixel/overall accuracy) following Xu et al. [166] and object level metrics mentioned in the previous section are employed. As shown in Tables 4.2 and 4.3, HigherNet-DST achieves the highest values on both pixel level metrics and object level metrics except for AP_L. This demonstrates the high performance of the models in image segmentation and boundary delineation. Specifically, the model obtains an AP of 38.4%, which is 9.4% higher than that of HiSup. The value of AR_L is

increased by more than 20% with HigherNet-DST model. And the values of AP_{50}, AP_{75}, AP_m, AR, AR_s, AR_m are also increased by more than 10% with HigherNet-DST. The experiment confirms the success of the proposal.

Table 4.2: Evaluation results on the Inria Building Dataset-object level

Methods	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_L	AR	AR_s	AR_m	AR_L	$AP^{boundary}$
HiSup[166]	29.0	50.0	29.8	17.8	41.6	49.8	34.0	21.5	45.7	59.2	24.2
*HigherNet-DST	38.4	64.2	40.8	26.8	52.2	40.2	46.1	34.0	58.1	59.6	34.0

Table 4.3: Evaluation results on the Inria Building Dataset-pixel level (in %)

Methods	IoU	Accuracy
FFL[44]	74.8	96.0
HiSup[11][166]	80.7	97.0
*HigherNet-DST	82.6	97.4

4.3.3 Results on the WHU Building Dataset

The WHU Building dataset was released with binary building masks [66]. To the best of my knowledge, the dataset has never been used in polygon delineation in literature. Therefore, in this chapter, binary building masks are first converted to polygon annotations in the MS COCO format. To test the performance of the method on building polygon delineation, PolyMapper [90] and HiSup [166] are selected for the comparative study.

As shown in Figure 4.6, the results from the HiSup and the HigherNet-DST are more accurate than those from the PolyMapper. Specifically, incomplete polygons or omitted polygons in the outputs of PolyMapper are completed and detected in the outputs of HiSup and HigherNet-DST. Improvement can be found in small objects. Small objects in the outputs of HiSup, especially in the first and last columns, have incomplete polygons, which are fixed when using the proposed new method. The visualization results confirm the evaluation results discussed above.

As shown in Table 4.4, the method achieves the best performance when compared to the PolyMapper and the HiSup. Specifically, the method obtains an AP of 60.1%, compared to 51.4% for PolyMapper and 58.3% for HiSup. The method obtains an AR

Figure 4.6: Rooftop delineation results on the WHU Building Dataset obtained using the PolyMapper, the HiSup and the HigherNet-DST (from top to bottom)

Table 4.4: Evaluation results on the WHU Building Dataset

Methods	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_L	AR	AR_s	AR_m	AR_L	$AP^{boundary}$
PolyMapper[90]	51.4	77.6	58.4	41.6	66.2	36.1	60.5	49.9	76.4	63.0	-
HiSup[166]	58.3	80.4	65.8	43.7	76.0	75.3	63.2	47.8	81.0	84.0	56.7
*HigherNet-DST	60.1	82.8	69.0	46.2	77.1	74.9	63.6	49.3	80.2	80.7	58.5

of 63.6%, compared to 60.5% for Polymapper and 63.2% for HiSup. For small objects, the method obtains an APs of 46.2%, compared to 41.6% for PolyMapper and 43.7% for HiSup. Although the ARs value of the proposed method is lower than PolyMapper, it is higher than that of HiSup. The performance on large objects drops, but the AP and AR increase. Therefore, with this experiment, it confirms the success of the proposal in terms of dealing with poor performance caused by small objects.

4.3.4 Results on the Waterloo Building Dataset

To test the method's robustness regarding different spatial resolutions, the Waterloo Building Dataset is down sampled to 0.3 m/pixel to test the performance. In this experiment, PolyMapper and HiSup are employed for a comparative study. By considering time limitation, PolyMapper in the comparative study on the 0.12 m/pixel dataset is omitted.

Figure 4.7: Rooftop delineation results on the Waterloo Building Dataset (top: results from HiSup, bottom: results from HigherNet-DST)

Figures 4.7 and 4.8 present the visualization results. Specifically, Figure 4.7 provides results predicted by HiSup on the 0.12 m/pixel Waterloo Building Dataset, followed by

Table 4.5: Evaluation results on the Waterloo Building Dataset (in %)

Pixel Size	Methods	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_L	AR	AR_s	AR_m	AR_L	$AP^{boundary}$
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
	*HigherNet-DST	66.5	82.5	74.0	31.5	76.2	82.8	70.8	37.4	78.7	87.4	51.0
0.30 m/pixel	PolyMapper	38.4	64.0	40.8	29.0	50.6	34.4	51.5	39.8	66.6	60.1	-
	*HiSup	42.8	62.3	47.7	30.7	56.1	70.7	48.0	33.3	61.0	78.2	40.2
	*HigherNet-DST	51.5	73.2	59.6	37.3	66.8	64.4	55.4	39.5	70.2	70.3	49.0

Figure 4.8: Rooftop delineation results on the Waterloo Building Dataset (top: results from HiSup, bottom: results from HigherNet-DST)

extraction results generated by HigherNet-DST on the same dataset. In Figure 4.8, results predicted by PolyMapper, HiSup and HigherNet-DST on the 0.3 m/pixel Waterloo Building Dataset are visualized. As shown in Figure 4.7, two methods can delineate building polygons with similar performance, but HigherNet-DST has high sensitivity when distinguishing rooftops from building walls. In addition, HigherNet-DST can extract small objects with higher performance, as shown in the third example in Figure 4.7. As shown in Figure 4.8, building polygons become more complete and accurate going from top to bottom. In addition, as shown in the second and last column, compared to HiSup, the advantage of the proposed method is apparent when segmenting buildings which are very close. It can also be attributed to the high performance when delineating small objects.

Table 4.5 provides quantitative evaluation results. HigherNet-DST has similar performance to HiSup on the 0.12 m/pixel dataset, and higher performance when delineating small and medium objects. On the 0.3 m/pixel dataset, HigherNet-DST achieves an AP of 51.5% and an AR of 55.4%. The values of AP_{50} and AP_{75} increased by more than 10%. By outperforming the previous state-of-the-art, these results demonstrate the effectiveness of the proposed method.

4.4 Discussion

4.4.1 Ablation Study

This section explores the effectiveness of each component of HigherNet-DST compared to the baseline. Specifically, on the WHU Building Dataset, HiSup is taken as the baseline to test the performance of automatic mixed precision, DST, the Higher Resolution Network, high spatial resolution supervision targets and the extra semantic segmentation branch (as shown in Figure 4.9). The extra semantic segmentation branch here refers to the branch with a semantic segmentation head taking the feature from backbone. In Table 4.6, their performance on the WHU Building Dataset is presented and denoted as "+amp", "+DST", "+HigherNet", "*HigherNet-DST", and "+extra branch", respectively. In addition, the performance of HigherNet-DST trained with full precision is provided and denoted as "HigherNet-DST".

As shown in Table 4.6, by adding DST, replacing High Resolution Network (HRNet v2) with the Higher Resolution Network, using higher spatial resolution supervision targets and adding extra semantic segmentation branch on the backbone, the delineation performance increased gradually. The increased performance confirms the effectiveness of

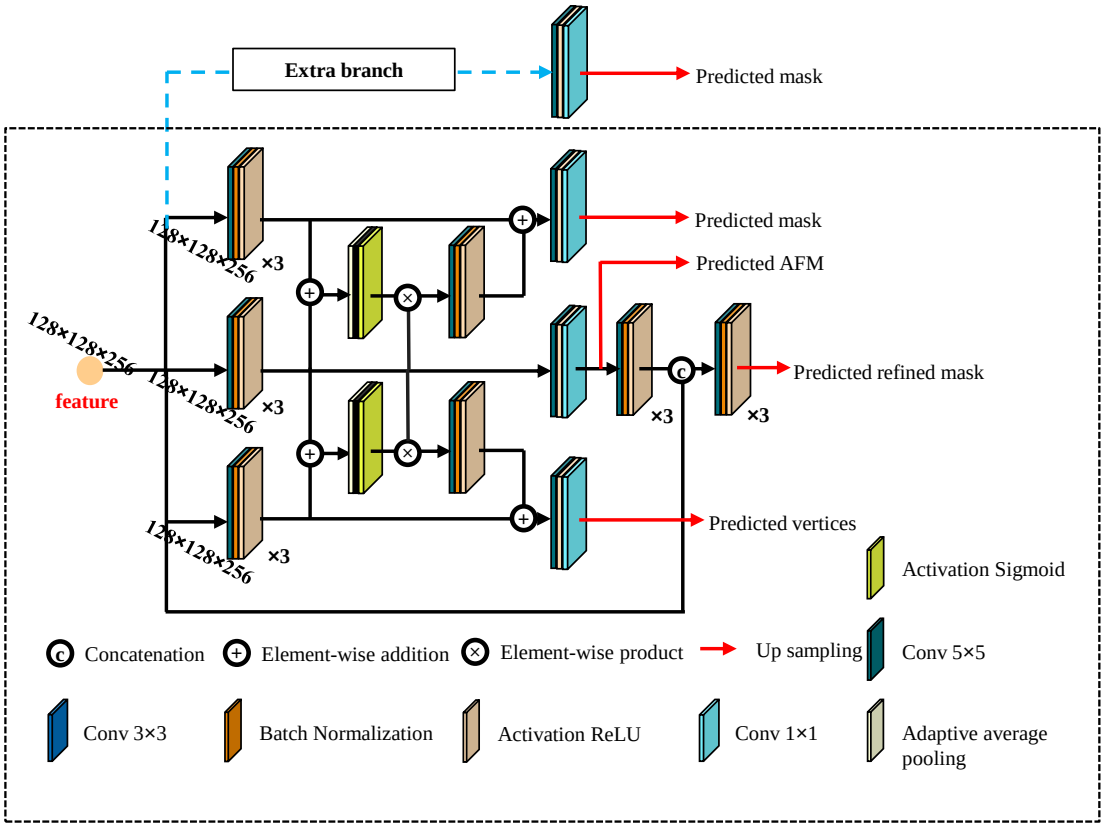

Figure 4.9: Architecture of the object extraction branch with the extra semantic segmentation branch

Table 4.6: Ablation study conducted on the WHU Building Dataset (in %)

Methods	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_L	AR	AR_s	AR_m	AR_L	$AP^{boundary}$
HiSup	59.1	80.6	67.2	43.8	77.0	77.5	63.4	47.7	81.5	83.2	57.5
+amp	58.3	80.4	65.8	43.7	76.0	75.3	63.2	47.8	81.0	84.0	56.7
+DST	59.4	80.6	67.1	44.4	77.4	77.6	63.7	48.3	81.6	83.3	57.6
+HigherNet	59.6	80.7	67.4	44.9	77.4	76.1	64.0	48.7	81.8	82.8	58.0
*HigherNet-DST	60.1	82.8	69.0	46.2	77.1	74.9	63.6	49.3	80.2	80.7	58.5
+extra branch	60.2	82.8	69.1	46.3	77.5	74.0	63.8	49.2	80.6	80.2	58.5
HigherNet-DST	61.4	83.8	70.9	47.3	78.1	75.2	64.8	50.6	81.3	81.1	59.6

each modification. It is worth noting that with the extra semantic segmentation branch, HigherNet-DST can be improved further. However, the increase is marginal compared to the increase of computational burden. Therefore, it is not included in HigherNet-DST.

4.4.2 MS Training/Testing

To further show the superior performance of the DST in rooftop delineation, this section compares the DST with multi-scale training and testing, which are commonly used to deal with scale-variance issues. The baseline architecture in this section is HiSup with auto mixed precision.

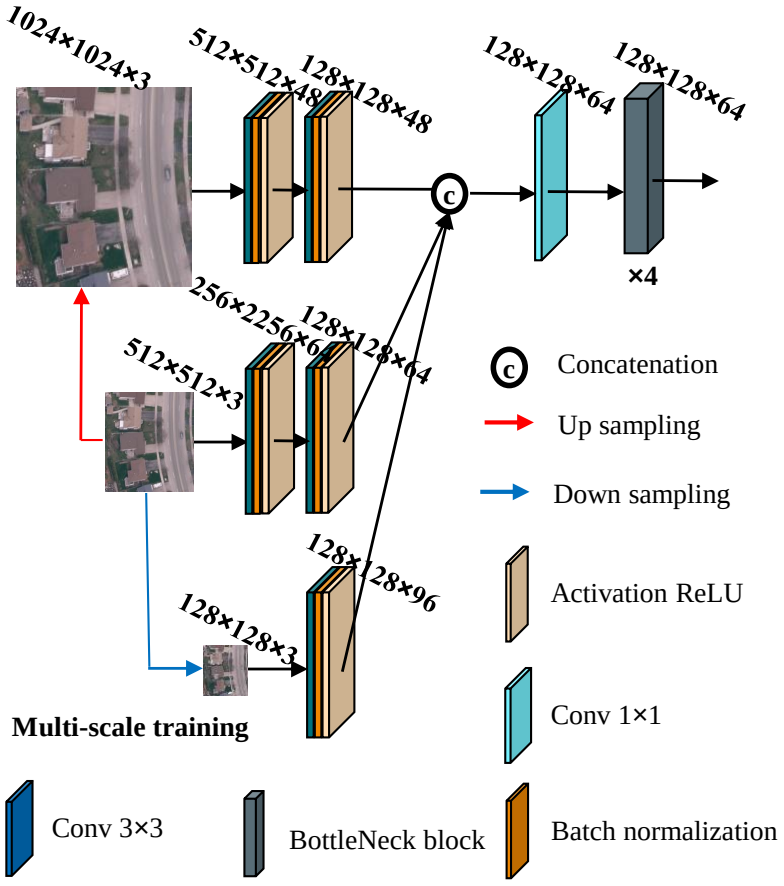

Figure 4.10: Feature extraction part of multi-scale training

Following Liu et al. [98], three scales, including 256×256 pixels, 512×512 pixels, and 1024×1024 pixels, are used for multi-scale training and testing. Specifically, in the training phase, input images are resized to three scales and then passed through a combination of convolution layers and batch normalization layers. The features generated by three scales are concatenated and used as the input of the first stage in the backbone (as shown in Figure 4.10). For multi-scale testing, input images are resized to three scales (2×, 1×, and 0.5×) before flowing into the deep network. In Table 4.7, the HiSup, the HiSup with auto mixed precision, the HiSup with auto mixed precision and multi-scale training, and the HiSup with auto mixed precision and the DST as "HiSup", "+amp", "+multi scale training", and "+DST". Multi-scale testing with different spatial resolution inputs is noted by the side length of the input. The performance of the output combination from different scales input and noted as "combination".

Table 4.7: Performance of MS training and testing on the WHU Building Dataset (in %)

Methods		AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_L	AR	AR_s	AR_m	AR_L	$AP^{boundary}$
HiSup		59.1	80.6	67.2	43.8	77.0	77.5	63.4	47.7	81.5	83.2	57.5
+amp		58.3	80.4	65.8	43.7	76.0	75.3	63.2	47.8	81.0	84.0	56.7
+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
	512	58.8	80.7	67.1	43.6	76.7	77.7	63.1	47.5	81.0	83.7	57.0
	256	28.6	56.5	26.0	9.9	48.9	66.9	32.2	12.3	54.5	74.7	26.3
	Combination	35.3	45.6	40.6	31.0	57.0	43.3	61.3	42.1	83.4	88.1	34.1
+DST	1024	38.8	60.9	43.2	29.4	52.1	46.7	44.1	32.7	57.4	54.5	37.0
	512	59.4	80.6	67.1	44.4	77.4	77.6	63.7	48.3	81.6	83.3	57.6
	256	30.0	58.4	28.4	10.6	51.1	67.3	33.2	12.9	56.1	73.3	27.9
	Combination	37.2	47.8	42.5	31.8	60.0	46.7	61.8	42.8	83.6	87.4	36.0

As shown in Table 4.7, multi-scale training with the 512×512 pixels size input indeed improves the whole performance while the improvement is less than that brought by employing the DST. For example, by applying multi-scale training with the input size of 512×512 pixels, the AP value is increased from 58.3% to 58.8%. However, by applying the DST with the 512×512 pixels size input, the AP value is increased from 58.3% to 59.4%. In addition, employing the multi-scale training consumes more computational resources than applying DST. Therefore, the multi-scale training is not effective compared to the DST in rooftop delineation. As for multi-scale testing, Table 4.7 shows that using both models with the input size of 512×512 pixels gives the best performance, which indicates that multi-scale testing has a negative impact on rooftop delineation.

4.4.3 Results with Different Backbones

Recently, the Vision Transformer based backbones show high performance in many computer vision benchmarks compared to CNNs. Therefore, the performance of state-of-the-art Transformer based backbones for building polygon delineation are also tested using the WHU Building Dataset. Specifically, the Pyramid Vision Transformer (PVT) v2 [154], Sequencer [143] and the High-Resolution vision Transformer (HRFormer) [177] were tested.

Table 4.8: Evaluation results on different backbones of HiSup on WHU Building Dataset (in %)

Backbone	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_L	AR	AR_s	AR_m	AR_L	$AP^{boundary}$
HRNet v2	59.1	80.6	67.2	43.8	77.0	77.5	63.4	47.7	81.5	83.2	57.5
*HRNet v2	58.3	80.4	65.8	43.7	76.0	75.3	63.2	47.8	81.0	84.0	56.7
$PVTv2\_b0$	50.2	75.4	56.9	34.3	69.2	65.2	54.2	37.6	73.5	72.6	48.0
$PVTv2\_b1$	52.6	75.6	59.9	37.4	71.3	66.7	56.7	40.5	75.5	73.7	50.7
$PVTv2\_b2$	53.2	76.7	60.2	37.7	72.0	68.3	57.3	41.0	76.1	75.4	51.3
$PVTv2\_b3$	53.4	76.7	61.1	37.8	72.0	69.1	57.4	41.1	76.4	75.6	51.5
$PVTv2\_b4$	54.8	77.8	62.4	39.0	73.5	71.4	58.7	42.4	77.6	77.6	52.9
$PVTv2\_b5$	54.8	78.7	62.2	39.1	73.2	71.6	59.0	42.9	77.6	77.3	52.9
$PVTv2\_b4 + FPN$	53.4	77.6	60.3	37.6	72.2	69.9	57.4	41.0	76.4	75.2	51.3
$PVTv2\_b5 + FPN$	55.1	78.6	62.2	39.4	73.6	72.5	59.1	42.9	77.7	78.1	53.1
$Sequencer\_small$	53.9	77.5	61.2	38.4	73.0	65.9	58.1	41.7	77.1	76.5	52.1
$Sequencer\_medium$	54.1	77.6	61.1	38.4	73.0	69.4	58.3	41.9	77.1	77.0	52.3
$*Sequencer\_large$	54.2	77.7	61.3	38.3	73.1	69.7	58.2	41.7	77.3	75.9	52.3
$*HRFormer\_small$	55.3	78.8	63.2	39.6	73.9	72.2	59.3	43.0	78.1	79.1	53.5
$*HRFormer\_base$	56.0	79.6	63.4	40.5	74.6	73.9	60.0	43.9	78.6	80.5	54.3
$*HRSeq\_medium$	54.2	78.3	61.7	39.0	72.7	69.5	58.6	42.7	77.0	75.6	52.1
$*HRSeq\_large$	55.3	79.4	62.9	39.6	74.0	71.2	59.5	43.3	78.4	77.4	53.4
$*HRNetv2\_2block$	57.4	79.4	64.9	42.5	75.6	76.3	62.0	46.3	80.2	82.6	55.8
$*HRNetv2\_1block$	56.3	79.5	63.9	41.0	74.9	72.9	60.8	44.9	79.3	79.1	54.5
$*HRFormer\_base\_1block$	57.3	80.5	65.2	42.0	75.6	73.6	61.5	45.6	79.7	80.2	55.6
$+dense\_connection$	55.8	79.6	63.2	40.4	74.2	72.4	60.0	44.0	78.4	79.1	54.0
$*HRSeq\_medium\_1block$	55.8	78.6	63.9	40.6	74.2	71.6	60.0	44.0	78.4	78.0	53.7

The PVT v2 is comprised of four stages of vision transformer with a linear complexity attention layer, overlapping patch embedding and convolutional feed-forward network. The official implementation is embedded into HiSup and all six versions of PVT v2 are used in this experiment. In Table 4.8, they are denoted as "$PVTv2\_b0$", "$PVTv2\_b1$", "$PVTv2\_b2$", "$PVTv2\_b3$", "$PVTv2\_b4$" and "$PVTv2\_b5$". FPN is also tested after "$PVTv2\_b4$" and "$PVTv2\_b5$", denoted as "$PVTv2\_b5+FPN$" and "$PVTv2\_b4+FPN$".

In Sequencer, Long Short-Term Memory (LSTM) instead of self-attention is used to model long range dependencies. Sequencer is also embedded into HRNet v2 as the counterpart of HRFormer. In the HRFormer, vision transformer block is used to replace CNN block in HRNet v2. The official implementation of Sequencer and HRFormer are employed in this chapter. Small and medium versions of Sequencer as backbones of HiSup are tested. They are denoted as "$Sequencer\_small$", and "$Sequencer\_medium$". By applying automatic mixed precision in model training, large version of Sequencer, small and base versions of HRFormer as the backbone of HiSup are tested and named them as "$*Sequencer\_large$", "$*HRFormer\_small$" and "$*HRFormer\_base$", respectively. In addition, medium and large versions of Sequencer are employed in HRNet v2 as basic blocks for this experiment and named as "$*HRSeq\_medium$" and "$*HRSeq\_large$" in Table 4.8. By reducing the number of blocks in each stage and branch of HRFormer and HRNet v2, new backbones are made and named as "$*HRNetv2\_2block$", "$*HRNetv2\_1block$", "$*HRFormer\_base\_1block$", "$*HRSeq\_medium\_1block$". Because of limited computational resources, "$*HRSeq\_large\_1block$" is omitted. Furthermore, dense connection scheme is also tested on top of "$*HRFormer\_base\_1block$" and named it "$+dense\_connection$".

As shown in Table 4.8, among different backbones, the original HRNet v2 obtains the best performance. Among different PVT v2 variations, larger backbones show better performance. Interestingly, adding the FPN does not result in a performance increase as expected. The best performance among different HRFormer based backbones comes from "$HRFormer\_base$". However, it is still inferior to the HRNet v2 based model. HRFormer and Sequencer based "HRSeq" include 2 blocks in each stage and each branch. By reducing the block number to one and reducing the number of HRNet v2 modules from 4 to 1, the HRFormer "$HRFormer\_base\_1block$" starts to surpass "$*HRNetv2\_1block$" and becomes comparable to the "$*HRNetv2\_2block$" backbone of HiSup. An attempt was made to improve "$*HRFormer\_base\_1block$" by applying dense connections to each branch, which resulted in lower performance. "$*HRSeq\_medium\_1block$" still shows poor performance compared to its counterparts. As described in Raghu et al. [121], increasing the size of the transformer network does not result in better performance in these circumstances, which is contrary to CNNs.

4.5 Chapter Summary

In this chapter, a new deep learning network named HigherNet-DST for rooftop delineation is proposed. By applying the DST, adopting the scale-aware Higher-Resolution Network, and using higher resolution supervision targets based on HiSup, the method can relieve the scale-variance issue and improve the performance of building boundaries delineation.

By conducting an extensive comparative study, HigherNet-DST showed competitive performance on the AICrowd Building Dataset and better performance on the Inria Building Dataset, the WHU Building Dataset and the Waterloo Building Dataset compared to other state-of-the-art methods. The ablation study further shows the effectiveness of each module of the proposed method. Experiments on the AICrowd Building Dataset show the competitive performance of HigherNet-DST with an AP of 68.5%. On the Inria Building Dataset, with an IoU of 82.6% and accuracy of 97.4%, HigherNet-DST achieves the best pixel classification performance among all benchmarked methods. In terms of rooftop delineation, it has 9.4%-27.2% higher values on all object level metrics, except for AR_L, compared to HiSup, the previous state-of-the-art method. On the WHU Building Dataset, it achieves 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 APboundary, respectively, which are higher than those of PolyMapper and HiSup, while other metrics are also competitive. On the 0.3 m/pixel Waterloo Building Dataset, it surpasses HiSup by 6.2%-11.9% on all metrics except for AP_L and AR_L. On the original Waterloo Building Dataset, it shows competitive performance compared to HiSup while better performance on small and medium size objects. The experiments show the effectiveness of the new network in dealing with scale variance issues, especially excelling at the small building's regime, which are long-standing problems in rooftop delineation.

Chapter 5

Weakly Supervised Rooftop Delineation[12]

5.1 Introduction

Different annotations have been applied in weakly supervised rooftop delineation. Among these, image tags are widely used as weak annotations [9, 40, 169, 180]. Rules, such as a certain ratio of building pixels to all pixels or the existence of building objects within per image patch, are employed to assign positive image tags or image-level labels to images for weakly supervised rooftop delineation. In literature, however, there is little research in weakly supervised instance segmentation methods for rooftop delineation, which may be more suitable for rooftop delineation as it can generate individual rooftop boundaries. In contrast, in computer vision, weakly supervised instance segmentation methods are well explored. Among different annotation types, box-supervised instance segmentation has attracted much attention in recent research. As reported in Li et al. [88], annotating an object with a bounding box takes around 11 times less time than that with a polygon on average (79.2 seconds for a polygon and 7 seconds for a bounding box), significantly alleviating the labeling cost in instance segmentation. Recent research, such as the DiscoBox [79], the BoxInst [145], and the Box2Mask [88], have achieved high performance in box-supervised instance segmentation, which are competitive with respect to fully supervised instance segmentation methods.

[12]The content of the chapter has been submitted to IEEE Transactions on Geoscience and Remote Sensing with the paper entitled "Box2Boundary: Delineation of building boundaries with box supervision"

This chapter introduces a novel weakly supervised method called Box2Boundary, which is inspired by the current state-of-the-art approach, Box2Mask. Specifically, the Box2Mask is refined by replacing the official backbone with the latest powerful backbone for feature extraction and adding a dynamic scale training strategy [23] to deal with scale variance issues for rooftop delineation. For feature learning, the InternImage [153] is employed as the backbone. To refine the generated building masks, the regularization method proposed in Wei et al. [158] is adopted as the post-processing step in the proposed method. The objectives of this chapter are as follows:

(1) presenting a new box-supervised instance segmentation method for rooftop delineation, and

(2) conducting the extensive experiments on the publicly available dataset and showing the competitive performance of the new method.

5.2 Datasets and Methods

5.2.1 Dataset Preparation

In literature, the WHU Building dataset is widely used in weakly supervised rooftop delineation [157, 16]. Therefore, in this chapter, it is used to evaluate the proposed Box2Boundary and other methods in the comparative study and the ablation study. Its training, validation and testing splits have 4736, 1036, and 2416 tiles with the size of 512×512 pixels and the spatial resolution of 0.3 m/pixel in RGB bands as described in sections 2.1, 3.2 and 4.2. For box-supervised rooftop delineation, in the training and testing phases, only box annotations are used, which are generated based on ground truth building masks in this chapter. The instance masks of buildings are used only for evaluation.

5.2.2 Box2Mask

Box2Mask combines a level-set evolution model with deep learning to achieve accurate mask prediction using only bounding box supervision. The method utilizes input images and generated features for level-set evolution. In addition, a local consistency module based on a pixel affinity kernel is used to mine the local context and spatial relations. The instance masks are iteratively optimized within their bounding boxes by minimizing the level-set energy. Li et al. [88] conducted experiments on box supervised instance segmentation with ResNet [54] and Shifted Windows (Swin) transformer [99] based Box2Mask. With Swin

transformer, the transformer based Box2Mask gave significantly better performance. In both CNN-based and transformer-based Box2Mask, three components make up the model, including an instance-aware decoder, the box-level matching assignment and the level-set evolution, as shown in Figure 5.1.

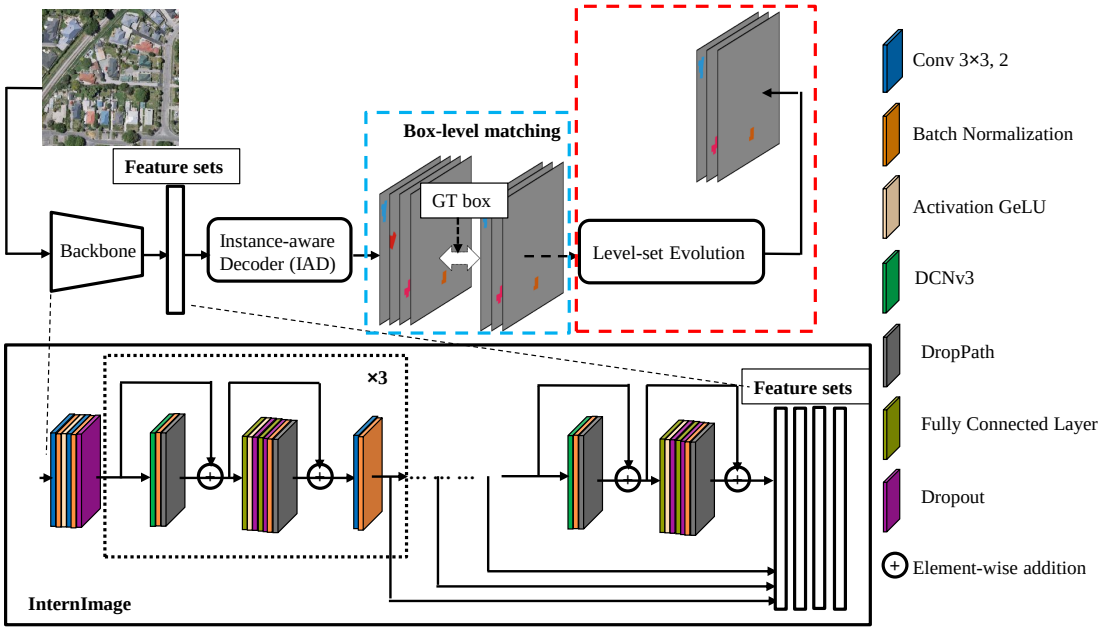

Figure 5.1: Architecture of Box2Boundary for rooftop delineation

Instance-Aware Decoder (IAD) consists of a pixel-decoder and a kernel-learning module. The kernel-learning module captures distinctive attributes of individual instances, such as intensity, appearance, shape, and location. With these attributes, the IAD module can generate instance-aware mask maps corresponding to objects in the input images. The CNN-based IAD utilizes dynamic convolution with learned kernels from a kernel learning network, along with unified mask features aggregated from multi-scale features extracted by the Feature Pyramid Network (FPN). In the Transformer-based IAD, instance-aware kernel vectors are learned from a transformer decoder, encoding instance-wise features. Instead of convolutional layers, a multi-scale deformable attention transformer is utilized as the pixel decoder to generate long-range feature maps. The final instance-aware mask maps are generated by multiplying the learned vectors with the feature maps.

Box-level Matching Assignment assigns labels to generated instance-aware mask maps. For the CNN-based method, each location of the image has a correspondent instance-

aware mask map. If the location falls into the center region of any ground truth bounding boxes, the mask map is assigned as positive; otherwise, it is assigned as negative. The center region of bounding boxes is designed, using a scale factor, to ensure each bounding box has on average 3 matched mask maps. For the Transformer-based method, the assignment is finished by using a matching cost calculated as follow:

$$\text{Cost} = 2 \cdot \text{Cost}_{\text{Cls}} + 5 \cdot \text{Cost}_{\text{Loc}} \tag{5.1}$$

$$\text{Cost}_{\text{Loc}} = P_{\text{dice}}(\text{mask}_x^p, \text{box}_x^p) + P_{\text{dice}}(\text{mask}_y^p, \text{box}_y^p) \tag{5.2}$$

where mask_x^p, mask_y^p, box_x^p, and box_y^p represent the coordinate projections of the masks and the boxes on the x and y axes, respectively. The 1-D dice coefficient measurement, represented by "dice," is employed. Here, P denotes projection. The cross-entropy loss is utilized to calculate the category cost. In the transformer-based method, the box level assignment is performed by assigning each bounding box to each mask using this scheme [88].

Level-set Evolution model is a mathematical framework used in image processing and computer vision to represent and evolve curves and surfaces in a continuous manner. As mentioned previously, both input images and deep features, generated from unified features in the IAD module, are utilized to evolve the instance contours initialized from instance mask maps. In addition, in each step, an initial contour is generated by using a box projection function as the projection defined location loss in box-level matching assignment. As a result, the contour for each object is optimized in each step and its accuracy is improved gradually. To further exploit the pixel affinity with the neighborhood, an affinity kernel function is defined with 8-way local neighbors. After 10 iterations, a robust level-set is obtained with local affinity consistency.

5.2.3 InternImage

The backbone plays a key role in feature extraction for various computer vision tasks. In this section, the InternImage model, depicted at the bottom of Figure 5.1, is taken as the feature extractor. The InternImage model is composed of a stem block, and four basic blocks separated by three down-sampling blocks. To enhance its capabilities, the model utilizes Deformable Convolutional Networks (DCN) v3. As a core operator proposed in Li et al. [88], the DCN v3 is good at capturing long-range dependencies and adaptive spatial aggregation compared to conventional CNNs. With the inductive bias inherited from convolutions, DCN v3 demonstrates efficiency in terms of training speed and data utilization compared to attention-based operators. Furthermore, its sparse sampling nature enables computational and memory efficiency, along with ease of optimization.

5.2.4 Dynamic Scale Training

The dynamic scale training strategy, proposed by Chen et al. [23], has been demonstrated to be effective in addressing the challenges of small object detection, thereby improving detection performance. Given that rooftops exhibit variations in object sizes and scale, this strategy is employed to enhance the accuracy of box-supervised instance segmentation.

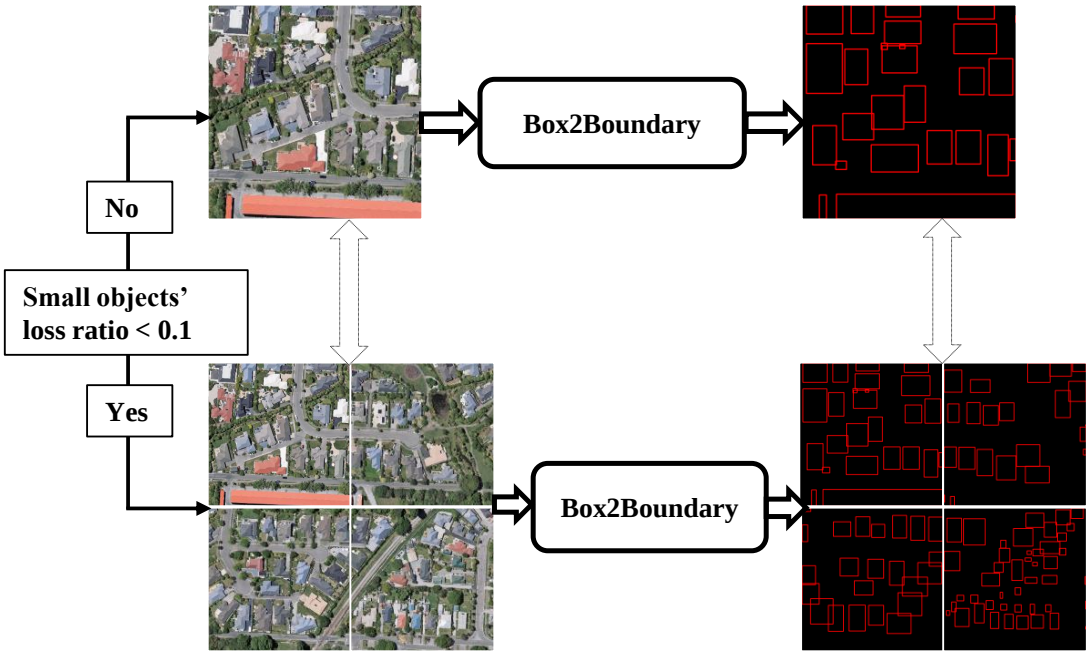

Figure 5.2: DST strategy in Box2Boundary for rooftop delineation

The principle behind the strategy is straightforward: if the ratio of small objects' loss to the total loss is smaller than a threshold, k images and their targets are collaged in next iteration. This process is illustrated in Figure 5.2. In this chapter, the threshold is set to 0.1 and the value of k is set to 4 following the empirical setting introduced by Chen et al. [23].

5.2.5 Post-processing

Although the building masks generated by the refined Box2Mask are accurate, the boundaries are irregular. In addition, the vectorized building boundaries are the required ultimate data. Therefore, a post-processing step needs to be conducted. In this chapter, two

algorithms [158] are employed to process each instance mask. Firstly, extremely small polygons, over-sharp angle edges, over-smooth edges and extremely short edges are removed after vectorizing the instance masks generated in previous steps. Secondly, the longest edge that constitutes the remaining edge is identified, and its direction is included in the main direction list. Then, other edges are traversed, and their directions are added to the main direction list based on an angle threshold between their directions and the directions already present in the list. Using this list and the angle adjustments, short and long edges are appropriately modified. Parallel edges are merged, resulting in the construction of the final polygons.

5.2.6 Evaluation Metrics

In weakly supervised rooftop delineation, the performance is commonly evaluated using metrics such as IoU, Precision, Recall, F1 score, and OA [16, 40, 169, 178]. These metrics are widely employed to assess the performance of different methods. Additionally, to facilitate comparison with other instance segmentation-based methods, object-level metrics, as used in the MS COCO [93] challenge, are also utilized to evaluate the extraction of rooftop delineation in this section.

5.2.7 Implementation Details

Following the Box2Mask, the level-set energy function is taken as the objective function in the proposed method. In addition, the cross-entropy loss is adopted to calculate the category loss, the dice coefficient-based projection loss is adopted as the bounding box detection loss and the differentiable level-set energy function is adopted as the mask segmentation loss. The weighted sum of these losses with a ratio of 2:5:1 is taken as the total loss for backpropagation. or the classification loss, a weight ratio of 10:1 is assigned to positive objects and negative objects, respectively. Data preparation involves dynamic scale training, random flipping, and large-scale jittering augmentation. The optimizer used for model training is AdamW, with an initial learning rate set to 2e-4 and weight decay set to 0.05. The models are trained for 200k iterations. The learning rate is decreased by a factor of 10 at iteration 160k and again at iteration 180K. The batch size and the input image size are set to 4 and 512×512 pixels. Unless stated otherwise, all backbones in this chapter are pretrained on ImageNet-1k, and all deep learning models share the same optimizer, the same learning rate and learning rate schedule, and the same number of total training iterations. All experiments in this section are conducted with Pytorch 1.10.0 and CUDA 11.8 on windows 11 with an Nvidia®RTX 3090 GPU.

5.3 Experimental Results and Analysis

5.3.1 Instance Level Extraction

In the comparative study, several methods are considered for instance segmentation. For fully supervised instance segmentation, the methods used include PolyMapper [90], HiSup [166], and HigherNet-DST (presented in Chapter 4) (as shown in Figure 4.6). For box supervised instance segmentation, methods from the computer vision field such as DiscoBox [79], BoxInst [145], and Box2Mask [88] are adopted. The qualitative evaluation results are provided in Figure 5.3, while the quantitative evaluation results are tabulated in Table 5.1.

Table 5.1: Instance segmentation results with the WHU Building Dataset (in %)

Methods	Supervision	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_L	AR	AR_s	AR_m	AR_L
PolyMapper	Mask	51.4	77.6	58.4	41.6	66.2	36.1	60.5	49.9	76.4	63.0
*HiSup	Mask	58.3	80.4	65.8	43.7	76.0	75.3	63.2	47.8	81.0	84.0
*HigherNet-DST	Mask	60.1	82.8	69.0	46.2	77.1	74.9	63.6	49.3	80.2	80.7
DiscoBox	Box	26.7	72.9	6.8	15.1	42.0	55.0	35.9	23.0	50.4	65.2
BoxInst	Box	37.4	79.0	29.9	21.7	52.7	69.8	44.8	30.1	61.2	79.8
Box2Mask	Box	46.7	83.5	49.2	29.3	64.4	70.9	50.8	35.0	68.8	75.4
Box2Boundary	Box	48.7	84.9	52.4	31.3	66.1	73.3	52.3	36.7	70.2	78.2

Figures 4.6 and 5.3, provides results generated by PolyMapper, HiSup, HigherNet-DST, DiscoBox, BoxInst, Box2Mask, Box2Boundary. As shown in Figure 4.6, mask supervised methods achieved better performance with accurate building boundaries, while box supervised methods can also detect building objects accurately and generate almost accurate building outlines. As discussed in section 4.3.4, among mask supervised methods, HigherNet-DST achieves the highest performance specifically on small objects in the first and last columns. Regarding the box supervised methods, there is a progression of increasingly accurate building boundaries from DiscoBox to the proposed Box2Boundary, with tighter and more regular outlines. Taking the first column as an example, DiscoBox and BoxInst cannot distinguish building pixels from neighboring pixels, resulting in larger masks that fully cover building objects. In contrast, Box2Mask and Box2Boundary classify building pixels and their neighborhood accurately but may generate incomplete building masks from some building objects. The difference between Box2Mask and Box2Boundary

Figure 5.3: Qualitative evaluation results of instance level rooftop delineation

is marginal in the first two columns but becomes more obvious in the last column. In the last column, Box2Boundary shows higher performance in delineating close buildings. Overall, by visually examining rooftop delineation results of the three examples, box supervised methods give competitive results with Box2Boundary demonstrating the best performance.

5.3.2 Pixel Level Extraction

Although instance segmentation is arguably more suitable for rooftop delineation, there have been numerous studies on semantic segmentation methods. Therefore, to compare with these methods, the evaluation results of both fully supervised and weakly supervised semantic segmentation methods, along with the proposed Box2Boundary, are tabulated in Figure 5.3 and Table 5.2 using abbreviated names. Firstly, the fully supervised semantic segmentation methods including the U-Net [123], the DeepLab v3+ [18] and the Deep Dual-Resolution Networks (DDRNet) [57] are selected for the comparison. Binary building masks are used in these methods as supervision targets.

Secondly, weakly supervised semantic segmentation with image level annotations, including the Explicit Pseudo-pixel Supervision (EPS) [83], the Class Activation Maps (CAM) [192] based method, a weakly supervised building segmentation method that combines the Superpixel Pooling and Multi-scale Feature fusion structures (SPMF-Net) [17], a self-supervised equivariant attention mechanism (SEAM) [156], the Weakly Supervised Feature-fusion NETwork (WSF-NET) [42], a weakly supervised network integrating MultiScale Generation and Superpixel Refinement (MSG-SR-Net) [168] and the Pixelwise Affinity Network (PANet) [169], the AdvCAM [83] and the Adversarial Climbing Gated Convolution (ACGC) [40] are selected for the comparison purpose.

Furthermore, weakly supervised methods with scribble annotations, including the Weakly-supervised Salient Object Detection (WSOD) [179], the Uncertainty Reduction and Self-Supervision (URSS) [115], the ScRoadExtractor [159] and the Structure-Aware Weakly Supervised Network (SAWSN) [16] are collected for the comparison.

Finally, to compare the proposed method Box2Boundary with existing box supervised methods,DiscoBox, BoxInst and Box2Mask are also included here by converting generated instance masks to binary masks. In addition, the weakly supervised semantic segmentation with the Background-Aware pooling and Noise-Aware loss (BANA) and the method with a multiscale feature retrieval (MFR), a pseudo-mask generation and correction (PGC) modules (MFR-PGC-Net) are collected from remote sensing field for the comparative study.The results for fully supervised methods and weakly supervised methods with image level annotations and scribble annotations are cited from [169] and Chen et al. [16]. Furthermore,

Table 5.2: Semantic segmentation results with the WHU Building Dataset (in %)

Methods	Supervision	Precision	Recall	F_1 score	OA	IoU
U-Net	Mask	95.11	93.83	94.47	-	89.51
DeepLab v3+	Mask	94.04	93.42	93.73	-	88.20
DDRNet	Mask	94.72	94.68	94.70	-	89.93
EPS	Image tag	60.06	47.44	53.01	-	36.06
CAM	Image tag	-	-	59.66	84.67	42.50
SPMF-Net	Image tag	-	-	53.23	85.60	36.26
SEAM	Image tag	-	-	68.82	89.74	52.47
WSF-Net	Image tag	-	-	62.24	88.99	45.18
MSG-SR-Net	Image tag	-	-	68.98	91.80	52.64
PANet	Image tag	-	-	74.59	93.68	59.48
AdvCAM	Image tag	79.13	62.98	70.14	-	54.01
ACGC	Image tag	77.86	66.70	71.85	-	56.06
WSOD	Scribble	88.25	89.42	88.83	-	79.90
URSS	Scribble	88.04	93.01	90.46	-	82.58
ScRoadExtractor	Scribble	67.30	79.88	73.05	-	57.55
SAWSN	Scribble	92.00	92.45	92.22	-	85.57
BANA	Box	87.24	85.21	86.22	-	75.77
MFR-PGC-Net	Box	85.03	89.07	87.00	-	76.99
DiscoBox	Box	76.21	95.55	84.79	96.19	73.60
BoxInst	Box	85.43	93.51	89.29	97.50	80.66
Box2Mask	Box	91.92	86.00	88.86	97.60	79.96
Box2Boundary	Box	92.05	87.48	89.71	97.77	81.34

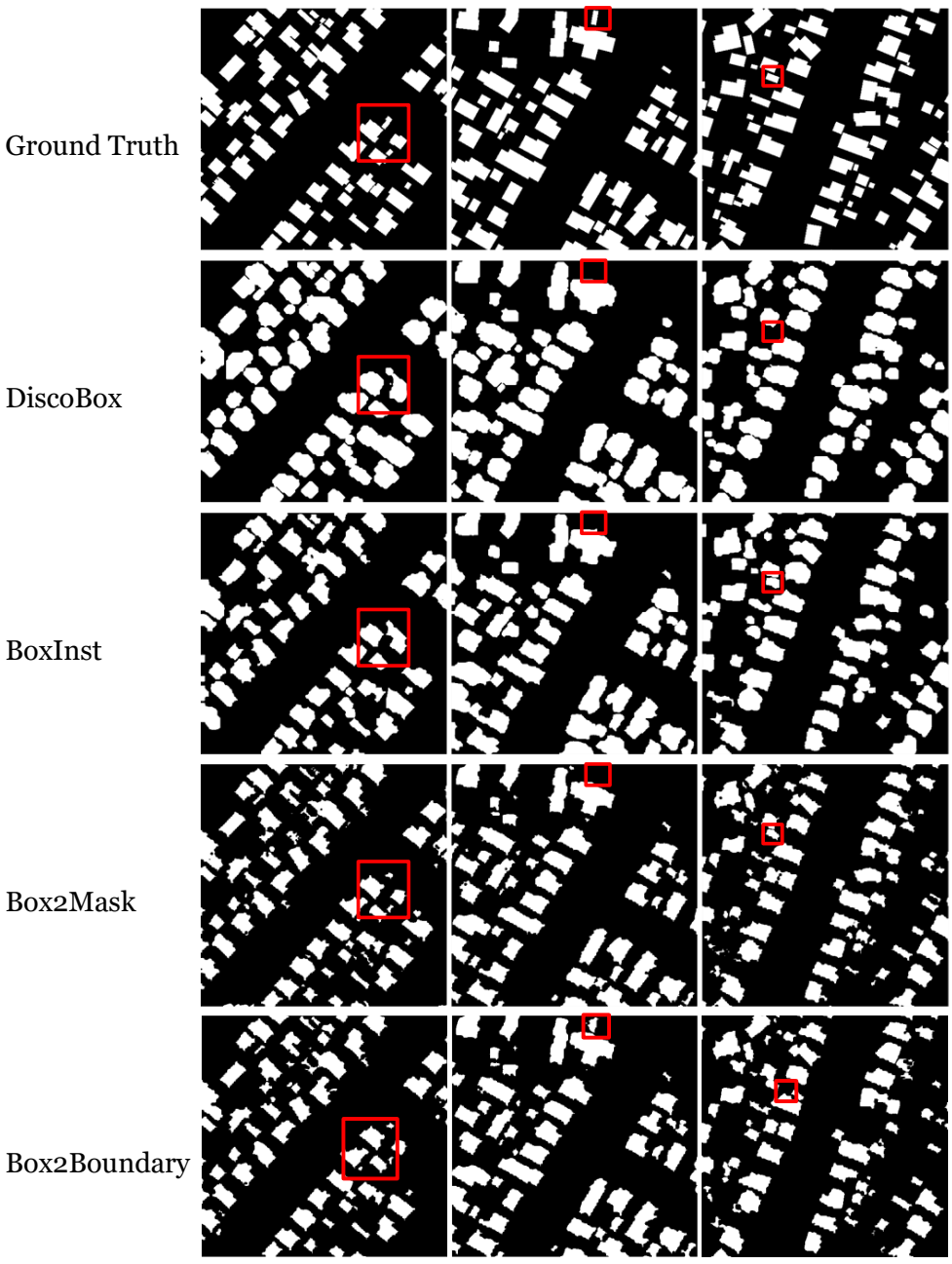

Figure 5.4: Qualitative evaluation results of pixel level rooftop delineation

this chapter collects results of AdvCAM, ACGC, BANA and MFR-PGC-Net from Zheng et al. [190].

Similarly, to evaluate the results of pixel level extraction, the qualitative evaluation results are provided in Figure 5.4. However, only box supervised methods are evaluated in the qualitative evaluation section. There are three reasons for this decision: (1) fully supervised, image tag supervised, and scribble supervised methods are not the focus in this thesis; (2) several methods do not have open-sourced code; and (3) several methods are hard to re-implement.

In Figure 5.4, the ground truth and extraction results from DiscoBox, BoxInst, Box2Mask and Box2Boundary are given from the top to the bottom row, respectively. As analyzed previously, DiscoBox and BoxInst give larger masks for building objects. In addition, DiscoBox and BoxInst miss the extraction of several small building objects in all three columns. As shown in the red boxes in Figure 5.4, Box2Mask and Box2Boundary show better performance compared to DiscoBox and BoxInst. In the second column, the building located in the middle top of the second example is missed in DiscoBox, BoxInst and Box2Mask based extraction. The successful extraction of the object by Box2Boundary confirms its superiority compared to other methods. Overall, Box2Boundary exhibits slightly better performance than Box2Mask and significantly outperforms DiscoBox and BoxInst.

To evaluate their performance quantitatively, Table 5.2 compares the fully supervised, the image tag supervised, the scribble supervised, and the box supervised methods as mentioned previously. As expected, fully supervised methods give the best performance. Scribble supervised and box supervised methods give similar performance, which is worse than that of mask supervised methods but better than that of image tag supervised methods. Based on the experiment results, it is promising that with higher performance deep learning models more accurate results can be achieved.

5.4 Post-processing

As shown in Figures 5.3 and 5.4, Box2Boundary can extract building boundaries accurately, but the boundaries are irregular which need to be refined further. By applying the method proposed in Wei et al.[158], the results generated by Box2Boundary are regularized and presented in this subsection. In the post-processing method, the key parameter is a threshold, epsilon, used in the Douglas–Peucker (DP) algorithm for removing redundant vertices. Specifically, given a line with multiple vertices, if the max length of all interval points to the line connecting the first and last vertices is smaller than the threshold, all

interval vertices are removed. If not, the point corresponding to the max length will be the starting point or the ending point for the next iteration. For each line, the iteration continues till no vertex can be removed. In Figure 5.5, a fictitious example is presented to illustrate DP algorithm. As shown in the figure, if the max length dmax is smaller than the threshold, epsilon, the vertices including B, C, D, E, and F can be removed. If not, the line will be split into two lines, from A to C and from C to G. And the vertex C will be preserved for the final polyline. The next iteration starts from line A to C or line C to G. The threshold, epsilon, is set to 6 by default, which means 6 pixels.

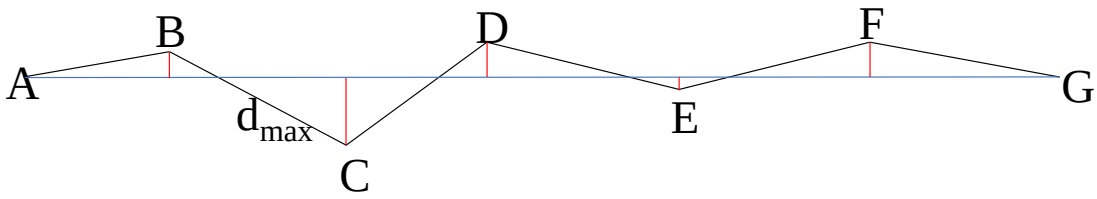

Figure 5.5: An example of the Douglas-Peucker (DP) algorithm

Table 5.3: Performance of Box2Boundary with different thresholds in post-processing (in %)

Methods	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_L	AR	AR_s	AR_m	AR_L
Box2boundary	48.7	84.9	52.4	31.3	66.1	73.3	52.3	36.7	70.2	78.2
Episilon=1	47.5	82.3	51.5	32.6	64.6	58.6	52.3	37.3	69.6	71.4
Episilon=3	47.2	82.3	51.0	32.5	64.4	54.0	52.2	37.3	69.4	68.8
Episilon=6	46.8	82.1	50.0	32.4	63.8	52.4	51.9	37.1	69.0	68.1

By simplifying polygons, the accuracy of rooftop delineation results will inevitably decrease, but more aesthetically pleasing rooftop delineations can be generated. Table 5.3 shows the decreased performance when applying post-processing and increasing the threshold. Figure 5.6 visualizes the post-processing results with different thresholds based on the outputs generated by Box2Boundary. Specifically, from top to bottom, the figure presents the rooftop delineation results without post-processing, with post-processing using thresholds of 1, 3 and 6, respectively. As shown in Figure 5.6, by post-processing extraction results, the building boundaries become more regular. However, increasing the threshold epsilon exacerbates the shifting issue. For instance, in the first example, the top right building exhibits a more regular boundary with increased threshold, but the bottom right

Figure 5.6: Qualitative evaluation results of Box2Boundary with different thresholds in post-processing

part of the building is excluded from the rooftop. Another example can be found in the top left of the second example close to the image boundary. The regular but inaccurate rooftop boundary is generated by increasing the threshold. The issue can be found in all three examples and all the extraction results. As manually checking and editing is inevitable, it is preferable to obtain aesthetically pleasing and easily editable results [134]. Therefore, the default epsilon value 6 is recommended to ease the manually editing work in practice.

5.5 Discussion

5.5.1 Ablation Study

As mentioned, Box2Boundary is developed on top of Box2Mask by replacing the backbone and applying DST in model training. To test the effectiveness of each modification, Box2Mask, Box2Mask with the tiny InternImage, and Box2Boundary are evaluated in this section. As shown in Table 5.4, the performance improvement from Box2Mask to Box2Boundary mainly comes from the backbone replacement, although applying DST also helps the improvement. For example, in terms of the AP value, the backbone replacement brings 1.8% increment, and the DST strategy improves the value by 0.2%. Applying DST may lead to a decrease in AR metrics except AR_L, but it helps the improvement in general. Therefore, both modifications in this chapter are effective regarding box supervised rooftop delineation.

Table 5.4: The ablation study of Box2Boundary in rooftop delineation (in %)

Methods	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_L	AR	AR_s	AR_m	AR_L
Box2Mask	46.7	83.5	49.2	29.3	64.4	70.9	50.8	35.0	68.8	75.4
+tiny InternImage	48.5	84.0	52.3	31.3	66.0	71.2	52.7	37.3	70.4	77.2
Box2boundary	48.7	84.9	52.4	31.3	66.1	73.3	52.3	36.7	70.2	78.2

5.5.2 Other Techniques Tried

Pretrain Models or Not

In computer vision, it is common to initialize the weights of deep learning models using pretrained networks trained on standard image datasets for general object tasks. For example, in this section, all backbones are pretrained on ImageNet 1K. In contrast, in rooftop delineation, deep learning models are generally trained from scratch. In Table 5.5, the performance difference between using pretrained models and training from scratch is explored. Specifically, Box2Mask with a tiny Swin transformer is applied as the baseline. The table indicates whether the model is "with pretrained" or "without pretrained", denoting whether the weights were initialized using a pretrained network or trained from scratch. The learning curves including training loss, bounding box mAP and mask mAP of validation dataset along iteration steps are provided in Figure 5.7.

Table 5.5: Performance of Box2Mask using pretrained weights (in %)

Methods	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_L	AR	AR_s	AR_m	AR_L
With pretrained	46.7	83.5	49.2	29.3	64.4	70.9	50.8	35.0	68.8	75.4
Without pretrained	45.5	82.9	47.2	28.3	62.6	64.9	49.5	34.2	67.0	70.3

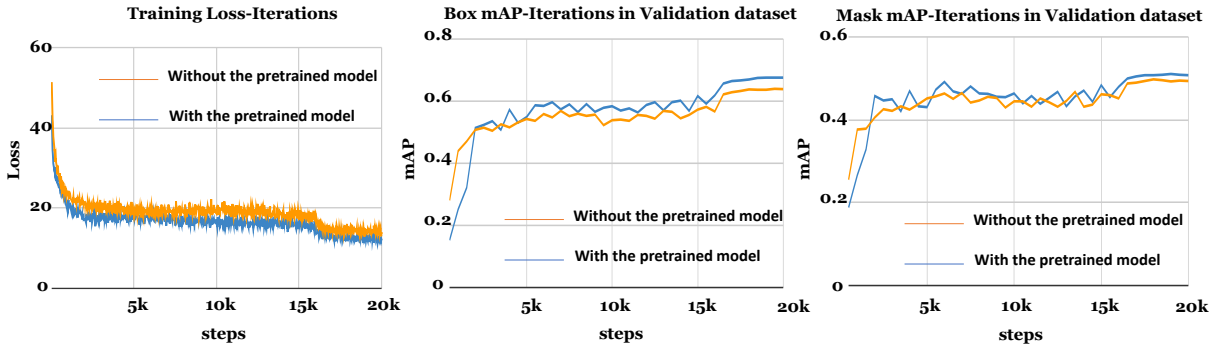

Figure 5.7: Learning curves with and without pretrained weights

As shown in Table 5.5, the utilization of a pretrained backbone resulted in a notable improvement in rooftop delineation performance. This positive impact of pretrained model weights on model training is further confirmed by the learning curves presented in Figure

5.7. Although there are initially different trends in bounding box AP and mask AP for the validation dataset, the positive impact becomes evident as the training progresses. In conclusion, the use of pretrained model weights is beneficial for effective model training.

Different Feature Extractors

The selection of a proper backbone for feature extraction is crucial for improving performance in Box2Boundary. This section evaluates the performance of a small InternImage and the state-of-the-art backbone, tiny ConvNeXt v2 [161]. With its simple convolutional architecture, ConvNeXt v2 demonstrates competitive performance compared to Vision Transformer-based methods in the computer vision field. Given limited computational resources, only tiny ConvNeXt is considered here. The performance comparison includes Box2Mask as the baseline with a tiny Swin transformer, followed by Box2Mask with small InternImage, tiny InternImage, and tiny ConvNeXt v2. Table 5.6 tabulates the performance of Box2Mask with different backbones.

Table 5.6: Performance of Box2Mask with different backbones (in %)

Methods	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_L	AR	AR_s	AR_m	AR_L
Tiny Swin	46.7	83.5	49.2	29.3	64.4	70.9	50.8	35.0	68.8	75.4
Small InternImage	45.5	81.9	48.0	28.2	63.0	70.9	49.7	34.0	67.7	76.8
Tiny InternImage	48.5	84.0	52.3	31.3	66.0	71.2	52.7	37.3	70.4	77.2
Tiny ConvNext v2	47.4	84.7	50.3	29.8	64.8	72.1	51.3	35.6	69.2	76.0

In Table 5.6, it can be observed that Box2Mask with tiny InternImage achieves the best performance, followed by tiny ConvNeXt v2, while small InternImage yields the lowest performance. Interestingly, the underperformance of small InternImage compared to tiny InternImage is unexpected. One possible explanation is that a larger feature extractor may overfit to a box supervised instance segmentation. Specifically, the memory consumption for the models as follows: tiny Swin transformer-17,434 MB, tiny InternImage-19,776 MB, small InternImage-21,780 MB, and tiny ConvNeXt v2-19,947 MB. The performance differences align with the memory consumption except for the tiny Swin transformer, which is understandable since it lacks certain recently developed techniques in deep learning. The variation in performance can be attributed to overfitting using different backbones.

Overfitting Issues

As discussed earlier, the overfitting issue probably is hindering the performance of box supervised instance segmentation. To address this issue, Liu et al. [100] proposed early dropout and late dropout techniques to mitigate underfitting and overfitting. Specifically, for underfitting models, dropout is applied only in the first 1000 iterations and disabled afterward, while for overfitting models, dropout is enabled only in the last 10% of iterations. Both strategies have been shown to be effective in their experiments. In addition to the early dropout and the late dropout, in this section, the impact of dropouts is further explored by making the dropout rate trainable in all dropout layers. In this section, the aim is to enhance box supervised rooftop delineation by testing these dropout strategies on Box2Mask with tiny InternImage as the baseline. Table 5.7 presents the evaluation and comparison of normal training and training with different dropout strategies. Among the various strategies, training with late dropout achieves performance comparable to normal training. This suggests that the tiny InternImage-based method indeed suffers from overfitting, and further fine-tuning of the training schedule is necessary for improved performance.

Table 5.7: Performance of Box2Mask with different dropout strategies (in %)

Methods	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_L	AR	AR_s	AR_m	AR_L
Normal	48.5	84.0	52.3	31.3	66.0	71.2	52.7	37.3	70.4	77.2
Early Dropout	47.6	83.9	51.0	30.6	64.7	71.7	51.7	36.3	69.3	76.3
Late Dropout	48.3	84.9	51.8	31.2	65.8	71.8	52.1	36.6	69.9	77.0
Trainable Dropout	48.0	84.7	51.5	30.4	65.7	73.7	51.7	35.9	69.8	78.1

Multi to Single Module for Scale-Variance Issue

As discussed in Chapters 2 and 4, scale-variance issues exist in rooftop delineation and impede performance. Various methods, including data preparation and model optimization techniques, have been employed to address these scale-variance issues, as discussed in Section 2.2. In this section, the recently developed model optimization method Multi to Single Module (M2S) [47] is embedded into Box2Mask with tiny InternImage to evaluate its performance on dealing with scale-variance issues.

There are two modules in M2S: the Cross-scale Feature Aggregation (CFA) and the Dual Relationship Module (DRM), which consider spatial and channel relationships. In

the CFA modules, every three adjacent features from the backbone output are taken as input to generate fused features. The fused features are taken as input for next level feature fusion. Finally, high, middle, and low-level features are created from a "V" shape CFA modules combo, which will be sent to DRM. In Box2Mask with tiny InternImage, four features (C1, C2, C3, and C4) are generated from the backbone. C2, C3, and C4 are used to create the high-level feature through a CFA module, while C1, C2, and the high-level feature are used to generate the middle-level feature in another CFA module. C1 serves as the low-level feature. The three level features and the decoded feature using C4 are then passed through the DRM. Afterward, the four features are integrated into a single output feature, which replaces the decoded feature in the original Box2Mask with tiny InternImage. This integration is expected to enhance small object extraction and overall performance. However, as indicated in Table 5.8, incorporating M2S results in a decline in rooftop delineation performance. This poor performance is likely due to severe overfitting caused by the addition of the M2S module, which consumes 20,496MB of memory in each iteration. This experiment confirms previous findings. Therefore, when aiming to enhance the performance of Box2Mask or Box2Boundary, it is advisable to avoid increasing the model complexity.

Table 5.8: Performance of Box2Mask with tiny InternImage using M2S module (in %)

Methods	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_L	AR	AR_s	AR_m	AR_L
Original	48.5	84.0	52.3	31.3	66.0	71.2	52.7	37.3	70.4	77.2
M2S	45.7	82.6	48.0	28.8	62.7	68.3	49.8	34.5	67.3	73.3

5.6 Chapter Summary

In this chapter, the costly annotation issue in rooftop delineation is taken as the core topic. Based on Box2Mask, a box supervised instance segmentation in computer vision, Box2Boundary is proposed for box supervised rooftop delineation. Box2Boundary improves upon Box2Mask by replacing the backbone with tiny InternImage and incorporating DST to handle scale-variance issues. By conducting an experiment on the WHU Building Dataset, Box2Boundary achieves promising performance compared to fully supervised methods, shows competitive performance compared to scribble-based methods, and surpasses image-tag supervised, and other box supervised rooftop delineation methods. Specifically, on the WHU Building Dataset, Box2Boundary surpasses its baseline

Box2Mask by 2% and 1.5% in terms of AP and AR, showing its superiority. It also shows significant improvement compared to BoxInst and DiscoBox in terms of all instance level metrics. In semantic segmentation-based methods, box supervised methods show promising performance compared to other methods using different supervision targets. Specifically, box supervised methods show promising performance compared to fully supervised methods. They also show significantly better performance compared to image tag supervised methods. To refine the rooftop delineation results, the post-processing proposed by Wei et al. [158] is employed in this chapter. However, it is observed that while post-processing effectively regularizes the rooftop shape, it adversely affects the accuracy score of the extraction.

In the discussion section, the success of the Box2Boundary proposal is confirmed through the ablation study, which demonstrates improved performance compared to its baseline, Box2Mask. The effectiveness of using a pretrained backbone is also validated as a helpful factor for rooftop delineation. Specifically, employing a pretrained tiny Swin transformer enhances the performance of Box2Mask compared to training from scratch. Different backbones are also tested for better performance. The results show that the tiny InternImage yields superior feature extraction and rooftop delineation. The experiments also highlight the presence of overfitting issues that hinder performance improvement in rooftop delineation. To address these issues, various dropout strategies are tested, which confirm the presence of overfitting in Box2Mask with the tiny InternImage backbone. Finally, to deal with scale variance issues, the M2S module is also tested as a model optimization method. However, it is observed that incorporating the M2S module leads to a decrease in performance for box supervised methods. Consequently, the experiments demonstrate that Box2Boundary, featuring a pretrained InternImage backbone, DST, and post-processing, is the most effective method and pipeline for box supervised rooftop delineation compared to other weakly supervised methods.

Chapter 6

Conclusions and Recommendations

6.1 Conclusions

Rooftop delineation plays a crucial role in various applications such as urban planning and management, cadastral management, urban geo-database update and smart city construction. It is also important for disaster management, epidemic control, population estimation and damage assessment. Leveraging remote sensing imagery, particularly aerial imagery, DL-based methods offer an efficient approach for generating HD building maps. However, automated rooftop delineation faces challenges related to generalization issues, scale-variance issues, and the costly annotation issues. Different geo-locations and spatial resolutions in images introduce variations in building characteristics, leading to reduced performance in DL-based extraction methods. Scale-variance issues, common in computer vision tasks, pose a specific challenge in accurately extracting rooftops, especially for smaller structures. DL methods heavily rely on extensive, high-quality training data, which necessitates labor-intensive and expensive human labeling efforts. To address these challenges, this thesis proposes three corresponding methods.

To address generalization issues, a data composition approach using the SISR method MSCA-RFANet is proposed. By employing high-performance SISR methods to process datasets with varying spatial resolutions and combining the processed datasets, positive results are achieved in mitigating generalization issues. The proposed MSCA-RFANet builds upon RFANet and introduces a dual attention scheme using SCA blocks in the trunk part. To enhance model training and improve performance, a momentum scheme is adopted for the skip connection between the basic modules of MSCA-RFANet. The proposed SISR method demonstrates superior performance in SISR. By employing the

proposed SISR method for data composition, better performance is achieved in addressing generalization issues for rooftop delineation. Experimental results on the Waterloo Building Dataset, the WHU Building Dataset, and the Massachusetts Building Dataset reveal the positive impact of the proposed SISR method in improving the performance of rooftop delineation by integrating the spatial resolution of different datasets. The proposed data composition with MSCA-RFANet shows better performance on rooftop delineation using unseen data as test data. Consequently, it can be concluded that the data composition with MSCA-RFANet is effective in dealing with generalization issues

To address scale-variance issues, HigherNet-DST is proposed. Building upon the HiSup method, which is a state-of-the-art end-to-end rooftop delineation approach, HigherNet-DST introduces scale-aware HigherNet and DST techniques to address the challenges posed by scale-variance. Notably, HigherNet-DST leverages high spatial resolution supervision targets instead of low spatial resolution ones to capture more detailed features and improve performance. By incorporating these innovative modifications, HigherNet-DST achieves impressive results in rooftop delineation. It outperforms other state-of-the-art methods, particularly in accurately delineating small objects, as demonstrated on publicly available building datasets. As a result, HigherNet-DST is regarded as an effective method for practical rooftop delineation, showcasing its potential in real-world applications.

To alleviate the costly annotation issues, box-supervised instance segmentation methods from computer vision tasks are innovatively introduced in rooftop delineation. For better performance, after examining state-of-the-art methods in box supervised methods, Box2Boundary is proposed for rooftop delineation with box supervision. State-of-the-art box supervised instance segmentation methods, including DiscoBox, BoxInst, and Box2Mask, are examined, and based on the best-performing method, Box2Mask, several enhancements are made. Firstly, the backbone of Box2Mask is replaced with the InternImage, which enhances feature extraction in deep learning. Additionally, to overcome the challenge of scale-variance, DST is incorporated into the Box2Boundary framework. Since the generated results may have irregular shapes, post-processing methods are implemented to refine and regularize the rooftop boundaries. With only box annotations, Box2Boundary shows promising performance compared to fully supervised methods, competitive performance compared to scribble supervised methods which require similar effort for generating training data, better performance compared to image tag supervised methods. With post-processing, the performance of Box2Boundary is inevitably decreased, but the generated rooftop boundaries become regular and easy to be refined. Overall, it can be concluded that the proposed Box2Boundary can alleviate the costly annotation issues with decent performance.

In conclusion, this doctoral thesis proposes three innovative solutions for addressing key

challenges in deep learning-based rooftop delineation: generalization issues, scale-variance issues, and costly annotation. Through extensive experiments, the effectiveness of each solution is demonstrated. Extensive experiments in this thesis prove the effectiveness of data composition with MSCA-RFANet in dealing with generalization issues in terms of different spatial resolutions and geo-locations. The experiments also show the high performance of HigherNet-DST in rooftop delineation, especially in delineating small buildings, compared to other state-of-the-art methods. The experiments on weakly supervised methods confirm the potential of using box supervised method for rooftop delineation with descent performance. All in all, these methods can facilitate automated and accurate rooftop delineation and improve the practical applicability of deep learning-based rooftop delineation.

6.2 Contributions

DL based rooftop delineation methods are widely used recently. However, generalization issues, scale variance issues and costly annotation issues still impede their practical use. In this thesis, the innovatively proposed methods provide feasible solutions for these challenges. Therefore, this thesis has three contributions:

(1) A novel SISR method, MSCA-RFANet, is proposed, which can be combined with data composition to overcome generalization errors in rooftop delineation when training data and test data have different spatial resolutions and characteristics. The proposed dual attention block, known as the SCA block, effectively captures both channel information and spatial details, enhancing the model's ability to focus on relevant features. Additionally, the introduction of a momentum scheme facilitates training and enables the network to be deeper. Experiments show the effectiveness of the proposed method and the data composition solution on dealing with generalization issues. On the Inria Building Dataset, the solution achieves high performance with an OA of 88.78% and an IoU of 39.93%. These results are promising when compared to extraction results obtained from in-distribution datasets and demonstrate superior performance compared to models trained on single datasets or simple mixed datasets.

(2) A high-performance end-to-end rooftop delineation method, HigherNet-DST, is proposed, which can generate accurate and regular building boundaries. By leveraging the scale-aware HigherNet, high spatial resolution supervision targets, and DST, the method excels in delineating small buildings while maintaining overall performance. Experimental results demonstrate the competitive performance of HigherNet-DST on the AICrowd Building Dataset, achieving an average precision (AP) of 68.5%. Moreover, the method ex-

hibits substantial improvement compared to its baseline, HiSup, on other building datasets and outperforms PolyMapper in terms of performance.

(3) Box2Boundary is proposed to alleviate costly annotation issues in DL based rooftop delineation. While box supervised instance segmentation methods have shown excellent performance in computer vision tasks, their potential in rooftop delineation remains unexplored. To fully take advantage of DL techniques, box supervised instance segmentation is introduced into rooftop delineation. By utilizing InternImage as a superior feature extractor instead of the one used in Box2Mask, Box2Boundary shows better performance in rooftop delineation. Further enhancement is achieved by incorporating DST into the framework. Box2Boundary with an AP value of 48.7% on the WHU Building Dataset achieves a competitive performance compared to PolyMapper, a fully supervised rooftop delineation method. This result surpasses the performance of its baseline and other box supervised methods. By comparing with other semantic segmentation-based rooftop delineation methods, Box2Boundary shows promising performance with an IoU of 81.34% compared to fully supervised methods. Notably, it outperforms image tag methods significantly and shows superiority over several scribble supervised methods while remaining competitive with other state-of-the-art methods. Although post-processing affects its performance, it leads to aesthetically pleasing building boundaries that are easier to refine. Therefore, Box2Boundary is an effective method capable of achieving decent performance in rooftop delineation using only box annotations.

6.3 Recommendations for Future Research

In this thesis, three challenges in rooftop delineation have been addressed. The proposed solutions have provided alleviation to some extent, although there is still potential for further performance improvement. Additionally, it is important to recognize that there are other challenges in rooftop delineation that require urgent attention. The following recommendations are suggested for future research:

Worldwide data composition: in the thesis, data composition has been proven effective for dealing with generalization issues. However, when evaluated on the Inria Building Dataset, there is still a noticeable performance gap in handling generalization challenges. The relatively poor performance of the models trained on the composited data can be attributed to the limited representation of building characteristics within the Inria Building Dataset. This dataset consists of aerial images captured in the USA and Austin, exhibiting diverse illumination conditions and distinct building characteristics. Given the

constraints of computational resources, this thesis focused on utilizing the Waterloo Building Dataset, the WHU Building Dataset, and the Massachusetts Building Dataset for data composition. While collecting training data specifically from the USA and Austin could potentially improve performance on the Inria Building Dataset, it is also important to consider the development of a more comprehensive dataset that encompasses diverse building types and global locations. Constructing such a dataset would enable training a rooftop delineation model capable of accurately delineating rooftops worldwide. In other words, it is beneficial to construct a dataset that supports training a rooftop delineation model with the ability to extract various types of buildings, akin to the Segment Anything Model [74] with respect to segmenting any objects in natural scenes.

Constructing a model with global generalizability requires the construction of a diverse training dataset comprising images from various regions worldwide. Utilizing the WHU Building Dataset and the SpaceNet Building Dataset as the foundation for a composite dataset is a promising solution. These datasets encompass aerial images obtained from Oceania, North America, Europe, Africa, and South America, providing a good representation of spatial variability. However, it is evident that the dataset volume would still be substantial, even without incorporating aerial images from Asia. One interesting area for future research is exploring effective strategies for creating training data that can enhance the model's generalizability. This involves considering how to maximize the dataset's diversity while managing its size and complexity. Techniques such as data augmentation, transfer learning, and domain adaptation could be employed to augment the existing dataset and simulate a more comprehensive global representation. Moreover, incorporating additional datasets from underrepresented regions, such as Asia, would further improve the model's ability to generalize across diverse geographies. Finding efficient ways to generate effective training data, considering factors like data volume, diversity, and representativeness, is an important avenue for further investigation. By addressing these challenges, researchers can advance the development of models with enhanced generalizability for rooftop delineation across a global scale.

Improve computational efficiency: for better performance, as discussed above, more training data are required, which extends the training time. For instance, training HigherNet-DST on the AICrowd Building Dataset using 2 Nvidia®RTX 3090 GPUs, even with the utilization of auto mixed precision, typically requires approximately three weeks. As model architectures become more complex to achieve better performance, training times are further prolonged. Therefore, it is essential to explore research directions that aim to accelerate model training while maintaining high performance in rooftop delineation. Given limited computational resources, there are several promising solutions to address this challenge. Firstly, developing or redesigning learning rate schedule strategies can help

optimize the training process and potentially reduce training time. Secondly, leveraging transfer learning techniques, such as using pretrained models as initializations, can enable faster convergence and more efficient training. Additionally, exploring knowledge distillation methods, which involve transferring knowledge from a larger, pretrained model to a smaller one, can lead to faster training while preserving performance. Efforts should be made to further investigate these strategies and explore other potential techniques to improve the computational efficiency of rooftop delineation models. By finding effective ways to accelerate model training without compromising performance, researchers can enhance the practical applicability of deep learning-based rooftop delineation in real-world scenarios.

Reduce the manual editing for post-processing: in instance segmentation or semantic segmentation-based methods, manual editing is often necessary to generate the final building maps. While hypothesis-based methods like Wei et al. [158], and learning-based methods such as those proposed by Zorzi and Fraundorfer [201] and Zori et al. [200] have demonstrated their effectiveness, there is a need for improved methods to minimize the extent of manual editing and ultimately reduce or eliminate human intervention in the map generation process.

Extraction from off-nadir imagery: in building footprints or rooftops mapping, it is commonly assumed that images are ortho-imagery or ortho-rectified, implying that the rooftops in the images align precisely with the building footprints or have minimal deviations. However, this is not true in general. The issue for high buildings is more obvious and severe. Recent studies [150, 166] and the proposed HigherNet-DST have introduced offset prediction as one of the tasks in rooftop delineation, resulting in noticeable performance improvements. However, there are still gaps that need to be addressed in order to further enhance the performance, particularly when dealing with off-nadir imagery.

3D building reconstruction: 3D building models present more comprehensive information compared to rooftops or building footprints in the previously mentioned applications. Previous research has explored various approaches for generating height information from single-view imagery with shadow information and other auxiliary data [116, 104], or by utilizing multi-view imagery to create stereo pair imagery and employing photogrammetry methods for 3D building model generation [173]. However, these methods often require specialized expertise and may rely on digital surface models (DSMs) or DSM-derived datasets to estimate height information. With the development of LiDAR technology, LiDAR data have become more accessible, providing effective data for 3D building reconstruction. LiDAR can be used to generate DSMs or normalized DSMs (nDSMs) for image-based reconstruction methods. It can also be directly used to generate 3D building models with rich spectral information. To address noise and data inconsistency issues in

LiDAR data, the fusion of images and LiDAR data has been explored for 3D building model reconstruction [152]. Research focusing on fully leveraging LiDAR data for more accurate and comprehensive 3D building model reconstruction holds significant interest and potential utility.

References

[1] Scott T Acton. Diffusion partial differential equations for edge detection. In *The Essential Guide to Image Processing*, pages 525–552. Elsevier, 2009.

[2] Salman Ahmadi, Mohammad Javad Valadan Zoej, Hamid Ebadi, Hamid Abrishami Moghaddam, and Ali Mohammadzadeh. Automatic urban building boundary extraction from high resolution aerial images using an innovative model of active contours. *International Journal of Applied Earth Observation and Geoinformation*, 12(3):150–157, 2010.

[3] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4981–4990, 2018.

[4] Mohammad Awrangjeb, Mehdi Ravanbakhsh, and Clive S Fraser. Automatic detection of residential buildings using LIDAR data and multispectral imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(5):457–467, 2010.

[5] Orsan Aytekin, Ilkay Ulusoy, Esra Zeynep Abacioglu, and Erhan Gokcay. Building detection in high resolution remotely sensed images based on morphological operators. In *2009 4th International Conference on Recent Advances in Space Technologies*, pages 376–379. IEEE, 2009.

[6] Mariana Belgiu and Lucian Drăguţ. Comparing supervised and unsupervised multiresolution segmentation approaches for extracting buildings from very high resolution imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 96:67–75, 2014.

[7] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The Lovász-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union mea-

sure in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4413–4421, 2018.

[8] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[9] Derrick Bonafilia, James Gill, Saikat Basu, and David Yang. Building high resolution maps for humanitarian aid and development with weakly-and semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1–9, 2019.

[10] Chin Wei Bong, Choong-Chin Liew, and Hong Yoong Lam. Ground-glass opacity nodules detection and segmentation using the snake model. In *Bio-Inspired Computation and Applications in Image Processing*, pages 87–104. Elsevier, 2016.

[11] Yuwei Cai, Hongjie He, Ke Yang, Sarah Narges Fatholahi, Lingfei Ma, Linlin Xu, and Jonathan Li. A comparative study of deep learning approaches to rooftop detection in aerial images. *Canadian Journal of Remote Sensing*, 47(3):413–431, 2021.

[12] Guo Cao and Xin Yang. Man-made object detection in aerial images using multistage level set evolution. *International Journal of Remote Sensing*, 28(8):1747–1757, 2007.

[13] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. GCNet: Non-local networks meet squeeze-excitation networks and beyond. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1971–1980. IEEE Computer Society, 2019.

[14] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via subcategory exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8991–9000, 2020.

[15] Bodhiswatta Chatterjee and Charalambos Poullis. On building classification from remote sensor imagery using deep neural networks and the relation between classification and reconstruction accuracy using border localization as proxy. In *2019 16th Conference on Computer and Robot Vision (CRV)*, pages 41–48. IEEE, 2019.

[16] Hui Chen, Liang Cheng, Qizhi Zhuang, Ka Zhang, Ning Li, Lei Liu, and Zhixin Duan. Structure-aware weakly supervised network for building extraction from remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2022.

[17] Jie Chen, Fen He, Yi Zhang, Geng Sun, and Min Deng. SPMF-Net: Weakly supervised building segmentation by combining superpixel pooling and multi-scale feature fusion. *Remote Sensing*, 12(6):1049, 2020.

[18] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.

[19] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5659–5667, 2017.

[20] Mengge Chen. Building detection from very high resolution remotely sensed imagery using deep neural networks. Master's thesis, University of Waterloo, 2019.

[21] Qi Chen, Lei Wang, Yifan Wu, Guangming Wu, Zhiling Guo, and Steven L. Waslander. Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings. *arXiv preprint arXiv:1807.09532*, 2018.

[22] Renxi Chen, Xinhui Li, and Jonathan Li. Object-based features for house detection from RGB high-resolution images. *Remote Sensing*, 10(3):451, 2018.

[23] Yukang Chen, Peizhen Zhang, Zeming Li, Yanwei Li, Xiangyu Zhang, Lu Qi, Jian Sun, and Jiaya Jia. Dynamic scale training for object detection. *arXiv preprint arXiv:2004.12432*, 2020.

[24] Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15334–15342, 2021.

[25] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5386–5395, 2020.

[26] Dominic Cheng, Renjie Liao, Sanja Fidler, and Raquel Urtasun. DARNet: Deep active ray network for building segmentation. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR), pages 7431–7439, 2019.

[27] Gong Cheng, Junyu Yang, Decheng Gao, Lei Guo, and Junwei Han. High-quality proposals for weakly supervised object detection. *IEEE Transactions on Image Processing*, 29:5794–5804, 2020.

[28] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1251–1258, 2017.

[29] Arie Croitoru and Yerach Doytsher. Monocular right-angle building hypothesis generation in regularized urban areas by pose clustering. *Photogrammetric Engineering & Remote Sensing*, 69(2):151–169, 2003.

[30] Shiyong Cui, Qin Yan, and Peter Reinartz. Graph search and its application in building extraction from high resolution remote sensing imagery. *Search Algorithms and Applications*, pages 133–150, 2011.

[31] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 764–773, 2017.

[32] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11065–11074, 2019.

[33] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. DeepGlobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops(CVPRW)*, pages 172–181, 2018.

[34] Barnali Dixon and Nivedita Candade. Multispectral landuse classification using neural networks and support vector machines: one or the other, or both? *International Journal of Remote Sensing*, 29(4):1185–1206, 2008.

[35] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2015.

[36] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[37] Shihong Du, Fangli Zhang, and Xiuyuan Zhang. Semantic classification of urban buildings combining VHR image and GIS data: An improved random forest approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 105:107–119, 2015.

[38] Grégory Duveiller and Pierre Defourny. A conceptual framework to define the spatial resolution requirements for agricultural monitoring using remote sensing. *Remote Sensing of Environment*, 114(11):2637–2650, 2010.

[39] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes Challenge: A retrospective. *International Journal of Computer Vision*, 111:98–136, 2015.

[40] Fang Fang, Daoyuan Zheng, Shengwen Li, Yuanyuan Liu, Linyun Zeng, Jiahui Zhang, and Bo Wan. Improved pseudomasks generation for weakly supervised building extraction from high-resolution remote sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:1629–1642, 2022.

[41] Jonas Franke, Michael Gebreslasie, Ides Bauwens, Julie Deleu, and Florian Siegert. Earth observation in support of malaria control and epidemiology: MALAREO monitoring approaches. *Geospatial health*, 10(1), 2015.

[42] Kun Fu, Wanxuan Lu, Wenhui Diao, Menglong Yan, Hao Sun, Yi Zhang, and Xian Sun. WSF-NET: Weakly supervised feature-fusion network for binary segmentation in remote sensing image. *Remote Sensing*, 10(12):1970, 2018.

[43] Syed Ali Naqi Gilani, Mohammad Awrangjeb, and Guojun Lu. An automatic building extraction and regularisation technique using lidar point cloud data and orthoimage. *Remote Sensing*, 8(3):258, 2016.

[44] Nicolas Girard, Dmitriy Smirnov, Justin Solomon, and Yuliya Tarabalka. Polygonal building extraction by frame field learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5891–5900, 2021.

[45] Haonan Guo, Bo Du, Liangpei Zhang, and Xin Su. A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 183:240–252, 2022.

[46] Haonan Guo, Xin Su, Shengkun Tang, Bo Du, and Liangpei Zhang. Scale-robust deep-supervision network for mapping building footprints from high-resolution remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:10091–10100, 2021.

[47] Xiaohui Guo. A novel multi to single module for small object detection. *arXiv preprint arXiv:2303.14977*, 2023.

[48] Shir Gur, Tal Shaharabany, and Lior Wolf. End to end trainable active contours via differentiable rendering. *arXiv preprint arXiv:1912.00367*, 2019.

[49] Ali Hatamizadeh, Debleena Sengupta, and Demetri Terzopoulos. End-to-end trainable deep active contour models for automated image segmentation: Delineating buildings in aerial imagery. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 730–746. Springer, 2020.

[50] Hongjie He, Kyle Gao, Weikai Tan, Lanying Wang, Nan Chen, Lingfei Ma, and Jonathan Li. Super-resolving and composing building dataset using a momentum spatial-channel attention residual feature aggregation network. *International Journal of Applied Earth Observation and Geoinformation*, 111:102826, 2022.

[51] Hongjie He, Zijian Jiang, Kyle Gao, Sarah Narges Fatholahi, Weikai Tan, Bingxu Hu, Hongzhang Xu, Michael A. Chapman, and Jonathan Li. Waterloo building dataset: A city-scale vector building dataset for mapping building footprints using aerial orthoimagery. *Geomatica*, 75(3):99–115, 2022.

[52] Hongjie He, Hongzhang Xu, Ying Zhang, Kyle Gao, Huxiong Li, Lingfei Ma, and Jonathan Li. Mask R-CNN based automated identification and extraction of oil well sites. *International Journal of Applied Earth Observation and Geoinformation*, 112:102875, 2022.

[53] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017.

[54] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[55] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 558–567, 2019.

[56] Francisco Herrera, Sebastián Ventura, Rafael Bello, Chris Cornelis, Amelia Zafra, Dánel Sánchez-Tarragó, Sarah Vluymans, Francisco Herrera, Sebastián Ventura, Rafael Bello, et al. *Multiple instance learning*. Springer, 2016.

[57] Yuanduo Hong, Huihui Pan, Weichao Sun, and Yisong Jia. Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. *arXiv preprint arXiv:2101.06085*, 2021.

[58] Cheng-Chun Hsu, Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, and Yung-Yu Chuang. Weakly supervised instance segmentation using the bounding box tightness prior. *Proceedings of the 33th International Conference on Neural Information Processing Systems (NeurIPS)*, 32, 2019.

[59] Jianfeng Huang, Xinchang Zhang, Qinchuan Xin, Ying Sun, and Pengcheng Zhang. Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 151:91–105, 2019.

[60] Xin Huang and Liangpei Zhang. Morphological building/shadow index for building extraction from high-resolution imagery over urban areas. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(1):161–172, 2011.

[61] Xin Huang and Liangpei Zhang. A multidirectional and multiscale morphological index for automatic building extraction from multispectral geoeye-1 imagery. *Photogrammetric Engineering & Remote Sensing*, 77(7):721–732, 2011.

[62] Andres Huertas and Ramakant Nevatia. Detecting buildings in aerial images. *Computer Vision, Graphics, and Image Processing*, 41(2):131–152, 1988.

[63] Jordi Inglada. Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features. *ISPRS Journal of Photogrammetry and Remote Sensing*, 62(3):236–248, 2007.

[64] R Bruce Irvin and David M McKeown. Methods for exploiting the relationship between buildings and their shadows in aerial imagery. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(6):1564–1575, 1989.

[65] Mohammad Izadi and Parvaneh Saeedi. Three-dimensional polygonal building model estimation from single satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, 50(6):2254–2272, 2011.

[66] Shunping Ji, Shiqing Wei, and Meng Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1):574–586, 2018.

[67] Brian Johnson and Zhixiao Xie. Classifying a high resolution image of an urban area using super-object information. *ISPRS Journal of Photogrammetry and Remote Sensing*, 83:40–49, 2013.

[68] Konstantinos Karantzalos and Nikos Paragios. Recognition-driven two-dimensional competing priors toward automatic and accurate building detection. *IEEE Transactions on Geoscience and Remote Sensing*, 47(1):133–144, 2008.

[69] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

[70] Antonis Katartzis and Hichem Sahli. A stochastic framework for the identification of building rooftops using a single remote sensing image. *IEEE Transactions on Geoscience and Remote Sensing*, 46(1):259–271, 2007.

[71] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1646–1654, 2016.

[72] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1637–1645, 2016.

[73] Taejung Kim and Jan-Peter Muller. Development of a graph-based approach for building detection. *Image and Vision Computing*, 17(1):3–14, 1999.

[74] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[75] Santhana Krishnamachari and Rama Chellappa. Delineating buildings by grouping lines with mrfs. *IEEE Transactions on Image Processing*, 5(1):164–168, 1996.

[76] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90, 2012.

[77] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 624–632, 2017.

[78] John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. MSeg: A composite dataset for multi-domain semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2879–2888, 2020.

[79] Shiyi Lan, Zhiding Yu, Christopher Choy, Subhashree Radhakrishnan, Guilin Liu, Yuke Zhu, Larry S Davis, and Anima Anandkumar. DiscoBox: Weakly supervised instance segmentation and semantic correspondence from box supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3406–3416, 2021.

[80] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4681–4690, 2017.

[81] D Scott Lee, Jie Shan, and James S Bethel. Class-guided building extraction from Ikonos imagery. *Photogrammetric Engineering & Remote Sensing*, 69(2):143–150, 2003.

[82] Dong Hyuk Lee, Kyoung Mu Lee, and Sang Uk Lee. Fusion of lidar and imagery for reliable building extraction. *Photogrammetric Engineering & Remote Sensing*, 74(2):215–225, 2008.

[83] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4071–4080, 2021.

[84] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. BBAM: Bounding box attribution map for weakly supervised semantic and instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2643–2652, 2021.

[85] Sébastien Lefèvre, Jonathan Weber, and David Sheeren. Automatic building extraction in VHR images using advanced morphological operators. In *2007 Urban Remote Sensing Joint Event*, pages 1–5. IEEE, 2007.

[86] Er Li, John Femiani, Shibiao Xu, Xiaopeng Zhang, and Peter Wonka. Robust rooftop extraction from visible band images using higher order CRF. *IEEE Transactions on Geoscience and Remote Sensing*, 53(8):4483–4495, 2015.

[87] Weijia Li, Wenqian Zhao, Huaping Zhong, Conghui He, and Dahua Lin. Joint semantic-geometric learning for polygonal building segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 1958–1965, 2021.

[88] Wentong Li, Wenyu Liu, Jianke Zhu, Miaomiao Cui, Risheng Yu, Xiansheng Hua, and Lei Zhang. Box2Mask: Box-supervised instance segmentation via level-set evolution. *arXiv preprint arXiv:2212.01579*, 2022.

[89] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-aware trident networks for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6054–6063, 2019.

[90] Zuoyue Li, Jan Dirk Wegner, and Aurélien Lucchi. Topological map extraction from overhead images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1715–1724, 2019.

[91] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 136–144, 2017.

[92] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2125, 2017.

[93] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects

in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755, 2014.

[94] Yuh-Tay Liow and Theo Pavlidis. Use of shadows for extracting buildings in aerial images. *Computer Vision, Graphics, and Image Processing*, 49(2):242–277, 1990.

[95] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2359–2368, 2020.

[96] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8759–8768, 2018.

[97] Tang Liu, Ling Yao, Jun Qin, Ning Lu, Hou Jiang, Fan Zhang, and Chenghu Zhou. Multi-scale attention integrated hierarchical networks for high-resolution building footprint extraction. *International Journal of Applied Earth Observation and Geoinformation*, 109:102768, 2022.

[98] Yuanyuan Liu, Dingyuan Chen, Ailong Ma, Yanfei Zhong, Fang Fang, and Kai Xu. Multiscale U-shaped CNN building instance extraction framework with edge constraint for high-spatial-resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7):6106–6120, 2020.

[99] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

[100] Zhuang Liu, Zhiqiu Xu, Joseph Jin, Zhiqiang Shen, and Trevor Darrell. Dropout reduces underfitting. *arXiv preprint arXiv:2303.01500*, 2023.

[101] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.

[102] Zhanming Ma, Min Xia, Liguo Weng, and Haifeng Lin. Local feature search network for building and water segmentation of remote sensing image. *Sustainability*, 15(4):3034, 2023.

[103] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3226–3229, 2017.

[104] Jisan Mahmud, True Price, Akash Bapat, and Jan-Michael Frahm. Boundary-aware 3D building reconstruction from a single overhead image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 441–451, 2020.

[105] Xiao-Jiao Mao, Chunhua Shen, and Yubin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 2802–2810, 2016.

[106] Diego Marcos, Devis Tuia, Benjamin Kellenberger, Lisa Zhang, Min Bai, Renjie Liao, and Raquel Urtasun. Learning deep structured active contours end-to-end. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8877–8885, 2018.

[107] Iman Marivani, Evaggelia Tsiligianni, Bruno Cornelis, and Nikos Deligiannis. Joint image super-resolution via recurrent convolutional neural networks with coupled sparse priors. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 868–872. IEEE, 2020.

[108] McGlone and Shufelt. Projective and object space geometry for monocular building extraction. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 54–61. IEEE, 1994.

[109] Volodymyr Mnih. *Machine learning for aerial image labeling*. University of Toronto (Canada), 2013.

[110] Sharada Prasanna Mohanty. CrowdAI Mapping Challenge 2018 : Baseline with Mask-RCNN. https://github.com/crowdai/crowdai-mapping-challenge-mask-rcnn, 2018.

[111] Yousif Abdul-kadhim Mousa. *Building footprint extraction from LiDAR data and imagery information*. PhD thesis, Curtin University, 2020.

[112] Youngmin Oh, Beomjun Kim, and Bumsub Ham. Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 6913–6922, 2021.

[113] Ali Ozgun Ok. Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts. *ISPRS Journal of Photogrammetry and Remote Sensing*, 86:21–40, 2013.

[114] Ali Ozgun Ok, Caglar Senaras, and Baris Yuksel. Automated detection of arbitrarily shaped buildings in complex environments from monocular VHR optical satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 51(3):1701–1717, 2012.

[115] Zhiyi Pan, Peng Jiang, Yunhai Wang, Changhe Tu, and Anthony G Cohn. Scribble-supervised semantic segmentation by uncertainty reduction on neural representation and self-supervision on neural eigenspace. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7416–7425, 2021.

[116] Hui En Pang and Filip Biljecki. 3D building reconstruction from single street view images using deep learning. *International Journal of Applied Earth Observation and Geoinformation*, 112:102859, 2022.

[117] Patrick Aravena Pelizari, Christian Geiß, Paula Aguirre, Hernán Santa María, Yvonne Merino Peña, and Hannes Taubenböck. Automated building characterization for seismic risk assessment using street-level imagery and deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 180:370–386, 2021.

[118] Martino Pesaresi and Jon Atli Benediktsson. A new approach for the morphological segmentation of high-resolution satellite imagery. *IEEE transactions on Geoscience and Remote Sensing*, 39(2):309–320, 2001.

[119] Zheng Qin, Zhaoning Zhang, Dongsheng Li, Yiming Zhang, and Yuxing Peng. Diagonalwise refactorization: An efficient training method for depthwise convolutions. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.

[120] Linhao Qu, Siyu Liu, Xiaoyu Liu, Manning Wang, and Zhijian Song. Towards label-efficient automatic diagnosis and analysis: a comprehensive survey of advanced deep learning-based weakly-supervised, semi-supervised and self-supervised techniques in histopathological image analysis. *Physics in Medicine & Biology*, 2022.

[121] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Proceedings of the 35st International Conference on Neural Information Processing Systems (NeurIPS)*, pages 12116–12128, 2021.

[122] Kriti Rastogi, Pankaj Bodani, and Shashikant A Sharma. Automatic building footprint extraction from very high-resolution imagery using deep learning techniques. *Geocarto International*, 37(5):1501–1513, 2022.

[123] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015.

[124] Ribana Roscher, Michele Volpi, Clément Mallet, Lukas Drees, and Jan Dirk Wegner. SemCity Toulouse: A benchmark for building instance segmentation in satellite images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 5:109–116, 2020.

[125] Franz Rottensteiner, Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, Sebastien Benitez, and Uwe Breitkopf. The ISPRS benchmark on urban object classification and 3D building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences; I-3*, 1(1):293–298, 2012.

[126] Lixiang Ru, Heliang Zheng, Yibing Zhan, and Bo Du. Token contrast for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR*, pages 3093–3102, 2023.

[127] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, 30, 2017.

[128] Liora Sahar, Subrahmanyam Muthukumar, and Steven P French. Using aerial imagery and GIS in automated building footprint extraction and shape recognition for earthquake risk assessment of urban inventories. *IEEE Transactions on Geoscience and Remote Sensing*, 48(9):3511–3520, 2010.

[129] Michael E Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Momentum residual neural networks. In *Proceedings of International Conference on Machine Learning*, pages 9276–9287. PMLR, 2021.

[130] Caglar Senaras, Mete Ozay, and Fatos T Yarman Vural. Building detection with decision fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(3):1295–1304, 2013.

[131] Aaron K Shackelford and Curt H Davis. A combined fuzzy pixel-based and object-based approach for classification of high-resolution multispectral data over urban areas. *IEEE Transactions on GeoScience and Remote sensing*, 41(10):2354–2363, 2003.

[132] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, 2016.

[133] Yilei Shi, Qingyu Li, and Xiao Xiang Zhu. Building segmentation through a gated graph convolutional neural network with deep structured feature embedding. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:184–197, 2020.

[134] Yuanming Shu. *Deep convolutional neural networks for object extraction from high spatial resolution remotely sensed imagery*. PhD thesis, University of Waterloo, 2014.

[135] Bharat Singh and Larry S Davis. An analysis of scale invariance in object detection-SNIP. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3578–3587, 2018.

[136] Bharat Singh, Mahyar Najibi, and Larry S Davis. SNIPER: Efficient multi-scale training. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS)*, page 9333–9343, 2018.

[137] Beril Sirmacek and Cem Unsalan. Urban-area and building detection using SIFT keypoints and graph theory. *IEEE Transactions on Geoscience and Remote Sensing*, 47(4):1156–1167, 2009.

[138] Gunho Sohn and Ian Dowman. Data fusion of high-resolution satellite imagery and LiDAR data for automatic building extraction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 62(1):43–63, 2007.

[139] Emre Sumer and Mustafa Turker. An adaptive fuzzy-genetic algorithm approach for building detection using high-resolution satellite images. *Computers, Environment and Urban Systems*, 39:48–62, 2013.

[140] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5693–5703, 2019.

[141] Ying Sun, Xinchang Zhang, Xiaoyang Zhao, and Qinchuan Xin. Extracting building boundaries from high resolution optical images and LiDAR data by integrating the convolutional neural network and the active contour model. *Remote Sensing*, 10(9):1459, 2018.

[142] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3147–3155, 2017.

[143] Yuki Tatsunami and Masato Taki. Sequencer: Deep LSTM for image classification. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 38204–38217, 2022.

[144] Jim Thomas, Ahsan Kareem, and Kevin W Bowyer. Automated poststorm damage classification of low-rise building roofing systems using high-resolution aerial imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 52(7):3851–3861, 2013.

[145] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 282–298. Springer, 2020.

[146] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4799–4807, 2017.

[147] Cem Ünsalan and Kim L Boyer. A system to detect houses and residential street networks in multispectral satellite images. *Computer Vision and Image Understanding*, 98(3):423–461, 2005.

[148] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018.

[149] Guanchun Wang, Xiangrong Zhang, Zelin Peng, Xiuping Jia, Xu Tang, and Licheng Jiao. MOL: Towards accurate weakly supervised remote sensing object detection via Multi-view nOisy Learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196:457–470, 2023.

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

[151] Jun Wang, Xiucheng Yang, Xuebin Qin, Xin Ye, and Qiming Qin. An efficient approach for automatic rectangular building extraction from very high resolution optical satellite imagery. *IEEE Geoscience and Remote Sensing Letters*, 12(3):487–491, 2014.

[152] Ruisheng Wang. 3D building modeling using images and LiDAR: A review. *International Journal of Image and Data Fusion*, 4(4):273–292, 2013.

[153] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, Xiaogang Wang, and Yu Qiao. InternImage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14408–14419, 2023.

[154] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. PVT v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022.

[155] Xinggang Wang, Jiapei Feng, Bin Hu, Qi Ding, Longjin Ran, Xiaoxin Chen, and Wenyu Liu. Weakly-supervised instance segmentation via class-agnostic learning with salient images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10225–10235, 2021.

[156] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12275–12284, 2020.

[157] Shiqing Wei and Shunping Ji. Graph convolutional networks for the automated production of building vector maps from aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2021.

[158] Shiqing Wei, Shunping Ji, and Meng Lu. Toward automatic building footprint delineation from aerial images using CNN and regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 58(3):2178–2189, 2019.

[159] Yao Wei and Shunping Ji. Scribble-based weakly supervised deep learning for road surface extraction from remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2021.

[160] Qi Wen, Kaiyu Jiang, Wei Wang, Qingjie Liu, Qing Guo, Lingling Li, and Ping Wang. Automatic building extraction from Google Earth images under complex backgrounds based on deep instance segmentation network. *Sensors*, 19(2):333, 2019.

[161] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNeXt V2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16133–16142, 2023.

[162] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[163] Yifan Wu, Linlin Xu, Yuhao Chen, Alexander Wong, and David A Clausi. TAL: Topography-aware multi-resolution fusion learning for enhanced building footprint extraction. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

[164] Jing Xiao, Markus Gerke, and George Vosselman. Building extraction from oblique airborne imagery based on robust façade detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 68:56–68, 2012.

[165] Yanhua Xie, Anthea Weng, and Qihao Weng. Population estimation of urban residential communities using remotely sensed morphologic data. *IEEE Geoscience and Remote Sensing Letters*, 12(5):1111–1115, 2015.

[166] Bowen Xu, Jiakun Xu, Nan Xue, and Gui-Song Xia. HiSup: Accurate polygonal mapping of buildings in satellite imagery with hierarchical supervision. *ISPRS Journal of Photogrammetry and Remote Sensing*, 198:284–296, 2023.

[167] Nan Xue, Song Bai, Fudong Wang, Gui-Song Xia, Tianfu Wu, and Liangpei Zhang. Learning attraction field representation for robust line segment detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1595–1603, 2019.

[168] Xin Yan, Li Shen, Jicheng Wang, Xu Deng, and Zhilin Li. MSG-SR-Net: A weakly supervised network integrating multiscale generation and superpixel refinement for building extraction from high-resolution remotely sensed imageries. *IEEE Journal of*

Selected Topics in Applied Earth Observations and Remote Sensing, 15:1012–1023, 2021.

[169] Xin Yan, Li Shen, Jicheng Wang, Yong Wang, Zhilin Li, and Zhu Xu. PANet: Pixel-wise affinity network for weakly supervised building extraction from high-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

[170] Chih-Yuan Yang, Chao Ma, and Ming-Hsuan Yang. Single-image super-resolution: A benchmark. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 372–386. Springer, 2014.

[171] Ke Yang, Dongsheng Li, and Yong Dou. Towards precise end-to-end weakly supervised object detection network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8372–8381, 2019.

[172] Rui Yang, Lin Song, Yixiao Ge, and Xiu Li. BoxSnake: Polygonal instance segmentation with box supervision. *arXiv preprint arXiv:2303.11630*, 2023.

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

[174] Yongtao Yu, Yongfeng Ren, Haiyan Guan, Dilong Li, Changhui Yu, Shenghua Jin, and Lanfang Wang. Capsule feature pyramid network for building footprint extraction from high-resolution aerial imagery. *IEEE Geoscience and Remote Sensing Letters*, 18(5):895–899, 2020.

[175] Jiangye Yuan. Learning building extraction in aerial scenes with convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2793–2798, 2017.

[176] Jiangye Yuan and Anil M Cheriyadat. Learning to count buildings in diverse aerial scenes. In *Proceedings of the 22nd ACM Special Interest Group on Spatial Information (SIGSPATIAL) International Conference on Advances in Geographic Information Systems*, pages 271–280, 2014.

[177] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. HRFormer: High-resolution vision transformer for dense predict. In *Proceedings of the 35th International Conference on Neural Information Processing Systems(NeurIPS)*, pages 7281–7293, 2021.

[178] Jiabin Zhang, Hu Su, Yonghao He, and Wei Zou. Weakly supervised instance segmentation via category-aware centerness learning with localization supervision. *Pattern Recognition*, 136:109165, 2023.

[179] Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-supervised salient object detection via scribble annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12546–12555, 2020.

[180] Jun Zhang, Yue Liu, Pengfei Wu, Zhenwei Shi, and Bin Pan. Mining cross-domain structure affinity for refined building segmentation in weakly supervised constraints. *Remote Sensing*, 14(5):1227, 2022.

[181] Qian Zhang, Xin Huang, and Guixu Zhang. A morphological building detection framework for high-resolution optical imagery over urban areas. *IEEE Geoscience and Remote Sensing Letters*, 13(9):1388–1392, 2016.

[182] Xiaopeng Zhang, Jiashi Feng, Hongkai Xiong, and Qi Tian. Zigzag learning for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4262–4270, 2018.

[183] Xueting Zhang, Wei Huang, Qi Wang, and Xuelong Li. SSR-NET: Spatial–spectral reconstruction network for hyperspectral and multispectral image fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7):5953–5965, 2020.

[184] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018.

[185] Yun Zhang. Optimisation of building detection in satellite images by combining multispectral classification and texture filtering. *ISPRS Journal of Photogrammetry and Remote Sensing*, 54(1):50–60, 1999.

[186] Yunbo Zhang, Pengfei Yi, Dongsheng Zhou, Xin Yang, Deyun Yang, Qiang Zhang, and Xiaopeng Wei. CSANet: Channel and spatial mixed attention cnn for pedestrian detection. *IEEE Access*, 8:76243–76252, 2020.

[187] Kang Zhao, Jungwon Kang, Jaewook Jung, and Gunho Sohn. Building extraction from satellite images using Mask R-CNN with building boundary regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 247–251, 2018.

[188] Peng Zhao, Jindi Zhang, Weijia Fang, and Shuiguang Deng. SCAU-Net: Spatial-channel attention U-Net for gland segmentation. *Frontiers in Bioengineering and Biotechnology*, 8:670, 2020.

[189] Wufan Zhao, Claudio Persello, and Alfred Stein. Building outline delineation: From aerial images to polygons with an improved end-to-end learning framework. *ISPRS Journal of Photogrammetry and Remote Sensing*, 175:119–131, 2021.

[190] Daoyuan Zheng, Shengwen Li, Fang Fang, Jiahui Zhang, Yuting Feng, Bo Wan, and Yuanyuan Liu. Utilizing bounding box annotations for weakly supervised building extraction from remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[191] Yanfei Zhong, Ailong Ma, Yew soon Ong, Zexuan Zhu, and Liangpei Zhang. Computational intelligence in optical remote sensing image processing. *Applied Soft Computing*, 64:75–93, 2018.

[192] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016.

[193] Dongzhan Zhou, Xinchi Zhou, Hongwen Zhang, Shuai Yi, and Wanli Ouyang. Cheaper pre-training lunch: An efficient paradigm for object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 258–274, 2020.

[194] Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Weakly supervised instance segmentation using class peak response. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3791–3800, 2018.

[195] Hegui Zhu, Tian Geng, Jiayi Wang, Qingsong Tang, and Wuming Jiang. Improved sub-category exploration and attention hybrid network for weakly supervised semantic segmentation. *Neural Computing and Applications*, pages 1–15, 2023.

[196] Qing Zhu, Cheng Liao, Han Hu, Xiaoming Mei, and Haifeng Li. MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7):6169–6181, 2020.

[197] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017.

[198] Yi Zhu, Yanzhao Zhou, Huijuan Xu, Qixiang Ye, David Doermann, and Jianbin Jiao. Learning instance activation maps for weakly supervised instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3116–3125, 2019.

[199] Stefano Zorzi, Shabab Bazrafkan, Stefan Habenschuss, and Friedrich Fraundorfer. PolyWorld: Polygonal building extraction with graph neural networks in satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1848–1857, 2022.

[200] Stefano Zorzi, Ksenia Bittner, and Friedrich Fraundorfer. Machine-learned regularization and polygonization of building segmentation masks. In *Proceedings of International Conference on Pattern Recognition (ICPR)*, pages 3098–3105, 2021.

[201] Stefano Zorzi and Friedrich Fraundorfer. Regularization of building boundaries in satellite images using adversarial and regularized losses. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 5140–5143. IEEE, 2019.

APPENDICES

Appendix A

Evaluation Metrics used for Evaluating Deep Learning Models

A.1 Pixel-level Metrics

The metrics derived from the confusion matrix may not be suitable for evaluating rooftop delineation methods using high spatial and very high spatial resolution imagery [134]. Previous studies have commonly used metrics such as IoU, mIoU, Precision, Recall, F1-score, and accuracy to evaluate the performance of rooftop delineation methods. Specifically, IoU represents the percentage of the overlap between ground truth and the prediction output. mIoU represents the average of positive objects IoU and negative objects IoU. Precision indicates how many predicted positive objects are correct compared to all predicted positive objects. Recall shows how many positive objects are predicted accurately compared to all positive objects from the ground truth. F1 score or F1 measure is the harmonic mean of precision and recall. The accuracy or average accuracy calculates the percentage of correctly classified pixels or other objects in all images. All these metrics are defined as follows.

$$IoU = \frac{TP}{TP + FP + FN} \tag{A.1}$$

$$mIoU = \frac{1}{2}\left(\frac{TP}{TP + FP + FN} + \frac{TN}{TN + FP + FN}\right) \tag{A.2}$$

$$Precision = \frac{TP}{TP + FP} \tag{A.3}$$

$$Recall = \frac{TP}{TP + FN} \tag{A.4}$$

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{A.5}$$

$$Pixel\,(Overall)\,Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{A.6}$$

where, TP indicates true positive, denoting a correct prediction of the positive class (presence of building); FP refers to False Positive which occurs when the model predicts positive class as negative class; FN stands for false negative (FN) in which the model classified positive class into negative class; and TN represents true negative in which the model predicted the negative class correctly in the output.

A.2 Object-level Metrics

The currently used object-level metrics were originally designed to evaluate methods for object detection and instance segmentation using the MS COCO dataset [93]. These metrics include Average Precision (AP) and Average Recall (AR) calculated at different IoU thresholds and for different object sizes. The subscripts following "AP" or "AR" indicate the averaged AP or AR value at specific IoU thresholds, such as 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90 and 0.95. For example, AP_{75} and AP_{50} represent the AP value at IoU thresholds of 0.75 and 0.50, respectively. The metrics are further divided into categories based on the size of the objects: 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). These categories represent the average AP or AR values for small, medium and large-sized buildings, defined based on their pixel sizes, with small buildings being smaller than 32×32 pixels, medium buildings between 32×32 and 96×96, and large buildings larger than 96×96 pixels. To illustrate the calculation of AP_{75}, a fictitious example is provided below.

With an IoU threshold of 0.75, 10 masks are detected for 6 objects. These masks can be sorted based on the predicted scores, ranging from the highest score to the lowest score. By varying the classification score threshold, the precision and recall can be recalculated at the object level, as shown in Table A.1. During the calculation, the predicted masks that overlap with the ground truth by more than 75% ($IoU > 0.75$) and have classification scores higher than the threshold are considered as "building", while the rest are labeled as "Other". Based on the values in Table A.1, the precision-recall curve can be plotted as shown in Figure A.1.

Table A.1: Precision and recall along with different prediction scores

Scores Rank	Ground Truth	Precision	Recall
1	Building	1.00	0.17
2	Building	0.50	0.17
3	Building	0.67	0.33
4	Building	0.75	0.50
5	Building	0.60	0.50
6	Building	0.50	0.50
7	Building	0.57	0.67
8	Building	0.63	0.83
9	Building	0.67	1.00
10	Other	0.60	1.00

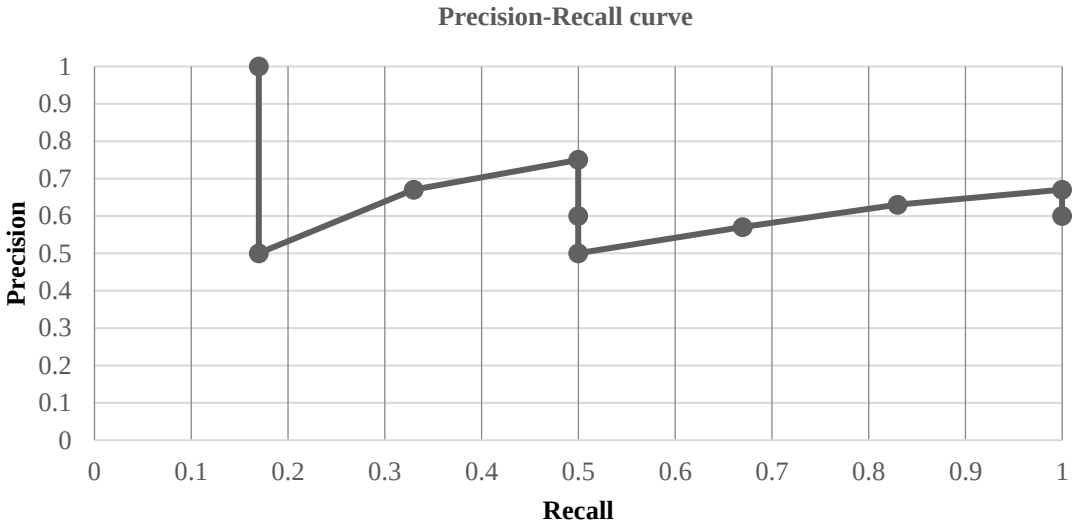

Figure A.1: The precision-recall curve of the fictious example

AP_{75} is calculated as the area under to the precision-recall curve. In the experiment, the equations used by the Pascal Visual Object Classes (VOC) 2010 [39] is adopted, which can be calculated as follows:

$$AP = \sum_{1 \leq i \leq n} (r_i - r_{i-1}) \cdot p_i \tag{A.7}$$

where r_i and p_i are recall and precision values under a certain classification score. r_1,r_n is the smallest and the largest recall value. If two precision values match one recall value, the larger precision will be preserved for the curve. In the example, $AP_{75} = (1-0.83) \times 0.67 + (0.83-0.67) \times 0.63 + (0.67-0.50) \times 0.57 + (0.5-0.333) \times 0.75 + (0.33-0.17) \times 0.67 = 0.55$. The AP values with other IoU thresholds can be calculated in the same way.

Appendix B

My Publications during the PhD Study

Hongjie He, Lingfei Ma, and Jonathan Li. Box2Boundary: Delineation of rooftops with box supervision. IEEE Transactions on Geoscience and Remote Sensing, submitted.

Hongjie He, Lingfei Ma, and Jonathan Li. HigherNet-DST: Higher resolution network with dynamic scale training for rooftop delineation. IEEE Transactions on Geoscience and Remote Sensing, under revision.

Hongjie He, Kyle Gao, Weikai Tan, Lanying Wang, Nan Chen, Lingfei Ma, and Jonathan Li. Super-resolving and composing building dataset using a momentum spatial-channel attention residual feature aggregation network. International Journal of Applied Earth Observation and Geoinformation, 111:102826, 2022.

Hongjie He, Hongzhang Xu, Ying Zhang, Kyle Gao, Huxiong Li, Lingfei Ma, and Jonathan Li. Mask R-CNN based automated identification and extraction of oil well sites. International Journal of Applied Earth Observation and Geoinformation, 112:102875, 2022.

Hongjie He, Zijian Jiang, Kyle Gao, Sarah Narges Fatholahi, Weikai Tan, Bingxu Hu, Hongzhang Xu, Michael A. Chapman, and Jonathan Li. Waterloo building dataset: a city-scale vector building dataset for mapping building footprints using aerial orthoimagery. Geomatica, 75(3):99–115, 2022.

Hongjie He, Ke Yang, Shusen Wang, Hasti Andon Petrosians, Ming Liu, Junhua Li, Jose Marcato Junior, Wesley Nunes Goncalves, Lanying Wang, and Jonathan Li. Deep learning approaches to spatial downscaling of GRACE terrestrial water storage products

using EALCO model over Canada. Canadian Journal of Remote Sensing, 47(4):657–675, 2021.

Hongjie He, Ke Yang, Yuwei Cai, Zijian Jiang, Qiutong Yu, Kun Zhao, Junbo Wang, Sarah Narges Fatholahi, Yan Liu, Hasti Andon Petrosians, Bingxu Hu, Liyuan Qing, Zhehan Zhang, Hongzhang Xu, Siyu Li, Kyle Gao, Linlin Xu, and Jonathan Li. The impact of data volume on performance of depp learning based building rooftop extraction using very high spatial resolution aerial images. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium(IGARSS), pages 1343–1346. IEEE, 2021.

Hongjie He, Hongzhang Xu, Ying Zhang, Michael A. Chapman, Yiping Chen, and Jonathan Li. Automated detection of oil/gas well sites detection from multi-source high spatial resolution images. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium(IGARSS), pages 4615–4618. IEEE, 2022.

Hongjie He, Kyle Gao, Weikai Tan, Lanying Wang, Sarah Narges Fatholahi, Nan Chen, Michael A. Chapman, and Jonathan Li. Impact of deep learning-based super-resolution on building footprint extraction. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 43:31–37, 2022.

Yuwei Cai, **Hongjie He**, Ke Yang, Sarah Narges Fatholahi, Lingfei Ma, Linlin Xu, and Jonathan Li. A comparative study of deep learning approaches to rooftop detection in aerial images. Canadian Journal of Remote Sensing, 47(3):413–431, 2021.

Qiutong Yu, Shusen Wang, **Hongjie He**, Ke Yang, Lingfei Ma, and Jonathan Li. Reconstructing GRACE-like TWS anomalies for the Canadian landmass using deep learning and land surface model. International Journal of Applied Earth Observation and Geoinformation, 102:102404, 2021.

Nan Chen, Lichun Sui, Biao Zhang, **Hongjie He**, Kyle Gao, Yandong Li, José Marcato Junior, and Jonathan Li. Fusion of hyperspectral-multispectral images joining spatial-spectral dual-dictionary and structured sparse low-rank representation. International Journal of Applied Earth Observation and Geoinformation, 104:102570, 2021.

Nan Chen, Lichun Sui, Biao Zhang, **Hongjie He**, José Marcato Junior, and Jonathan Li. Single satellite imagery superresolution based on hybrid nonlocal similarity constrained convolution sparse coding. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14:7489–7505, 2020.

Kyle Gao, Yina Gao, **Hongjie He**, Denning Lu, Linlin Xu, and Jonathan Li. NeRF: Neural radiance field in 3D vision, a comprehensive review. arXiv preprint arXiv:2210.00379, 2022.

Kyle Gao, Mengge Chen, Sarah Narges Fatholahi, **Hongjie He**, Hongzhang Xu, José Marcato Junior, Wesley Nunes Gonçalves , Michael A. Chapman, and Jonathan Li. A

region-based deep learning approach to instance segmentation of aerial orthoimagery for building rooftop extraction. Geomatica, 75(1):148–164, 2022.

Kyle Gao, **Hongjie He**, Dening Lu, Linlin Xu, Lingfei Ma, and Jonathan Li. Optimizing and evaluating Swin transformer for aircraft classification: Analysis and generalizability of the MTARSI dataset. IEEE Access, 10:134427–134439, 2022.

Diogo Nunes Gonçalves, Mauro dos Santos de Arruda, Hemerson Pistori, Vanessa Jordão Marcato Fernandes, Ana Paula Marques Ramos, Danielle Elis Garcia Furuya, Lucas Prado Osco, **Hongjie He**, Jonathan Li, José Marcato Junior, Wesley Nunes Gonçalves. A deep learning approach based on graphs to detect plantation lines. arXiv preprint arXiv:2102.03213, 2021.

Zhimeng He, **Hongjie He**, Jonathan Li, Michael A. Chapman, and Haiyong Ding. A short-cut connections-based neural network for building extraction from high resolution orthoimagery. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 43:39–44, 2022.

Sarah Narges Fatholahi, **Hongjie He**, Lanying Wang, Awase Syed, and Jonathan Li. Monitoring surface deformation over oilfield using MT-Insar and production well data. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium(IGARSS), pages 2298–2301. IEEE, 2021

Lanying Wang, Weikai Tan, Hongzhang Xu, **Hongjie He**, Nan Chen, Dilong Li, Michael A. Chapman, and Jonathan Li. Deep learning-based method to extend the time series of global annual VIIRS-like nighttime light data. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 43:73–78, 2022

Appendix C

Waiver of Copyright

Elsevier, as the publisher of the two manuscripts fully or partly adopted in Chapter 2, Chapter 3 and Chapter 4 allow the reuse of published papers in the thesis without formal permissions. Thus, the waivers of copyright from Elsevier are achieved by the following statement:

Policy Regarding Thesis/Dissertation Reuse in Elsevier Copyright: "Authors can use their articles, in full or in part, for a wide range of scholarly, non-commercial purposes in a thesis or dissertation (provided that this is not to be published commercially)."