UNIVERSITÀ DEGLI STUDI
DI TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE
ICT International Doctoral School

Novel methods for the Semantic Segmentation of Remote Sensing Images

Lei Ding

Advisor

Prof. Lorenzo Bruzzone

Università degli Studi di Trento

May 2021

Acknowledgements

I am grateful for this PhD experience through which I gained not only research experience but also a vision of the career. Back to January 2018 upon the arrival of Italy, I was still confused about the future and was obsessed to temporary joys. It is through the bricking of experiments, the solving of scientific puzzles where I learned what interests me most.

I would like to sincerely thank my advisor Lorenzo Bruzzone. This PhD study in Trento is not possible without your help since the very beginning. You have inspired me in not only the research methodologies but also the general scientific principles. I am grateful for your lead, following which I have entered the academic world.

I would also like to thank my family for their understanding and supports. I never felt being along through this journey. Thank Jing for accompanying me in the final year. Special thanks to Enzo for coming into my life. Thanks also to my domestic advisors, especially Prof. Zhang, Prof. Guo and Prof. Lin who have offered me data and hardware supports.

Thanks to my colleagues, classmates and friends in Trento. Thank Massimo for fixing the hardware issues and answering my doubts every now and then. Thank Hao and Yahui for their helps to my experiments. Thank the RSLab Chinese squad members for the sharing of research experiences. Thank Paolo, Jianbo, Ruozhou and Kun for travelling and hiking with me.

Thank Trento for offering the mountains and lakes where I can release anxiety. May we meet again years later.

Abstract

With the development of Earth observation technologies, there is a tremendous increase in the volume of available remote sensing images (RSIs), and subsequently a growing need for the automatic analysis of the collected data. The pixel-wise classification, i.e., the semantic segmentation of RSIs, is important for a variety of land-cover and land-use mapping applications. Recent studies on the semantic segmentation of RSIs have achieved great progress with the use of Convolutional Neural Networks (CNNs). However, they suffer from some common problems such as fragmentation errors, boundary ambiguity, and the need for optimization of the results. In this thesis, we address these problems and propose methods to improve the segmentation accuracy in the context of i) The semantic segmentation of very high resolution (VHR) RSIs; ii) The semantic segmentation of High-Resolution (HR) Synthetic Aperture Radar (SAR) images; iii) The segmentation of roads in VHR RSIs; iv) The segmentation of buildings in VHR RSIs.

Through research activities conducted under these sub-topics, this dissertation presents four novel contributions.

First, we propose a Local Attention Network (LANet) for the semantic segmentation of VHR RSIs. Conventional CNN models extract features within a limited Receptive Field (RF) due to their local information aggregation mechanism. In the proposed LANet, we design a patch attention module to enhance the embedding of context information, as well as an attention embedding module to enrich the semantic information in low-level features. Experimental results show that these designs reduce fragmentation errors and improve segmentation accuracy.

Second, we present a novel CNN architecture for the semantic segmentation of HR SAR images. SAR images contain intense speckle noise which affects the segmentation algorithms. To alleviate its impact, we design a Multi-Path Residual Network (MPResNet) that contains three parallel feature embedding branches. Compared to other CNN architectures, it has wider RF, thus being able to exploit better the local discriminative features.

Third, we propose a Direction-aware Residual Network (DiResNet) for the

segmentation of roads in VHR RSIs. State-of-the-art methods for road segmentation suffer from discontinuity problems (affected by occlusions and redundant spatial information). In the DiResNet we introduce the supervision of road directions to improve the detection of linear features, as well as several auxiliary designs to improve the road structure and completeness. These lead to significant improvements in precision and connectivity of the results.
Last, we introduce an adversarial training strategy to model the shape information for building segmentation in VHR RSIs. Common CNNs cannot model the shape of objects of interest. We propose an Adversarial Shape Learning Network (ASLNet) to learn explicitly the shape constraints that data exhibit, which is beneficial for inpainting the missing building parts and regularizing the building contours. This approach improves the results in both pixel-based accuracy and object-based metrics.
The effectiveness of the proposed approaches has been tested with both ablation studies and comparative experiments on the corresponding benchmark datasets. The quantitative and qualitative results are presented together with a comprehensive performance analysis.

Keywords remote sensing, deep learning, semantic segmentation, road extraction, building extraction

Contents

List of Tables

List of Figures

List of Abbreviations

AEM Attention Embedding Module

ASLNet Adversarial Shape Learning Network

BCE Binary Cross-Entropy

BEP Break Even Point

EO Earth Observation

FLOPS FLoating Point Operations Per Second

CNN Convolutional Neural Network

CRF Conditional Random Field

DiResNet Direction-aware Residual Network

DC Dilated Convolution

DFC DeFormable Convolution

DSM Digital Surface Model

GAN Generative Adversarial Network

GSD Ground Sampling Distance

GT Ground Truth

HR High-Resolution

IR Infra-Red

IoU Intersection over Union

OA Overall Accuracy

OSM Open Street Map

PAM Patch Attention Module

PolSAR Polarimetric Synthetic Aperture Radar

LANet Local Attention Network

LC Land Cover

LU Land Use

MP-ResNet Multi-Path Residual Network

RF Receptive Field

RSI Remote Sensing Image

SAR Synthetic Aperture Radar

SD Shape Discriminator

SDM signed distance map

SR Shape Regularizer

SVM support vector machine

TP True Positive

VHR Very High Resolution

VRF Valid Receptive Field

Chapter 1

Introduction

This chapter introduces the background of this PhD thesis and gives an overview of the conducted research works. First the importance and general challenges for the semantic segmentation in RSIs are illustrated. Then the objectives and novel contributions in the conducted works are introduced. Finally the structure of this thesis is reported.

1.1 Background and Motivations

Remote sensing Images (RSIs) are a powerful source of data containing rich information on the Earth surface. Due to the rapid progress in data collection technologies, large volumes of RSIs are available at lower costs. Modern Earth Observation (EO) systems are able to collect information at a spatial resolution superior to 1m and with a temporal resolution up to days, yielding petabytes of raw (unprocessed) images per year. However, the raster data produced by EO platforms cannot be utilized directly without proper interpretation of the contained information.

The term of semantic segmentation originates from the computer vision community, which refers to the dense classification of image pixels based on their semantic context in the image. Semantic segmentation of RSIs refers to the pixel-wise classification based on the spectral and spatial information. Differently from the semantic segmentation of natural images, the semantic segmentation of RSIs usually aims at segmenting ground objects based on their Land Cover (LC) or Land Use (LU) types, as well as extracting certain classes of interesting targets (roads, buildings, water

bodies, etc.).

Despite of the rich information potentially available in HR RSIs, semantic segmentation of RSIs remains to be a challenging task. Several major barriers exist: i) Intra-class variance and inter-class similarity problems that are common in large-scale RSI datasets; ii) The complexity of LCLU mapping, which requires not only a high-level abstraction of local pixel information, but also the representation of object contours; ii) Occlusions caused by ground objects such as trees, shadows and vehicles; and iv) redundancy of the spatial information contained in HR RSIs. Traditional machine learning methods based on pre-defined feature extraction algorithms can hardly address these problems. Moreover, they are mostly parameter-dependent and cannot deal with large volumes of data collected in different areas.

Since the breakthrough made by Hinton and Salakhutdinov in 2006 [40], convolutional neural networks (CNNs) have quickly become a popular technology in a wide range of computer vision applications. By simulating the connection of neurons in human brain with a deep multi-layer structure, CNNs exhibit the ability of extracting features at various levels, which is beneficial for the full exploitation of spectral and spatial information present in images. Compared with traditional approaches, deep learning-based semantic segmentation methods have the advantage of automatically extracting highly discriminative features and of exhibiting higher generalization capability [64]. They have been widely used for the semantic segmentation of RSIs in recent years [155].

However, present CNN-based methods still exhibit several major problems in the segmentation results. First, fragmentation errors are common especially for VHR RSIs. This is caused by the limited receptive field (RF) in CNNs, where the local image pixels are classified isolated from their surroundings. Second, inaccurate segmentation of object boundaries. Since typical CNNs down-sample features to abstract the semantic information, many spatial details (that are crucial for finding object boundaries) are lost. Third, lack of optimization of the segmentation results. The CNN-segmented ground objects commonly have uneven contours or show geometric errors, thus further post-processing operations are often required [116]. To overcome these limitations, it is important to consider the

specific context of LCLU mapping in HR RSIs (instead of using directly the well-known CNN models developed in the computer vision community). The RSIs have many distinct characteristics that are different from natural images, including: i) imaging angle (bird-view), ii) rich spectral representations, iii) large spatial size. Considering these characteristics while developing CNN-based semantic segmentation methods can lead to better performance.

Taking use of the emerging techniques in the computer vision field and exploiting the special properties of RSIs, we conduct this thesis study to improve the segmentation accuracy in LCLU mapping applications. We are particularly interested in developing CNN-based approaches and technologies that can i) exploit better the context information in RSIs (without compromising spatial accuracy), ii) model the geometric patterns (e.g., shape and structure) of ground objects, and iii) directly optimize the segmentation results.

1.2 Objectives

In this PhD thesis, we aim to develop a set of novel CNN-based methods to improve the semantic segmentation of HR RSIs. According to the number of target classes, semantic segmentation can be roughly divided into multi-class segmentation and single-class segmentation (i.e., object extraction). Under the multi-class semantic segmentation, different kinds of data are studied, including optical data and Synthetic Aperture Radar (SAR) data. Under the single-class semantic segmentation, we consider two kinds of important ground objects: buildings and roads. An illustration of the relationship between the studied subtopics is presented in Fig.1.1.

In greater detail, the main objectives of this thesis are:

- Development of methodologies to improve the accuracy of semantic segmentation in VHR RSIs. Improvements are expected for both the aggregation of context information and the preservation of spatial information.
- Design of a method for the semantic segmentation of HR SAR images that mitigate the effect of speckle noise.

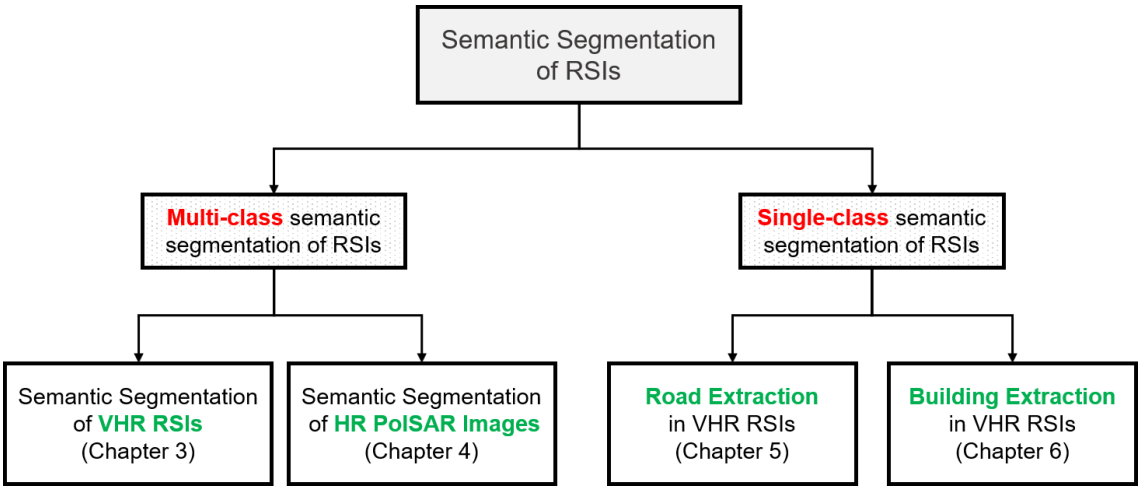


Figure 1.1: The relationship between the studied sub-topics in this thesis.

- Design of a method for road extraction in VHR RSIs that can detect better the occluded and non-salient roads.
- Design of a method for building extraction in VHR RSIs that can model the shape patterns of buildings to reduce the geometric errors.

1.3 Novel Contributions

The major contributions in this thesis are four CNN-based methods developed considering the aforementioned different objectives. They are listed as follow:

1. Proposing a Local Attention Network (LANet) for the semantic segmentation in VHR RSIs, which includes two local attention-based modules that are beneficial for improving the context representations of extracted features.
2. Designing a Multi-path Residual Network (MP-ResNet) for the semantic segmentation of SAR images, which can aggregate wider image context to alleviate the impact of speckle noise.
3. Proposing a Direction-aware Residual Network (DiResNet) that utilizes direction supervisions to improve the segmentation of roads in VHR RSIs.

4. Introducing an Adversarial Shape Learning Network (ASLNet) that learns shape constraints to improve the segmentation of buildings in VHR RSIs.

In the following, we briefly introduce the context and novelties of each contributions.

1.3.1 LANet: local attention embedding for the Semantic Segmentation of VHR RSIs

The trade-off between feature representation and discrimination power and spatial localization accuracy is crucial for the dense classification/semantic segmentation of RSIs. High-level features extracted from the late layers of a neural network are rich in semantic information, yet have blurred spatial details; low-level features extracted from the early layers of a network contain more pixel-level information, but are isolated and noisy. It is therefore difficult to bridge the gap between high and low-level features due to their difference in terms of physical information content and spatial distribution. To address this issue, we propose to enhance the feature representation in two ways. On the one hand, a patch attention module (PAM) is proposed to enhance the embedding of context information based on a patch-wise calculation of local attention. On the other hand, an attention embedding module (AEM) is proposed to enrich the semantic information of low-level features by embedding local focus from high-level features. Both proposed modules are light-weight and can be applied to process the extracted features of CNNs. Experiments show that, by integrating the proposed modules into a baseline Fully Convolutional Network (FCN), the resulting local attention network (LANet) greatly improves the performance over the baseline and outperforms other attention based methods on two RSI datasets.

1.3.2 Multi-path Residual Network for the Semantic segmentation of HR PolSAR Images

There are limited studies on the semantic segmentation of high-resolution Polarimetric Synthetic Aperture Radar (PolSAR) images due to the scarcity

of training data and the complexity of mitigating the effects of speckle noise. The Gaofen contest has provided open access a high-quality PolSAR semantic segmentation dataset. Taking this opportunity, we propose a Multi-path ResNet (MP-ResNet) architecture for the semantic segmentation of high-resolution PolSAR images. Compared to conventional U-shape encoder-decoder convolutional neural network (CNN) architectures, the MP-ResNet learns semantic context with its parallel multi-scale branches, which greatly enlarges its Valid Receptive Fields (VRFs) and improves the embedding of local discriminative features. In addition, MP-ResNet adopts a multi-level feature fusion design in its decoder to effectively exploit the features learned from its different branches. Comparisons with the baseline method (FCN with ResNet34) show that the MP-ResNet has achieved significant accuracy improvements. It also surpasses several state-of-the-art methods in terms of overall accuracy (OA), mF_1 and fwIoU, with only a limited increase of computational costs. This CNN architecture can be used as a baseline method for future studies on the semantic segmentation of PolSAR images.

1.3.3 Direction-aware Residual Network for Road Extraction in VHR RSIs

The binary segmentation of roads in Very High Resolution (VHR) RSIs has always been a challenging task due to factors such as occlusions (caused by shadows, trees, buildings, etc.) and the intra-class variances of road surfaces. The wide use of CNNs has greatly improved the segmentation accuracy and made the task end-to-end trainable. However, there are still margins to improve in terms of the completeness and connectivity of the results. In this framework, we consider the specific context of road extraction and present a direction-aware residual network (DiResNet) that includes three main contributions: i) An asymmetric residual segmentation network with deconvolutional layers and a structural supervision to enhance the learning of road topology (DiResSeg); ii) A pixel-level supervision of local directions to enhance the embedding of linear features; iii) A refinement network to optimize the segmentation results (DiResRef). Ablation studies on two benchmark datasets (the Massachusetts dataset and

the DeepGlobe dataset) have confirmed the effectiveness of the presented designs. Comparative experiments with other approaches show that the proposed method has advantages in both OA and F_1 score.

1.3.4 Adversarial Shape Learning for Building Extraction in VHR RSIs

Building extraction in VHR RSIs remains a challenging task due to occlusion and boundary ambiguity problems. Although conventional CNN-based methods are capable of exploiting local texture and context information, they fail to capture the shape patterns of buildings, which is a necessary constraint in the human recognition. To address this issue, we propose an adversarial shape learning network (ASLNet) to model the building shape patterns that increase the accuracy of building segmentation. In the proposed ASLNet, we introduce the adversarial learning strategy to explicitly model the shape constraints, as well as a CNN shape regularizer to strengthen the embedding of shape features. To assess the geometric accuracy of building segmentation results, we introduced several object-based quality assessment metrics. Experiments on two open benchmark datasets show that the proposed ASLNet improves both the pixel-based accuracy and the object-based quality measurements by a large margin.

1.4 Structure of the Thesis

The thesis is organized in seven chapters. Chapter 1 introduces the background, objectives and novel contributions in this thesis. Chapter 2 provides an overview to the literature works related to the semantic segmentation of RSIs. First we briefly review the existing techniques and state-of-the-art models for processing natural images; then we introduce the segmentation of RSIs, SAR images, roads and buildings, which are related to the studied sub-topics in the following chapters.

The major contributions in this thesis are presented in Chapters 3, 4, 5 and 6. These contributions are developed based on different types of data and different targets of interests. Firstly in Chapter 3, we focus on the general problems in the semantic segmentation of VHR RSIs and

present a local attention mechanism to improve the segmentation accuracy. Then, in Chapter 4 we focus on the semantic segmentation of HR SAR images, whose signal representations are different from optical RSIs. In the next two chapters we move to the segmentation of two kinds of interesting specific classes of ground objects: roads and buildings. In Chapter 5 we propose the DiResNet that strengthens the learning of linear features for road extraction. In Chapter 6 we present the ASLNet that models the shape patterns for building extraction.

Finally, in Chapter 7 we draw the conclusions of this thesis and summarize the possible developments in the studied areas in future works.

Chapter 2

Semantic Segmentation of RSIs with CNNs: State-of-the-Art

This chapter provides a review of the literature works related to the CNN-based semantic segmentation of RSIs. Since this thesis includes research works covering different types of data and different targets of interests, we introduce sequentially the literature papers related to each sub-topic. First we provide a general overview of the development of the CNN-based semantic segmentation. Then we introduce the multi-class semantic segmentation of RSIs, which is widely used in LCLU applications. Surveys to the segmentation of both optical RSIs and SAR data are given. Finally we move to the segmentation of specific classes of interesting targets and introduce separately the existing works for the segmentation of roads and buildings in RSIs.

2.1 Semantic Segmentation of Natural Images

The deep CNNs were developed for image classification tasks at their emergence in the last decade. They have been widely used for the semantic segmentation of images since the introduction of Fully Convolutional Network (FCN) [64]. In image classification tasks the last layers of CNNs are fully connected layers, which transform the 2D feature maps into 1D feature vectors. In FCN the fully connected layers are replaced with convolutional layers to obtain segmentation masks for dense classification tasks. In this section we review the development of semantic segmentation networks since

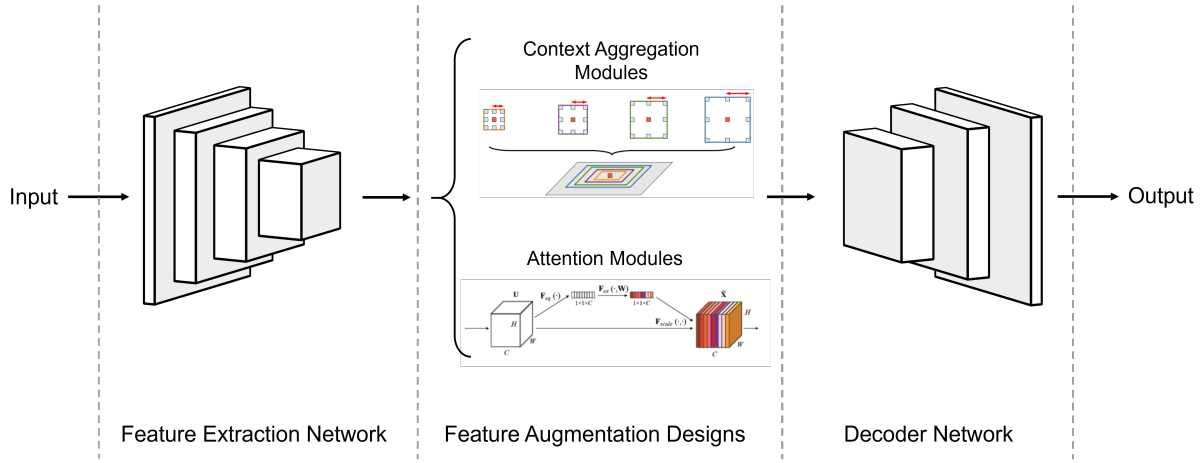


Figure 2.1: The structure of a typical semantic segmentation network.

the emergence of FCN.

2.1.1 Structure of Semantic Segmentation Networks

As illustrated in Fig.2.1, a typical semantic segmentation model consists of three components:

1. Feature extraction network. This is the backbone of a semantic segmentation model and often takes up most of the computational costs. It is often referred 'encoder' of the model, since it embeds semantic features from the input images.

2. Feature augmentation designs. Due to the locality problems, the feature maps produced by the feature extraction networks are often fragmented. To augment their semantic representations and to exploit better the context information, feature augmentation modules are often designed on top of the encoders. There are two types of feature augmentation designs: context aggregation modules and attention modules. The former aggregates local pixel information, while the latter calculates and embeds global scene focuses.

3. Decoder network. This is to recover the spatial size of embedded features before producing the results. It often consists of several deconvolution layers or up-sample operations.

In many cases the above-mentioned components can be connected (instead of strictly divided). For example, in [119] the attention designs are integrated into the feature extraction network. The skip-connections between the encoder and decoder networks are also common, resulting in various forms of 'encoder-decoder' structures. In the following sections we introduce sequentially the development of feature extraction networks, the encoder-decoder structures, as well as the context aggregation designs and the attention mechanism.

2.1.2 Feature Extraction Networks

A feature extraction network consists of stacked convolutional and down-sampling operations. The encoder in a semantic segmentation model is often selected from the advanced CNNs in image classification tasks. Here we briefly review the development of the CNNs for image classification tasks.

AlexNet [48], as the first network trying to increase the depth of CNNs, gained great research interest since its success in ImageNet Classification in 2012. It down-samples the feature map with 32 strides to reduce computation and increase the VRF. VGGNet [95] stacks 3x3 convolution operation to build a deeper network, while also involves 32 strides in feature maps. GoogLeNet [99] includes an inception block to include more diverse features. ResNet [38] adopts a 'bottleneck' design with a 'shortcut' connection in each stage, so that the network can be easily optimized even with a high depth. DenseNet [43] densely concatenates features to alleviate the vanishing-gradient problem and encourage feature reuse. HRNet [110] is a recent CNN architecture that contains multi-scale feature embedding branches to preserve the spatial information.

Among these networks, VGGNet, ResNet, DenseNet and HRNet are commonly employed as encoders in semantic segmentation networks. They all contain numerous 3x3 convolutions, which can reduce over-fitting risks. The VGGNet and ResNet are also frequently connected with decoders to form encoder-decoder structures.

2.1.3 Encoder-Decoder Structures

Although a plain FCN [64] can also produce 2D segmentation maps, there is tremendous loss of spatial information due to the multiple down-sampling operations. To increase the segmentation accuracy, the output of encoders must be recovered to the spatial size that is close to the input. To this end, unpooling and deconvolution operations are introduced in [77] to enlarge features. Since that, this symmetric encoder-decoder structure has been widely adopted.

A typical encoder-decoder network contains an encoder that gradually reduces the spatial size of feature maps and captures higher semantic information, as well as a decoder that gradually recovers the spatial size. SegNet [8] and U-Net [89] are typical networks using this architecture. They both contain a contracting path to capture context and a symmetric expanding path that enables precise localization. In SegNet, the pooling indices in down-sampling operations are recorded to perform non-linear upsampling. In UNet, the multi-scale features extracted by the encoder are reused in the decoder. Since the loss of spatial information is minimized in UNet, it has been widely used for medical image segmentation tasks. RefineNet further develops the encoder-decoder structure into a multi-path network and uses ResNet as its encoding module [56]. In DeepLabv3+ [16], a simple decoder is designed by transforming and concatenating the low-level features from the encoder.

However, the connections between encoder and decoder networks may introduce spatial noise and weaken the embedded semantic information. This limitation of the encoder-decoder structure is analyzed in Chapters 3 and 5.

2.1.4 Context Aggregation designs

The use of context information is essential to determine the object categories due to common intra-class similarity and inter-class variance problems. Many approaches add modules/blocks at the top of encoders to enlarge their VRFs and to integrate more context information. In [59], the importance of RF is discussed and global pooling operations are introduced to learn scene-level global context. PSPNet [39] extends the use of

global pooling to image sub-regions and proposes a parallel spatial pooling design to aggregate multi-scale context information. Dilated convolution is another design that enlarges the range of convolutional operations to extrac more information without increasing the risk to over-fitting [133, 18]. Combining dilated convolutions and multi-level pooling designs in PSPNet, the atrous spatial pyramid pooling (ASPP) module is proposed in [17] and is improved in [15, 16, 128].

2.1.5 Attention mechanism

The attention mechanism refers to the strategy of allocating biased computational resources to the processed signal to highlight its informative parts. In the tasks related to the understanding of image content, a typical solution for generating attention statistics is to gather information from a global scale, namely to exploit the scene or image-level information. This is because the scene information may provide clues about the possible contents in an image. In [109], the attention of the feature map is aggregated using an hourglass module in a residual manner. This residual attention network introduced a chunk-and-mask module, where the global attention is aggregated in the Soft Mask Branch through stacked down-sampling convolutions. In [41], a Squeeze-and-Excitation (SE) block is proposed, which uses global-pooling to generate channel-wise attention. In this way, spatial-irrelevant information can be learned to emphasize the scene-relevant feature channels. The design of 'squeezing' spatial information and the parallel connection of attention branch introduced in this work have been widely adopted in subsequent studies. In EncNet [140], a context encoding module is proposed to capture the scene-dependent global context as channel-wise attention. CBAM [119] introduced a spatial attention module to highlight the informative spatial regions. The spatial attention maps are generated by using pooling operations along the channel axis. BAM [82] has a similar module to exploit spatial correlations but it is implemented by applying dilated convolutions. PSANet [149] introduced the modelling of long-range correlation for each spatial position, but the channels of its inner layers are related to the input image size and cannot be applied to the prediction of full-size RSIs. A parallel de-

sign that models both channel-wise and point-wise attention is introduced in DANet [76]. A limitation of the non-local reasoning based networks is that the reasoning of global spatial correlation is calculation intensive. A light-weight graph-based module for reasoning latent correlations has been presented in [19].

2.2 Semantic Segmentation of RSIs

With their success in computer vision, CNNs quickly drew research interest in the field of remote sensing. The use of CNNs results in a breakthrough in the semantic segmentation of RSIs. First we briefly recall some relevant open datasets. Then we introduce the multi-class semantic segmentation of both optical RSIs and SAR data, respectively.

2.2.1 Available Open Datasets

The commonly used benchmark datasets for the semantic segmentation of RSIs are listed as follows:

- ISPRS Benchmarks for 2D semantic labeling[1].
  This is a VHR dataset for segmentation of objects in urban scenes. Multi-spectral aerial image data and the corresponding DSM data are included. Six object categories are defined.

- Gaofen Image Dataset (GID)[2] [103].
  This is an HR LC classification dataset collected by the Gaofen-2 (GF-2) satellite. It contains 150 GF-2 images collected from over 60 cities in China. It consists of a large-volume classification set with 5 LC classes, and a relatively small-volume classification set with 15 LC classes.

- Beijing Land-Use (BLU) dataset[3]. This is a LU dataset collected by the Beijing-2 satellite in Beijing. It consists of 4 tiles of large RSIs

[1] www2.isprs.org/commissions/comm3/wg4/semantic-labeling
[2] x-ytong.github.io/project/GID
[3] rslab.disi.unitn.it/dataset/BLU

collected in 4 sub-urban regions in Beijing, each one with a pixel size of 15680 × 15680.

- Gaofen PolSAR segmentation Challenge A PolSAR dataset for LC classification is provided in this contest. The PolSAR images are collected by the Gaofen-3 satellite. There are a total of 500 images with 5 classes of LC annotations.

- DeepGlobe LC Classification Challenge[4]. This contest provides a benchmark dataset for LC classification. Around 1150 satellite images are provided, each has 2448 × 2448 pixels. 7 LC classes are coarsely annotated.

- Zurich Summer Dataset[5] [107].
  These are 20 pan-sharpened 4 channel satellite images taken from a QuickBird acquisition of the city of Zurich at the spatial resolution of 0.62 m. Including 8 object categories. The average tile size is of about 800 × 800 pixels.

2.2.2 The Semantic Segmentation of RSIs

Studies on CNN-based semantic segmentation of RSIs begin to thrive after the emergence of several open datasets and contests. An early research was given on the generalization performance of pretrained CNNs applied to RSIs [83]. In [136], a multitask joint sparse representation is proposed to deal with the high inter-band correlation problem of hyperspectral RSIs. Object level segmentation has also been investigated to cope with the intra-class variability problem for classifying VHR RSIs[4][152][144].

One of the focuses in the semantic segmentation of RSIs is the collaborative use of CNNs and statistical modeling methods to improve the accuracy. In [78], CNN and crafted features are fused according to the pixel values on their probability maps to yield dense prediction. A hybrid MRF-CNN model was proposed in [139] using a fusion strategy that quantitatively calculates the uncertainty distribution of CNN outputs. Another research focus is the extraction of multi-scale features. In [132], the multi-scale

[4] competitions.codalab.org/competitions/18468
[5] sites.google.com/site/michelevolpiresearch/data/zurich-dataset

pyramid pooling module has been introduced to the semantic segmentation of RSIs. In [27], a two-stage design operating on seven different scales is presented to enlarge the RF of the network. The multi-scale alignment of edges and outputs are introduced in [58] and [84], respectively. To solve the problem of scale variation of RSIs, a multi-scale training strategy is proposed in [61].

Numerous studies put efforts on exploiting potential information from other types of data that are spatially registered to RSIs to improve the segmentation performance. In [46], satellite images and the corresponding OpenStreetMap (OSM) data were used in a pre-training phase to alleviate the data hungry problem of CNNs. OSM data have also been used as an input to SegNet in [6], as well as Lidar and multispectral data in [5]. A multiple-task feature fusion network integrating RGB bands, Digital Surface Model(DSM) and an Infrared (IR) band using U-Net has shown better performance compared with any single type of input data [97]. The information in DSM has also been separately exploited in [98][36] and [13]. In [7] signed distance maps are calculated and introduced in the segmentation network as an extra supervision to learn spatial-regularized results.

There are also works that use the attention mechanism for the segmentation of RSIs. In [81], a channel attention block is designed to enhance the decoding branch of the CNN. In [21], the attention mechanism is used to match the caption nouns with the objects in RSIs. The Global Attention Upsampling module [50] is introduced in [96] to provide global guidance from high-level features to low-level ones. In [74], the attention-based reasoning of both positional and channel-wise relations and their integration in serial and parallel manners have been studied. In [141], a multi-scale design has been introduced to aggregate context information through different branches.

From a general perspective, existing works for semantic segmentation on RSIs mostly employ multiple models to get hybrid results, or fuse different types of data to better model the ground information. However, limited attention has been paid to the special properties of RSIs, such as their large spatial size and low number of object categories (compared with natural images).

2.2.3 The Semantic Segmentation of SAR Images

Most of the literature works on the semantic segmentation of SAR images are designed under the assumption that there are not enough labeled samples to train a deep CNN. In [54], a relatively small portion of pixels of the image to be classified is used for the training phase. The FCN has been used in a pipeline that contains sliding-crop and sparse coding operations. In [115], the FCN is combined with the sparse and low-rank subspace representations to alleviate the problem of insufficient training data. In [30], a multi-scale CNN has been proposed for the semantic segmentation of SAR images. To solve the problem of lacking Ground Truth (GT) segmentation maps, it uses image scene labels for the training. In [12], the FCN is used in the complex domain of SAR data to include the phase information. In [120], extra datasets are used to pre-train CNN models for the semantic segmentation of SAR images. Since all these studies have been conducted on small datasets (most datasets contain only a single image), their objectives are mainly to reduce the dependence on training data keeping high generalization capabilities.

Recently, with the emergence of several benchmark datasets, new developments on the CNN-based semantic segmentation of SAR images have been proposed. In [73], a 'encoder-decoder' CNN network with inception modules and skip connections is introduced for the semantic segmentation of wetland PolSAR images. In [137], a multi-scale attention based FCN (MANet) is presented combining multi-scale feature extraction with the attention mechanism. In [114], a small yet efficient network (HR-SARNet) is proposed for the semantic segmentation of high-resolution SAR images. However, most of these literature studies are based on shallow CNNs to avoid over-fitting problems. Since shallow CNNs are not powerful enough to extract high-level semantic information, their accuracy is limited.

2.3 Segmentation of Roads in RSIs

Roads are one of the most important and challenging ground objects in RSIs. Different from many other objects that have compact contours, roads are usually extended through whole of a RSI. Literature works on auto-

matic road extraction can be divided into two categories: expert knowledge-based methods and CNN-based ones. Although the CNN-based methods have advantages in accuracy and generalization capabilities, previous works provide inspiration on how to utilize the spatial and spectral properties of roads. In this section we briefly review these two categories of methods.

2.3.1 Benchmark datasets

The available benchmark datasets for road extraction are listed as follow.

- Massachusetts Road Dataset [71]. This is an aerial datasets containing 1171 images, each of which has 1500×1500 pixels. The dataset was collected in Massachusetts, US. The GSD is 1.2m.
- DeepGlobe Road Extraction Challenge [6] This contest provides a large satellite dataset for developing road extraction methods. A total of 8570 images are provided, each has 1024×1024 pixels with the GSD of 0.5m.
- SpaceNet Road Network Detection Challenge[7] This Contest provides a dataset for road centerline detection. It includes 2780 aerial images. The image resolution is 1300×1300 and the GSD is 0.3m.

2.3.2 Expert Knowledge-Based Methods for Road Extraction

Previous works on road extraction before the emergence of CNNs generally consist of two essential steps, including i) the segmentation of roads, and ii) the refinement of the classified road segments.

Segmentation of Roads

Literature methods on road segmentation are based on either the local spectral homogeneity or the intensity contrast of road surfaces. There are supervised methods that require training samples and unsupervised ones that operate without any labeled data. Typical supervised segmentation

[6] competitions.codalab.org/competitions/18467
[7] spacenet.ai/spacenet-roads-dataset/

methods employ classifiers like the support vector machine (SVM) to label pixels based on the spectral values [22]. Some early works used simple neural networks as feature extractor [72]. However, these networks only contained a single hidden layer thus with limited capabilities to capture the problem complexity. Unsupervised segmentation methods can be further divided into edge-detection based and object-based ones.

The edge detection based methods are suitable for detecting ridge-like linear features. The Canny detector is one of the most widely used algorithms to extract road candidates [105][131][90]. In [22], the gradients obtained by Canny detectors are followed by a singular value decomposition to extract the road boundaries. In [92], 1-dimensional filtering operators are used to detect edges. The detection of local directionality is presented in [138] and shows better performance compared with the Sobel operator. Hough transform is another commonly used algorithm to detect the dominant linear features in an image [31].

Meanwhile, object-based methods are applicable to the extraction of ribbon-like structures. A typical strategy is to employ clustering techniques (based on spectral and texture features) to obtain candidate super-pixels, after which applying tracking or grouping algorithms to obtain road segments [135]. The clustering methods are usually bottom-up pixel-merging algorithms [44], some of which are implemented by the eCognition commercial software [131]. There are also plenty of works that employ angular operators to extract roads. In [42], the concept of road footprint is introduced to measure the shape of neighbourhood pixels and track the road directions. The work in [28] further merges the direction-homogeneous pixels into candidate road segments.

Refinement of Road Segments

After the coarse segmentation of roads, candidate road-like objects are presented in the binary maps. Typically, a filtering operation is applied to these maps to remove the false alarms. Several works employ geometric calculations to discriminate the shape of candidate regions. In [70], candidate road segments are classified based on the length-width ratio of their minimum bounding rectangles. In [69], the second-order moments of

segments are used to filter the non-road ones. Angular operators have also been used to measure the shape of segments in binary maps based on their circularity and rectangularity [142][29].

Another refinement process of the results is the optimization of the extracted road segments. This process generally includes thresholding calculations based on the geometric parameters of segments (e.g., length, distance, orientation) to simplify the road chains [31], merge the overlaps [122] and connect the adjacent regions and junctions [22]. Tensor voting is also a frequently used method to link road segments [70]. It is based on a geometric analysis of the differential information of local pixels [68].

2.3.3 CNN-based Methods for Road Extraction

Although the expert knowledge-based road extraction methods can achieve satisfactory results on some RSIs, they heavily rely on the setting of the values of many parameters. In this context, the use of CNNs brings an increase in both feature representation power and generalization ability at the cost of having a huge database of annotated samples for the training of the network. Here we briefly review the literature works on CNN-based road extraction in terms of two aspects: network designs and supervisions.

Network Designs

Most existing works are derived from UNet, a CNN originally designed for medical image processing [89]. Since it has a symmetric encoder-decoder design and concatenation operations between the encoder and decoders, it has the ability to preserve spatial details and is suitable for processing large scale images. In [129], the convolutional units in UNet are changed to recurrent ones, which contain summation operations between the convolutional layers to better preserve spatial information. In [146], the UNet is combined with the residual design in the ResNet [38] architecture. The resulting ResUNet shows a better performance compared with the original UNet. A similar design is introduced in [124] by combining UNet and the Dense block [43]. In [126], two attention units are incorporated into the DenseUNet network to introduce skip-layer attentions at both the global and local levels. Dilated convolutions have also been used to enlarge the

receptive filed of the CNN [147]. There are also CNNs designed for multi-task learning. In [20], two encoder-decoder CNNs are cascaded to perform the task of road segmentation and road centerline extraction, respectively. In [60], two parallel branches are added after the road segmentation network to learn road edges and centerlines simultaneously. Additionally, the generative adversarial network is introduced for the segmentation of roads in [143]. It includes a discriminator to improve the generation of road maps.

Supervisions

Additional supervision or the variation of loss functions can affect the learning of features. In [60], a multi-scale supervision is introduced to supervise each decoding layer. It also introduces human interactions to fix the incomplete predictions. In [145], the topology supervision (by centerline maps) is introduced to enable the network to better deal with occlusions. To emphasize the pixels close to road regions, a weighted loss function based on the calculation of euclidean distance has been introduced in [117]. While the binary cross-entropy (BCE) loss function is commonly used for binary segmentation, the structure similarity (SSIM) loss has been adopted in [37] to enhance the quality of the segmentation. In [9], a parallel branch that learns the orientation of roads is added as an auxiliary supervision to improve the connectivity of the road features.

To conclude, although there are numerous works on CNN-based road extraction in VHR RSIs, most of them are simple extensions of the commonly used CNN architectures without considering the specific context of road extraction. Thus, there are still margins to improve the accuracy of road segmentation in terms of completeness and smoothness.

2.4 Segmentation of Buildings in RSIs

Buildings are one of the most interested targets in LCLU mapping applications. In this chapter we briefly review the previous CNN-based works for building extraction. Since the proposed method in chapter 6 adopts an adversarial training strategy to model the shape of buildings, we also review

the existing works related to shape modeling and adversarial learning.

2.4.1 Benchmark Datasets

Below we briefly introduce the available benchmark datasets for building segmentation.

- Inria Building Dataset[8]. This dataset provides aerial images collected in a variety of cities, including Austin, Chicago, Kitsap, Tyrol and Vienna. Each image has 5000×5000 pixels and the GSD of 0.3m.
- Massachusetts Building Dataset [71]. The Massachusetts Building Dataset contains a total of 151 aerial images each with 1024×1024 pixels. The images are acquired in Boston, US. The GSD is 1.2m.
- DeepGlobe Building Detection Challenge [9] This contest provides a building footprint dataset. It contains 240,586 images and includes a total of 3,020,701 buildings. The images are collected in Las Vegas, Paris and Shanghai and Khartoum. Each image has 650×650 pixels.
- WHU Building Dataset[10] This is an aerial dataset with 8,189 images, each has 512×512 pixels. The raster labels are generated from vector data after manual editing. The GSD is 0.3m.

2.4.2 CNN-based Methods for Building Extraction

Literature works focusing on CNN for building extraction can be roughly divided into three types based on the studied perspectives: supervisions, architecture designs and post-processing algorithms. To begin with, while binary GT maps are widely used to compute the segmentation losses, several papers have explored the use of other kinds of supervisions. In [134], the supervision of signed distance map (SDM) is introduced to highlight the difference between building boundaries and inner structures. In [127] signed distance labels are also introduced but in the form of classification

[8] project.inria.fr/aerialimagelabeling/
[9] competitions.codalab.org/competitions/18467
[10] http://gpcv.whu.edu.cn/data/

supervision. This SDM has also been used in [94] as an auxiliary supervision.

Most CNN models for building extraction are variants of the well-known architectures for image classification and semantic segmentation. In [125], the ResUNet has been introduced for building extraction from VHR RSIs, which combines ResNet [38] with the UNet [89] structure. The MFCNN in [123] is also a symmetric CNN with ResNet as the feature extractor, whereas it contains more sophisticated designs (such as dilated convolution units and pyramid feature fusion). In [45], a Siamese UNet with two branches is designed to extract buildings from different spatial scales. In [11] a hybrid network with multiple sub-nets is introduced to exploit information from the multi-source input data. In [154], the MAPNet is proposed, which is a HRNet-like architecture with multiple feature encoding branches and channel attention designs. In [65], the global multi-scale encoder-decoder network (GMEDN) is proposed, which consists of a UNet-like network and a non-local modelling unit.

Since conventional CNN models only produce coarse segmentation results, post-processing operations are often required to obtain detailed results. In [125], guided filters are used to optimize the segmented building boundaries and to remove noise. In [116] and [151], regularization algorithms are developed to refine the segmentation maps. These algorithms perform object-based analysis on the edges and junction points to generate building-like polygons. In [123], a regularization algorithm is designed based on morphological operations on the rotated segmentation items. In [52], a graph-based conditional random field (CRF) model is combined with the segmentation network to refine the building boundaries.

2.4.3 CNN-based Methods for Shape modelling

There is a limited number of papers on CNN-based modelling of 2D shapes. To begin with, the work in [2] shows that CNNs can recognize shapes in binary images with high accuracy. In [88], the modelling of shape information is studied for the segmentation of kidneys from ultrasound scan images. In this work, a CNN auto-encoder is introduced to regularize the CNN output, which is pre-trained to recover the intact shape from ran-

domly corrupted shapes. The shape regularization network is trained by three loss terms that measure the distance between the input segmentation map, regularized segmentation map, and the ideal segmentation map. In [100], a gated shape CNN is proposed for the semantic segmentation. It contains an explicit shape stream that deals with the object boundary information.

Several works use binary mask features to preserve and model the shape information. In [49], the shape priors are modeled to improve the instance segmentation. The label masks are cluttered to generate class-wise shape priors. These priors are then weighted by a learnt vector of parameters to estimate the coarse instance region. In [24], a shape-variant convolution is proposed for the semantic segmentation. It uses a novel paired convolution to learn context-dependent masks to limit the RFs on interested image regions. In [55], the modeling of object contour polygons is studied for the instance segmentation. The polygons are first generated with a segmentation CNN and then transformed into a transformer network to fit to the object contours.

To the best of our knowledge, there is no existing work that explicitly models shape constraints for the segmentation of RSIs.

2.4.4 Adversarial Learning

Generative Adversarial Networks (GANs) [35]

GANs typically consist of two important components: a generator and a discriminator. The aim of the generator is to generate realistic results from the input data, while the discriminator is used to distinguish between the real data and the generated one. Since the discriminator is also a CNN, it is capable of learning the intrinsic differences between the real and fake data, which can hardly be modeled by human-defined algorithms. Therefore, the GANs have been widely used for a variety of complex tasks in the computer vision field, such as image generation [47, 102, 91, 101], semantic segmentation [106, 108], object detection [51, 113], depth estimation [3], and image/action recognition [104, 79].

Adversarial Learning for Building Extraction

Several literature works have introduced the adversarial learning strategy for building extraction. The segmentation model can be seen as a generative network, thus the building segmentation results can be learned in an adversarial manner by employing a CNN discriminator. The work in [53] is an early attempt on using the adversarial learning for building extraction. It forwards the masked input RSIs to the discriminator and uses an auto-encoder to reconstruct it. In [10], the GAN has been used to generate synthetic depth maps, thus improving the accuracy of building segmentation. In [1], the generative adversarial learning is introduced to improve the accuracy of building segmentation by employing a discriminator to distinguish whether the segmentation map is the GT map or the segmentation results. In [80], a multi-scale L1 loss is calculated from the discriminator to train the segmentation network. In [93], a conditional Wasserstein GAN with gradient penalty (cwGAN-GP) is proposed for building segmentation, which combines the conditional GAN and Wasserstein GAN.

In general, the literature papers on using adversarial learning for building extraction mostly combine segmentation maps and RSIs as input data to the discriminator, whereas the shape of segmented items is not modeled.

Chapter 3

Local Attention Embedding to Improve the Semantic Segmentation of VHR RSIs

This chapter [1] *presents a novel LANet that introduces a local attention mechanism to the semantic segmentation of VHR RSIs. This network bridges the semantic gap between high and low-level features, thus improving the embedding of context information while preserving important spatial details. Building on top of a feature encoder (the ResNet), first we designe a patch attention module (PAM) to enhance the semantic representations of extracted features. Then, we introduce an attention embedding module (AEM) to embed semantic focus across different levels of features. The resulting network (LANet) can integrate better the multi-scale features, thus obtaining accuracy improvements on two benchmark RSI datasets.*

3.1 Introduction

The dense classification of RSIs, which is often referred as semantic segmentation, is a crucial step for the automatic analysis of remote sensing data. It is widely used in a variety of applications, such as land-use and land-change mapping, urban management, environment monitoring, etc. With the development of CNNs and their application to dense classification

[1] This chapter appears in:
[J2] L. Ding, H. Tang, L. Bruzzone, "LANet: Local Attention Embedding to Improve the Semantic Segmentation of Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 59, No. 1, pp. 426-435, January 2021.

(introduced in FCN [64]), the accuracy of semantic segmentation on RSIs has been greatly improved [5]. A commonly used design in CNNs is based on stacked convolutions and pooling operations, which constantly reduce the spatial size of features to enhance their semantic representations [38]. Although this feature embedding design (referred as 'encoders') has the benefits of enlarging the RF and learning more intrinsic feature representations, it has the cost of losing detailed spatial information. Thus, the semantic segmentation results are generated by considering a large area as a whole instead of precisely classifying each pixel. As a result, small objects may be neglected and the contours of objects are ambiguous. To conquer this problem, 'decoders' are introduced, which typically employ the low-level features from 'encoders' to retrieve the lost spatial information [8, 89, 56]. However, the low-level and high-level features have significant differences in both semantic information and spatial distributions (e.g. low-level feature are more sensitive to gradient changes and distinct points, while the high-level features have stronger activation in the center of objects), thus the fusion of them does not bring significant improvements to the segmentation accuracy [148].

This trade-off between feature embedding power and spatial localization accuracy is crucial for the semantic segmentation of RSIs. On the one hand, different categories of the ground objects may share similar spectral features, thus requiring for an aggregation of context information [74]. On the other hand, many applications of the analysis of the RSIs require high precision in mapping contours of ground objects. Therefore, detailed spatial information is needed for identifying accurately both the boundary of regions and small objects.

The introduction of the attention mechanism is an effective strategy to reduce the confusion in predicted categories without losing spatial information. With the global statistics aggregated from the whole image, scene information can be embedded to highlight (or suppress) the features with strong correlations [41]. However, the spatial size of RSIs is usually much larger than that of natural images, whereas the number of object categories is often smaller. For example, each image in the ISPRS semantic labelling dataset (Potsdam area) has 6000×6000 pixels divided into 6 object categories in this dataset. As a result, almost every image contains all the

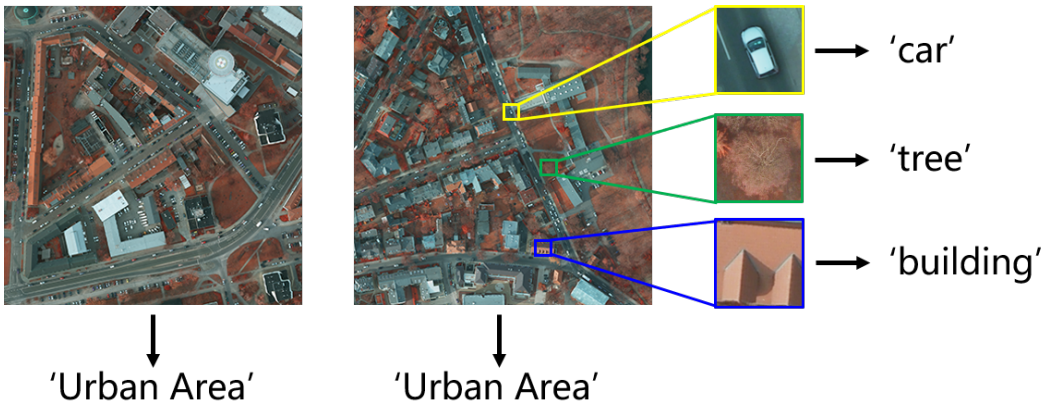


Figure 3.1: Examples of the image-level information of RSIs. The information of a whole RSI cannot be deduced more specifically than just 'urban area', but the information of image patches can be easily attributed to classes like 'car, 'tree' and 'building'.

object categories, and no clear global scene information can be embedded at the global level. In other words, we argue that the typical attention-based techniques cannot be directly applied to the semantic segmentation of large-size RSIs.

In This chapter, we propose the generation of patch-level local attention to improve the semantic segmentation of RSIs. The proposed approach is based on the finding that the image-level semantic information of RSIs is not clear, whereas the local image patches have clear semantic references (an illustration example of this observation is given in Fig. 3.1). Therefore, we propose a novel Patch Attention Module (PAM) to exploit patch-wise local attention. This module operates on extracted feature maps and can aggregate context information from the local patch to reduce confusions. In our model, the PAM is appended after both the high-level and low-level features to enhance their representation. Moreover, to bridge the gap between high-level and low-level features, an Attention Embedding Module (AEM) is proposed to embed semantic focus from high-level features into low-level features. This module can greatly improve the semantic representation of low-level features without losing their spatial details, thus improving the effectiveness of the fusion between high-level and low-level features. The proposed modules are light-weight and can be incorporated into existing CNN architectures to improve the segmentation accuracy. The experiments on two RSI datasets have validated the effectiveness of

the proposed architecture.

To summarize, the novel contributions in this chapter are as follows:

1. Proposing a PAM to embed scene information from local patches, as well as an AEM to enhance the semantic representation of low-level features by introducing attention from high-level features;
2. Proposing a local attention network (LANet) to improve the semantic segmentation of RSIs by enhancing the scene-related representation in both encoding and decoding phases;
3. Performing extensive ablation studies on two RSI datasets by incorporating the proposed modules into the baseline FCN in sequence. The resulting LANet is further compared with other methods that contain decoding or attention-based designs to evaluate its performance.

This chapter is organized as follows. First we describe the proposed LANet in Section 3.2. Then, we introduce the experimental datasets and implementation settings in Section 3.3. In Section 3.4 we present the detailed experimental evaluation and discussions. Finally, we conclude this chapter in Section 3.5.

3.2 Proposed Approach

In this section we present the proposed LANet devised for improving semantic segmentation of RSIs. Firstly, an overview of the network is given to introduce the general motivation and architecture. After that, the proposed modules are described in detail. Finally, a further explanation is given on integration of the proposed modules into the baseline network (the FCN).

3.2.1 Overview of the Proposed LANet

Contextual information is known to be crucial for the semantic segmentation of RSIs. Global pooling is an effective operation to aggregate contextual information, since it utilizes the scene information to learn biased focus on object categories. However, this approach is less effective on RSIs,

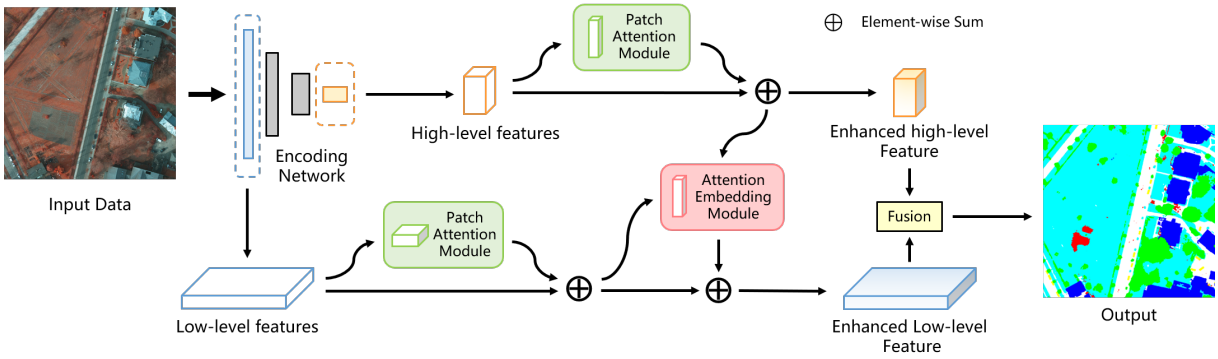


Figure 3.2: Architecture of the proposed local attention network (LANet). The patch attention module (PAM) generates attention maps to highlight patch-wise focus in feature maps. The attention embedding module (AEM) embeds semantic information from high-level features to low-level ones.

since the image-level information is not clear, as discussed in Section 3.1. To address this problem, we propose the LANet to utilize patch-based scene information on RSIs.

The motivations in this chapter is twofold: 1) employing patch-based attention to enhance the embedding of contextual information; 2) enriching the semantic representation of low-level features to better utilize the spatial information. To achieve this goal, two separate modules are introduced in LANet: i) a PAM to enhance the embedding of local context information, and ii) an AEM to improve the use of spatial information. Specifically, we designed two parallel branches to process features from different layers. As shown in Fig. 3.2, in the upper branch, high-level features (produced by late layers of a CNN) go through a PAM to enhance their feature representation; in the lower branch, low-level features (produced by early layers of a CNN) are first enhanced by PAM, then embedded with semantic information from high-level through AEM. The final segmentation results are produced by the fusion of the features from both branches.

3.2.2 Patch Attention Module

Semantic segmentation of RSIs suffers greatly from the problem of intra-class inconsistence, since the discrimination of object categories is a comprehensive task affected by both the surface type and the context of an

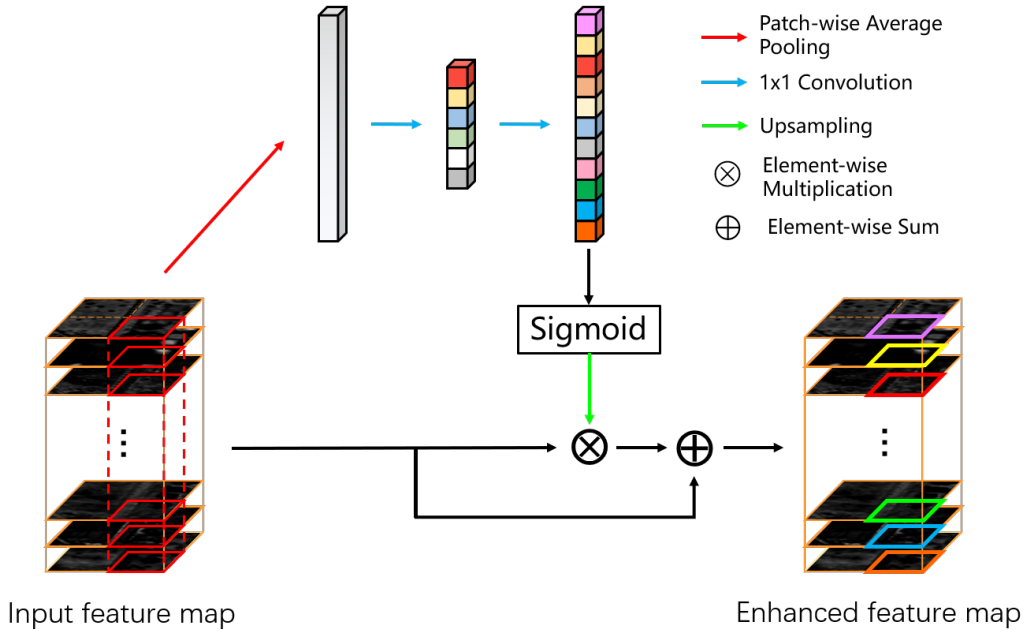


Figure 3.3: Detailed design of the PAM. Descriptors are calculated patch-wisely to aggregate local context information.

image. To alleviate this problem, we propose a patch attention module to enhance the aggregation of context information in the extracted features.

Fig. 3.3 shows the design of the PAM. Our work is inspired by the design of the SE-block [41]. The original SE-block introduced global average pooling to generate one single descriptor for each feature channel. However, as discussed in Section 3.1, this cannot be applied to the processing of large-size RSIs. In our approach, we limit the generation of descriptors to local patches, so that each descriptor contains meaningful information of the local context. Let us first consider a single patch. The descriptor z_c for the c-th channel of a generic patch is calculated as:

$$z_c = \frac{1}{h_p w_p} \sum_{i=1}^{h_p} \sum_{j=1}^{w_p} x_c(i,j), \tag{3.1}$$

where h_p and w_p denote the horizontal and vertical spatial size of the pooling window, respectively, and x_c denotes a pixel at c-th channel. In this way, a c-channel vector $\mathbf{z}_p$ can be generated, which contains the statistics describing the patch p. After that, we follow the bottleneck gating design in [41] to learn an attention vector $\mathbf{a}_p \in \mathbb{R}^{c \times h_p \times w_p}$ for the patch p. In-

stead of using fully connected layers, we employ convolutional operations so that they can be applied to process other patches without assigning extra weights. The gating operation to generate attention maps can be symbolized as:

$$\mathbf{a}_p = F_U\{\sigma[H_i\delta(H_r\mathbf{z}_p)]\}, \tag{3.2}$$

where σ and δ denote sigmoid and ReLU functions [75], respectively; H_r represents the 1×1 dimension-reduction convolution with the reduction ratio r, H_i denotes the 1×1 dimension-increasing convolution that recovers the feature dimension back to c. F_U is the upsampling operation.

Let us now extend the case of a single local patch to the global level. Given a feature map $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, maps of descriptors $\mathbf{Z} \in \mathbb{R}^{C\times H'\times W'}$ can be generated. Here, H' and W' are determined by the size of each patch (pooling window) as:

$$H' = \frac{H}{h_p}, W' = \frac{W}{w_p}, \tag{3.3}$$

where h_p and w_p are set according to the spatial reduction ratio of the corresponding encoding layer to ensure a remarkable enlargement of the RF. An alternative is to use a sliding window for generating the descriptors, so that the descriptor maps have the same size of input images. However, this option will tremendously increase the calculation, thus, it is not adopted in our implementation. After the convolutional layers, attention maps $\mathbf{A} \in \mathbb{R}^{C\times H\times W}$ can be produced. Finally, the original input features $\mathbf{X}$ are multiplied element-wisely with $\mathbf{A}$ to enhance their representation. A residual design is adopted to ensure the stable back-propagation of gradients.

3.2.3 Attention Embedding Module

An effective exploitation of low-level features is difficult due to their difference with high-level features in terms of spatial distribution and physical meaning. The most frequently used way of employing low-level features is to concatenate them with high-level features, which brings only slight improvement in performance (refer to discussion in Section3.3). To make the best use of low-level features, we propose an attention embedding module

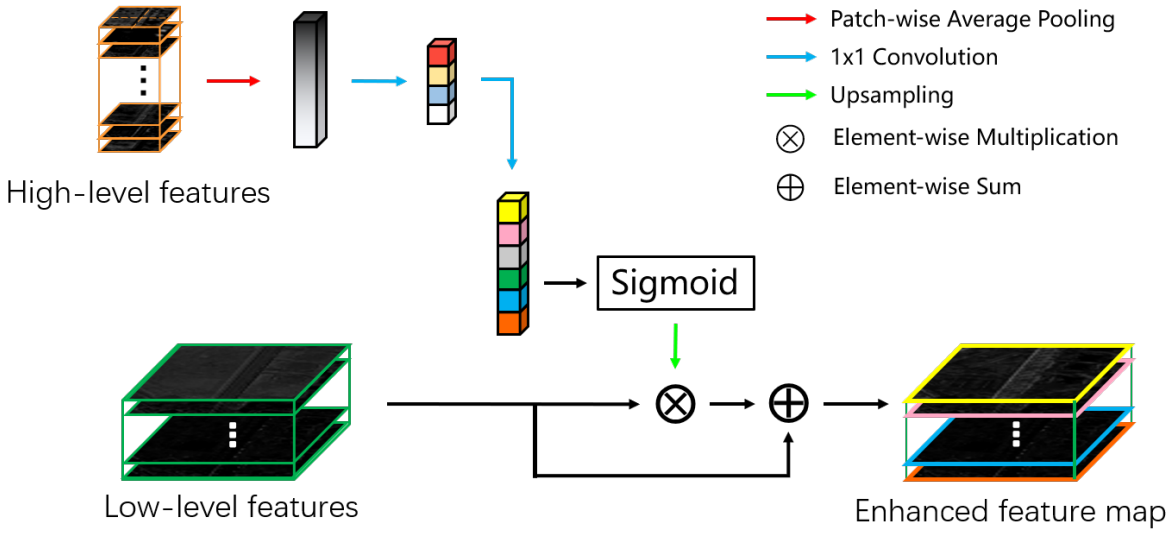


Figure 3.4: Detailed design of the AEM. Low-level features are semantically enriched by embedding local focus from high-level features.

to enrich their semantic meaning. This operation bridges the gap between high-level and low-level features without sacrificing the spatial details of the latter.

Fig.3.4 shows the design of the proposed AEM. The intuition of this approach is to embed local attention from high-level features into the low-level features. In this way, low-level features are embedded with context information that goes beyond the limitation of their RFs, while their spatial details are kept. First, we generate descriptors from high-level features through the same calculation as in (3.1). Let us denote these maps of descriptors as $\mathbf{Z}_h \in \mathbb{R}^{C_h \times H' \times W'}$, and the low-level features as $\mathbf{X}_l \in \mathbb{R}^{C_l \times H_l \times W_l}$. We generate attention maps for the low-level features $\mathbf{A}_l$ by transforming $\mathbf{Z}_h$ through bottleneck convolutions as:

$$\mathbf{A}_l = F_U\{\sigma[H_l\delta(H_r\mathbf{Z}_h)]\}, \tag{3.4}$$

where H_r is a dimension reduction convolution and H_l changes the number of channels to be the same as $\mathbf{X}_l$. To avoid excessive interference of high-level features, we add a residual design to emphasize the importance of low-level features. The enhanced low-level features are calculated as:

$$\mathbf{X}_l = \mathbf{X}_l + \mathbf{X}_l\mathbf{A}_l \tag{3.5}$$

3.2.4 Feature Fusion between Different Layers

After being processed by AEM, low-level features are semantically enriched and can potentially give a higher contribution to the prediction of the pixel class. Both the high-level and low-level features keep their dimensions after the processing of PAM and AEM. Accordingly, classic feature fusion operations (e.g., concatenation) can be applied to the outputs of the two branches. Since the specific feature fusion operation is not the focus of this chapter, also considering the interest in validating the output from each branch, we simply train two separate classifiers for each branch, and perform an element-wise sum to generate the final results.

3.3 Dataset Description and Design of Experiments

To assess the effectiveness of the proposed method, experiments have been conducted on two RSI datasets, i.e, the Potsdam dataset and the Vaihingen dataset. In this section we provide a short description of both datasets and then present the design of experiments providing implementation details.

3.3.1 Descriptions of Datasets

We employ two public available datasets to evaluate the proposed methods. The first dataset is the Potsdam dataset, which consists of 38 true orthophoto (TOP) tiles and the corresponding DSMs collected from a historic city with large building blocks. 24 imageries are used for training and the remaining 14 for testing. There are four spectral bands in each TOP image (red, green, blue and near infrared) and one band in each DSM. All data files have the same spatial size, equal to 6000 $\times$ 6000 pixels. The ground sampling distance (GSD) of this dataset is 5cm. The reference data are labeled according to six land-cover types: impervious surfaces, building, low vegetation, tree, car and clutter/background.

The second dataset is the Vaihingen dataset contains 33 TOP tiles and the corresponding DSMs collected from a small village. 16 images are used for training and the remaining 17 ones for testing. Differently from the Potsdam dataset, each TOP in the Vaihingen dataset contains three

spectral bands (near infrared, red and green bands) and one DSM band. The spatial size of the images varies from 1996 × 1995 pixels to 3816 × 2550 pixels. The GSD of this dataset is 9 cm. The reference data are divided into the same six categories as the Potsdam dataset.

3.3.2 Design of Experiments

Following the evaluation method provided by the data publisher and used in literature [67, 132, 62], three evaluation metrics are used to evaluate the performance of methods, i.e., OA, per-class F_1 score and average F_1 score. OA is calculated by dividing the correctly classified number of pixels with the total number of pixels. The F_1 score for a certain class is defined as the harmonic mean of precision and recall:

$$\mathrm{F}_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{3.6}$$

The same preprocessing, data augmentation and weight initialization settings have been used in all the experiments. The DSMs are concatenated with TOPs as input data, so that we obtain five channels for the Potsdam dataset and four channels for the Vaihingen dataset. Due to the limitation of computational resources, the input data are cropped using a 512 × 512 window during the training phase. The training batch size is 8 and the initial learning rate is set to 0.1. However, the prediction for the test set is performed whole-image-wise to obtain an accurate evaluation of the compared methods. Random-flipping and random-cropping operations are conducted during each iteration of the training phase as an augmentation approach. We use ResNet50 as the backbones for all compared networks with the pretrained weights for Pascal VOC dataset loaded from the PyTorch library. Since these weights are related to optical data that have only 3 channels, the weights in the first convolutional layers are duplicated in the channel dimension so that they can be applied to the Potsdam dataset (5 input channels) and the Vaihingen dataset (4 input channels). Following the design of DeepLabv3+ [16], we choose the output features of the first convolutional block of ResNet50 as the low-level features in the implementation. This has been done considering as empirical criterion a spatial scaling rate of the features equal to 1/4. Considering

Patch size	40	80	160	320
OA (%)	90.78	90.84	90.78	90.74

Table 3.1: The OA obtained with different patch sizes tested on the Potsdam dataset.

the different GSD of the two datasets, the down-sampling stride for the Potsdam dataset is set to 32, while for the Vaihingen dataset it is set to 16. The networks are implemented with PyTorch and the experiments are conducted on a server with a NVIDIA Quadro P6000 23GB GPU.

3.4 Experimental Results

In this section we present the tests of the proposed modules through an ablation study. Then, we compare the proposed LANet with state-of-the-art methods and draw the conclusion of our experimental validation.

3.4.1 Ablation Study

Impact of Patch Size. The patch size for calculating the local descriptors is an important hyper-parameter in the proposed LANet. To find out the appropriate patch size, we conduct an ablation study by gradually increasing its value. The results tested on the Potsdam dataset are reported in Table.3.1. The LANet achieves the best accuracy when the patch size is 80, thus this value is set as a default hyper-parameter.

Quantitative Results. In order to verify the effectiveness of the proposed modules, ablation studies have been conducted on the two datasets. FCN (ResNet-50) is used as the baseline network for comparison. Since the proposed LANet uses low-level features, the effect of considering low-level features has also been measured.

Table 3.2 shows the results of the ablation study on the Potsdam dataset. Three groups of observations can be done from the results. When no low-level features are involved in the decoding stage, the use of only one PAM (added on top of the FCN) increases the OA of 0.19%. With the inclusion of low-level features (concatenated with high-level features), the OA of the baseline FCN increases by only 0.16%. However, when two PAMs are

added to process the high-level and low-level features separately, the OA increases by another 1.07%. When the proposed AEM is used instead to enhance low-level features, the OA increases by 1.02%. With the use of both PAM and AEM, the proposed LANet increases the OA and average F_1 compared with the baseline FCN (with the use of low-level features) of 1.26% and 0.72%, respectively.

The results of the ablation study on the Vaihingen dataset are presented in Table 3.3. The inclusion of low-level features improves the OA of the baseline FCN of 0.18%. However, the use of both low-level features and PAM brings an increase of 0.7% on OA and 1.35% on average F_1. The use of low-level features and AEM brings an increase of 0.39% on OA and 0.63% on average F_1. Under the condition that low-level features are considered, the proposed LANet improves the average F_1 score and OA of 1.57% and 0.99%, respectively.

Table 3.2: Results of the ablation study on the Potsdam dataset. (*) low-feat indicates the use of low-level features.

Method	low-feat*	PAM	AEM	mean F_1	OA
FCN				88.66	89.42
FCN+PAM		√		89.03	89.61
FCN	√			91.23	89.58
FCN+PAM	√	√		91.76	90.65
FCN+AEM	√		√	91.78	90.60
LANet	√	√	√	**91.95**	**90.84**

Table 3.3: Results of the ablation study on the Vaihingen dataset. (*) low-feat indicates the use of low-level features.

Method	low-feat*	PAM	AEM	mean F_1	OA
FCN				86.14	88.66
FCN+PAM		√		86.42	88.68
FCN	√			86.52	88.84
FCN+PAM	√	√		87.49	89.36
FCN+AEM	√		√	86.80	89.05
LANet	√	√	√	**88.09**	**89.83**

Qualitative Analysis of Features. To visually confirm the effectiveness

of the proposed modules, we present comparisons of the segmented features generated independently before and after the use of the proposed modules. Fig. 3.5 shows the effect of applying the PAM module on high-level features. Since high-level layers already have relatively large RF before using the PAM, the enhancement is not significant. However, one can still observe that some of the meaningless small segments are removed, and the segmentation of easily-confused areas is improved.

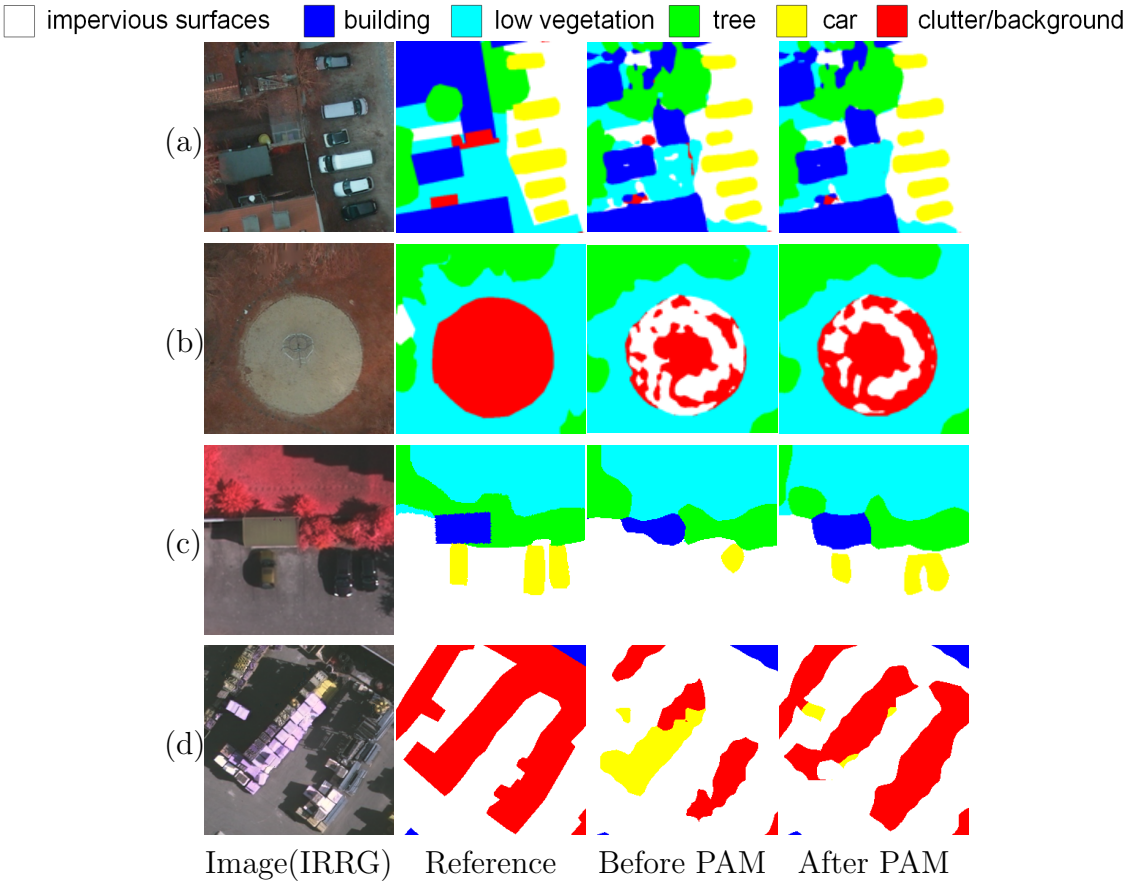


Figure 3.5: Comparison of segmented high-level features before and after the use of PAM. (a), (b) are selected from the Potsdam dataset, (c), (d) are selected from the Vaihingen dataset.

Fig. 3.6 shows changes of the segmented low-level features before and after the sequenced use of PAM and AEM. In the original low-level feature maps, pixels are only related to their neighborhoods due to the limitation of small RF. This leads to fragmented results and confusion of object classes. However, after the enhancement obtained with the proposed modules, the

semantic representation of low-level features is significantly improved. The pixels are classified based on not only the surface type of objects but also the context information. Moreover, one can verify from the clearly segmented boundaries that the spatial details of low-level features are kept.

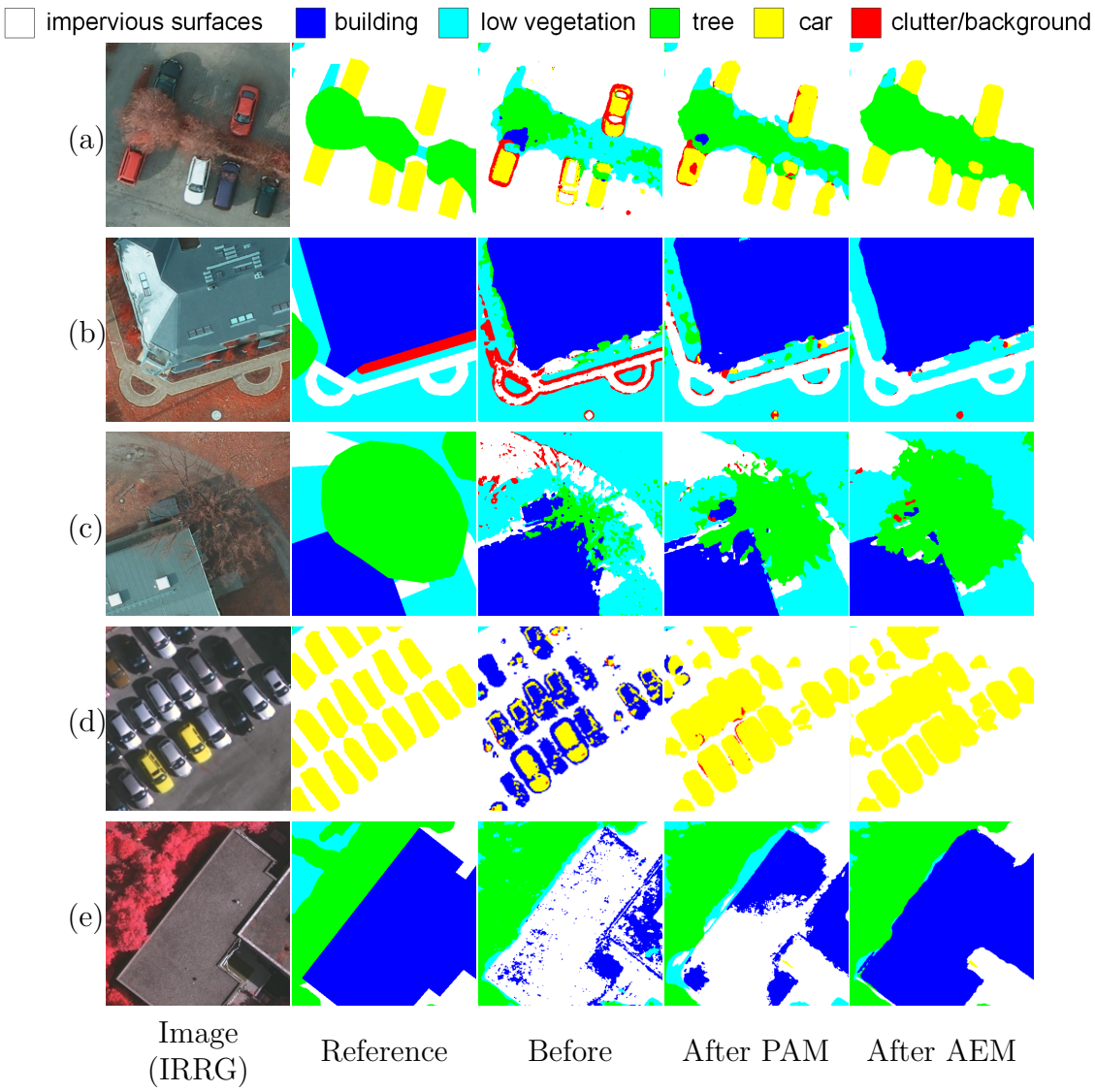


Figure 3.6: Comparison of segmented low-level features before and after the use of PAM and AEM. (a), (b), (c) are selected from the Potsdam dataset, (d), (e) are selected from the Vaihingen dataset.

Table 3.4: Results in terms of per-class F_1 score, average F_1 score and OA (Potsdam dataset).

Method	Per-class F_1 Score (%)					Average F_1 (%)	OA (%)
	Impervious Surface	Building	low vegetation	Tree	Car		
FCN	91.46	96.63	85.99	86.94	82.28	88.66	89.42
FCN+SE [41]	91.47	96.57	86.21	87.51	81.07	88.56	89.55
FCN+BAM [82]	90.43	94.97	85.84	87.47	85.63	88.87	88.83
FCN+CBAM [119]	91.37	96.49	86.00	87.40	83.22	88.89	89.46
FCN+GloRe [19]	91.55	96.54	86.17	87.42	82.69	88.87	89.57
DANet [76]	91.61	96.44	86.11	**88.04**	83.54	89.14	89.72
PSPNet [39]	91.61	96.30	86.41	86.84	91.38	90.51	89.45
DeepLabv3+ [16]	92.35	96.77	85.22	86.79	93.58	90.94	89.74
Proposed LANet	**93.05**	**97.19**	**87.30**	**88.04**	**94.19**	**91.95**	**90.84**

Table 3.5: Results in terms of per-class F_1 score, average F_1 score and OA (Vaihingen dataset).

Method	Per-class F_1 Score (%)					Average F_1 (%)	OA (%)
	Impervious Surface	Building	low vegetation	Tree	Car		
FCN	90.98	94.10	81.25	87.58	76.80	86.14	88.66
FCN+SE [41]	90.43	93.95	81.33	87.50	63.33	83.31	88.27
FCN+BAM [82]	90.77	94.01	81.54	87.78	71.76	85.17	88.62
FCN+CBAM [119]	90.86	94.03	81.16	87.63	76.26	85.99	88.61
FCN+GloRe [19]	90.57	93.99	81.28	87.49	70.09	84.68	88.41
DANet [76]	90.78	94.11	81.40	87.42	75.85	85.91	88.59
PSPNet [39]	91.44	94.38	81.52	87.91	78.02	86.65	88.99
DeepLabv3+ [16]	91.35	94.34	81.32	87.84	78.14	86.60	88.91
Proposed LANet	**92.41**	**94.90**	**82.89**	**88.92**	**81.31**	**88.09**	**89.83**

3.4.2 Quantitative Comparison with State-of-the-Art Methods.

Comparisons are made between the proposed LANet and approaches presented in the literature. All the tested approaches use the same backbone network (resnet50) and conduct the prediction on full-size test data. The experiments consider several recent works that have used the attention mechanism, including the SE block [41], the BAM [82], the CBAM[119], the GloRe [19] and the DANet [76]. The PSPNet [39] and DeepLabv3+ [16] with receptive-field-enlarging designs are also included in the comparisons. Table 3.4 and 3.5 report the quantitative results on the Potsdam dataset and the Vaihingen dataset, respectively. Compared with the baseline FCN, the use of most attention-based modules such as SE, BAM and CBAM do not involve noticeable performance improvement. The use of the SE-block even causes decreases in terms of F_1 scores, especially for the car class. This is because the channel-wise descriptors are calculated on the whole feature map, and the classes that account for a small portion of total pixels are suppressed. This proves our assumption that the global-level calculation of attention descriptors is not suitable for processing large-size RSIs. The DANet with a spatial dependency modelling design improves the OA of 0.3% on the Potsdam dataset, but there is a decrease of OA on the Vaihingen dataset. DeepLabv3+, which uses both low-level features and dilated convolutions, has good performance in F_1 scores. The proposed LANet, with the use of both context aggregation and attention embedding strategies, shows significant advantages over the compared methods. It shows the best performance in terms of both average F_1 score and OA, and obtains better F_1 scores in all the predicted categories.

To evaluate the required amount of calculation resources of the compared models, Table 3.6 represents the values of two metrics, i.e. the size of parameters and the floating point operations per second (FLOPS) (for processing each batch of data). The calculations are based on the input channels and pooling stride of processing the Potsdam dataset. Overall, the attention based methods (SE, BAM, CBAM and GloRe) are light-weight, whereas the context-aggregation based methods (PSPNet and DeepLabv3+) require more calculations. The proposed LANet does not significantly increase the calculations compared to the baseline FCN.

Table 3.6: Comparison of model size and calculations expressed in terms of params (Mb) and FLOPs (Gbps), respectively.

Method	FCN	FCN+SE	FCN+BAM	FCN+CBAM	FCN+GloRe	DANet	PSPNet	DeepLabv3+	Proposed LANet
Params (Mb)	23.79	23.80	24.15	23.97	23.81	47.73	46.94	39.73	23.80
FLOPS (Gbps)	21.95	21.95	22.38	21.95	21.95	28.01	31.67	30.72	21.98

3.4.3 Qualitative Analysis of the Semantic Segmentation Results.

Examples of the predicted patches on the two datasets are shown in Fig. 3.7. The segmentation maps provided by FCN are ambiguous (especially at the contours of objects) due to the loss of spatial information. The direct use of attention-based methods (e.g., SE and DANet) brings limited improvements. The context-aggregation based approaches (e.g. PSPNet and Deeplabv3+) show improvements in segmenting confusing areas, but also produce many fragmented segments. With the aggregation of local contextual information, the proposed LANet not only significantly reduces the errors, but also better preserves the spatial details. Specifically, the discrimination between cars and impervious surface, as well as between buildings and clutters has been greatly improved. There are also noticeable improvements in preserving the boundaries of objects. Figs. 3.8 and 3.9 show the large-size predictions on the Potsdam dataset and the Vaihingen dataset, respectively. Observing from a larger scale, the results of DANet are more reliable compared to FCN, but still suffer from low spatial accuracy; the results of PSPNet and Deeplabv3+ are more fragmented. As a comparison, in the predicted maps of the proposed LANet there are less false alarms in the surrounding areas of buildings, which can be attributed to the embedding of contextual information. Meanwhile, the segmentation of small objects (e.g., cars, paths, small clutters) is more accurate, which is due to the incorporation of enhanced low-level features. This points out that, the proposed method improves both the discrimination of critical categories and the preservation of spatial details.

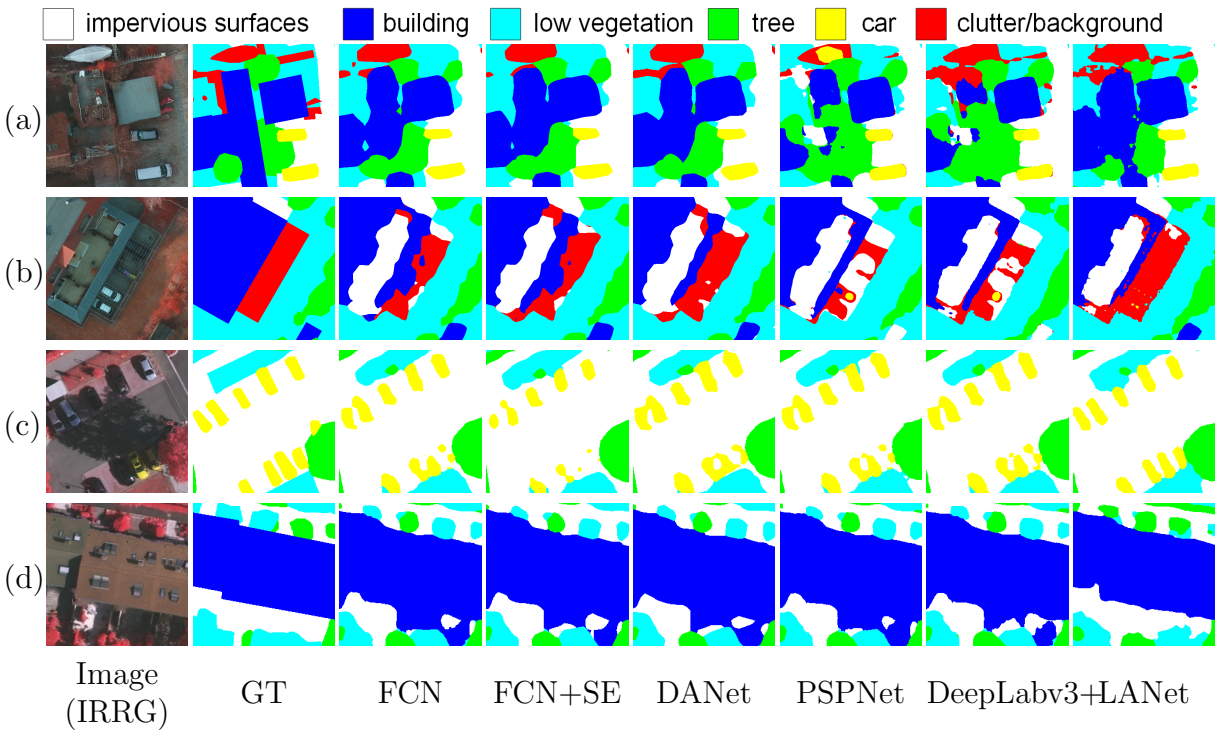


Figure 3.7: Examples of semantic segmentation results. (a), (b) are selected from the Potsdam dataset, (c), (d) are selected from the Vaihingen dataset.

3.5 Conclusions

Attention mechanism is a commonly used strategy in CNNs for aggregating context information in images. However, RSIs have large spatial size and a relatively small number of classes with respect to natural images do not express clear image-level scene information, which limits the use of the attention mechanism. In This chapter, we present a LANet that employs patch-level scene information to improve semantic segmentation of RSIs. Specifically, two modules are proposed for enhancing the representation of features based on the exploitation of local attention: i) The PAM enhances encoding of context information based on patch-wise calculation of local descriptors, and ii) the AEM embeds attention from high-level layers into low-level ones to enrich their semantic information.

Experimental results on two benchmark RSI datasets (Potsdam and Vaihingen datasets) show that the proposed approach greatly improves the representation of extracted features. The aggregation of local atten-

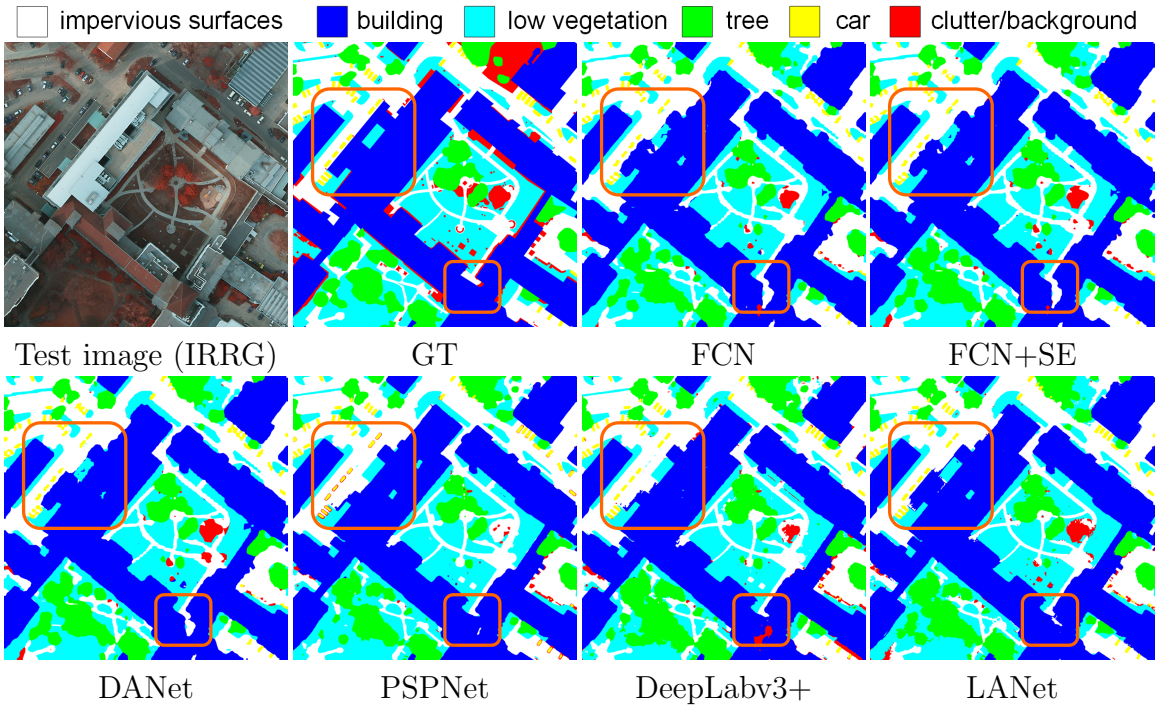


Figure 3.8: Example of large-size semantic segmentation results (Potsdam dataset). Major differences are marked with orange squares (zoom in for more details).

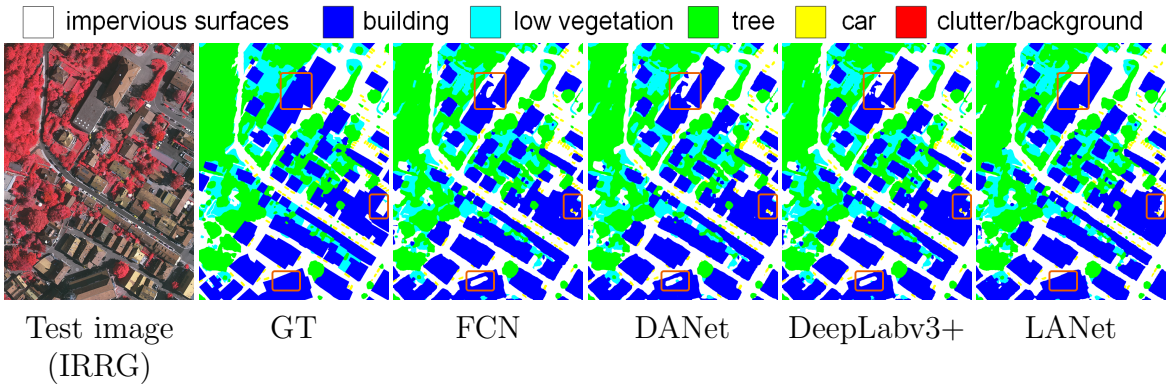


Figure 3.9: Example of large-size semantic segmentation results (Vaihingen dataset). Major differences are marked with orange squares (zoom in for more details).

tion (using the PAM) is beneficial for classifying the easily-confused areas, while the embedding of attentions from high-level features to low-level ones improves the preservation of spatial details. Comparative results show that the proposed LANet outperforms other global-attention and receptive-field-enlarging based approaches. However, one of the remaining problems in semantic segmentation of RSIs is that the objects in segmented maps are still more-or-less fragmented, especially at the boundaries. To conquer this limitation, as further development of this work we plan to study feature encoding strategies to improve the embedding of high-level features in the network.

Chapter 4

Multi-path Residual Network for the Semantic segmentation of HR PolSAR Images

This chapter [1] *presents a novel CNN architecture for the semantic segmentation of HR Polarimetric Synthetic Aperture Radar (PolSAR) images named Multi-path Residual Network (MP-ResNet). Differently from conventional U-shape encoder-decoder CNN architectures that contain only a major feature embedding branch, the proposed MPResNet has three parallel branches. This design greatly enlarges its VRF and improves its embedding of local discriminative features. Additionally, the MP-ResNet has feature fusion designs to integrate the encoded multi-branch features. The accuracy obtained by the MP-ResNet are tested by experimental comparisons with both the baseline method and several state-of-the-art CNN models in the semantic segmentation of SAR images. It shows improvements in terms of OA, mF_1 and fwIoU, with only a limited increase of computational costs.*

4.1 Introduction

SAR has been widely used in Earth observation applications due to its capability to work under all weather and daylight conditions. The semantic segmentation of PolSAR images, namely the pixel-wise classification of

[1] This chapter appears in:
[J4] L. Ding, K. Zheng, D. Lin, Y. Chen, B. Liu, J. Li, L. Bruzzone, "MP-ResNet: Multi-path Residual Network for the Semantic segmentation of High-Resolution PolSAR Images," *IEEE Geoscience and Remote Sensing Letters*, In press, 2021.

PolSAR images according to ground surface types, is beneficial to a large number of remote sensing applications (e.g., urban area management, disaster monitoring and land-cover mapping).

In recent years, with the rise of CNNs many methods have been developed for the semantic segmentation of natural images [150][16] and RSIs [26][74]. However, a limited number of studies has been conducted on the semantic segmentation of PolSAR images based on deep CNNs [33]. There are two major barriers: i) PolSAR images contain intense speckle noise due to the coherent imaging mechanism of SAR systems. This speckle noise is a severe challenge for the automatic segmentation algorithms. ii) Large amount of training data is required to train an effective deep CNN. The manual annotation of SAR data is not only labor-intensive but also difficult, since the ground objects in SAR images can hardly be recognized by human observation without assisting data and expert knowledge [30].

Recently, a large benchmark dataset for the semantic segmentation of PolSAR images have been made available by the Gaofen contest committee. Taking this opportunity, we aim to improving the semantic segmentation of HR PolSAR images by proposing a deep CNN architecture, i.e., the Multi-path Residual Network (MP-ResNet). This architecture enables a multi-scale modeling of high-level semantic features through its parallel branches, which strengthens the learning of local discriminative features and reduces the effects of speckle noise. Comparisons with its baseline method show that the proposed approach achieves an increase of 0.36% and 0.64% in terms of average OA and fwIoU, respectively, compared to the baseline FCN. Its average accuracy also surpasses several literature works in the comparative experiments.

In the following sections in this chapter, we introduce by sequence the methodology, experimental results and conclusions to this chapter.

4.2 Proposed Approach

The key to improve the semantic segmentation of PolSAR images is to learn discriminative features from a larger image context, so that the effects of speckle noise can be mitigated. To this end, we propose the MP-ResNet

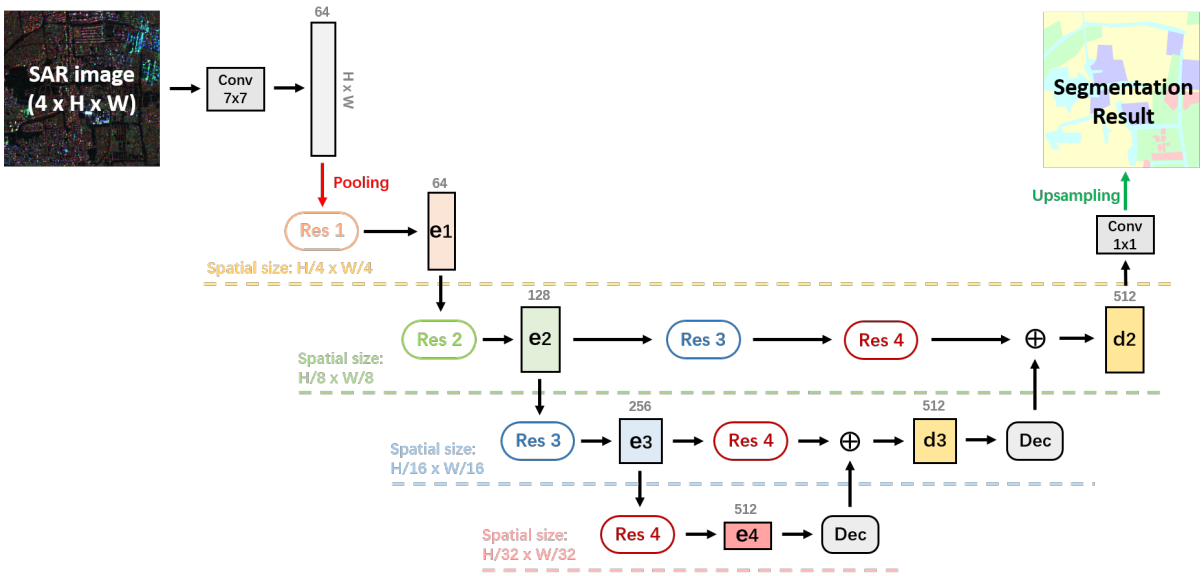


Figure 4.1: Architecture of the proposed Multi-path Residual Network (MP-ResNet). Res1-4: residual blocks from the original ResNet; e1-3: encoded features at each convolutional stage; d2-3: decoded features at each stage; $\oplus$: pixel-wise summation.

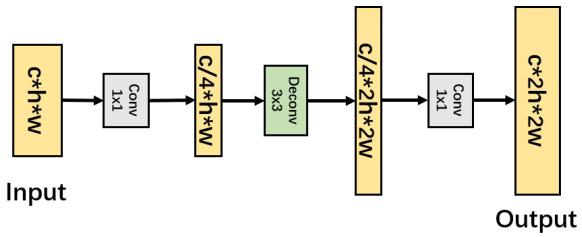


Figure 4.2: Deconvolution block of the proposed MP-ResNet.

shown in Fig.4.1. Compared to typical CNNs, which contains only a major encoding branch, the MP-ResNet employs 3 parallel encoding branches to aggregate context information at multiple scales. In this section we introduce first the choice of encoding network, then the multi-path feature aggregation design in the MP-ResNet, and finally the fusion of the encoded multi-scale features.

4.2.1 Choice of the Feature Extraction Network

PolSAR data generally contain 4 channels which are related to the 4 combination of linear polarizations (i.e. the HH and VV co-polarized channels and the HV and VH cross-polarized channels). The corresponding 4 im-

ages are stacked and given as input to a CNN. Since CNNs are capable of extracting semantic features from raw input data, no extra filtering operations have been applied to the input images. Instead, we merely apply maximum-suppression and normalization operations to stabilize and squeeze the value range of input data.

To alleviate the impact of speckle noise in PolSAR images, it is necessary to exploit discriminative features (e.g., local texture patterns, context information) from a wider image range. Therefore, we adopt the ResNet as the backbone feature extraction network. The ResNet contains strided convolution and max-pooling operations in its early layers, which quickly decrease the scaling rate of features to 1/4. Additionally, the kernel size of the first convolutional layer in the ResNet is 7×7, which is larger than the more commonly used 3×3 kernels. Both these designs enable the network to aggregate features from a larger area and thus be less sensitive to speckle noise. After the first two residual modules of ResNet, the scaling rate of features is decreased to 1/8. This feature size is large enough for balancing the accuracy and the spatial consistency of the results. Therefore, it is set as the fundamental feature size for the aggregation of context information in the proposed network.

4.2.2 Multi-path Semantic Information Embedding

The size of valid receptive fields (VRFs) is known to be crucial to the embedding of context information in CNNs [140]. In the semantic segmentation of remote sensing data, the size of VRFs determines the spatial range from which the CNN can exploit discriminative features, which is related to the granularity of semantic segmentation results [27]. Although the serial connection of pooling and strided convolutions can greatly enlarge the VRFs, it also brings the problem of losing spatial information. Therefore, how to simultaneously enlarge VRFs and preserve spatial information is one of the most crucial bottleneck problems in semantic segmentation tasks. PSPNet [150] and DeepLab [16] managed to enlarge VRFs without severe loss of spatial information through the use of additional context aggregation modules in the late layers of the CNNs. However, the context information is aggregated through pooling and dilated convolutions

in these designs, which are less effective than stacked convolutional layers. The dilated convolutions may also cause gridding effects and enlarge computational costs.

Alternatively, HRNet[110] presents a multi-path architecture that organizes multi-scale convolutional layers in a parallel manner. The highest scaling rate of HRNet is 1/32, which ensures a large VRF of the network. However, the serial convolutions in its parallel branches greatly increases the computational costs of this network. They may also cause over-fitting problems on small datasets. In addition, in HRNet the multi-scale feature branches are concatenated together to generate the semantic segmentation results, thus the features are not fully fused and utilized.

Inspired by the multi-path feature embedding design of the HRNet, in the proposed MP-ResNet we designed three parallel feature encoding branches after the second convolutional block (whereas the baseline ResNet contains only a single branch). Therefore, the features are encoded both forwardly (size remain the same) and downstream-wise (size reduced). Each of these parallel encoding branches is equivalent to the ResNet34[38]. They consist of the same amount of convolutional layers and residual blocks, except for the different scaling rate. In this way, each encoding branch contains rich semantic information. The number of parallel branches is designed taking into account the efficiency. Although 4 parallel branches (starting from a lower spatial scale) may potentially improve the accuracy, it will require much more computations. On the contrary, the computational costs of MP-ResNet is only slightly higher than ResNet and is far less than the HRNet (see discussion in Section 4.4).

Differently from the HRNet, the parallel branches in MP-ResNet focus on the embedding of high-level features (the lowest scaling rate is 1/8). Although the parallel embedding of larger feature maps are potentially feasible in other applications, for the semantic segmentation of PolSAR images our objective is to exploit discriminative features from a wider range. In this way, the segmentation results become less sensitive to pixel noise. Another significant difference of the MP-ResNet is that a decoder network is employed to fuse the features learned from its parallel branches.

4.2.3 Fusion of Multi-scale Features

Decoder networks are commonly used in semantic segmentation to recover the spatial details of encoded features. A common design is to concatenate or add the encoded features with the features from early layers of the encoder networks (e.g. UNet[89] and SegNet[8]). Although this design can aggregate spatial information from low-level features, it also introduces redundant information (minor details and noise). For the segmentation of high-resolution PolSAR images this problem can be critical. However, in the proposed MP-ResNet, the fused multi-scale features are the ones encoded by the parallel branches of the encoder. These features contain rich semantic information, thus their fusion does not lead to noise problems. A feature deconvolution module [14] is introduced to enlarge the spatial size of features from higher branches before the fusion. Fig.4.2 shows this spatial deconvolution block. It adopts a channel-wise 'Bottleneck' design to reduce the computation and refine the crucial semantic information. In this way, the multi-branch features are fused in a coarse-to-fine manner in the decoder.

4.3 Dataset and Experimental Settings

4.3.1 Dataset Descriptions

The experiments dataset of this study are developed on the Gaofen dataset provided by the '2020 Gaofen contest on automated high-resolution earth observation image interpretation'. The PolSAR images are collected from the Gaofen-3 satellite. Their ground sampling distance is between 1m and 3m. The GT maps are annotated according to 5 land-cover types: background, built-up area, vegetation, water and bare soil. The accessible training data are 500 pairs of PolSAR images and label maps each with 512×512 pixels. The testing data are not visible to users, but a scoring system is provided to evaluate the uploaded algorithms.

4.3.2 Evaluation Metrics

We adopt 3 metrics for the evaluation of semantic segmentation results. They are OA, F_1 score and frequency weighted Intersection over Union (fwIoU). OA is the ratio of the number of correctly classified pixels among all pixel numbers. F_1 is the harmonic mean of precision and recall values of the results. FwIoU is the evaluation metric suggested by the contest organizer. It is calculated as:

$$fwIoU = \frac{1}{\sum_{i=0}^{N}\sum_{j=0}^{N} s_{ij}} \sum_{i=0}^{N} \frac{\sum_{j=0}^{N} s_{ij} s_{ii}}{\sum_{j=0}^{N} s_{ij} + \sum_{j=0}^{N} s_{ji} - s_{ii}} \tag{4.1}$$

where N is the number of total classes, s_{ij} denotes the number of i-th class pixels that are classified into the j-th class.

4.3.3 Implementation Settings

The proposed method is implemented with the Pytorch library and is trained on a server with a NVIDIA GTX 1080Ti GPU (12GB memory). Random-flipping is performed to augment the dataset during the training. The training batch size is 24 and the initial learning rate is set to 0.1. The number of training epochs is 200. Pretrained weights of the ResNet34 are used to initialize the network. To adapt to the 4-channel SAR data, the weights in the first convolutional layer are duplicated in the input channel dimension.

4.4 Experimental Results

4.4.1 Multi-fold comparisons with the baseline method

To quantitatively evaluate the improvement of the proposed MP-ResNet over the baseline method (FCN), we conduct multi-fold comparative experiments on the Gaofen dataset. The training and validation sets are randomly divided from all available training data (500 image pairs) with a numeric ratio of 9:1. In this way, a total of 10 training and validation sets are obtained. To reduce the effects of random factors, the experiments are conducted on all the 10 training-validation sets. The results

Table 4.1: Results of the multi-fold experiments. 'mF_1' refers to the mean value of F_1 scores of all the classes.

Datasets	FCN (Baseline)			MP-ResNet (Proposed)		
	OA(%)	mF_1(%)	fwIoU(%)	OA(%)	mF_1(%)	fwIoU(%)
Val 1	93.08	91.33	88.14	93.95(+0.87)	92.47(+1.14)	89.63(+1.49)
Val 2	93.41	91.81	89.04	93.88(+0.47)	92.58(+0.77)	89.88(+0.84)
Val 3	91.70	88.20	86.35	91.86(+0.16)	88.24(+0.04)	86.71(+0.36)
Val 4	90.08	85.52	84.65	90.22(+0.14)	86.11(+0.59)	84.58(+0.24)
Val 5	93.06	89.64	88.34	93.45(+0.39)	90.46(+0.82)	88.97(+0.63)
Val 6	91.96	89.28	86.37	92.85(+0.89)	90.76(+1.48)	87.81(+1.44)
Val 7	92.97	89.22	88.25	93.33(+0.36)	89.20(-0.02)	88.90(+0.65)
Val 8	91.43	87.21	86.09	91.60(+0.17)	87.58(+0.37)	86.52(+0.43)
Val 9	93.11	90.66	88.66	93.24(+0.13)	90.72(+0.06)	88.91(+0.25)
Val 10	93.40	91.59	89.50	93.62(+0.22)	91.76(+0.17)	89.85(+0.35)
Test set	-	-	69.42	-	-	70.65 (+1.23)
Average Improvements				+0.36	+0.54	+0.64

are reported in table 4.1. Compared to its baseline method (FCN), the proposed MP-ResNet shows average improvements of 0.36%, 0.54% and 0.64% in OA, mean F_1 and fwIoU, respectively. The improvements have also been verified on the contest test set which is not directly available. According to the results obtained from the contest, the fwIoU scores (the only provided metric) on the test set is significantly lower than that on the validation set, which may be attributed to domain difference. Under this circumstance, the proposed MP-ResNet still obtains an increase of 1.23% in fwIoU compared to the FCN.

4.4.2 Comparative Experiments

To further assess the improvements brought by the proposed MP-ResNet, we compared it with several literature works. Apart from the baseline method FCN, several well-established methods in the field of semantic segmentation have been considered, including the SegNet [8], the UNet [89], the PSPNet [150] and the DeepLabv3+ [16]. Since the proposed MP-ResNet is inspired by the multi-path architecture of HRNet[110] and the feature fusion design of LinkNet[14], we also included these networks in comparison. Moreover, several literature methods presented for the

semantic segmentation of SAR images have also been tested, including the multi-scale FCN (MS-FCN) in [120], the Inception FCN (Inc-FCN) in [73] and the HR-SARNet [114]. The training and validation sets used in the comparison are these related to the first row of table 4.1. For fairness, the same parameter settings have been applied during the training process (e.g. training epochs, batch size, learning rate).

Table 4.2 presents the quantitative results obtained on the Gaofen dataset. Due to the effects of speckle noise, the performances of shallow networks (SegNet, HR-SARNet, Inc-FCN and UNet), which rely heavily on low-level features, is unsatisfactory. Although there is also a multi-scale design in the MS-FCN, there is no enhancement in its branches, thus its accuracy is lower than that of the FCN. The simple FCN without any sophisticated design ranks at the 3rd place. Although HRNet shows better performance in the semantic segmentation of VHR optical RSIs [141], its size is too large to be fully trained on the Gaofen dataset, thus its accuracy is far lower than that of the FCN. LinkNet also has a deep encoding network (ResNet34) but it has skip connections with low-level features, which introduce noise and degrades its accuracy. The PSPNet has an additional multi-scale average pooling head compared with the FCN. However, this design does not improve the accuracy metrics on the Gaofen dataset. The multi-scale atrous convolution head in DeepLabV3+ increases the mean F_1, OA and fwIoU by 0.24%, 0.25% and 0.15%, respectively, compared to the FCN. The proposed MP-ResNet achieves the best accuracy in nearly all the metrics except the F_1 of the 'others' class. Its improvements over the DeepLabV3+ are 0.60%, 0.62% and 1.32% in mean F_1, OA and fwIoU, respectively. This proves the effectiveness of the proposed network with which we have won the 2nd place in the 'Gaofen Challenge' contest.

Fig.4.3 shows a comparison of the segmentation results on several sample areas. The first column in this figure shows the PolSAR images, whose land-cover types can hardly be distinguished by only considering the local patterns in pixel neighborhoods. Due to its multi-path modeling and multi-scale feature fusion design, the proposed MP-ResNet can model context information from a wider image range. Therefore, some critical areas for other networks are correctly segmented and the detected object boundaries are more continuous.

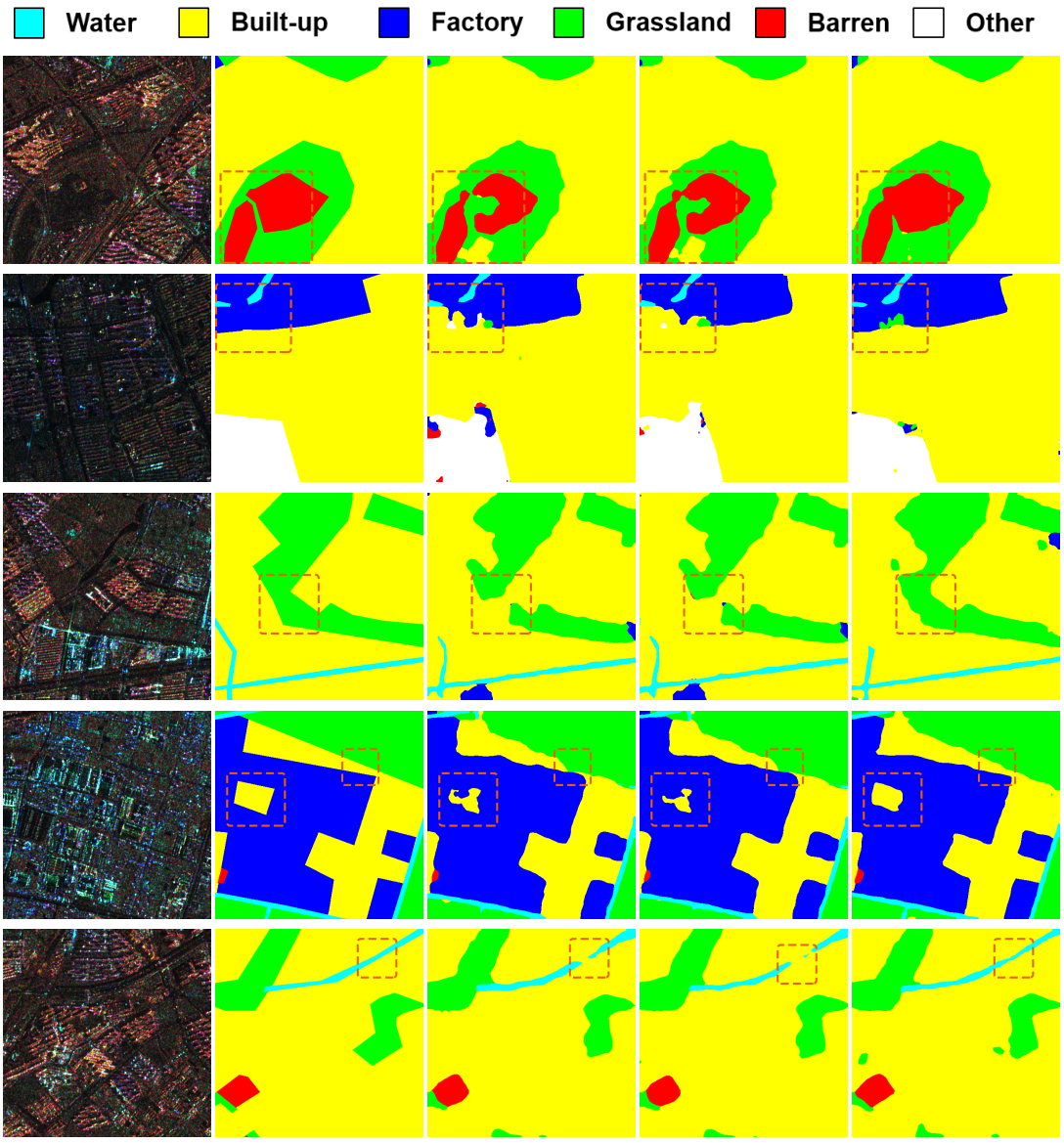


Figure 4.3: Comparison of segmented maps obtained by different methods on sample testing areas.

Table 4.2: Quantitative Results of the comparative study.

Methods	F_1(%)						mean F_1(%)	OA(%)	fwIoU(%)▼
	Water	Built-up area	Industrial area	Grassland	Barren	Others			
SegNet[8]	45.78	24.13	36.42	65.80	0.00	0.00	28.69	45.79	36.25
HR-SARNet[114]	73.10	76.29	57.31	78.67	0.00	0.10	47.58	74.01	62.87
MS-FCN[120]	69.68	77.14	60.02	79.13	0.00	0.01	47.66	74.63	63.09
Inc-FCN[73]	83.62	74.90	52.92	78.58	4.41	15.05	51.58	73.74	62.97
UNet[89]	83.58	77.54	59.63	79.67	0.02	22.36	53.80	76.14	65.43
HRNet[110]	91.30	90.19	85.09	90.34	73.96	92.98	87.31	90.02	83.41
LinkNet[14]	90.18	91.63	88.43	91.11	74.95	92.90	88.20	91.17	85.19
PSPNet[150]	92.71	93.36	89.71	93.12	84.03	94.91	91.31	92.98	88.05
FCN[64]	91.89	93.53	90.26	93.20	84.79	94.28	91.33	93.08	88.14
DeepLabV3+[16]	93.44	93.67	90.37	93.41	85.15	**95.32**	91.90	93.38	88.50
MP-ResNet (proposed)	**94.85**	**94.29**	**90.78**	**93.93**	**86.03**	94.96	**92.47**	**93.95**	**89.63**

Table 4.3: Comparison of model size and computational cost expressed in terms of params and FLOPS, respectively.

Methods	HR-SARNet	MS-FCN	LinkNet	FCN	HRNet	DeeplabV3+	PSPNet	MP-ResNet (proposed)	UNet	Inc-FCN	SegNet
Params (Mb)	0.06	23.54	21.65	21.35	65.85	22.29	22.73	54.97	9.16	6.14	3.51
FLOPS (Gbps) ▼	6.12	21.73	27.76	90.97	93.64	94.84	95.55	115.93	221.68	329.37	341.06

Table 4.3 presents the sizes and computational costs of the compared methods. The FLOPS are calculated based on the input size [4, 512, 512] of the images in the Gaofen dataset. The literature methods for the semantic segmentation of SAR images (HR-SARNet and Inc-FCN) are generally shallow, thus and their sizes are relatively small. UNet, Inc-FCN and SegNet need the most FLOPS since they apply many convolution operations on the early-layer features. Although the parameter size of the MP-ResNet is much larger than the FCN, its FLOPS do not increase significantly.

4.5 Conclusion

The semantic segmentation of PolSAR images is challenging due to the intense speckle noise and the lack of large training datasets. Taking advantage of the open dataset from the Gaofen contest, we propose a MP-ResNet for the semantic segmentation of high-resolution PolSAR images. Compared to the baseline FCN, the MP-ResNet has three parallel semantic embedding branches to strengthen the aggregation of context information. It also adopts a multi-scale feature fusion design in its decoder to take advantage from each encoding branch. As a result, the VRF of

the MP-ResNet is significantly enlarged, thus allowing the aggregation of discriminative features from a wider range and alleviating the impact of noise. The multi-fold comparisons with the baseline method conducted on the Gaofen dataset has proved the effectiveness of our designs. The comparative experiments with several state-of-the-art methods show that the proposed method has significant improvements in all accuracy metrics.

Since the objectiveness of this chapter is to propose a general architecture for the semantic segmentation of PolSAR images, we did not add sophisticated designs. In future works we plan to integrate context aggregation designs (e.g. attention modules in [26] and [74]) into the higher branches of the MP-ResNet to model the long-range context information. Moreover, the polarimetric characteristics of PolSAR images are not modelled in the proposed method. Incorporating structural and texture features in PolSAR images [86] may lead to further accuracy improvements, which is left for future studies.

Chapter 5

Direction-aware Residual Network for Road Extraction in VHR RSIs

This chapter [1] *presents a novel CNN model for road extraction in VHR RSIs, which considers the geometric characteristics of roads to better detect the non-salient and occluded ones. The proposed DiResNet includes three novel designs, including: i) an asymmetric residual segmentation network with deconvolutional layers and a structural supervision to enhance the learning of road topology (DiResSeg); ii) a pixel-level supervision of local directions to enhance the embedding of linear features; and iii) a refinement network to optimize the segmentation results (DiResRef). Ablation studies and comparative experiments on two benchmark datasets have confirmed the effectiveness of the presented model. It improves in particular the continuity of segmented road networks.*

5.1 Introduction

Road extraction from VHR RSIs is essential for the mapping and updating of Geographic Information Systems (GIS). This task has been studied for decades but we have not satisfactory automatic solution yet. This is due to the special characteristics of roads. Compared with other compact ground objects (such as buildings and water), roads in VHR RSIs appear to be elongated regions with similar spectral and texture patterns. Additionally,

[1] This chapter appears in:
[J3] L. Ding, L. Bruzzone, "DiResNet: Direction-aware Residual Network for Road Extraction in VHR Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 59, In press, 2021.

roads have fixed width and limited curvatures, and they are not suddenly interrupted [112]. To model these geometric features, the road extraction algorithms are expected to have a certain level of optimization and regularization of the results to reduce the discontinuities and false alarms.

Conventional expert-knowledge based methods for road extraction usually combine multiple edge detection, tracking, region clustering and filtering algorithms to obtain integrated results, since any single algorithm cannot model the complex structure of roads [112]. This often makes the results parameter-dependent and leads to error accumulation problems. The rise of CNNs makes it possible to model roads in an end-to-end manner and generalize the results to large volumes of data. Accordingly, due to the great feature embedding power of CNNs, the accuracy of road extraction has been significantly improved. In consequence, since around 2017 the CNN-based methods have been the mainstream in road extraction [20].

In the CNN-based approaches, road extraction is viewed as a binary segmentation problem. Cascaded convolutional layers are employed to model the spectral and spatial distribution of roads, followed by a classifier to densely discriminate the pixel categories (roads or non-roads). Compared to conventional methods based on hand-crafted features, if properly trained on a large number of representative annotated samples, CNNs are able to learn high-level semantic features of the roads automatically, and thus can be considered as a powerful feature extractor and classifier.

Some remaining problems in CNN-based road extraction methods are the recognition of the spectral outliers and the recovering of occluded areas (e.g. caused by shadows, trees, buildings and vehicles). These problems have been alleviated due to the encoding design of CNNs that aggregates local contextual information. However, there are still discontinuities in road segmentation maps. A possible solution to these problems is to enhance the embedding of linear features within the CNN architectures.

In This chapter we address the above-mentioned problems by proposing a direction-aware residual network that adds a supervision to force the network to learn directional features. In this way, the learned network is direction-sensitive and the linear features are strengthened. Moreover, most literature works employ UNet-like architectures [89]. They typically contain symmetric designs with connections to the low-level features to re-

cover the spatial details. Although this skip connection design can provide spatial details, we argue that it has the side effect of aggravating the occlusions and fragmenting the results. By contrast, we employ ResNet as the backbone network with two additional designs: a structural loss function to enhance the learning of the road topology, and a decoder network to smoothly enlarge the feature maps. Additionally, a refinement sub-net is designed to optimize the segmentation results.

To summarize, the main contributions in This chapter are as follow:

1. Designing a residual segmentation network (DiResSeg) with deconvolutional layers and structural supervision for the task of road extraction. This network design is aimed at enhancing the structural completeness of the road networks.

2. Introducing a direction supervision to the network. This enables the learned model to be direction-aware, thus strengthens the detection of linear features.

3. Introducing a refinement sub-net (DiResRef) to optimize the road extraction results.

4. Performing ablation studies and comparative experiments on two benchmark datasets (the Massachusetts dataset and the DeepGlobe dataset) to verify the effectiveness of the introduced designs and the overall architecture (DiResNet).

The remainder of this chapter is organized as follows. Section 5.2 illustrates the proposed method in details. Section 5.3 describes the implementation details and the experimental settings. Section 5.4 presents the results and analyzes the effect of the proposed method. Section 5.5 draws the conclusions of this chapter.

5.2 Proposed Direction-aware Residual Network

In this section we present a direction-aware residual network (DiResNet) integrating several network designs and auxiliary supervisions. We illustrate first the network designs and then the auxiliary supervisions.

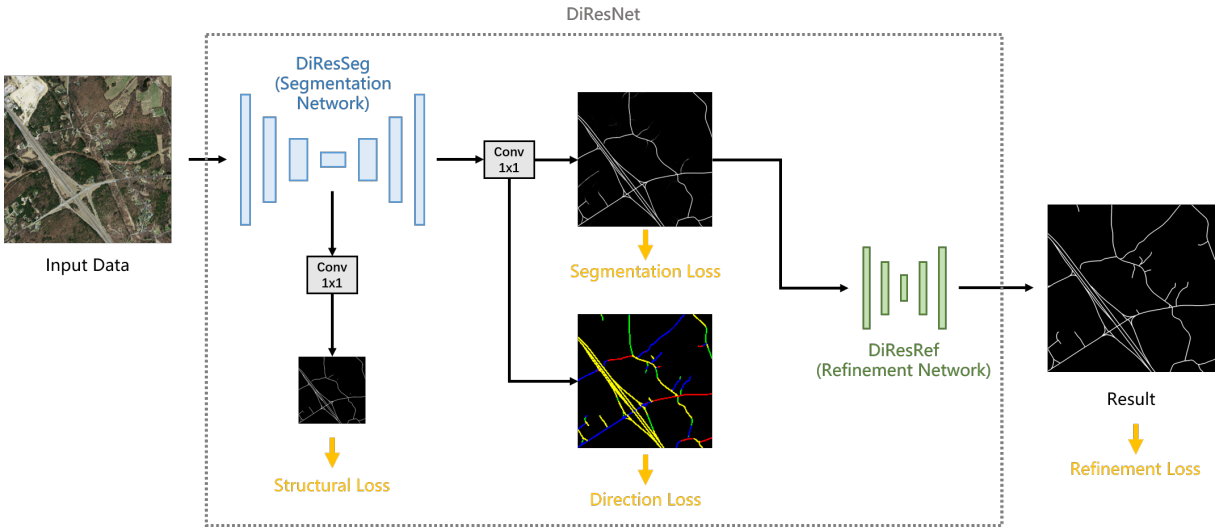


Figure 5.1: The proposed direction-aware residual network (DiResNet).

5.2.1 Network architecture

DiResSeg: Designed Segmentation Network

An overview of the designed direction-aware residual network (DiResNet) is shown in Fig.5.1. The network consists of two sub-nets: i) a segmentation network (DiResSeg) for the coarse segmentation of roads, and ii) a refinement network (DiResRef) to optimize the segmentation results. There are also two auxiliary supervisions in the network: i) a structure supervision in the middle of the segmentation network, and ii) a direction supervision from a parallel branch of the decoder module.

Most literature works on road extraction adopt UNet-like architectures. It has been believed that the multi-scale concatenation of low-level features can improve the performance. However, recent studies found that the integration of low-level features contributes little to binary segmentation tasks, whereas it does pass through noisy information [111] and increases the computational costs [121]. For the task of road extraction, we expect the extracted objects to be continuous and smooth elongated regions with fixed width, while the pixel-level accurate segmentation of road boundaries is not necessary. The low-level features are generally more noisy (due to occlusions and spectral outliers). The skip connections with them may lead to uneven boundaries and interruptions. Therefore, we argue that the

multi-level concatenation operations in UNet-like structures is unnecessary, if not disadvantageous, for road extraction. On the contrary, we emphasize on the embedding power of encoder networks, and present an asymmetrical encoder-decoder design with strengthened encoder and simplified decoder.

Fig.5.2 shows a comparison of our network design versus the UNet-like networks. They both contain an encoder network and a decoder module. Fig.5.2(a) shows the case of a UNet-like CNN with 4 encoding layers. Let us denote the encoded features as $\{E1, E2, E3, E4\}$, their corresponding spatial scaling ratios are 1, 1/2, 1/4 and 1/8, respectively. Accordingly, the decoder module contains 3 levels of upsampling and feature fusion operations. Each level of the decoded feature $D_i \in \{D_3, D_2, D_1\}$ is calculated as:

$$D_i = U\{F_U(E_{i+1}), E_i\}, \tag{5.1}$$

Where F_U denotes an upsampling or deconvolution operation and U denotes a convolution operation.

In our design in Fig.5.2(c), the encoder network is replaced to a layer-rearranged version of ResNet. Compared with the original FCN [64] with ResNet backbone (see Fig.5.2(b)), the striding operations in our network are designed inside three convolutional blocks so that the feature size is reduced gradually. It also contains a simple decoder with 3 serial deconvolutional layers to enlarge the feature map and smooth the boundaries. In this way, the predicted road maps are closely related to the high-level features of the encoder network. Compared with UNet-like architectures, the designed segmentation network (DiResSeg) gives more focus on the completeness of the road topology, rather than the extraction of road boundaries. An auxiliary supervision is further added to improve the training of the encoder network (see 5.2.2 for more details).

DiResRef: Designed Refinement Network

The coarsely segmented road maps may still contain interruptions and errors. An additional refinement process helps to optimize the segmentation results. An approach used in the literature work employs the tensor voting algorithm as a post-processing to the output of the CNN [32]. This algorithm is able to model the underlying spatial distribution pattern of images

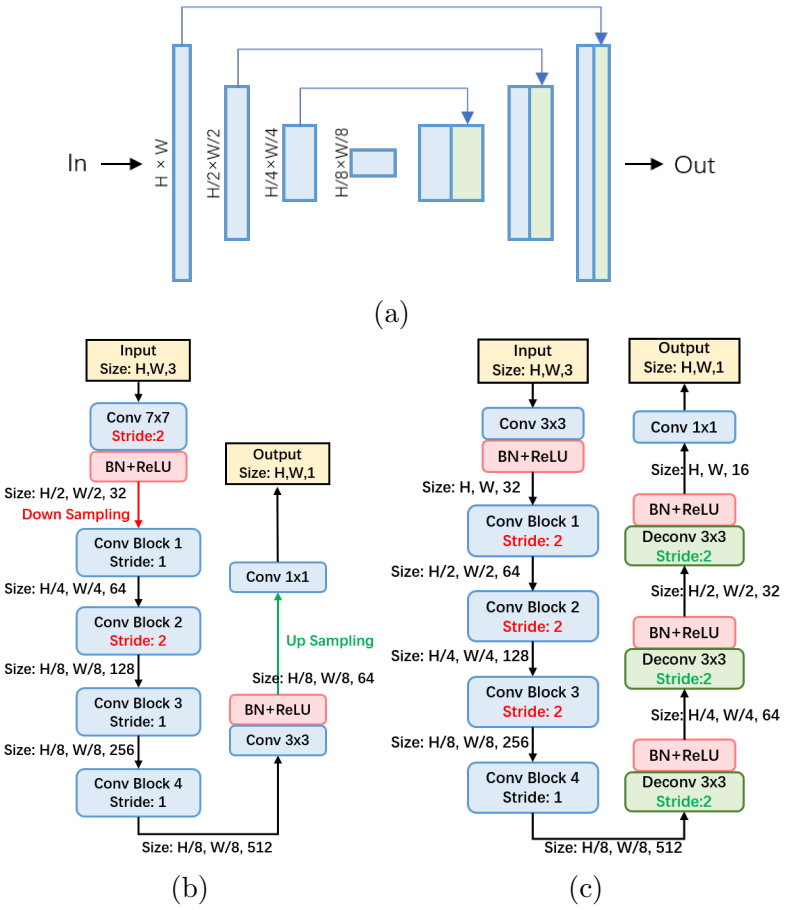


Figure 5.2: Illustration of the segmentation networks. (a) Traditional UNet-like Architecture, (b) FCN, (c) DiResSeg: the designed segmentation network.

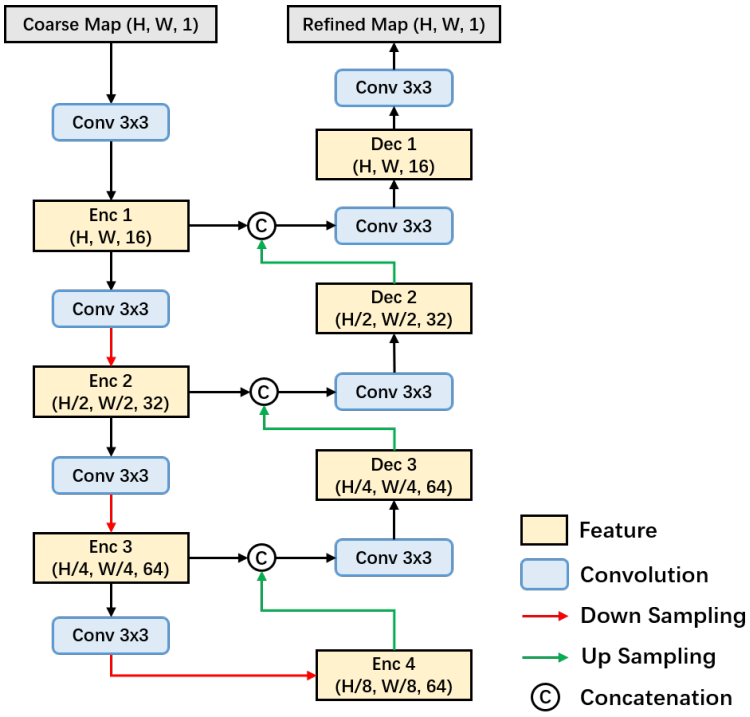


Figure 5.3: DiResRef: the designed refinement network.

and thus to connect the broken road segments. However, a limitation of the tensor voting algorithm is that it is based on a parameter-dependent deduction of the binary results, thus it is not stable and may produce false alarms. In this work we introduce a sub-net to perform the refinement, named as the DiResRef. It produces more stable results and makes the whole segmentation process end-to-end trainable.

The designed DiResRef is a UNet-like CNN inspired by the refine module presented in [87]. We change the striding rate and the number of channels per-layer to adapt the network to the task of road extraction. Fig.5.3 shows the design of the DiResRef. It operates on 4 sequential encoding levels with an increasing number of channels. The input of the network is the probability map produced by the segmentation network. It produces a residual feature map which strengthens the road-like regions and suppresses the non-road ones. This network optimizes the results in various aspects, including linking the possible interruptions, removing the isolated false alarms and increasing the probability salience of the road features.

5.2.2 Supervisions and Loss Functions

The proposed direction-aware network contains different supervisions: two segmentation supervisions, a structure supervision and a direction supervision. A hybrid loss $\mathcal{L}$ is calculated based on these supervisions as follow:

$$\mathcal{L} = \alpha\mathcal{L}_{seg} + \beta\mathcal{L}_{struct} + \gamma\mathcal{L}_{direct} + \theta\mathcal{L}_{ref} \tag{5.2}$$

where $\mathcal{L}_{struct}$, $\mathcal{L}_{direct}$, $\mathcal{L}_{seg}$ and $\mathcal{L}_{ref}$ denote the losses for the structure supervision, the direction supervision, the segmentation network (DiResSeg) and the refinement network (DiResRef), respectively. $\alpha, \beta, \gamma, \theta$ are 4 weight variables for the different losses.

Segmentation Supervisions

The segmentation results are expected to be probability maps, while the reference maps are binary. BCE loss is the most widely used function to measure the differences between predictions and targets, calculated as:

$$\mathcal{L}_{bce} = -\sum_{(r,c)} T(r,c)log[P(r,c)] + [1 - T(r,c)]log[1 - P(r,c)] \tag{5.3}$$

where $T(r,c) \in \{0,1\}$ is the target value of pixel (r,c) and P(r,c) is the predicted probability value.

Structure Supervision

This is an auxiliary supervision added at the highest level of the encoder network, related to the down-scaled feature maps. The reference data for the structure supervision are generated by down sampling the GT maps (using the area interpolation). In the down-scaled maps, the width of roads is greatly reduced, thus the road boundaries are obscure. Therefore, a supervision at this level gives more attention to the center road pixels, thus strengthening the geometric structure of roads. Additionally, this auxiliary supervision is beneficial for improving the training stability of the encoder network. We deem the embedding of road structures as a regression problem and use the L1 loss to measure the structural differences:

$$\mathcal{L}_{struct} = \sum_{(r,c)} |P_s(r,c) - T_s(r,c)| \tag{5.4}$$

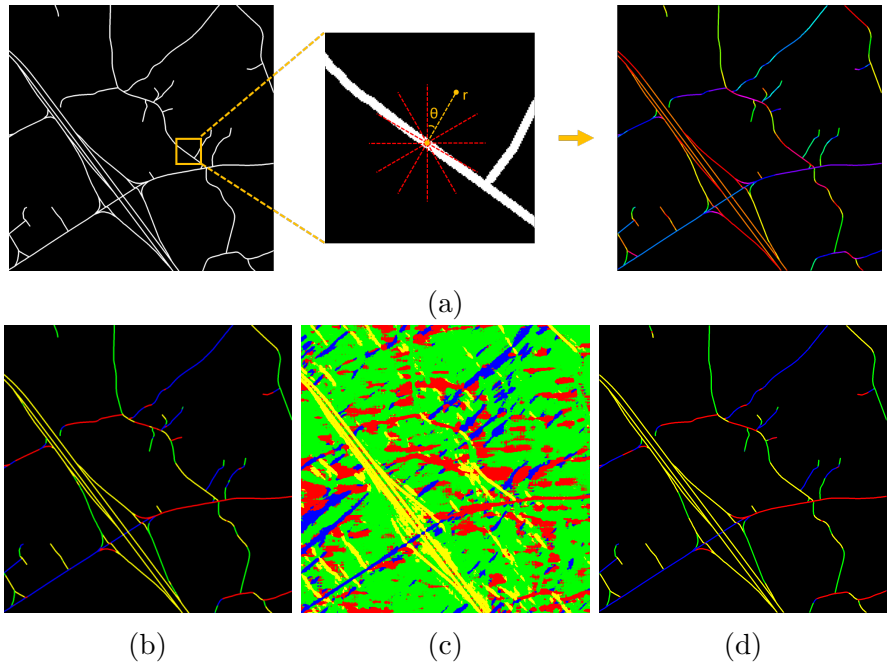


Figure 5.4: Generation of the reference direction map. (a) Calculation of the local direction, (b) Generated reference direction map (4 main directions only), (c) Predicted direction map, (d) Predicted direction map (without background pixels).

where $P_s(r,c) \in [0,1]$ and $P_s(r,c) \in [0,1]$ are the pixel values at the scaled target and prediction maps, respectively.

Direction Supervision

A previous study has found that learning the road orientations is beneficial to improve the connectivity of road segmentation results [9]. However, this study calculates the road orientations based on vector data, which does not apply to raster reference data. We extend this idea to common road segmentation tasks by generating the reference direction maps from the binary GT maps.

Fig.5.4 illustrates the algorithm to generate reference direction maps. The angular operators are used to measure the local direction of road pixels [28]. The algorithm is illustrated in Algorithm 1.

The parameters r and Δ_θ are selected based on the minimum and maximum pixel width of the roads in T. The algorithm is implemented using convolutional layers with fixed weights, so that the reference maps can be

Algorithm 1 Algorithm for Generating the Direction Map.

Input: Binary GT road map T;
Parameters: detecting radius r, angle step Δ_θ;
Output: Reference road direction map T_d;

for $T(i,j)$ in T **do**
  if $(T(i,j)=1)$ **then**
    for $\theta = 0$ to π step Δ_θ **do**
      $d_\theta(i,j) = \sum_{\rho=1}^{r} T(\rho\sin\theta, \rho\cos\theta) + T(-\rho\sin\theta, -\rho\cos\theta)$
    end for
    find θ_{max} that:
      $d_{\theta_{max}}(i,j) = max\{d_\theta(i,j)\}, \theta \in [0,\pi]$
    $T_d(i,j) = \theta_{max}$
  else
    $T_d(i,j) = invalid$
  end if
end for
return T_d

generated dynamically during the training phase. To avoid under-fitting problems, we clip the target direction map T_d to a 5 channel map representing 4 major directions and the non-road label ($T_d(r,c) \in \{0,1,2,3,4\}$). The multi-class cross entropy loss is used to evaluate the predicted direction map P_d:

$$\mathcal{L}_{direct} = \sum_{(r,c)} \sum_{i=1}^{N_d} \{-P_d(r,c)[d_i] * log[\sum_{j=1}^{N_d} exp(P_d(r,c)[j])]\} \tag{5.5}$$

where N_d is the number of road directions. To encourage the modeling of linear features over all areas, the non-road labels are neglected in l_{direct}.

This direction supervision is connected to a parallel branch of the segmentation network, thus the learned direction features can contribute to the segmentation results.

5.3 Dataset Description and Design of Experiments

In this section we describe the experimental dataset, the implementation details and the evaluation metrics.

5.3.1 Datasets Descriptions

Two datasets are selected for experiments: the Massachusetts Dataset and the DeepGlobe Dataset. These are by far the two largest datasets openly available for road segmentation in VHR RSIs.

Massachusetts Dataset [71]

This is an aerial dataset collected in Massachusetts, US., covering an area of 2.25 square kilometers. The GSD of this dataset is 1.2m per pixel. There are 1171 images in total, among which 1108 images are for training, 14 ones for validation and the remaining 49 ones for testing. Each image has 1500 $\times$ 1500 pixels. The imaged regions include urban, suburban and rural scenes. The reference maps are generated by rasterizing the vector data of road centerlines, so each road has a fixed width.

DeepGlobe Dataset [23]

This is a satellite dataset containing images collected in Thailand, Indonesia and India. It covers a total land area of 2220 km^2, containing both urban and suburban areas. The GSD of this dataset is 50cm per pixel. Each image has a size of 1024 $\times$ 1024 pixels. The original dataset contains 8570 images, among which 6226 training images are openly available with GT data. Therefore, we further divide the accessible data into the training and testing set with a ratio of 5:1. 5189 images are selected for training, the remaining 1037 images for testing. Compared to the Massachusetts dataset, more types of road surfaces are contained, thus road extraction in this dataset is more challenging.

5.3.2 Implementation Details

All the experiments are performed on a workstation with 32 GB RAM and a NVIDIA Quadro P6000 GPU (23GB). The designed networks are implemented using the PyTorch library. Due to the limitation of GPU memory, the training is performed using cropped images with the spatial size of 320 $\times$ 320 pixels. To avoid the over-fitting problem, dynamic cropping and flipping operations are performed as augmentations to the dataset. The

Figure 5.5: Illustration of the considered dynamic data augmentation process.

input images are first loaded and stored in the memory. During each training iteration, they are randomly cropped and flipped before being loaded to CNNs. Fig.5.5 shows our data augmentation strategy. In our implementation, each training image generates 10 cropped patches during each epoch, while the training takes 50 epochs. The training batch size is set to 16. During the validation and testing phase, full-size input images are used instead (without the cropping operation) to avoid the impact of cropping parameters. The parameters $\alpha, \beta, \gamma, \theta$ in formula 5.2 are empirically set to 1.0, 0.5, 0.1 and 1.0, respectively. The l_{direct} is assigned with a lower weight since its values are bigger.

Since the experimental datasets have different GSDs, we empirically chose different scaling rates for them in the implemented networks. The minimum down sampling rates are 1/8 and 1/16 for the Massachusetts and DeepGlobe datasets, respectively. We adopt ResNet34 as the backbone network to perform experiments on both datasets. Since road segmentation is a single-class segmentation problem that does not include complex modeling of the semantic information, the number of layers in ResNet34 is powerful enough to embed road features. The chosen down-sampling rates and ResNet backbone are implemented in all the compared methods to ensure fairness.

5.3.3 Evaluation Metrics

To give a comprehensive evaluation on the considered methods, seven evaluation metrics are adopted: Precision (P), Recall (R), F_1 score, Break Even Point (BEP), mean IoU, OA and connectivity. These are the most widely used measurements in both road extraction and other binary segmentation

tasks [60]. They are calculated as follows:

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN} \tag{5.6}$$

$$F_1 = 2 \times \frac{P \times R}{P + R}, OA = \frac{TP + TN}{TP + FP + TN + FN} \tag{5.7}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{5.8}$$

where TP, FP, TN and FN represent true positive, false positive, true negative and false negative, respectively. Since there is a negative correlation between the values of precision and recall (under different thresholds), we also use the break-even point as a measurement. The BEP is defined as the intersection point on the precision-recall curve, where the values of precision and recall are equal. Apart from the measures based on pixel statistics, a road connectivity measure (denoted as Conn) has also been adopted to evaluate the topology of road extraction results [118]. It is calculated as follows:

$$Conn = \frac{2N_{conn}}{N_{GT} + N_{pred}} \tag{5.9}$$

where N_{conn}, N_{GT} and N_{pred} are the number of connected segments, the total number of segments on GT map and the total number of segments on the prediction map, respectively.

5.4 Experimental Results

This section presents the experimental results obtained on the Massachusetts roads dataset. First an ablation study is performed to test the modules and auxiliary supervisions. Then the effects of the direction supervision and the refinement network are analyzed. Finally we compare the proposed direction-aware residual network with several literature works and analyze the results.

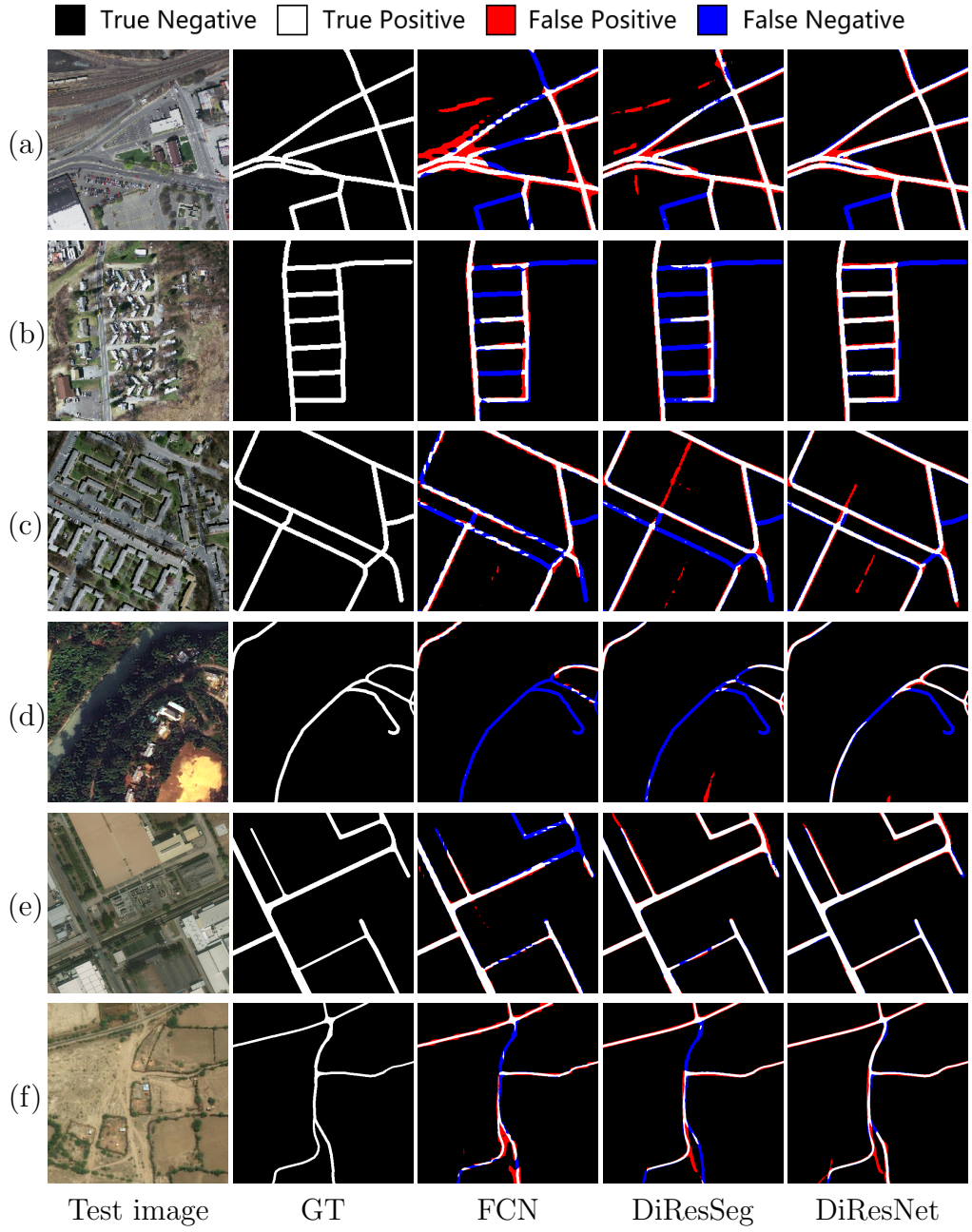


Figure 5.6: Example of the segmentation results (ablation study). (a)-(c) Results selected from the Massachusetts dataset, (d)-(f) Results selected from the DeepGlobe dataset.

Hyper-parameters	F1 under different weights			
	0	0.2	0.5	1.0
$\beta(\gamma = 0)$	78.35	78.43	**78.55**	78.49
$\gamma(\beta = 0)$	78.35	**78.69**	78.47	78.29
$\beta = 0.5, \gamma = 0.2$		**79.09**		

Table 5.1: The F_1 under different hyper-parameters tested on the DeepGlobe dataset.

Table 5.2: Results of the ablation study related to the proposed DiResNet (Massachusetts dataset).

	Method	Components		Supervisions		OA(%)	P(%)	R(%)	F_1(%)	Conn(%)
		DiResSeg	DiResRef	Structure	Direction					
	FCN [64]					97.76	74.10	80.68	77.09	80.47
proposed	FCN-R		√			97.91	77.26	79.50	78.19	80.57
proposed	DiResSeg	√				98.07	79.05	80.48	79.61	81.01
proposed	DiResNet-R	√	√			98.08	79.03	**80.80**	79.77	**81.35**
proposed	DiResNet-S	√		√		98.08	79.12	80.52	79.67	81.25
proposed	DiResNet-D	√			√	98.11	79.20	80.65	79.73	81.29
proposed	DiResNet	√	√	√	√	**98.13**	**80.12**	80.29	**80.06**	81.20

5.4.1 Ablation Study

Influence of Hyper-Parameters. The hyper-parameters α, β, γ and θ in Eq.(5.2) balance between the segmentation losses and the auxiliary losses. Since α and θ are directly related to the training of the DiResSeg and the DiResRef, we set their values to 1 to ensure that they are the primary losses. The weight of auxiliary losses, i.e. β, γ, are scaled from 0 to 1.0 to find the best values. The tests are performed on the DeepGlobe dataset and are reported in Table.5.1. We find that the set of hyper-parameters $\{1.0, 0.5, 0.2, 1.0\}$ leads to the best accuracy, which is fixed to train the DiResNet.

Quantitative results. To test the effectiveness of the proposed direction-aware residual network including both the sub-nets and the auxiliary supervisions, we performed an ablation study. Since the DiResSeg is a modified version of the baseline FCN [64], the latter is also used in the comparison to evaluate the improvements.

Table 5.2 reports the quantitative results of the ablation study on the Massachusetts dataset. DiResNet-R, DiResNet-S and DiResNet-D refer to the DiResSeg with the DiResRef, the structural supervision and the direc-

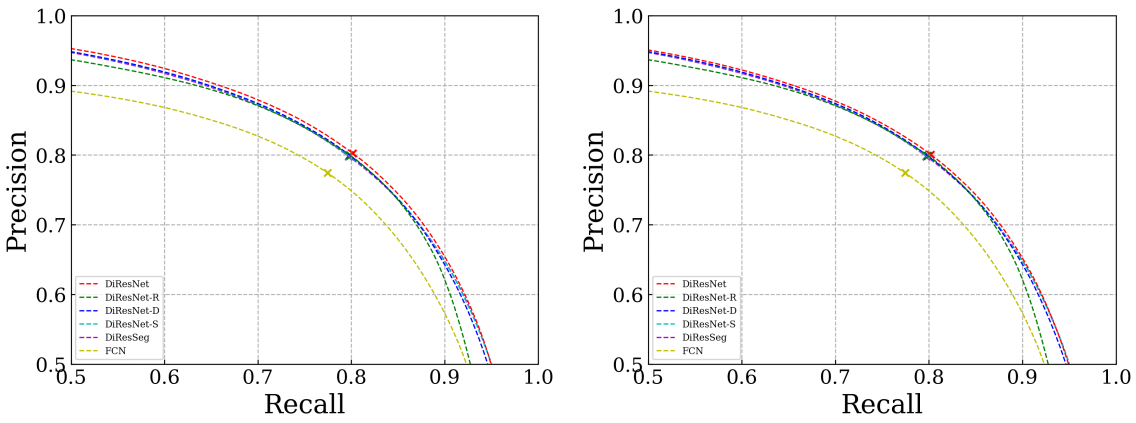


Figure 5.7: Accuracy curves of the ablation study (Massachusetts dataset). (a) Precision-recall curves, (b) OA curves.

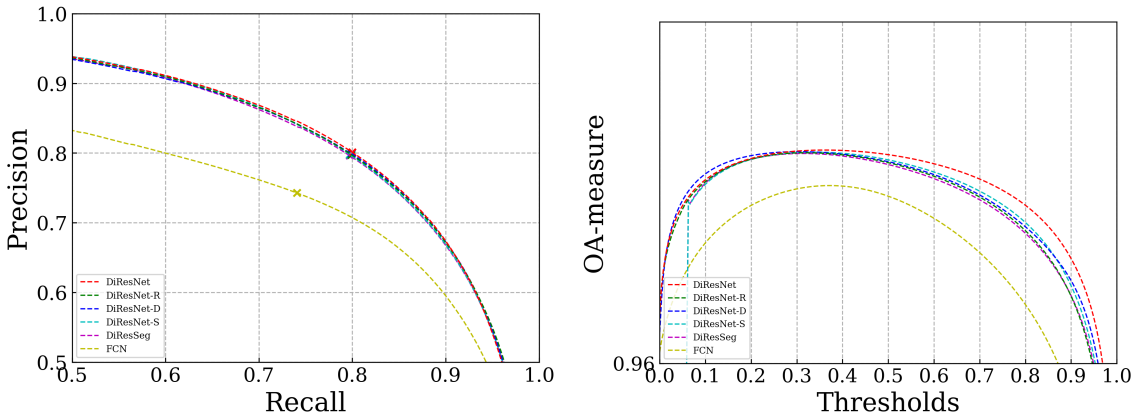
tion supervision, respectively. We also add the DiResRef to the baseline FCN (referred as FCN-R) to test its improvements.

Compared to the baseline FCN, DiResSeg shows a significant improvements in both OA and F_1 measures. The use of DiResRef in DiResNet-R improves the results of 0.16% in F_1 and 0.34% in Conn. Its use in the FCN-R also improves the accuracy of FCN (0.15% in OA and 1.1% in F_1). Compared to the DiResSeg with only the segmentation network, DiResNet-S and DiResNet-D have advantages of 0.06% and 0.12% in the F_1 measure, respectively. The DiResNet with all auxiliary designs shows an increase of 0.45% in the F_1 measure and 0.06% in the OA. Compared with the baseline FCN, the DiResNet has an advantage of 0.37% in OA, 2.97% in F_1 and 0.73% in Conn. Figure 5.7 shows the accuracy curves of the ablation study in the Massachusetts dataset. Figure 5.7(a) shows the precision-recall curve, while Figure 5.7(b) shows the calculated OA under different thresholds. One can observe that the designed DiResNet has a great advantage compared to the baseline FCN. The red curve represents the DiResNet, which has the biggest areas in both graphs.

Table 5.3 reports the quantitative results of the ablation study on the DeepGlobe dataset. Compared to the baseline FCN, the designed DiResSeg shows an improvement of 4.52% and of 0.37% in the F_1 measure and the OA, respectively. The FCN-R also has notable improvements over the FCN

Table 5.3: Results of the ablation study related to the proposed DiResNet (DeepGlobe dataset).

	Method	Components		Supervisions		OA(%)	P(%)	R(%)	F_1(%)	Conn(%)
		DiResSeg	DiResRef	Structure	Direction					
	FCN [64]					97.95	69.95	81.05	73.83	68.16
proposed	FCN-R		√			98.31	77.85	79.59	77.43	75.64
	DiResSeg	√				98.32	75.77	83.61	78.35	75.46
	DiResNet-R	√	√			98.35	76.25	83.49	78.56	**76.49**
	DiResNet-S	√		√		98.38	77.53	82.08	78.55	75.52
	DiResNet-D	√			√	98.36	76.29	**83.62**	78.69	75.70
	DiResNet	√	√	√	√	**98.44**	**78.76**	81.46	**79.09**	75.90



(a) (b)

Figure 5.8: Accuracy curves of the ablation study (DeepGlobe dataset). (a) Precision-recall curves, (b) OA curves.

(3.6% in F_1 and 0.36 in OA). Compared to the DiResSeg, the DiResNet-R, DiResNet-S and DiResNet-D have increased the F_1 measure of 0.21%, 0.20% and 0.34%, respectively. DiResNet shows an increase of 0.45% in the F_1 measure and 0.06% in the OA compared to the DiResSeg. Its improvement over the baseline FCN is 3.35% in F_1 measure, 0.43% in OA and 7.74% in Conn. The precision-recall and OA curves of the compared methods are reported in Fig.5.8. The DiResNet and its variations achieve significant advantages over the baseline FCN.

Qualitative results. Fig.5.6 shows the segmentation results on several testing areas. Compared with the baseline FCN, the segmentation maps of DiResSeg are smoothed and less fragmented due to its deconvolutional layers. However, there are still many false alarms and interruptions in the segmented road maps. With the use of auxiliary designs in DiResNet, the false alarms are reduced. Additionally, the connectivity of roads is greatly improved.

As a conclusion, the ablation study has proved the effectiveness of the proposed designs. The segmentation network (DiResSeg) has made significant improvements over the baseline ResNet. The designed auxiliary components have brought improvements in different dimensions. Specifically, the DiResNet-S emphasizes the structural completeness of road networks and improves the precision of results. The DiResNet-D explicitly uses local direction supervision to force the network to capture more linear features, which is beneficial to the recognition of roads. The DiResNet-R learns to repair some of the broken road segments, which greatly improves the road connectivity. Combining these advantages, the accuracy of the DiResNet is greatly increased, and it achieves the best F_1 and OA on both datasets.

5.4.2 Analysis of the Effect of DiResRef

As presented in Table 5.2, the DiResRef increases F_1 measure of DiResSeg by 0.16% in the Massachusetts dataset and by 0.21% in the DeepGlobe dataset. The DiResNet-R has the best Conn measures on both datasets, which proves its capability to improve the road topology. To qualitatively evaluate the DiResRef, in Fig.5.9 we compare the obtained results before and after its use. The optimization of DiResRef is based on the road

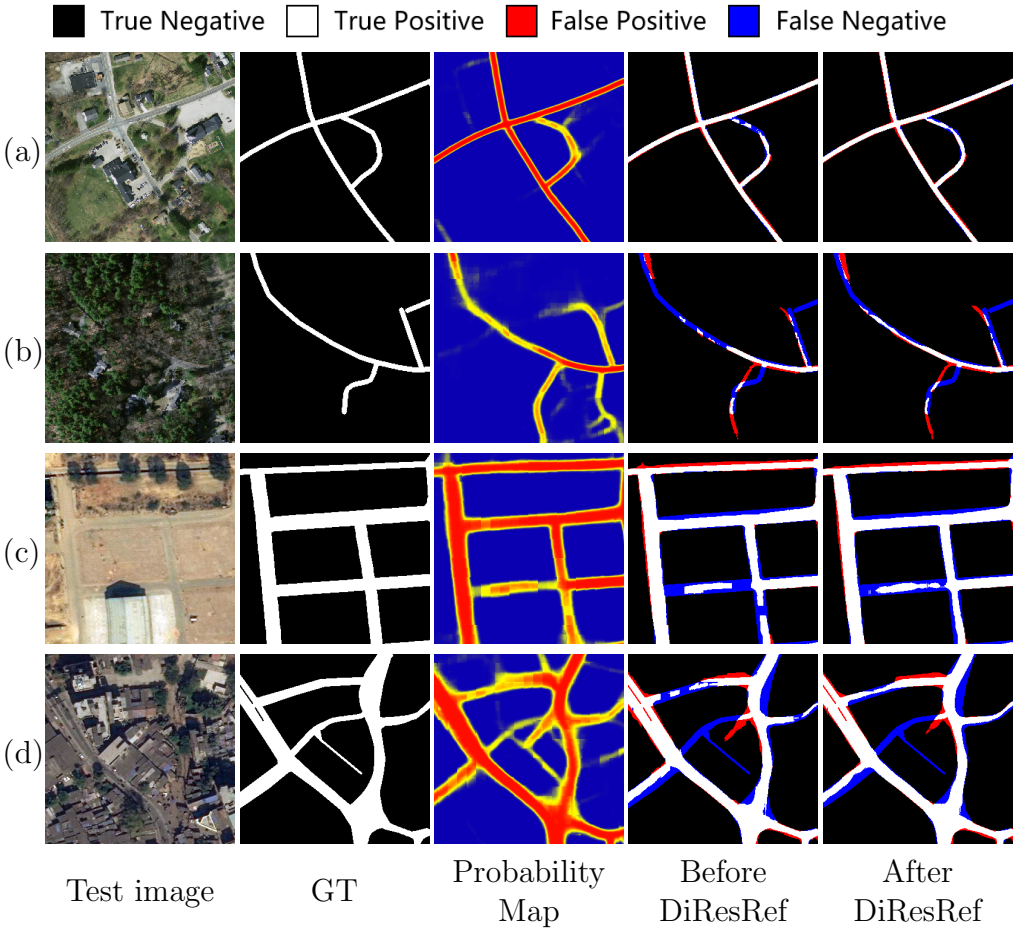


Figure 5.9: Examples illustrating the effect of DiResRef on sample testing areas. (a), (b) Results selected from the Massachusetts dataset, (c), (d) Results selected from the DeepGlobe dataset.

probability maps produced by the network. Fig.5.9(a) and (b) show two sample areas in the Massachusetts dataset affected by occlusions (caused by trees). In the original road maps produced by the DiResSeg, there are interruptions on the roads. After using the DiResRef, some of the interrupted segments have been connected and the results are more complete. Fig.5.9(c) and (d) show two cases of wide roads in the DeepGlobe dataset. One can observe that the use of DiResRef not only connects some of the interruptions, but also smooths the road boundaries.

5.4.3 Analysis of the Effects of Auxiliary Supervisions

To visually assess the effects of auxiliary supervisions, in this section we compare the results obtained with and without their use. Fig.5.10 presents several examples of the effects of the structural supervision. This supervision emphasises the structural completeness of the road network. In some cases it improves the segmentation of road center pixels (e.g., Fig.5.10 (a),(b)). In other cases when there are uneven pixels near the road boundaries, it also reduces the false alarms and makes the results more smooth (e.g., Fig.5.10 (b),(c)).

In Fig.5.11 we compare the segmentation results with and without the use of the direction supervision. The direction salience maps are generated by adding the 4 channel outputs of the predicted directions, which imply the linear features learned by the network. One can observe that the use of direction supervision enables the network to better embed the linear features. Some of the road features are not obvious due to their spectral characteristics. After considering their linear salience, these features become more distinguishable and are recognized by the DiResNet-D. One of the side effects of the direction supervision is that it also enhances some of the non-road linear features.

5.4.4 Comparative Experiments

In this section we compare the proposed method with literature works. The compared methods include the baseline FCN [38], the CasNet [20], the original UNet [89]. CasNet is an early work on road segmentation us-

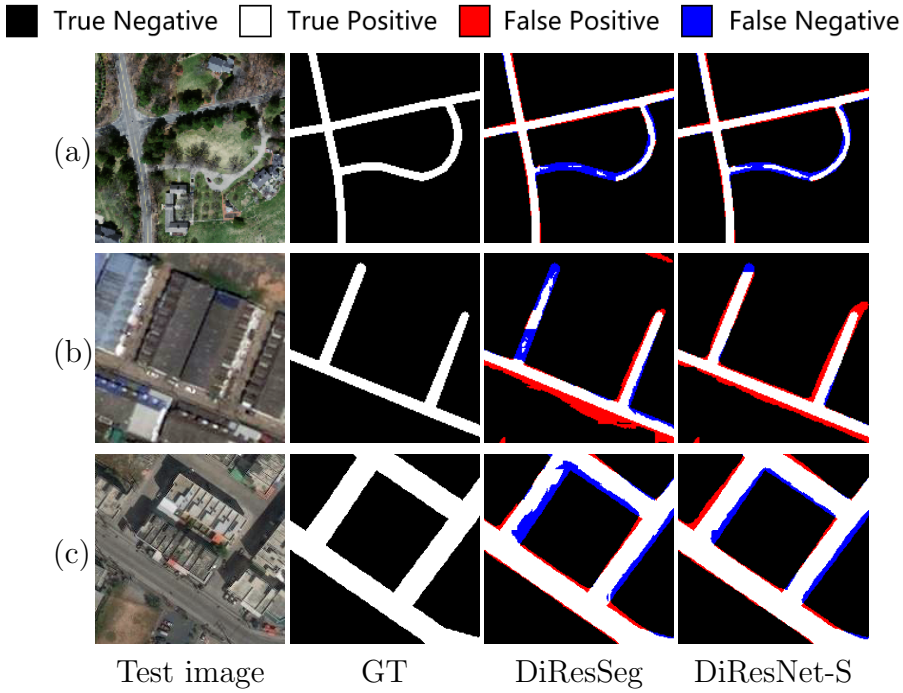


Figure 5.10: Examples illustrating the effect of structural supervision on sample testing areas. (a) Results selected from the Massachusetts dataset, (b), (c) Results selected from the DeepGlobe dataset.

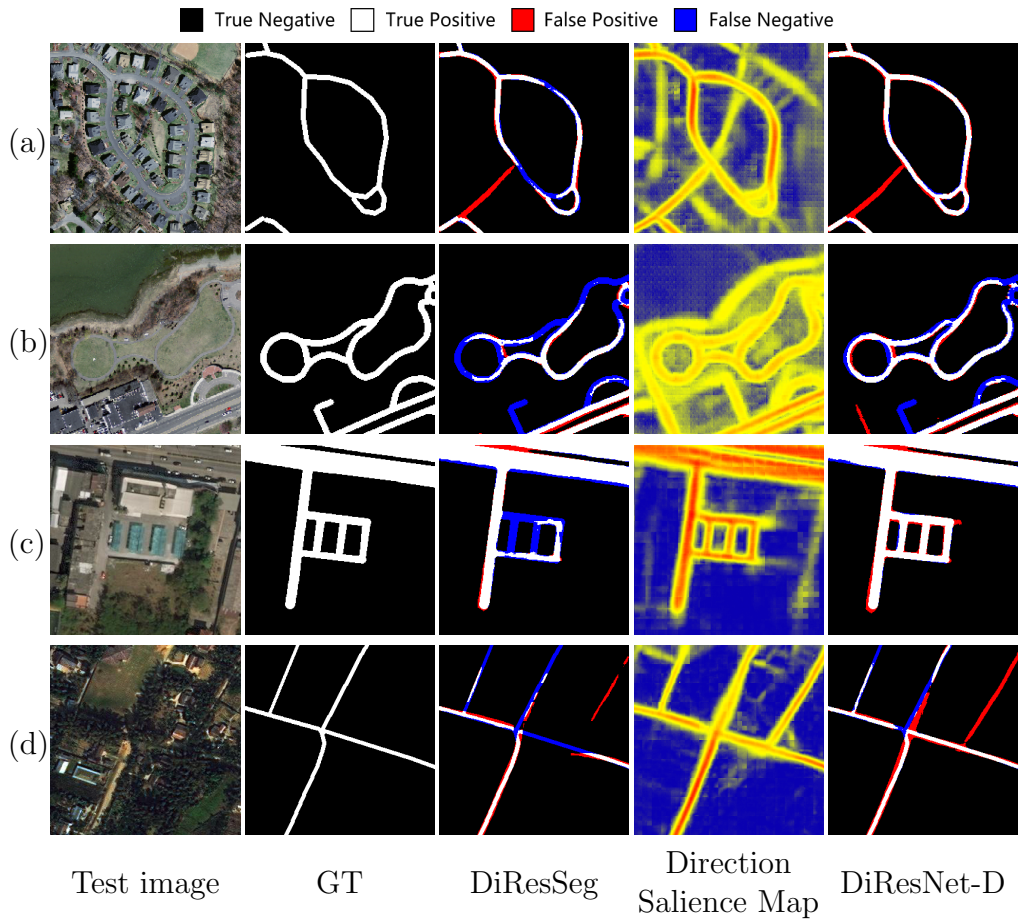


Figure 5.11: Examples illustrating the effect of direction supervision on sample testing areas. (a), (b) Results selected from the Massachusetts dataset, (c), (d) Results selected from the DeepGlobe dataset.

Table 5.4: Comparison of model size and calculations expressed in terms of params (Mb) and FLOPS (Gbps), respectively.

Methods	Params (Mb)	FLOPS (Gbps)
DLinkNet [153]	31.30	33.56
CasNet [20]	3.83	39.02
FCN [64]	21.32	90.63
Res-UNet [146]	8.22	182.21
UNet [89]	9.16	221.43
ASPP-UNet [146]	45.87	300.64
RCNN-UNet [146]	20.24	385.40
DiResSeg	21.32	115.31
DiResNet	21.49	121.20

ing a modified version of VGG-Net [95] as encoder. Several variants of the UNet are also used in the comparison, including the residual UNet (ResUNet) [146], the ASPP-UNet [37] and the recurrsive UNet (RCNN-UNet) [129]. The DLinkNet [153], which has achieved good performance in the DeepGlobe competition is also compared. The same data pre-processing procedures and parameter settings are used during the training for all networks. The FCN, DLinkNet and the proposed DiResNet and DiResSeg are all implemented using ResNet34 as the backbone network.

To fairly evaluate different methods, we first present a comparison of their computational costs in table 5.4. The FLOPS are calculated based on an input size of [3, 320, 320]. Among the compared methods, DLinkNet takes the least FLOPS since most of its convolutional layers operate at the scale of below 1/4 (width and height). The number of parameters in CasNet is the smallest. The FLOPS of DiResSeg and DiResNet are slightly more than those of the FCN. The UNet-like architectures require the highest number of FLOPS due to their multi-level concatenation designs, which include many large-size low-level features. The ASPP-UNet and RCNN-UNet require about three times the FLOPS of FCN and DiResNet.

Table 5.5 reports the quantitative results of the compared methods on the Massachusetts dataset. The accuracy provided by the FCN is lower than those of other methods. This is mainly due to the sequential down-sampling operations in its early layers. The UNet has an advantage of 0.07% in BEP compared with CasNet, but its computational cost is signif-

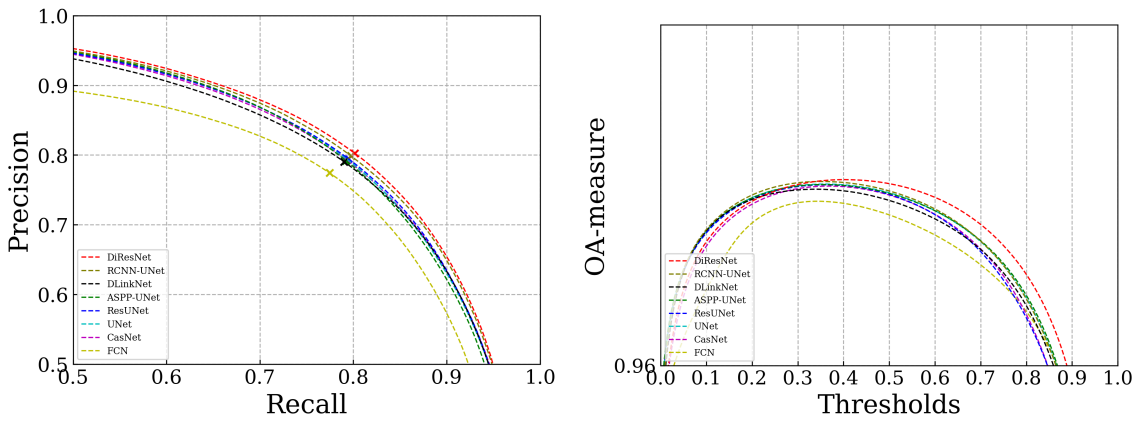
Table 5.5: Results of the comparative experiments (Massachusetts dataset).

Method	OA(%)	P(%)	R(%)	BEP(%)	F_1(%)	mIoU(%)	Conn.(%)
FCN [64]	97.76	74.10	80.68	77.44	77.09	62.97	80.47
CasNet [20]	97.99	77.65	80.87	79.29	79.06	65.63	78.27
UNet [89]	98.02	78.20	80.46	79.36	79.10	65.69	78.77
ResUNet [146]	98.00	77.69	81.14	79.47	79.23	65.85	78.51
ASPP-UNet [37]	98.01	78.19	80.13	79.21	78.97	65.50	78.32
DLinkNet [153]	97.93	76.42	81.54	79.03	78.76	65.23	80.57
RCNN-UNet [129]	98.04	77.78	**81.71**	79.82	79.56	66.31	81.12
DiResSeg	98.07	79.05	80.48	79.77	79.61	66.38	81.01
DiResNet	**98.13**	**80.12**	80.29	**80.21**	**80.06**	**67.00**	**81.20**

icantly higher. The residual design in ResUNet improves the F_1 of 0.13% and the BEP of 0.11% compared with the original UNet. DLinkNet and RCNN-UNet have achieved great improvement in Conn. The RCNN-UNet also achieves the best recall scores in all the compared methods. This can be attributed to its recursive design that greatly preserves spatial details.

Compared with the literature works, the designed DiResSeg obtains better results in terms of both F_1 and OA. Considering that its computation cost is significantly lower than those of UNet, ResUNet and RCNN-UNet, these improvements are remarkable. After adding the auxiliary supervisions and the DiResRef, the proposed DiResNet achieves the best performance in OA, BEP, F_1, mIoU and Conn. In greater detail, it has advantages of 0.50% in F_1, 0.09% in the OA and about 0.7% in mIoU compared with the RCNN-UNet. The precision-recall and OA curves of the compared methods are presented in Fig.5.12. The curves of the DiResNet cover the biggest areas, proving the highest accuracy performance of the proposed method.

Table 5.6 reports the quantitative results obtained on the DeepGlobe dataset. This dataset is more challenging since there are more types of road surfaces included. As a result, the UNet-like architectures (including UNet, ResUNet ASPP-UNet and RCN-UNet) show disadvantages compared with other methods that use powerful encoder networks. Compared with UNet, the CasNet with VGG-Net as its encoder has a advantage of 1.78% in the F_1 measure and 0.14% in the OA. The DLinkNet with both UNet and ResNet designs achieves the best accuracy among the literature



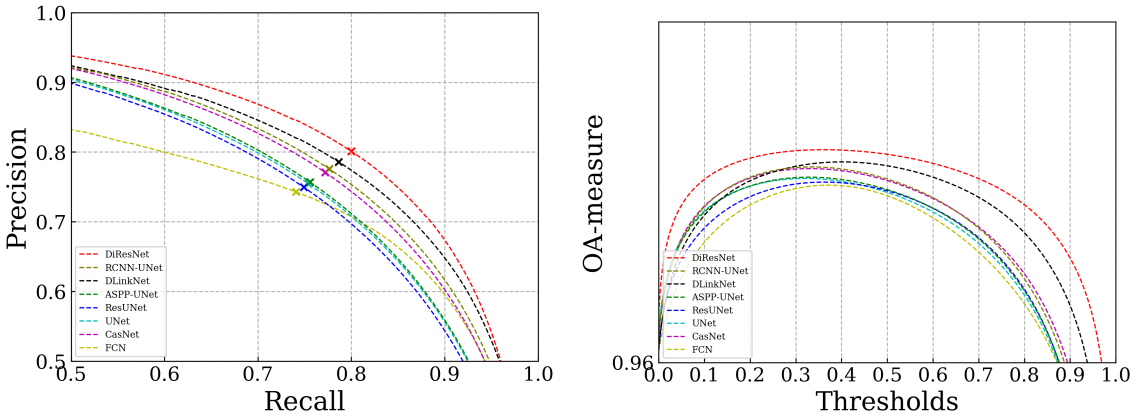
(a) (b)

Figure 5.12: Accuracy curves of the comparative study (Massachusetts dataset). (a) Precision-recall curves, (b) OA curves.

Table 5.6: Results of the comparative experiments (DeepGlobe dataset).

Method	OA(%)	P(%)	R(%)	BEP(%)	F_1(%)	mIoU(%)	Conn.(%)
FCN [64]	97.95	69.95	81.05	74.17	73.83	59.95	68.16
CasNet [20]	98.13	73.66	80.63	77.24	75.73	62.23	67.36
UNet [89]	97.99	71.82	79.07	75.43	73.95	60.07	60.66
ResUNet [146]	98.01	73.69	76.24	74.93	73.49	59.56	56.14
ASPP-UNet [37]	98.02	72.82	78.49	75.64	74.17	60.36	60.67
DLinkNet [153]	98.30	77.48	79.90	78.62	77.38	64.54	74.21
RCNN-UNet [129]	98.15	74.54	80.78	77.61	76.03	62.68	69.93
DiResSeg	98.32	75.77	**83.63**	79.67	78.35	65.73	75.46
DiResNet	**98.44**	**78.76**	81.46	**80.06**	**79.09**	**66.80**	**75.90**

methods. However, the proposed DiResSeg and DiResNet obtain significant improvements in all the measures. Their advantages in OA, F_1 and Conn are especially noticeable. The DiResNet also obtains great improvements in precision. We attribute these improvements to two factors: i) The DeepGlobe dataset has a relatively smaller GSD, thus the low-level concatenation designs in UNet-like architectures propagate more noise, which degrades the accuracy; ii) The auxiliary designs in DiResNet enhance the embedding of both road typologies and linear features, which greatly improves the precision. Fig.5.13 presents different accuracy curves of the compared methods. The OA of the proposed method is higher under all the thresholds.



(a) (b)

Figure 5.13: Accuracy curves of the comparative study (DeepGlobe dataset). (a) Precision-recall curves, (b) OA curves.

Fig.5.14 presents examples qualitative of the results obtained by different methods. The results of UNet and ResUNet are quite close. They both contain many interruptions and false alarms. In the areas with wide roads (e.g. Fig.5.14(e)), this disadvantage is more severe. The CasNet obtains smoother results on the DeepGlobe dataset (Fig.5.14(d)-(f)), but with more false alarms than on the Massachusetts dataset (Fig.5.14(d)-(f)). The DLinkNet shows better road connectivity compared with other methods, but it also results in more false alarms. The RCNN-UNet achieves very good results on the Massachusetts dataset, but it fails to segment many road segments on the DeepGlobe dataset. Compared with its competitors, the proposed DiResNet shows two major advantages: i) It produces less false alarms, as the auxiliary supervisions enhance the precision of the results. ii) It produces more complete and smooth results. Indeed, the DiResNet significantly reduces the broken segments and strengthens the linear features.

5.5 Conclusions

In This chapter we have studied the CNN-based extraction of roads in VHR RSIs and presented a DiResNet. The literature works mostly use UNet-like symmetric architectures, which are time-consuming and pass

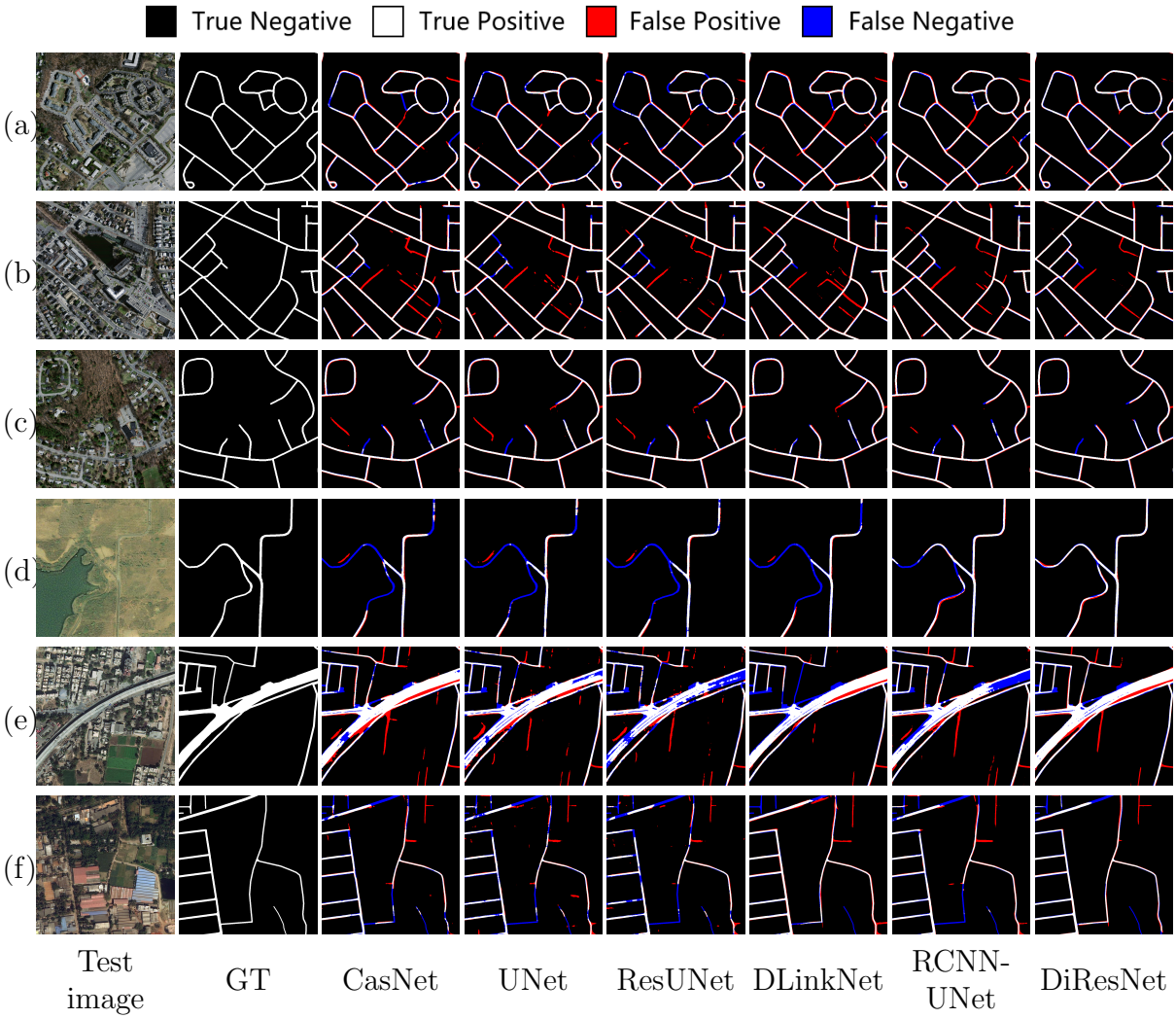


Figure 5.14: Examples of the segmentation results produced by the proposed method and the other used methods in the comparison. (a)-(c) Results selected from the Massachusetts dataset, (d)-(f) Results selected from the DeepGlobe dataset.

through more noises. We have argued and shown that the concatenation with low-level features is unnecessary and introduced an asymmetric network. This network is an extension of the Resnet, where its encoding layers are re-arranged to suit the task of road extraction. Experimental results on two benchmark datasets (the Massachusetts dataset and the DeepGlobe dataset) show that the designed segmentation network (DiResSeg) outperforms competitors in OA and F_1 measures, whereas its computational cost is significantly smaller.

Moreover, we have presented three auxiliary designs to improve the road segmentation accuracy: i) a structure supervision to emphasize the preservation of the road topology, ii) a direction supervision, where the angular operators are used to generate reference direction maps, thus supervising the network to learn directional linear features, and iii) a refinement sub-net to improve the smoothness and connectivity of the generated road maps. Combining these designs, the proposed DiResNet obtains sharp improvements in OA, BEP and F_1 measures. It is worth noting that the precision of the proposed method is particularly high.

One of the remaining problems in road extraction is that there are a certain types of road surfaces that are neglected by all the considered network models. In the future, we will investigate to add nodes and length based analysis of the road networks to improve the generalization ability of the segmentation network. In addition, the direction information can be refined and alternative ways can be studied to suppress the remaining side effects. This is left for future studies, where we plan both to use direction features as gating attention maps and design other forms of direction losses.

Chapter 6

Adversarial Shape Learning for Building Extraction in VHR RSIs

This chapter [1] *presents an Adversarial Shape Learning Network (ASLNet) for building extraction in VHR RSIs. Unlike existing adversarial learning methods, it explicitly models the shape patterns of buildings to improve the segmentation accuracy. The novel architecture includes i) a shape regularizer to model the building shape features; and ii) a shape discriminator to learn shape constraints to guide the segmentation network. Additionally, we designed 3 object-based metrics to evaluate the geometric properties of the building extraction results. Experimental results show that the shape learning leads to significant improvements in terms of both pixel-based accuracy and object-based metrics.*

6.1 Introduction

Shape is an important pattern in the process of visual recognition. Direct modeling of shape patterns in images is challenging since it requires a high-level abstract on the object contours. Among the real-world applications of image recognition techniques, building extraction in VHR RSIs is one of the most interesting and challenging tasks that can benefit greatly from learning the shape patterns. It is important for a wide variety of applications, such as land-cover mapping, urban resources management, detection

[1] This chapter appears in:
[J5] L. Ding, H. Tang, Y. Shi, X. Zhu, L. Bruzzone, "Adversarial Shape Learning for Building Extraction in VHR Remote Sensing Images," *IEEE Transactions on Image Processing*, Under Review.

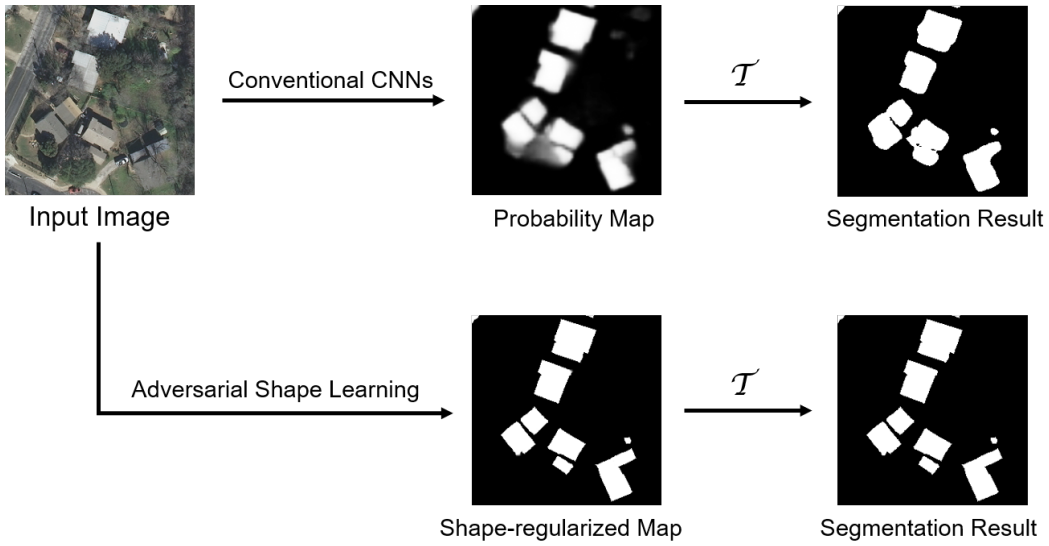


Figure 6.1: Illustration of the benefits of the proposed shape learning. Conventional CNN models lead to boundary ambiguity problems, whereas the proposed method produces shape-regularized results.

of illegal constructions, etc.

Conventional building extraction algorithms are based on handcrafted features that often fail to model high-level context information and are highly dependent on parameters. Recently, with the emergence of CNNs and their applications in semantic segmentation tasks (e.g., vehicle navigation [8], scene parsing [64], medical image segmentation [89]), a large research interest has been focused on adapting these CNN models to building extraction in VHR RSIs. The CNN-based building extraction methods employ stacked convolution operations to extract the intrinsic content information of images, thus they are both more effective in exploiting the context information and less sensitive to domain changes. A variety of CNN designs for the semantic segmentation of buildings have been introduced with good results [125, 154].

However, some critical challenges in building extraction remain unsolved even with the use of the recent CNN-based methods. First, occlusions (caused by trees and shadows) and intra-class diversity are common problems in VHR RSIs, which often cause fragmentation and incomplete segmentation. Second, it is common to have boundary ambiguity problems. Due to the effects of shadows and building profiles, an accurate localiza-

tion of the building boundaries is difficult (especially in the low-contrast areas). Conventional CNN-based methods produce ambiguous probability values in these areas, which often cause rounded or uneven building boundaries after thresholding. Last but not least, generally results are affected by over-segmentation and under-segmentation of the building objects due to these fragmentation and boundary-adhesion problems. Due to these limitations, post-processing algorithms are often required to optimize the building extraction results [123, 116].

Another important issue is that previous works on CNN-based building extraction pay more attention to the extraction of texture and context information in RSIs, whereas the explicit modeling of building shapes has rarely been studied. In most cases, buildings in VHR RSIs are compact and rectangular objects with sharp edges and corners. Their rectangularity is very discriminative compared to other ground objects. Learning this shape prior is beneficial for not only inpainting the occluded building parts but also reducing the boundary ambiguities and regularizing the segmentation results. An example is shown in Fig. 6.1 to illustrate the limitations of conventional CNNs and the benefits of the shape modelling.

In this chapter, we aim to address the previously mentioned issues and to improve the extraction of buildings by introducing an adversarial learning of their shape information. In greater detail, the main contributions of this chapter are as follows:

1. Proposing an adversarial shape learning network (ASLNet) to learn shape-regularized building extraction results. It includes a shape discriminator (SD) to exclude redundant information and focus on modelling the shape information, as well as a shape regularizer (SR) to enlarge the RFs and explicitly model the local shape patterns.

2. Designing three object-based quality assessment metrics to quantitatively evaluate the geometric properties of the building extraction results. These metrics take into account both the under-segmentation and over-segmentation problems and the shape errors of the predicted building items.

3. Achieving the state-of-the-art performance on the Inria and Mas-

sachusetts building extraction benchmark datasets. Without using sophisticated backbone CNN architectures or post-processing operations, the proposed ASLNet outperforms all the compared literature methods in both pixel-based and object-based metrics.

The remainder of this chapter is organized as follows. Section 6.2 illustrates the proposed ASLNet. Section 6.3 describes the implementation details and the experimental settings. Section 6.4 presents the results and analyzes the effect of the proposed method. Section 6.5 draws the conclusions of this study.

6.2 Adversarial Shape Learning Network

Typical CNN models [125, 123] for building segmentation exploit only the local texture and context information, thus the fragmentation and boundary ambiguity problems remain unsolved. Since buildings in VHR RSIs usually have clear shape patterns, it is meaningful to use the shape constraints to alleviate these problems. To this end, we propose the adversarial shape learning network (ASLNet) to explicitly model these shape constraints. In this section, we describe in detail the architecture, loss functions, and the CNN modules of our ASLNet.

6.2.1 Network Architecture

Fig. 6.2 illustrates the architecture of the proposed ASLNet for building extraction, which consists of a segmentation network and a discriminator network. The segmentation network itself is capable of segmenting buildings, while the discriminator is employed to guide the training of the segmentation network. The segmentation network follows the classic encoder-decoder structure in literature papers [89, 26, 16]. The encoder network contains down-sampling operations to extract high-level semantic features from image local patches, whereas the decoder network recovers the spatial resolution of encoded features. The choice of the encoder network is not the focus of this work, thus we simply adopt the ResNet [38] as the feature encoder. It has been widely used for feature extraction in

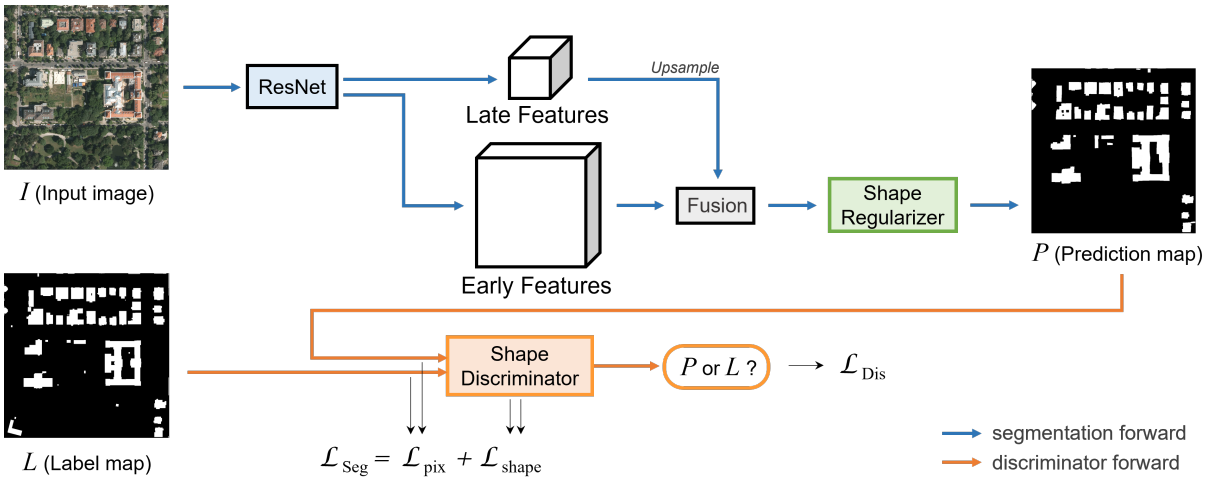


Figure 6.2: Architecture of the proposed Adversarial Shape Learning Network (ASLNet) for building extraction. We designed an explicit shape regularizer to model the shape features, and a shape discriminator to guide the segmentation network.

building segmentation [57], road segmentation [25], and other semantic segmentation related tasks [27]. The selected ResNet version is ResNet34, which can be replaced by other versions based on the complexity of the dataset.

Apart from the output features from the late layers of the ResNet (with 1/8 of the original GSD), the early features (with 1/4 of the original GSD) are also employed in the decoder to learn finer spatial details. This is a commonly adopted design in segmentation networks [16, 26]. This ResNet with encoder-decoder structure is a modified version of FCN [64], denoted as ED-FCN. Compared with the plain FCN, the ED-FCN models the spatial features at a finer resolution, which is essential for the segmentation of VHR RSIs. It is therefore set as the baseline method of our segmentation network. Building on top of the ED-FCN, we further designed a SR at the end of the segmentation network in the proposed ASLNet to produce shape-refined outputs.

6.2.2 Shape Regularizer

Although using a simple ResNet as the segmentation network is feasible for the adversarial shape learning, it is beneficial to model the shape features at finer spatial scales. Therefore, we design an explicit shape regularizer in

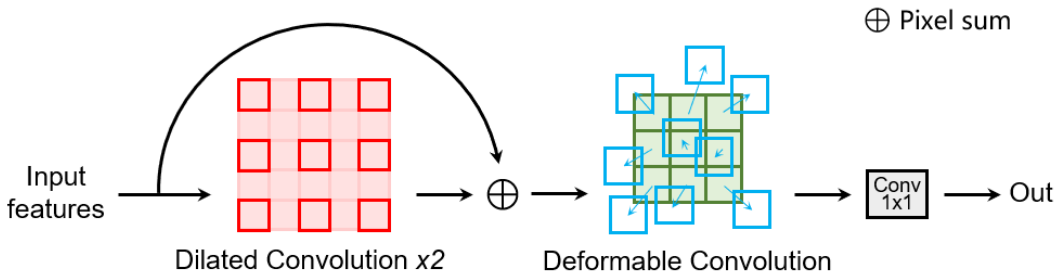


Figure 6.3: The designed shape regularizer. Dilated convolutions and deformable convolutions are employed to enlarge the RFs and learn the shape features.

the decoder of the segmentation network to enable a better adaptation to the shape constraints (see Fig. 6.3). The SR is placed at the spatial scale of 1/4 of the GSD, whose input features are fused multiscale features in the ED-FCN. This spatial resolution for shape modeling is adopted following the practice in [16] and [26], which is a balance between accuracy and computational costs. At this spatial scale, a conventional 3×3 convolutional kernel has the RF of around 12×12 pixels, which is too small for modelling the local shape patterns. Therefore, we introduce the dilated convolution (DC) and deformable convolution (DFC) [156] layers to enlarge the RFs and to learn shape-sensitive transformations.

Both the DC and DFC are based on the idea of enlarging the coverage of convolutional kernels. Let us consider a convolutional operation for pixel $x(r, c)$ as:

$$U(r, c) = \sum_{i,j} x_{r+i,c+j} \cdot k_{i,j}, \tag{6.1}$$

where $k_{i,j}$ denotes the kernel weight. In a standard 3×3 convolution, $i, j \in \{-1, 0, 1\}$. However, in a 3×3 DC, $i, j \in \{-r, 0, r\}$ where r is the dilation rate. In the designed SR we connected two 3×3 DCs as a residual block [38], which enlarges the RF to over 36×36 pixels. A DFC is further employed to exploit the shape information, defined as:

$$U_{df}(r, c) = \sum_{i,j} x_{r+i+u(r,c),c+j+v(r,c)} \cdot k_{i,j}, \tag{6.2}$$

where $u(r, c)$ and $v(r, c)$ are position parameters learned by the additional

convolutions, as follow:

$$u(r,c) = \sum_{i,j} x_{r+i,c+j} \cdot k'_{i,j}, v(r,c) = \sum_{i,j} x_{r+i,c+j} \cdot k''_{i,j}. \tag{6.3}$$

This enables the SR to perceive and adapt to the local shape patterns. Finally, a 1×1 convolution is followed to merge the features into a segmentation map.

6.2.3 Shape Discriminator

Although several literature works have introduced the adversarial learning for building extraction, most of them combine CNN outputs and input RSIs to train the discriminators [93, 10, 1, 80]. Under this condition, the discriminators are unlikely to learn the shape information, since they are affected by the redundant information in input RSIs. In the proposed ASLNet, the discriminator focuses only on the shape features, thus we exclude the use of input RSIs.

Training a shape discriminator with only binary inputs is challenging. Let I denote an input image, P be its corresponding prediction output and L be the GT map. Since in I there are usually mixed pixels (due to the sensor resolution) and discontinuities in objects representations (due to occlusions and low illumination conditions), it is common to have fuzzy areas in especially the building contours in the normalized prediction map $\sigma(P)$, where σ is the Sigmoid function. However, in L the human-annotated building contours have 'hard' edges, i.e. $L \in \{0,1\}$. Mathematically, let $\sigma(P) \in [0,1]$ be a smooth/fuzzy representation of the contours. This difference between $\sigma(P)$ and L can be easily captured by the discriminator and causes failure to the shape modelling. In some literature works [53] a thresholding (or argmax) function $\mathcal{T}$ is employed to binarize $\sigma(P)$ as:

$$R = \mathcal{T}[\sigma(P)] \tag{6.4}$$

where R is the binary segmentation map. Although the obtained $R \in \{0,1\}$, the $\mathcal{T}$ is non-differential in most cases, thus training the segmentation network with R and L will lead to zero-gradient problems.

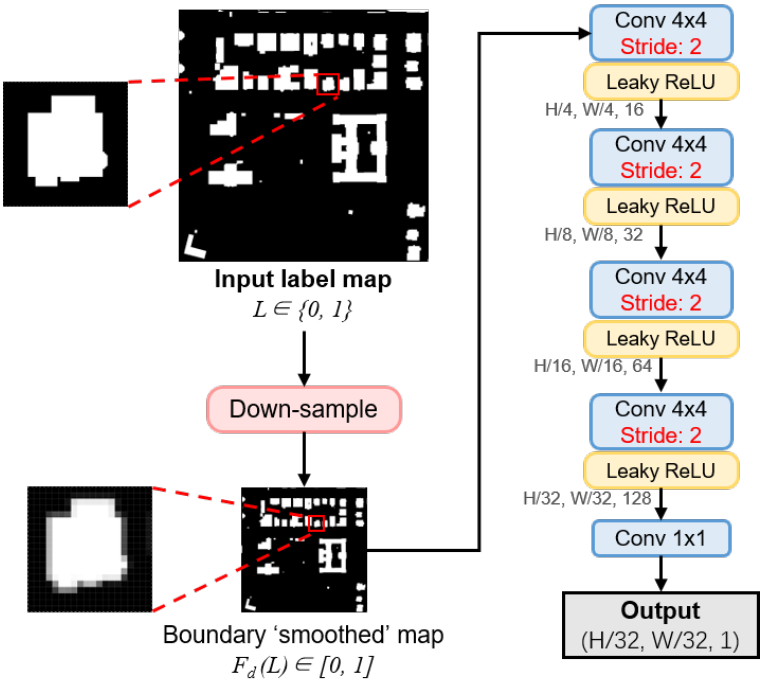


Figure 6.4: The designed shape discriminator. The input maps are down-scaled to exclude the impact of 'hard' building boundaries in reference maps.

In the designed SD we managed to eliminate this boundary difference and model only the shape information by adding a down-sampling operation F_d in the discriminator $\mathcal{D}$. Fig. 6.4 illustrates the designed SD. After applying F_d, the building boundaries in $F_d(L)$ are 'softened' ($F_d(L) \in [0, 1]$) and the boundary difference between $F_d(\sigma(P))$ and $F_d(L)$ is excluded. Specifically, four layers of strided convolution and activation functions are then employed to reduce the spatial size of feature maps and learn the local discriminative shape information. The output results are related to 1/32 of the original GSD.

The discriminator is trained with the BCE loss function. It is calculated as:

$$\begin{aligned} \mathcal{L}_{Dis} =& \mathbb{E}_{L\sim p_{\text{data}}(L)}[\log \mathcal{D}(L)] \\ &+ \mathbb{E}_{P\sim p_{data}(P)}[\log(1-\mathcal{D}(\sigma(P)))] \\ =& -y\log(p) - (1-y)\log(1-p), \end{aligned} \tag{6.5}$$

where $\mathbb{E}$ is the expected value for different types of input samples, y is the encoded signal that depending on the input map to the discriminator

can be L or $\sigma(P)$ ('1' and '0', respectively), and p is the output of the discriminator. In typical GANs the BCE loss is also used to supervise the generator network. However, in experimental practice, we found that this leads to training insatiability problems, since this encourages the segmentation network to generate fake predictions unrelated to the GT situations. As an alternative, we employ the Mean Squared Error (MSE) loss function to calculate the $\mathcal{L}_{Shape}$ as:

$$\mathcal{L}_{Shape} = \{\mathcal{D}(L) - \mathcal{D}[\sigma(P)]\}^2, \tag{6.6}$$

where $\mathcal{D}$ is the SD. In this way, the $\mathcal{L}_{Shape}$ is related to the L, thus the segmentation network is constrained by the GT conditions.

6.2.4 Optimization Objective of ASLNet

Let $\mathcal{L}_{Seg}$ be the loss function for the CNN-based segmentation of buildings. In conventional CNNs, $\mathcal{L}_{Seg}$ is only related to the pixel-wise accuracy, which does not consider the image context. At the end of shape learning via CNNs, it is essential to define a shape-based loss function $\mathcal{L}_{Shape}$. Previous works on shape analysis are often object-based [130, 63]. They include non-differential operations to calculate the shape measures, which are difficult to be incorporated into CNNs. Although there are also literature papers that use CNNs to regularize the shape of predictions [88], pre-training is often required and the regularization is limited to certain functions (e.g., inpainting of object contours). Since CNNs themselves can be trained to discriminate different shapes, we introduce the idea of adversarial learning to learn the $\mathcal{L}_{Shape}$ to guide the segmentation network.

$$\begin{aligned} \mathcal{L}_{Seg} =& \alpha \cdot \mathcal{L}_{Pix} + \beta \cdot \mathcal{L}_{Shape} \\ =& \alpha \cdot [L - \sigma(P)]^2 + \beta \cdot \{\mathcal{D}(L) - \mathcal{D}[\sigma(P)]\}^2, \end{aligned} \tag{6.7}$$

where $\mathcal{L}_{Pix} = [L - \sigma(P)]^2$ is the supervised pixel-based reconstruction loss, α and β are two weighting parameters. The first term in this formula drives the segmentation network to segment pixel-based P in order to fit L, while the second term strengthens the local shape similarities between P and L.

6.3 Dataset descriptions and design of Experiments

In this section, we describe the experimental dataset, the implementation details, and the considered evaluation metrics.

6.3.1 Dataset Descriptions

We conduct building extraction experiments on two VHR RSI datasets, i.e., the Inria dataset [66] and the Massachusetts Building dataset [71]. These are two of the most widely studied building extraction datasets in the literature [123, 53, 65, 57].

Inria Dataset [66]

This is an aerial dataset with the GSD of 0.3 m per pixel, covering 810 km^2. Each image has 5,000 $\times$ 5,000 pixels. There is a total of 360 images in this dataset, among which 180 are provided with the GT labels. These 180 images were collected in five different cities: Austin (U.S.), Chicago (U.S.), Kitsap (U.S.), Tyrol (Austria), and Vienna (Austria). Following the practice in [123, 65], we use the first 5 images in each city for testing and the rest 31 images for training.

Massachusetts (MAS) Building Dataset [71]

This is an aerial dataset collected on the Boston area. It has a GSD of 1.2 m per pixel, covering around 340 km^2. The imaged regions include urban and suburban scenes where there are buildings with different sizes. This dataset consists of a training set with 137 images, a validation set with 4 images, and a test set with 10 images. Each image has 1,500 $\times$ 1,500 pixels.

6.3.2 Implementation Details

The experiments were conducted on a workstation with 32 GB RAM and a NVIDIA Quadro P6000 GPU (23GB). Since it is impossible to train directly the large RSIs, they are randomly cropped into 512 $\times$ 512 patch images during the training process. The performed data preprocessing

and augmentation operations include data normalization, random cropping, and image flipping. The training batch size is set to 8 and the number of training epochs is 50. The validation and test sets are evaluated on the original size RSIs to avoid the impact of cropping parameters. The hyper-parameters α, β in the Eq. (6.7) are set to 5.0, 1.0, respectively. The choice of hyper-parameters is discussed in Section 6.4.1.

6.3.3 Evaluation Metrics

Pixel-based Evaluation Metrics

We adopt several commonly used evaluation metrics in building extraction [123, 94] and other binary segmentation tasks [25] to assess the accuracy of the results. These metrics are based on statistical analysis of the classified pixels, including: OA, Precision (P), Recall (R), F_1 score, and mean IoU. The calculations are:

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}, \tag{6.8}$$

$$F_1 = 2 \times \frac{P \times R}{P + R}, \quad OA = \frac{TP + TN}{TP + FP + TN + FN}, \tag{6.9}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{6.10}$$

Object-based Evaluation Metrics

Although the pixel-based evaluation metrics present the overall classification accuracy of the results, they fail to consider the thematic and geometrical properties of the segmented units [130]. To overcome this limitation, we designed three object-based evaluation metrics, including the matching rate (MR), the curvature error (E_{curv}), and the shape error (E_{shape}). These metrics are variants of the literature works [85, 63] to adapt to the assessment of building extraction results.

In order to compare the geometric quality of a segmented object S_j on the prediction map P and a reference object O_i on the GT map L, it

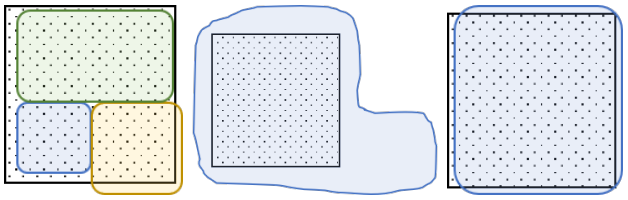


Figure 6.5: Illustration of three overlapping relationships between a segmented object S_j (colored region) and a reference object O_i (dotted region). (a) Over-segmentation, (b) Under-segmentation, and (c) Matching.

is essential to first discriminate if they are representing the same physical object. For each O_i $(i = 1, 2, 3, \cdots, n)$ and S_j $(j = 1, 2, 3, \cdots, n')$, their matching relationship $M(O_i, S_j)$ is calculated based on the over-segmentation error (E_{os}) and under-segmentation error (E_{us}) [85]:

$$M(O_i, S_j) = \begin{cases} 0, & E_{os}(O_i, S_j) > T \,||\, E_{us}(O_i, S_j) > T \\ 1, & E_{os}(O_i, S_j) \leqslant T \,\&\, E_{us}(O_i, S_j) \leqslant T \end{cases} \tag{6.11}$$

$$E_{os}(O_i, S_j) = 1 - \frac{|S_j \cap O_i|}{|O_i|}, \quad E_{us}(O_i, S_j) = 1 - \frac{|S_j \cap O_i|}{|S_j|}, \tag{6.12}$$

where T is a threshold value (empirically set to 0.3). The matching rate (MR) of P is the numeric ratio between the matched objects in L and all the O_i in L:

$$MR = \frac{\sum_{i,j} M(O_i, S_j)}{N_{O_i}} \tag{6.13}$$

After finding the matched item M_i in P for O_i, two geometric measurements are further calculated to measure the differences between M_i and O_i. First, E_{curv} is introduced to measure the differences in object boundaries. It is calculated as:

$$E_{curv}(O_i, M_i) = ||f_c(M_i) - f_c(O_i)||, \tag{6.14}$$

where f_c denotes the contour curvature function [34]. Since O_i is human-annotated, $f_c(O_i)$ is usually small. A large $E_{curv}(O_i, M_i)$ indicates that the boundary of $f_c(M_i)$ is uneven. The second measurement E_{shape} is introduced to assess the difference in shape, calculated as:

$$E_{shape}(O_i, M_i) = ||f_s(M_i) - f_s(O_i)||, \quad f_s(M_i) = \frac{4\pi|M_i|}{p_{M_i}^2}, \tag{6.15}$$

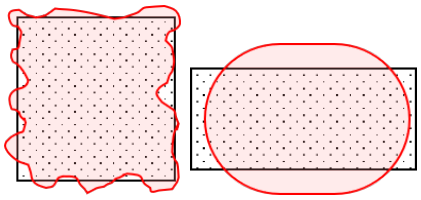


Figure 6.6: Examples of the reference object O_i (dotted region) and its matched segmented object M_i (colored region) that have: (a) high curvature error (E_{curv}), and (b) high shape error (E_{shape}).

where p_{M_i} is the perimeter of M_i. The value of $f_s(M_i)$ is 1 for a circle and $pi/4$ for a square [63, 34].

6.4 Experimental Results

This section presents the experimental results obtained on the two VHR building datasets. First, we present the ablation study to quantitatively evaluate the improvements brought by the proposed method. Then the effects of the SR and the SD are analyzed in greater detail on some significant sample areas. Finally, the proposed ASLNet is compared with several state-of-the-art CNN models for building extraction.

6.4.1 Ablation Study

Influence of Hyper-Parameters. The hyper-parameters α and β in Eq.(6.7) balance $\mathcal{L}_{pix}$ and $\mathcal{L}_{shape}$. To find which set of hyper-parameters leads to the best performance, we conduct an experiment on the Inria dataset. We set the value of one of the parameters to 1 and change the other one. The mIoU obtained with different hyper-parameter values are reported in Table 6.1. We find that setting $\mathcal{L}_{pix}$ as the primary loss (i.e., $\alpha > \beta$) leads to higher accuracy. The ASLNet obtains the best accuracy when $\alpha = 5, \beta = 1$. Therefore, these hyper-parameters are fixed in adversarial training of the ASLNet in all the experiments.

Quantitative Results. We conduct extensive ablation studies to assess the effectiveness of the proposed ASLNet. To compare the results before and after the use of SR and SD, the original FCN [64] and the baseline

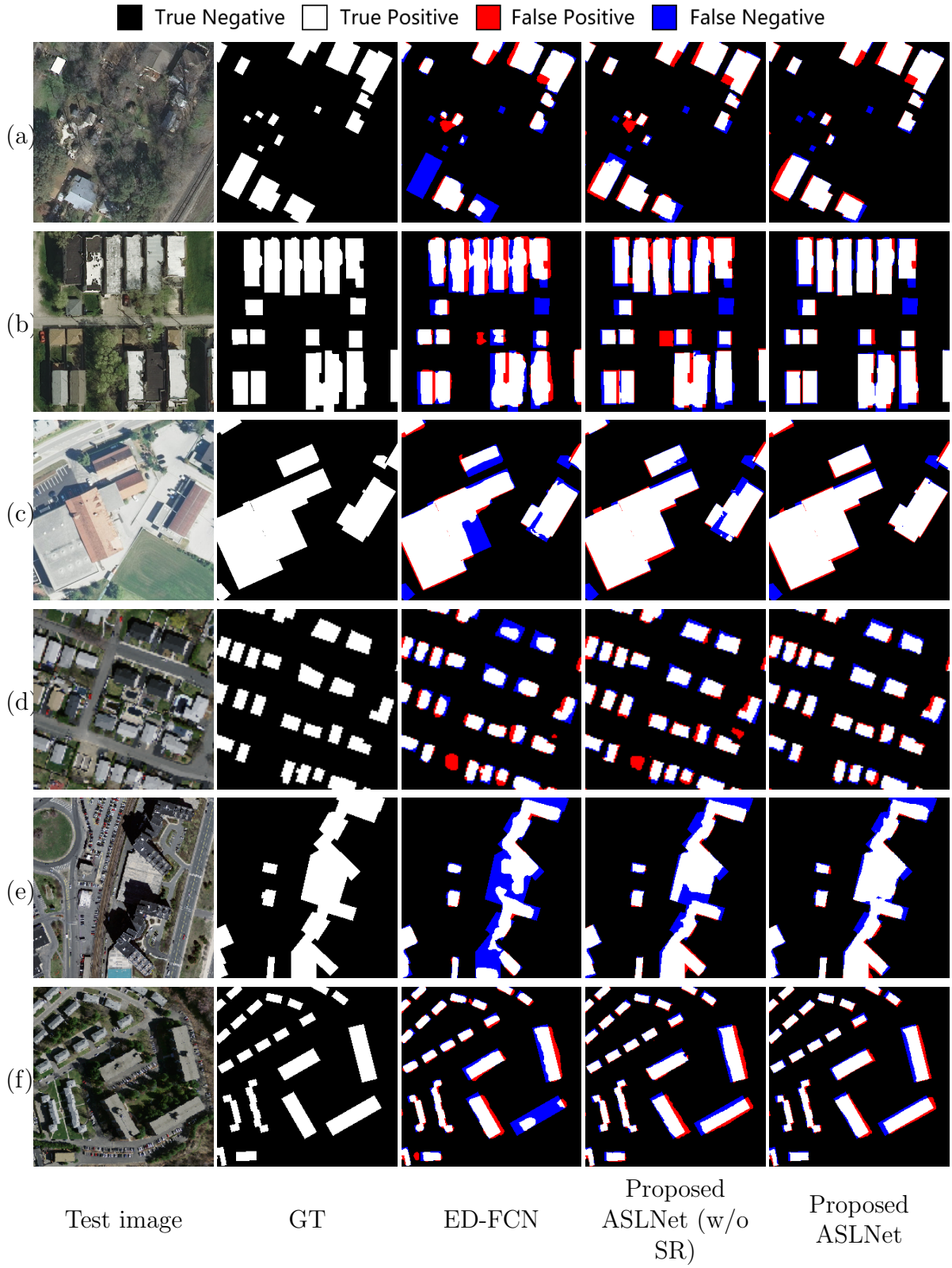


Figure 6.7: Examples of segmentation results obtained by the different methods (ablation study). (a)-(c) Results selected from the Inria dataset, (d)-(f) Results selected from the Massachusetts dataset.

Hyper-parameter	1	3	5	10
$\alpha(\beta = 1)$	77.56	78.58	**79.30**	78.82
$\beta(\alpha = 1)$	77.56	76.00	75.21	65.97

Table 6.1: The mIoU under different hyper-parameters tested on the Inria dataset.

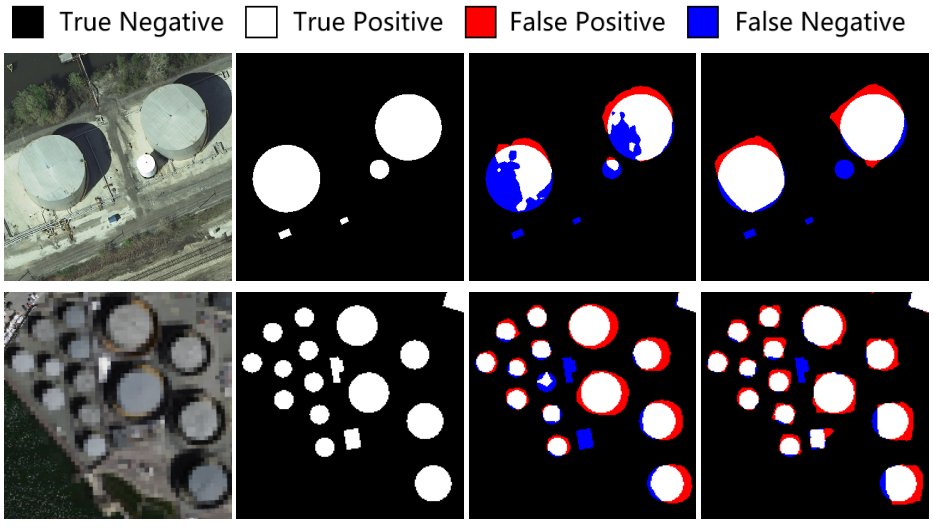


Figure 6.8: Examples of the failure cases. The ASLNet segments rectangular items for even the round objects, given its building-shape driven training.

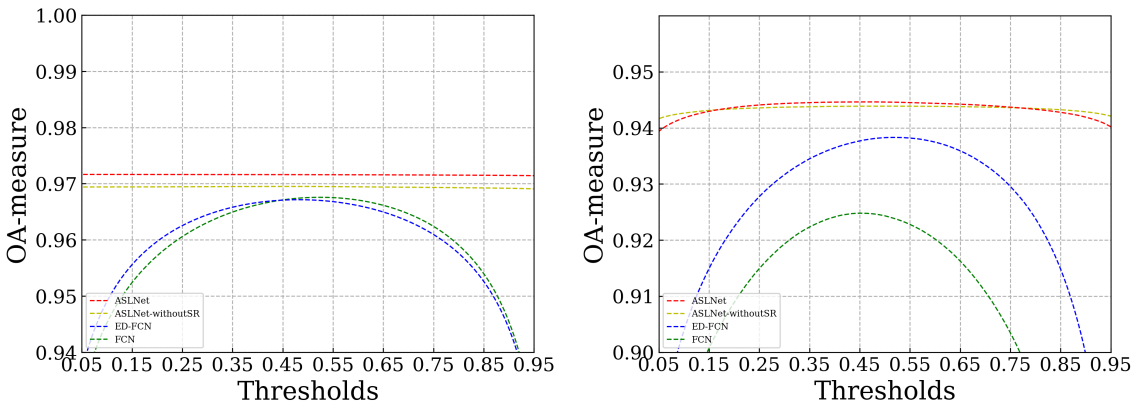
Table 6.2: Results of the ablation study on the two considered data sets.

Dataset	Method	Components		Pixel-based Metrics					Object-based Metrics		
		SR	SD	OA(%)	P(%)	R(%)	F_1(%)	mIoU(%)	MR(%)	E_{curv}	E_{shape}
Inria	FCN [64]			96.72	89.41	83.78	86.33	76.36	55.37	7.66	6.63
	ED-FCN			96.69	87.87	85.29	86.46	76.57	60.38	7.26	6.29
	Proposed ASLNet (w/o SR)		√	96.94	88.98	86.32	87.50	78.13	60.36	3.86	4.36
	Proposed ASLNet	√	√	**97.15**	**90.00**	**86.85**	**88.27**	**79.30**	**64.46**	**3.53**	**3.66**
MAS	FCN [64]			92.39	78.46	78.73	78.56	64.82	26.87	11.56	7.79
	ED-FCN			93.81	84.83	79.57	82.09	69.69	53.62	8.78	7.45
	Proposed ASLNet (w/o SR)		√	94.38	85.70	81.17	83.91	72.32	62.39	7.36	4.30
	Proposed ASLNet	√	√	**94.51**	**85.92**	**82.83**	**84.32**	**72.95**	**67.28**	**7.19**	**4.01**

method ED-FCN are also included in the comparison. The quantitative results are reported in Table 6.2. The baseline ED-FCN outperforms the FCN in terms of mean IoU by 0.21% and 4.87%, respectively in the Inria and the MAS dataset, which is attributed to the concatenation of low-level features in its decoder. Since the MAS dataset has lower spatial resolution, the improvements of the ED-FCN is more noticeable. After introducing the adversarial shape learning, the ASLNet (without the SR) has the mean IoU improvements of 1.56% and 2.63% on the two datasets. The complete ASLNet with both the SR and the SD provides improvements of 2.73% and 3.26% in mean IoU compared to the baseline ED-FCN. Fig. 6.9 shows a comparison of the OA values of the segmented probability maps versus different binarization ($\mathcal{T}$ in Formula (6.4)) thresholds. Since the ASLNet directly segments near-binary regularized results, its OA curves are close to horizontal, and are sharply above the baseline methods.

The improvements are even more significant in terms of object-based metrics. The baseline FCN encountered severe over-segmentation problems, which lead to low MR values. The ED-FCN slightly improves the three object-based metrics. The ASLNet (without the SR) has improvements of around 3% in both E_{curv} and E_{shape} in the two datasets. The ASLNet (with the SR) further improves the MR values of around 4% on the two datasets.

Qualitative Results. Fig. 6.7 shows the results of the ablation study on several sample areas. The segmentation results of the ED-FCN are generally round-edged. However, after adding the SD, the building edges became sharper and the object shapes became more rectangular. Moreover, the object shapes are modelled in a wider image range, thus the edges are more



(a) (b)

Figure 6.9: Accuracy curves versus different binarization threshold of (a) Inria dataset, and (b) Massachusetts dataset.

straight and some missing parts are inpainted. More specifically, Fig. 6.7(a) and (e) show two cases of occlusions caused by trees and shadows, respectively. Fig. 6.7(c) shows a case of under-segmentation. In these cases the ASLNet has successfully recovered the complete buildings. Fig. 6.7(b), (d), and (f) show several examples of the improvements in shapes. It is worth noting that the ASLNet managed to improve the segmented shape of compact small objects (e.g., houses), irregular large object (e.g., factories), and long bar-like objects (e.g., residential buildings). However, a side-effect of the ASLNet is that it fails to segment some round objects (e.g., oil tanks), since it is trained to optimize the rectangular contour of buildings. Some of examples of these cases are shown in Fig. 6.8. Considering the objective of the proposed method, this drawback has minor impacts. Note that the proposed shape-driven training could also be adapted to other general shapes to suit different applications

As a conclusion of the ablation study, the modeling of shape features in the ASLNet leads to three significant benefits: 1) inpainting of the missing parts of buildings; 2) providing a joint segmentation and regularization of the building contours; 3) mitigating the under-segmentation and over-segmentation problems. These advantages are verified by both the accuracy metrics and visual observation.

6.4.2 Comparative Experiments

We further compare the proposed ASLNet with several literature works to assess its effectiveness. Three classic models for the semantic segmentation are compared, including the UNet [89], the baseline method FCN [64] and the Deeplabv3+ [16]. The cwGAN-gp [93] that uses GAN for building extraction is also compared. Moreover, we compare the proposed method with several state-of-the-art methods for building extraction, including the ResUNet [125], the MAPNet [154], the GMEDN [65] and the FC-DenseNet+FRCRF [52] (which includes a CRF-based post-processing step). The quantitative results on the Inria dataset and the MAS dataset are reported in Table 6.3 and Table 6.4, respectively.

Let us first analyze the pixel-based metrics. The ResUNet, which is a variant of UNet for the building extraction, outperforms the classic semantic segmentation models (UNet, FCN and Deeplabv3+) by a large margin on the MAS dataset. The accuracy of cw-GAN-gp is higher than that of the FCN on the MAS dataset but it is lower on the Inria dataset. on the The MAPNet obtains competitive results on the Inria dataset, whereas its performance is inferior to the ResUNet and the Deeplabv3+ on the MAS dataset. On the contrary, the GMEDN obtains better accuracy on the MAS dataset. The FCN-DenseNet+FRCRF achieves the second best accuracy on the MAS dataset. The proposed ASLNet outperforms all the compared methods in almost all the metrics (except for the precision and recall on the MAS dataset), although its baseline method (the ED-FCN) is inferior to some of them. The advantages of the ASLNet are particularly noticeable on the Inria dataset, where the ASLNet improves the mean IoU of 1.51% with respect to the second best method. The reason for which the ASLNet has higher improvements on the Inria dataset can be attributed to the higher GSD of this dataset, where the building shape information is more discriminative.

In terms of object-based metrics, there are remarkable differences in the MR values. The cw-GAN-gp and the ResUNet obtained the third best MR values among the literature methods on the Inria dataset and the MAS dataset, respectively. The FCN-DenseNet+FRCRF obtained the second-best accuracy in all the object-based metrics due to its boundary-

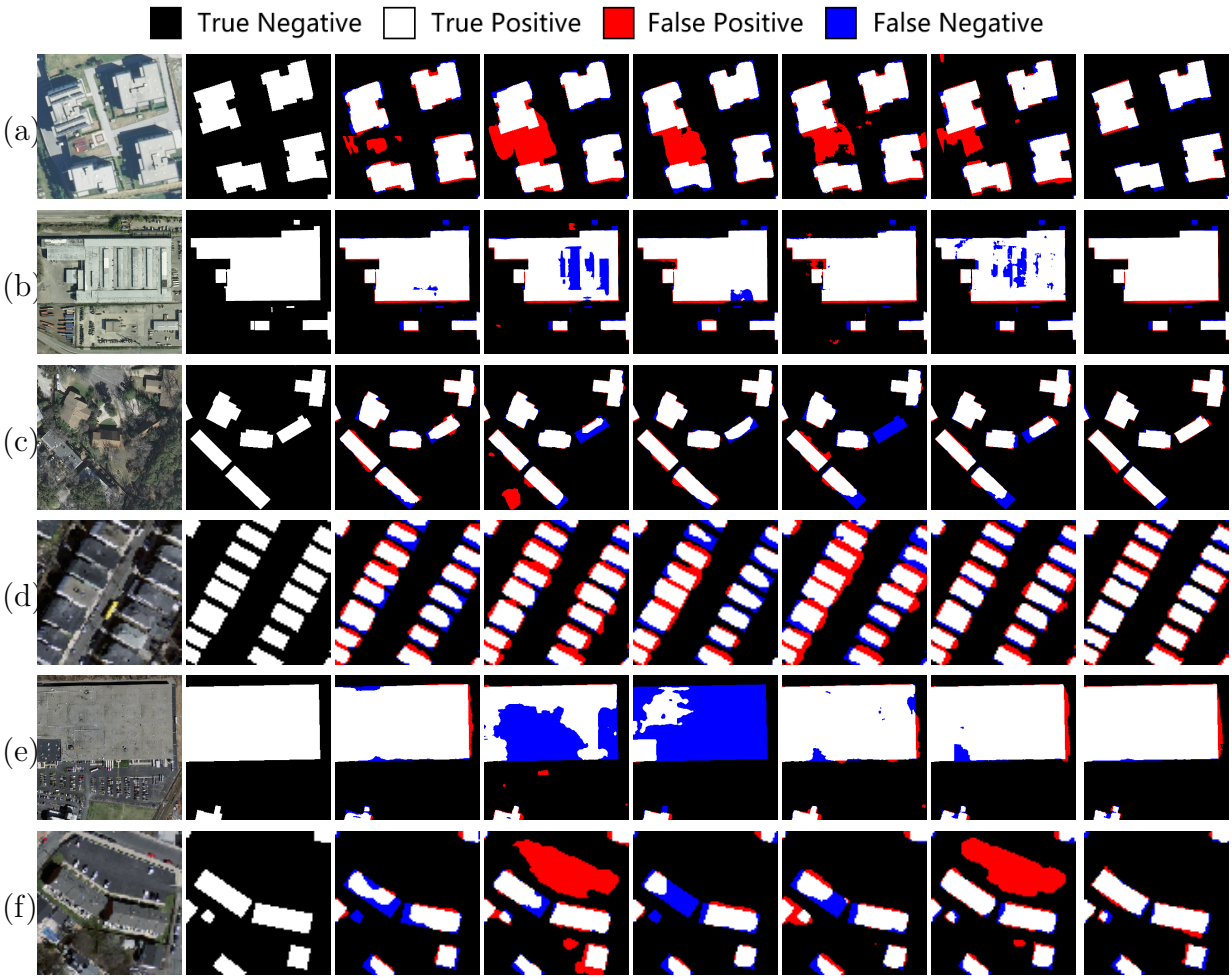


Figure 6.10: Examples of segmentation results obtained by the different methods (comparative experiments). (a)-(c) Results selected from the Inria dataset, (d)-(f) Results selected from the Massachusetts dataset.

Table 6.3: Results of the comparative experiments on the Inria dataset.

Method	Pixel-based Metrics					Object-based Metrics		
	OA(%)	P(%)	R(%)	F_1(%)	mIoU(%)	MR(%)	E_{curv}	E_{shape}
UNet [89]	95.52	81.76	82.76	82.03	70.03	43.87	10.89	7.84
FCN [64]	96.72	89.41	83.78	86.33	76.36	55.37	7.66	6.63
ED-FCN	96.69	87.87	85.29	86.46	76.57	60.38	7.26	6.29
Deeplabv3+ [16]	96.85	89.17	85.09	86.97	77.30	58.63	7.12	6.29
ResUNet [125]	96.50	88.33	83.60	85.68	75.41	55.72	7.47	6.50
cwGAN-gp [93]	96.54	86.43	85.61	85.94	75.76	61.51	7.10	5.47
MAPNet [154]	96.96	88.58	86.04	87.24	77.79	59.75	6.26	6.16
GMEDN [65]	96.23	87.03	81.37	83.88	72.95	52.65	8.43	5.54
FC-DenseNet+FRCRF [52]	96.74	89.55	83.68	86.36	76.34	63.43	4.31	4.13
ASLNet (proposed)	**97.15**	**90.00**	**86.85**	**88.27**	**79.30**	**64.46**	**3.53**	**3.66**

refinement CRF operations. All the other compared literature methods obtained very high E_{curv} and E_{shape} values. This indicates that they all suffer from irregular shapes and uneven boundaries problems. On the contrary, the proposed ASLNet shows significant advantages in terms of all these three metrics. Due to its learned shape constraints that regularize the segmented items and sharpen the building boundaries, the ASLNet exhibits great advantages in E_{shape} and E_{curv} in both datasets.

Fig.6.10 shows comparisons of the segmentation results obtained by the compared methods. One can observe that the proposed ASLNet exhibits several advantages in different scenes. It is capable of accurately segmenting the individual buildings in Fig.6.10(a), the occluded houses in Fig.6.10(c) and the large-size factories/supermarkets in Fig.6.10(b) and (e). When it deals with dense residential buildings as shown in Fig. 6.10(d), the over-segmentation and under-segmentation errors are reduced. It also excludes some uncertain areas by considering the shape patterns (e.g., the colored opening space in Fig.6.10(a) and the parking lot in Fig.6.10(f)).

6.5 Conclusions

Recent works on CNN-based building extraction exhibit severe limitations resulting in two main issues: 1) incomplete segmentation of objects due to occlusions and intra-class diversity; 2) geometric regularization of the building extraction results. To address these issues we introduce the ad-

Table 6.4: Results of the comparative experiments on the Massachusetts dataset.

Method	Pixel-based Metrics					Object-based Metrics		
	OA(%)	P(%)	R(%)	F_1(%)	mIoU(%)	MR(%)	E_{curv}	E_{shape}
UNet [89]	92.18	84.71	70.29	76.75	62.34	40.02	10.23	7.10
FCN [64]	92.39	78.46	78.73	78.56	64.82	26.87	11.56	7.79
ED-FCN	93.81	84.83	79.57	82.09	69.69	53.62	8.78	7.45
Deeplabv3+ [16]	93.27	82.28	78.95	80.53	67.52	47.15	9.82	7.67
ResUNet [125]	94.32	86.16	81.25	83.59	71.87	60.22	7.91	7.16
cw-GAN-gp [93]	93.00	81.03	79.64	80.29	67.15	51.94	9.37	6.74
MAPNet [154]	93.47	**87.88**	72.77	79.50	66.20	53.70	8.05	7.63
GMEDN [65]	93.29	84.09	77.49	80.63	67.61	51.20	9.20	7.26
FC-DenseNet+FRCRF [52]	94.48	85.28	**83.16**	84.18	72.77	67.21	7.92	6.66
ASLNet (proposed)	**94.51**	85.92	82.83	**84.32**	**72.95**	**67.28**	**7.19**	**4.01**

versarial training strategy to learn the shape of buildings and propose an ASLNet. Specifically, we designed a shape regularizer with shape-sensitive convolutional layers (DCs and DFCs) to regularize the feature maps, as well as a shape discriminator to learn the shape constraints to guide the segmentation network. To the best of our knowledge, this is the first work that learns adversarial shape constraints for the segmentation of RSIs. To quantitatively evaluate the thematic properties of the building extraction results, we also designed three object-based metrics: the matching rate, the curvature error and the shape error.

Experimental results on two VHR building datasets show that the proposed ASLNet has obtained significant improvements over the conventional CNN models in both pixel-based metrics and object-based metrics. These improvements can be attributed to two factors. First, learning the shape priors is beneficial to inpaint the missing building parts. Second, the shape constraints force the ASLNet to produce shape-regularized results, thus the segmented objects have rectangular shape and smooth boundaries. Additionally, we observed that the ASLNet greatly reduces the over-segmentation and under-segmentation errors (proved by the higher MR values). One of the limitation of the ASLNet is that it reduces its accuracy on the segmentation of objects with shape that are not rectangular (e.g., round buildings), which is due to its learned shape constraints.

The adversarial shape learning is potentially beneficial for other segmentation-related tasks with the RSIs, where the ground objects exhibit certain geo-

metric patterns. In future studies, we will investigate to use the adversarial shape learning to model other types of object shapes in different tasks (e.g., road extraction, change detection and land-cover mapping in RSIs).

Chapter 7

Conclusions

This chapter draws conclusions of the thesis. First, Section 7.1 summarizes the conducted research works and reports the achieved major contributions. Then, Section 7.2 analyzes the remaining problems and discusses the possible future developments.

7.1 Summary and Discussion

In this thesis we have studied how to improve the semantic segmentation of HR RSIs with the use of CNNs. We: 1) introduced the main open challenges and reviewed the recent developments and SOTA techniques related to the semantic segmentation of RSIs; 2) developed a local attention-based approach to the semantic segmentation of VHR RSIs; 3) proposed a multi-path CNN architecture for the semantic segmentation of SAR images; 4) proposed a direction-aware network for road extraction in VHR RSIs; and 5) developed an adversarial shape learning framework for building extraction in VHR RSIs. In the following, we summarize the conducted research works and discuss the achieved progress.

For CNN-based semantic segmentation of VHR RSIs, there is a dilemma between the aggregation of context information and the preservation of spatial details. In Chapter 3, we proposed a LANet to address this problem. Specifically, we designed i) a PAM that calculates descriptors in patches to embed the local scene information, and ii) an AEM that embeds the semantic information from high-level features to low-level features. The PAM embeds local focus into features, while the AEM bridges the seman-

tic gap between high-level and low-level features. The resulting LANet improves both the discrimination of critical categories and the preservation of spatial details. Ablation studies on two benchmark datasets have proved the effectiveness of the proposed LANet. Compared with several recent attention-based approaches, the LANet was proved more effective for the semantic segmentation of VHR RSIs.

Aiming at alleviating the impact of speckle noise in SAR images, in Chapter 4 we proposed a novel CNN architecture (the MP-ResNet) for the semantic segmentation of PolSAR images. Differently from its baseline method (the ResNet), the MP-ResNet embeds multi-scale high-level semantic features through its parallel branches. This greatly enlarges its VRF and allows it to learn more the discriminative local features. The parallel feature embedding design also narrows the semantic gap between multi-scale features. Comparative experiments show that the proposed MP-ResNet obtains stable improvements over its baseline method and surpasses several literature works in terms of segmentation accuracy. It obtains smoother object boundaries and segments better the critical areas.

The above-mentioned methodologies are developed for the semantic segmentation of multiple LU/LC classes. In practical applications the segmentation of certain specific targets of interest is also important. In this context, first we investigated the extraction of roads in VHR RSIs. Road extraction methods commonly suffers from occlusions and lack of contrast (compared to the surrounding environment). To conquer these limitations, in Chapter 5 we presented a DiResNet. It consists of several components that are specially designed for road extraction, including i) a segmentation network with deconvolutional layers and structural supervision, ii) a novel direction supervision function, and iii) a refinement sub-net. These designs strengthen the detection of road topology and optimize the road boundaries. An ablation study shows that the proposed DiResNet improves greatly the connectivity, precision and OA of results. Compared with several recent literature methods, the proposed DiResNet shows advantages in particularly the on DeepGlobe dataset, which includes more diverse road types.

Encouraged by the progress of modelling linear features for road extraction, we further proceeded to model a more challenging geometric pattern

in RSIs, i.e., the shape of buildings. The SOTA CNN-based building extraction methods show fragmentation errors and boundary ambiguity problems. To obtain more complete and edge-sharpened building extraction results, in Chapter 6 we proposed an ASLNet. In the ASLNet we designed i) a shape regularizer that employs dilated convolutions and deformable convolutions to perceive the local shape patterns presented in the data, and ii) a shape discriminator that learns shape constraints to guide the segmentation network. To quantitatively measure the geometric features of segmented objects, three object-based metrics were proposed. Experimental results on two benchmark datasets show that the ASLNet improves greatly the accuracy and geometric metrics. Qualitative results show that these improvements are obtained by inpainting the missing building parts and regularizing the building contours. The proposed ASLNet achieves the SOTA accuracy in both pixel-based and object-based metrics.

In a more general perspective, the contributions in this thesis related to the existing problems in semantic segmentation tasks mentioned in Section 1.1 can be summarised as follow:

- Reducing the fragmentation errors. In the proposed methods, this is achieved by increasing the VRF in various ways. The LANet learns local scene information through the PAM. The MP-ResNet employs multiple embedding branches to see wider the image context. The ASLNet learns local shape patterns with the shape regularizer. All of these contributions lead to better perception of the context information in RSIs, thus reducing the fragmentation errors.

- Improving the localization accuracy in object boundaries. In the LANet, the AEM augments the semantic representation of low-level features, which results in better preservation of the object edge information. In the ASLNet, the building boundaries are sharpened by learning the shape priors. These approaches all contribute to solving the boundary ambiguity problem.

- Optimization of the segmentation results. In the proposed DiResNet, a refinement subnet (the DiResRef) is designed to connect the broken segments and to smooth the segmented road boundaries. In the

proposed ASLNet, adversarial learning is introduced to learn shape-optimized segmentation results. These designs allow the CNNs to produce optimized segmentation results in an end-to-end manner.

7.2 Future Developments

Although many common problems related to the semantic segmentation in RSIs are studied in this thesis, some of them are not completely addressed. Many of these remaining problems are caused by the cognitive limitation of CNNs. Due to the intrinsic perception mechanism of CNNs, the image information is gathered, transformed and combined, but it is not grouped and reasoned in a structured way. In the following, we analyze the remaining problems in greater detail and suggest possible research developments in the future.

- Irregular contours of the segmented objects. Similar as roads and buildings, many other ground objects exhibit certain geometric patterns in LCLU mapping applications. However, the shape of CNN-segmented objects is often distinctively different from the GT annotations. This problem has not been addressed in existing works. It is possible to extend the modeling of object shape patterns to the multi-class semantic segmentation in HR RSIs.

- Wrong semantic correlations between the segmented ground objects. The studied approaches in this thesis focus more on the exploitation of local context information, whereas the modelling of long-range dependencies between objects is uncovered. In LCLU mapping applications, some segmented objects may exhibit wrong semantic correlations (e.g., cars on top of buildings, rivers connected with roads). These semantic dependencies are included in the reference annotations but are not learned by the CNNs. It is possible to combine non-local modelling designs [74] and adversarial learning to model the long-range context information in the future.

- Object-level fragmentation errors. Although the research proposed in the thesis has reduced the fragmentation errors in the pixel-level, there

are still significant object-level fragmentation errors. Many isolated segments are produced in the critical areas. A possible approach to solve this problem is to develop CNNs or other types of neural networks that model directly the objects or their structures (e.g., nodes, lines and polygons).

- Domain gaps between benchmark datasets and real-world data. Most of the experiments in this thesis are conducted on benchmark datasets with detailed annotations. However, in practical applications, human annotations are often not available. Thus, it is important to transfer the knowledge learned from labelled data to unseen data. Due to domain gaps in sensors, acquisition time and observed regions, this task remains to be challenging. Domain adaptation techniques are urgently needed to boost the wide application of AI techniques in LCLU mapping.

List of Publications

International Journals

[J1] **L. Ding**, J. Zhang, L. Bruzzone, "Semantic Segmentation of Large-Size VHR Remote Sensing Images Using a Two-Stage Multi-scale Training Architecture," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 8, pp. 5367-5376, August 2020

[J2] **L. Ding**, H. Tang, L. Bruzzone, "LANet: Local Attention Embedding to Improve the Semantic Segmentation of Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 59, No. 1, pp. 426-435, January 2021.

[J3] **L. Ding**, L. Bruzzone, "DiResNet: Direction-aware Residual Network for Road Extraction in VHR Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 59, In press, 2021.

[J4] **L. Ding**, K. Zheng, D. Lin, Y. Chen, B. Liu, J. Li, L. Bruzzone, "MP-ResNet: Multi-path Residual Network for the Semantic segmentation of High-Resolution PolSAR Images," *IEEE Geoscience and Remote Sensing Letters*, In press, 2021.

[J5] **L. Ding**, H. Tang, Y. Shi, X. X. Zhu, L. Bruzzone, "Adversarial Shape Learning for Building Extraction in VHR Remote Sensing Images," *IEEE Transactions on Image Processing*, Under Review.

[J6] **L. Ding**, D. Lin, S. Lin, J. Zhang, X. Cui, Y. Wang, H. Tang, L. Bruzzone, "Looking Outside the Window: Wider-Context Transformer for the Semantic Segmentation of High-Resolution Remote Sensing Images," *ISPRS Journal of Photogrammetry and Remote Sensing*, Under Review.

[J7] J. Zhang, S. Lin, **L. Ding**, L. Bruzzone, "Multi-scale Context Aggregation for Semantic Segmentation of Remote Sensing Images," *Remote Sensing*, Vol. 12, no. 4, pp. 701, February 2020.

[J8] Y. Cao, Y. Wang, J. Peng, C. Qiu, **L. Ding**, X. X. Zhu, "SDFL-FC: Semisupervised Deep Feature Learning With Feature Consistency for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 59, In press, 2021.

[J9] B. Liu, A. Yu, P. Zhang, **L. Ding**, W. Guo, K. Gao, X. Zuo, "Active deep densely connected convolutional network for hyperspectral image classification," *International Journal of Remote Sensing*, Vol. 42, no. 15, pp. 5905-5924, June 2021.

International Conferences

[C1] **L. Ding**, L. Bruzzone, "A Deep Architecture Based on A Two-stage Learning for Semantic Segmentation of Large-size Remote Sensing Images," *2019 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Yokohama, Japan, 2019, pp. 5228-5231.

References

[1] Abolfazl Abdollahi et al. "Building Footprint Extraction from High Resolution Aerial Images Using Generative Adversarial Network (GAN) Architecture". In: *IEEE Access* 8 (2020), pp. 209517–209527.

[2] Habibollah Agh Atabay. "Binary shape classification using convolutional neural networks". In: *IIOAB J* 7.5 (2016), pp. 332–336.

[3] Amir Atapour-Abarghouei and Toby P Breckon. "Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer". In: *CVPR*. 2018.

[4] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefevre. "How useful is region-based classification of remote sensing images in a deep learning framework?" In: *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE. 2016, pp. 5091–5094.

[5] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 140 (2018), pp. 20–32.

[6] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. "Joint learning from earth observation and openstreetmap data to get faster better semantic maps". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2017, pp. 67–75.

[7] Nicolas Audebert et al. "Distance transform regression for spatially-aware deep semantic segmentation". In: *Computer Vision and Image Understanding* 189 (2019), p. 102809.

[8] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation". In: *TPAMI* 39.12 (2017), pp. 2481–2495.

[9] Anil Batra et al. "Improved road connectivity by joint learning of orientation and segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 10385–10393.

[10] Benjamin Bischke et al. "Overcoming missing and incomplete modalities with generative adversarial networks for building footprint segmentation". In: *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE. 2018, pp. 1–6.

[11] Ksenia Bittner et al. "Building footprint extraction from VHR remote sensing images combined with normalized DSMs using fused fully convolutional networks". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11.8 (2018), pp. 2615–2629.

[12] Yice Cao et al. "Pixel-wise PolSAR image classification via a novel complex-valued deep fully convolutional network". In: *Remote Sensing* 11.22 (2019), p. 2653.

[13] Zhiying Cao et al. "End-to-End DSM Fusion Networks for Semantic Segmentation in High-Resolution Aerial Images". In: *IEEE Geoscience and Remote Sensing Letters* 16.11 (2019), pp. 1766–1770.

[14] Abhishek Chaurasia and Eugenio Culurciello. "Linknet: Exploiting encoder representations for efficient semantic segmentation". In: *2017 IEEE Visual Communications and Image Processing (VCIP)*. IEEE. 2017, pp. 1–4.

[15] Liang-Chieh Chen et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs". In: *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2018), pp. 834–848.

[16] Liang-Chieh Chen et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 801–818.

[17] Liang-Chieh Chen et al. "Rethinking atrous convolution for semantic image segmentation". In: *arXiv preprint arXiv:1706.05587* (2017).

[18] Liang-Chieh Chen et al. "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015.

[19] Yunpeng Chen et al. "Graph-based global reasoning networks". In: *CVPR*. 2019.

[20] Guangliang Cheng et al. "Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network". In: *IEEE Transactions on Geoscience and Remote Sensing* 55.6 (2017), pp. 3322–3337.

[21] Wei Cui et al. "Multi-Scale Semantic Segmentation and Spatial Relationship Recognition of Remote Sensing Images Based on an Attention Model". In: *Remote Sensing* 11.9 (2019), p. 1044.

[22] Sukhendu Das, TT Mirnalinee, and Koshy Varghese. "Use of salient features for the design of a multistage framework to extract roads from high-resolution multispectral satellite images". In: *IEEE transactions on Geoscience and Remote sensing* 49.10 (2011), pp. 3906–3931.

[23] Ilke Demir et al. "Deepglobe 2018: A challenge to parse the earth through satellite images". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE. 2018, pp. 172–17209.

[24] Henghui Ding et al. "Semantic correlation promoted shape-variant context for segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 8885–8894.

[25] Lei Ding and Lorenzo Bruzzone. "DiResNet: Direction-Aware Residual Network for Road Extraction in VHR Remote Sensing Images". In: *IEEE Transactions on Geoscience and Remote Sensing* (2020).

[26] Lei Ding, Hao Tang, and Lorenzo Bruzzone. "LANet: Local Attention Embedding to Improve the Semantic Segmentation of Remote Sensing Images". In: *IEEE Transactions on Geoscience and Remote Sensing* (2020).

[27] Lei Ding, Jing Zhang, and Lorenzo Bruzzone. "Semantic segmentation of large-size VHR remote sensing images using a two-stage multiscale training architecture". In: *IEEE Transactions on Geoscience and Remote Sensing* 58.8 (2020), pp. 5367–5376.

[28] Lei Ding et al. "Road Extraction Based on Direction Consistency Segmentation". In: *Chinese Conference on Pattern Recognition*. Springer. 2016, pp. 131–144.

[29] Lei Ding et al. "Using neighborhood centroid voting to extract road centerline from high resolution image". In: *J Image Graph* 20.11 (2015), pp. 1526–1534.

[30] Yiping Duan et al. "Multi-Scale Convolutional Neural Network for SAR Image Semantic Segmentation". In: *2018 IEEE Global Communications Conference (GLOBECOM)*. IEEE. 2018, pp. 1–6.

[31] Paolo Gamba, Fabio Dell'Acqua, and Gianni Lisini. "Improving urban road extraction in high-resolution images exploiting directional filtering, perceptual grouping, and simple topological concepts". In: *IEEE Geoscience and Remote Sensing Letters* 3.3 (2006), pp. 387–391.

[32] Lin Gao et al. "Road extraction from high-resolution remote sensing imagery using refined deep residual convolutional neural network". In: *Remote Sensing* 11.5 (2019), p. 552.

[33] Jie Geng, Wen Jiang, and Xinyang Deng. "Multi-scale deep feature learning network with bilateral filtering for SAR image classification". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 167 (2020), pp. 201–213.

[34] Rafael C Gonzalez and Richard E Woods. "Digital image processing. upper saddle River". In: *J.: Prentice Hall* (2002).

[35] Ian Goodfellow et al. "Generative adversarial nets". In: *NeurIPS*. 2014.

[36] Florent Guiotte et al. "Semantic segmentation of lidar points clouds: rasterization beyond digital elevation models". In: *IEEE Geoscience and Remote Sensing Letters* 17.11 (2020), pp. 2016–2019.

[37] Hao He et al. “Road extraction by using atrous spatial pyramid pooling integrated encoder-decoder network and structural similarity loss”. In: *Remote Sensing* 11.9 (2019), p. 1015.

[38] Kaiming He et al. “Deep residual learning for image recognition”. In: *CVPR*. 2016.

[39] Kaiming He et al. “Spatial pyramid pooling in deep convolutional networks for visual recognition”. In: *TPAMI* 37.9 (2015), pp. 1904–1916.

[40] Geoffrey E Hinton and Ruslan R Salakhutdinov. “Reducing the dimensionality of data with neural networks”. In: *science* 313.5786 (2006), pp. 504–507.

[41] Jie Hu, Li Shen, and Gang Sun. “Squeeze-and-excitation networks”. In: *CVPR*. 2018.

[42] Jiuxiang Hu et al. “Road network extraction and intersection detection from aerial images by tracking road footprints”. In: *IEEE Transactions on Geoscience and Remote Sensing* 45.12 (2007), pp. 4144–4157.

[43] Gao Huang et al. “Densely connected convolutional networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[44] Xin Huang and Liangpei Zhang. “Road centreline extraction from high-resolution imagery based on multiscale structural features and support vector machines”. In: *International Journal of Remote Sensing* 30.8 (2009), pp. 1977–1987.

[45] Shunping Ji, Shiqing Wei, and Meng Lu. “Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set”. In: *IEEE Transactions on Geoscience and Remote Sensing* 57.1 (2018), pp. 574–586.

[46] Pascal Kaiser et al. “Learning aerial image segmentation from online maps”. In: *IEEE Transactions on Geoscience and Remote Sensing* 55.11 (2017), pp. 6054–6068.

[47] Tero Karras, Samuli Laine, and Timo Aila. “A style-based generator architecture for generative adversarial networks”. In: *CVPR*. 2019.

[48] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[49] Weicheng Kuo et al. “Shapemask: Learning to segment novel objects by refining shape priors”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 9207–9216.

[50] Hanchao Li et al. “Pyramid attention network for semantic segmentation”. In: *arXiv preprint arXiv:1805.10180* (2018).

[51] Jianan Li et al. “Perceptual generative adversarial networks for small object detection”. In: *CVPR*. 2017.

[52] Qingyu Li et al. “Building footprint generation by integrating convolution neural network with feature pairwise conditional random field (FPCRF)”. In: *IEEE Transactions on Geoscience and Remote Sensing* 58.11 (2020), pp. 7502–7519.

[53] Xiang Li, Xiaojing Yao, and Yi Fang. “Building-a-nets: Robust building extraction from high-resolution remote sensing images with adversarial networks”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11.10 (2018), pp. 3680–3687.

[54] Yangyang Li et al. “A novel deep fully convolutional network for PolSAR image classification”. In: *Remote Sensing* 10.12 (2018), p. 1984.

[55] Justin Liang et al. “Polytransform: Deep polygon transformer for instance segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 9131–9140.

[56] Guosheng Lin et al. “Refinenet: Multi-path refinement networks for high-resolution semantic segmentation”. In: *CVPR*. 2017.

[57] Penghua Liu et al. “Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network”. In: *Remote Sensing* 11.7 (2019), p. 830.

[58] Shuo Liu et al. “ERN: Edge Loss Reinforced Semantic Segmentation Network for Remote Sensing Images”. In: *Remote Sensing* 10.9 (2018), p. 1339.

[59] Wei Liu, Andrew Rabinovich, and Alexander C Berg. “Parsenet: Looking wider to see better”. In: *arXiv preprint arXiv:1506.04579* (2015).

[60] Yahui Liu et al. “Roadnet: Learning to comprehensively analyze road networks in complex urban scenes from high-resolution remotely sensed images”. In: *IEEE Transactions on Geoscience and Remote Sensing* 57.4 (2018), pp. 2043–2056.

[61] Yan Liu et al. “Efficient Patch-Wise Semantic Segmentation for Large-Scale Remote Sensing Images”. In: *Sensors* 18.10 (2018), p. 3232.

[62] Yu Liu et al. “Hourglass-shapenetwork based semantic segmentation for high resolution aerial imagery”. In: *Remote Sensing* 9.6 (2017), p. 522.

[63] Ivan Lizarazo. “Accuracy assessment of object-based image classification: another STEP”. In: *International Journal of Remote Sensing* 35.16 (2014), pp. 6135–6156.

[64] Jonathan Long, Evan Shelhamer, and Trevor Darrell. “Fully convolutional networks for semantic segmentation”. In: *CVPR*. 2015.

[65] Jingjing Ma et al. “Building extraction of aerial images by a global and multi-scale encoder-decoder network”. In: *Remote Sensing* 12.15 (2020), p. 2350.

[66] Emmanuel Maggiori et al. “Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark”. In: *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE. 2017, pp. 3226–3229.

[67] Dimitrios Marmanis et al. “Classification with an edge: Improving semantic image segmentation with boundary detection”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 135 (2018), pp. 158–172.

[68] Gérard Medioni, Chi-Keung Tang, and Mi-Suen Lee. “Tensor voting: Theory and applications”. In: *Proceedings of RFIA*. Vol. 2000. 2000.

[69] Zelang Miao et al. “A method for accurate road centerline extraction from a classified image”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7.12 (2014), pp. 4762–4771.

[70] Zelang Miao et al. “Road centerline extraction from high-resolution imagery based on shape features and multivariate adaptive regression splines”. In: *IEEE geoscience and remote sensing letters* 10.3 (2012), pp. 583–587.

[71] Volodymyr Mnih. “Machine Learning for Aerial Image Labeling”. PhD thesis. University of Toronto, 2013.

[72] Volodymyr Mnih and Geoffrey E Hinton. “Learning to detect roads in high-resolution aerial images”. In: *European Conference on Computer Vision*. Springer. 2010, pp. 210–223.

[73] Fariba Mohammadimanesh et al. “A new fully convolutional neural network for semantic segmentation of polarimetric SAR imagery in complex land cover ecosystem”. In: *ISPRS journal of photogrammetry and remote sensing* 151 (2019), pp. 223–236.

[74] Lichao Mou, Yuansheng Hua, and Xiao Xiang Zhu. “A relation-augmented fully convolutional network for semantic segmentation in aerial scenes”. In: *CVPR*. 2019.

[75] Vinod Nair and Geoffrey E Hinton. “Rectified linear units improve restricted boltzmann machines”. In: *ICML*. 2010.

[76] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. “Dual attention networks for multimodal reasoning and matching”. In: *CVPR*. 2017.

[77] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. “Learning deconvolution network for semantic segmentation”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1520–1528.

[78] Sakrapee Paisitkriangkrai et al. “Effective semantic pixel labelling with convolutional networks and conditional random fields”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2015, pp. 36–43.

[79] Boxiao Pan et al. “Adversarial cross-domain action recognition with co-attention”. In: *AAAI*. 2020.

[80] Xuran Pan et al. “Building extraction from high-resolution aerial imagery using a generative adversarial network with spatial and channel attention mechanisms”. In: *Remote Sensing* 11.8 (2019), p. 917.

[81] Teerapong Panboonyuen et al. “Semantic segmentation on remotely sensed images using an enhanced global convolutional network with channel attention and domain specific transfer learning”. In: *Remote Sensing* 11.1 (2019), p. 83.

[82] Jongchan Park et al. “Bam: Bottleneck attention module”. In: *arXiv preprint arXiv:1807.06514* (2018).

[83] Otávio AB Penatti, Keiller Nogueira, and Jefersson A dos Santos. “Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?” In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2015, pp. 44–51.

[84] Daifeng Peng, Yongjun Zhang, and Haiyan Guan. “End-to-end change detection for high resolution satellite images using improved unet++”. In: *Remote Sensing* 11.11 (2019), p. 1382.

[85] Claudio Persello and Lorenzo Bruzzone. “A novel protocol for accuracy assessment in classification of very high resolution images”. In: *IEEE Transactions on Geoscience and Remote Sensing* 48.3 (2009), pp. 1232–1244.

[86] Minh-Tan Pham and Sébastien Lefèvre. “Very high resolution Airborne PolSAR Image Classification using Convolutional Neural Networks”. In: *EUSAR 2021; 13th European Conference on Synthetic Aperture Radar*. VDE. 2021, pp. 1–4.

[87] Xuebin Qin et al. “Basnet: Boundary-aware salient object detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 7479–7489.

[88] Hariharan Ravishankar et al. “Learning and incorporating shape models for semantic segmentation”. In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2017, pp. 203–211.

[89] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015.

[90] Moslem Ouled Sghaier and Richard Lepage. “Road extraction from very high resolution remote sensing optical images based on texture analysis and beamlet transform”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9.5 (2015), pp. 1946–1958.

[91] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. “SinGAN: Learning a Generative Model from a Single Natural Image”. In: *ICCV*. 2019.

[92] Yuanzheng Shao et al. “Application of a fast linear feature detector to road extraction from remotely sensed imagery”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 4.3 (2010), pp. 626–631.

[93] Yilei Shi, Qingyu Li, and Xiao Xiang Zhu. “Building footprint generation using improved generative adversarial networks”. In: *IEEE Geoscience and Remote Sensing Letters* 16.4 (2018), pp. 603–607.

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

[95] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[96] Yanzhou Su et al. "Semantic Segmentation of High Resolution Remote Sensing Image Based on Batch-Attention Mechanism". In: *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2019, pp. 3856–3859.

[97] Shihao Sun et al. "Feature Fusion through Multitask CNN for Large-scale Remote Sensing Image Segmentation". In: *2018 10th IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS)*. IEEE. 2018, pp. 1–4.

[98] Weiwei Sun and Ruisheng Wang. "Fully Convolutional Networks for Semantic Segmentation of Very High Resolution Remotely Sensed Images Combined With DSM". In: *IEEE Geoscience and Remote Sensing Letters* 15.3 (2018), pp. 474–478.

[99] Christian Szegedy et al. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[100] Towaki Takikawa et al. "Gated-scnn: Gated shape cnns for semantic segmentation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 5229–5238.

[101] Hao Tang et al. "Gesturegan for hand gesture-to-gesture translation in the wild". In: *ACM MM*. 2018.

[102] Hao Tang et al. "Xinggan for person image generation". In: *ECCV*. 2020.

[103] Xin-Yi Tong et al. "Land-cover classification with high-resolution remote sensing images using transferable deep models". In: *Remote Sensing of Environment* 237 (2020), p. 111322.

[104] Luan Tran et al. "Gotta Adapt'Em All: Joint Pixel and Feature-Level Domain Adaptation for Recognition in the Wild". In: *CVPR*. 2019.

[105] John C Trinder and Yandong Wang. "Automatic road extraction from aerial images". In: *Digital Signal Processing* 8.4 (1998), pp. 215–224.

[106] Yi-Hsuan Tsai et al. "Learning to adapt structured output space for semantic segmentation". In: *CVPR*. 2018.

[107] Michele Volpi and Vittorio Ferrari. "Semantic segmentation of urban scenes by learning local class interactions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2015, pp. 1–9.

[108] Tuan-Hung Vu et al. "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation". In: *CVPR*. 2019.

[109] Fei Wang et al. “Residual attention network for image classification”. In: *CVPR*. 2017.

[110] Jingdong Wang et al. “Deep high-resolution representation learning for visual recognition”. In: *IEEE transactions on pattern analysis and machine intelligence* (2020).

[111] Tiantian Wang et al. “Detect globally, refine locally: A novel approach to saliency detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3127–3135.

[112] Weixing Wang et al. “A review of road extraction from remote sensing images”. In: *Journal of traffic and transportation engineering (english edition)* 3.3 (2016), pp. 271–282.

[113] Xiaolong Wang, Abhinav Shrivastava, and Abhinav Gupta. “A-fast-rcnn: Hard positive generation via adversary for object detection”. In: *CVPR*. 2017.

[114] Xiaying Wang et al. “HR-SAR-net: A deep neural network for urban scene segmentation from high-resolution SAR data”. In: *2020 IEEE Sensors Applications Symposium (SAS)*. IEEE. 2020, pp. 1–6.

[115] Yan Wang et al. “A hierarchical fully convolutional network integrated with sparse and low-rank subspace representations for PolSAR imagery classification”. In: *Remote Sensing* 10.2 (2018), p. 342.

[116] Shiqing Wei, Shunping Ji, and Meng Lu. “Toward automatic building footprint delineation from aerial images using CNN and regularization”. In: *IEEE Transactions on Geoscience and Remote Sensing* 58.3 (2019), pp. 2178–2189.

[117] Yanan Wei, Zulin Wang, and Mai Xu. “Road structure refined CNN for road extraction in aerial image”. In: *IEEE Geoscience and Remote Sensing Letters* 14.5 (2017), pp. 709–713.

[118] Yao Wei, Kai Zhang, and Shunping Ji. “Simultaneous Road Surface and Centerline Extraction From Large-Scale Remote Sensing Images Using CNN-Based Segmentation and Tracing”. In: *IEEE Transactions on Geoscience and Remote Sensing* (2020).

[119] Sanghyun Woo et al. “Cbam: Convolutional block attention module”. In: *ECCV*. 2018.

[120] Wenjin Wu et al. “PolSAR image semantic segmentation based on deep transfer learning—Realizing smooth classification with small training sets”. In: *IEEE Geoscience and Remote Sensing Letters* 16.6 (2019), pp. 977–981.

[121] Zhe Wu, Li Su, and Qingming Huang. “Cascaded partial decoder for fast and accurate salient object detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3907–3916.

[122] Fanghong Xiao, Ling Tong, and Shiyu Luo. "A Method for Road Network Extraction from High-Resolution SAR Imagery Using Direction Grouping and Curve Fitting". In: *Remote Sensing* 11.23 (2019), p. 2733.

[123] Yakun Xie et al. "Refined Extraction Of Building Outlines From High-Resolution Remote Sensing Imagery Based on a Multifeature Convolutional Neural Network and Morphological Filtering". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13 (2020), pp. 1842–1855.

[124] Jiang Xin et al. "Road Extraction of High-Resolution Remote Sensing Images Derived from DenseUNet". In: *Remote Sensing* 11.21 (2019), p. 2499.

[125] Yongyang Xu et al. "Building extraction in very high resolution remote sensing imagery using deep learning and guided filters". In: *Remote Sensing* 10.1 (2018), p. 144.

[126] Yongyang Xu et al. "Road extraction from high-resolution remote sensing imagery using deep learning". In: *Remote Sensing* 10.9 (2018), p. 1461.

[127] Hsiuhan Lexie Yang et al. "Building extraction at scale using convolutional neural network: Mapping of the united states". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11.8 (2018), pp. 2600–2614.

[128] Maoke Yang et al. "Denseaspp for semantic segmentation in street scenes". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3684–3692.

[129] Xiaofei Yang et al. "Road detection and centerline extraction via deep recurrent convolutional neural network u-net". In: *IEEE Transactions on Geoscience and Remote Sensing* 57.9 (2019), pp. 7209–7220.

[130] Su Ye, Robert Gilmore Pontius Jr, and Rahul Rakshit. "A review of accuracy assessment for object-based image analysis: From per-pixel to per-polygon approaches". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 141 (2018), pp. 137–147.

[131] Dandong Yin et al. "A direction-guided ant colony optimization method for extraction of urban road information from very-high-resolution images". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8.10 (2015), pp. 4785–4794.

[132] Bo Yu, Lu Yang, and Fang Chen. "Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 99 (2018), pp. 1–10.

[133] Fisher Yu and Vladlen Koltun. "Multi-Scale Context Aggregation by Dilated Convolutions". In: *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2016.

[134] Jiangye Yuan. "Learning building extraction in aerial scenes with convolutional networks". In: *IEEE transactions on pattern analysis and machine intelligence* 40.11 (2017), pp. 2793–2798.

[135] Jiangye Yuan et al. "LEGION-based automatic road extraction from satellite imagery". In: *IEEE transactions on geoscience and remote sensing* 49.11 (2011), pp. 4528–4538.

[136] Yuan Yuan, Jianzhe Lin, and Qi Wang. "Hyperspectral image classification via multitask joint sparse representation and stepwise MRF optimization". In: *IEEE transactions on cybernetics* 46.12 (2015), pp. 2966–2977.

[137] Zhenyu Yue et al. "A Novel Attention Fully Convolutional Network Method for Synthetic Aperture Radar Image Segmentation". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13 (2020), pp. 4585–4598.

[138] Yu Zang et al. "Road network extraction via aperiodic directional structure measurement". In: *IEEE Transactions on Geoscience and Remote Sensing* 54.6 (2016), pp. 3322–3335.

[139] Ce Zhang et al. "VPRS-based regional decision fusion of CNN and MRF classifications for very fine resolution remotely sensed images". In: *IEEE Transactions on Geoscience and Remote Sensing* (2018).

[140] Hang Zhang et al. "Context encoding for semantic segmentation". In: *CVPR*. 2018.

[141] Jing Zhang et al. "Multi-Scale Context Aggregation for Semantic Segmentation of Remote Sensing Images". In: *Remote Sensing* 12.4 (2020), p. 701.

[142] Qiaoping Zhang and Isabelle Couloigner. "Benefit of the angular texture signature for the separation of parking lots and roads on high resolution multi-spectral imagery". In: *Pattern recognition letters* 27.9 (2006), pp. 937–946.

[143] Xiangrong Zhang et al. "Aerial image road extraction based on an improved generative adversarial network". In: *Remote Sensing* 11.8 (2019), p. 930.

[144] Xiaodong Zhang et al. "An object-based supervised classification framework for very-high-resolution remote sensing images using convolutional neural networks". In: *Remote Sensing Letters* 9.4 (2018), pp. 373–382.

[145] Yang Zhang et al. "Topology-aware road network extraction via Multi-supervised Generative Adversarial Networks". In: *Remote Sensing* 11.9 (2019), p. 1017.

[146] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. "Road extraction by deep residual u-net". In: *IEEE Geoscience and Remote Sensing Letters* 15.5 (2018), pp. 749–753.

[147] Zhengxin Zhang and Yunhong Wang. "JointNet: A common neural network for road and building extraction". In: *Remote Sensing* 11.6 (2019), p. 696.

[148] Zhenli Zhang et al. "Exfuse: Enhancing feature fusion for semantic segmentation". In: *ECCV*. 2018.

[149] Hengshuang Zhao et al. “Psanet: Point-wise spatial attention network for scene parsing”. In: *ECCV*. 2018.

[150] Hengshuang Zhao et al. “Pyramid scene parsing network”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2881–2890.

[151] Kang Zhao et al. “Building extraction from satellite images using mask R-CNN with building boundary regularization”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018, pp. 247–251.

[152] Wenzhi Zhao, Shihong Du, and William J Emery. “Object-based convolutional neural network for high-resolution imagery classification”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 10.7 (2017), pp. 3386–3396.

[153] Lichen Zhou, Chuang Zhang, and Ming Wu. “D-LinkNet: LinkNet With Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction.” In: *CVPR Workshops*. 2018, pp. 182–186.

[154] Qing Zhu et al. “MAP-Net: Multiple Attending Path Neural Network for Building Footprint Extraction From Remote Sensed Imagery”. In: *IEEE Transactions on Geoscience and Remote Sensing* (2020).

[155] Xiao Xiang Zhu et al. “Deep learning in remote sensing: A comprehensive review and list of resources”. In: *IEEE Geoscience and Remote Sensing Magazine* 5.4 (2017), pp. 8–36.

[156] Xizhou Zhu et al. “Deformable convnets v2: More deformable, better results”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 9308–9316.