Road Topology Extraction from Satellite images by Knowledge Sharing

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computer Science and Engineering
by Research

by

Anil Kumar Batra
201550847
anil.kumar@research.iiit.ac.in

International Institute of Information Technology
Deemed University
Hyderabad - 500 032, INDIA
July 2019

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "**Road Topology Extraction from Satellite images by Knowledge Sharing**" by **Anil Kumar Batra**, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Adviser: Prof. C.V. Jawahar

To

My wife *Ankita Batra* and sister *Jyoti*

Acknowledgments

There are many people I must thank for contributing to wonderful years of my experience as a Master student at IIIT.

First, I would like to thank my thesis adviser, *Prof. C.V. Jawahar*, for taking me in his research lab and supporting me throughout my journey of research.. I learned the importance of dedication, consistency, and patience required in the field of research. I'd also like to thank my co-advisers *Guan Pang* and *Saikat Basu*, for all of their advice and knowledge they have shared with me.

I would like to thank my close collaborator *Suriya Singh* for continuously helping me with unique perspectives and insightful comments, feedback and advice. My thinking and research philosophy has been shaped by thoughtful discussions and interactions with him.

A big thank to my wife *Ankita Batra* for all her support and understanding, and the many sacrifices she has made to let me continue pursuing my research. She always believes in me more than I do. Also, I would like to thanks my parents for having trust in me, showering their wholehearted love upon me and to allow me to pursue my dreams in all possible ways at all times.

Finally, I would like to thank all my friends for supporting me throughout.

Abstract

Motivated by human visual system machine or computer vision technology has been revolutionized in last few decades and make a strong impact in wide range of applications such as object recognition, face recognition and identification etc. However, despite much encouraging advancement, there are still many fields which lack to utilize the full potential of computer vision techniques. One such field is to analyze the satellite images for geo-spatial applications.

In past, building and launching the satellites in to space was expensive, and was big hurdle in acquiring low cost images from satellites. However, with technological innovations, the inexpensive satellites are capable of sending terabytes of images of our planet on the daily basis that can provide insights on global-scale economic, social and industrial processes. The significant applications of satellite imagery are urban planning, crop yield forecasting, mapping and change detection. The most obvious application of satellite imagery is to extract topological road network from the satellite images, as it plays an important role in planning the mobility between multiple geographical locations of interest. The extraction of road topology from the satellite images is formulated as binary segmentation problem in vision community. Despite of huge satellite imagery, the fundamental hurdle in applying computer vision algorithms based on deep learning architectures is unavailability of labeled data and causes the poor results. Another challenge in extraction of the roads from satellite imagery is visual ambiguity in identifying the roads and occlusion by various objects. This challenge causes many standard algorithms in computer vision research to perform poorly and is the major concern. In this thesis we develop deep learning based models and techniques that allows us to address the above challenges.

In the first part of our work, we make an attempt to perform road segmentation with the less labeled data and existing unsupervised feature learning techniques. In particular, we use self-supervised technique to learn visual representations with an artificial supervision, followed by fine tuning of model with labeled dataset. We use semantic image in-painting as an artificial/auxiliary task for supervision. The enhancement of road segmentation is in direct relation with the features captured by model through inpainting of the erased regions in the image. To further enhance the feature learning, we propose to inpaint the difficult regions of the image and develop a novel adversarial training scheme to learn mask used for erasing the image. The proposed scheme gradually learns to erase regions, which are difficult to inpaint. Thus, this increase in difficulty level of image in-painting leads to better road segmentation. Additionally, we study the proposed approach on scene parsing and land classification in satellite images.

In the second part of our work, we study the ineffectiveness of naive segmentation based approaches to extract connected road topology from satellite images. To learn the connected road representation, we develop a novel task of learning orientation of road segments at each pixel. The proposed task is capable to capture the relational information in road pixels encoded in the road orientation. We then pose the road extraction as a multi-task learning problem where model learns the representation of connected road segments. We demonstrate the significance of our task on state-of-art deep learning architectures. We also propose to use connectivity refinement as an additional layer to enhance the road network. Finally, we propose a stacked multi-branch convolution neural network, an integrated system, which is capable to refine and effectively utilize the mutual information between road orientation and segmentation task. We show that the proposed end-to-end learning framework is able to improve the road connectivity by overcoming the challenges of visual ambiguity and occlusion.

We evaluate the proposed techniques on two diverse road datasets using the pixel based metrics. To measure the road connectivity we use *Avereage Path Length Similarity* [80] metric and outperform state-of-art techniques by significant margin.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The visual system of human beings acquire the information of an object or phenomena without making any physical contact using the photo-receptive cells, such a system is called remote sensing system. This biological remote sensing system captures the information in the form of images, finest way of transmitting the data, and further processed by the brain. The brain analyzes the information and provide decision capabilities to humans, which effectively evolve the world around humans. Inspired from this biological system, humans develop artificial systems to understand the various complex phenomena such as urban growth, environment monitoring and climate change. The process of collecting data and analysis of these complex phenomena of our world/planet via remote sensing technologies is referred to as *Earth Observation*.

Earth Observation data is collected in the form of images through electronic sensors located on satellites. The imaging satellites revolve around the earth regularly to provide constant updates and latest changes happening on the ground. The obtained imagery is capable to provide the wider view of the geographical system. Recent technological developments in the micro-electronic industry leads to fast and large availability of very high resolution (VHR) satellite images. In turn, these imagery led to increase the research in developing fast and accurate automatic systems for various applications such as disaster management, mapping the road network, urban planning and crop yield forecasting. One such prominent research is to develop a system to extract the terrestrial objects such as roads.

The precise and up-to-date road network has a strong positive correlation with economic development and the growth of nation. It is essential in variety of applications such as navigation system, traffic management and advanced driver assistance systems. In the past, human experts manually analyze the satellite imagery or use GPS navigation vehicles to extract the road network. However, the manual process is laborious, expensive and requires intensive human efforts. To mitigate this challenge, we need to develop an automatic systems with computer algorithms utilizing high resolution satellite images. The difficulty in building such system comes from several aspects. The well known problem is occlusion of roads with shadows of trees, clouds, and buildings which leads to poor extraction of road network. Such a disconnected road network is not useful for practical applications. However, the human visual system is efficient in identifying most of the occluded roads. Another problem is visual ambiguity in differentiating the roads from the similar textured objects available in the images. For example, un-paved roads

has similar texture with the paths in farming lands. Such challenges make the road extraction problem as a non-trivial task and open research problem.

Figure 1.1: Few examples of challenges in the satellite imagery, (a) Occlusion by shadows of trees and buildings, and (b) Visual ambiguity in identifying the roads.

1.1 The Context

Satellite imagery is the rich and structured data source to provide accurate, easily accessible, and reliable spatial information for Geographical Information System (GIS). These images provide promising avenues for automatic mapping of the road network on earth and update the existing ones. With recent advancement in technology satellite images are highly improved in terms of spatial, spectral and temporal resolutions and geomatic communities are overwhelmed by the sheer volume of collected images.

Roads are the significant man made structure to be extracted automatically. Roads can be extracted as the process of identification and accurate localization of road pixels in the image so that when the transformation mapping from image to ground systems is performed, the true road network is obtained. In high resolution satellite images, roads are considered as elongated homogeneous regions that contrast from background with distinct visual characteristics. Automatic road extraction problem from these images is divided into three steps as (i) road detection/segmentation, (ii) center line extraction via skeletonization of road segments, and (iii) finally vectorization of extracted road line segments. Road segmentation is defined as the process of labeling each pixel in the image as `road` and `background`. This process classifies the entire image into two regions and has a major influence on the success of next stages. Automatic road segmentation is approached with machine learning or image processing based algorithms. The segmented image, usually contain some false and missed road pixels, is then processed

to remove false road pixels and connect the broken road segments using image processing algorithms. The processed segmentation image is converted into road center lines. Finally, the extracted road center line is vectorized and transformed into format required for GIS applications.

Object (roads) segmentation of large scale data has been tackled through various methods [84], [27], [39], [4], [55], in the computer vision community and machine learning techniques are most suited for such problems. At the time of writing, the state of the art techniques for automatic labeling of each pixel in image is demonstrated by the success of deep learning research methodologies [48], [32], [12], [54], [5], [11]. *Deep learning* is a subset of machine learning, and refers to the application of a set of algorithms called neural networks, and their variants. In such methods, the network/model learns to approximate the unknown mapping function with experience in the form of labeled examples. In road segmentation unknown function is transformation of satellite imagery to road segmentation. A specific designed neural network, named as *Convolution Neural Network (CNN)*, are most popular to solve the problems in the image domain. However, the performance of deep learning methods is significant effected by the number of available labeled data to train the model, that is to learn the appropriate mapping function *CNN* require huge amount of data. Creating the pixel based annotated data, as required for training the model is tedious and time consuming task. Thus, due to lack of huge annotated data, the deep learning algorithms face challenges for road segmentation from satellite images. We also observe that segmentation results obtained using the pixel based classification techniques are impractical for real world usage. As the generated road network has disconnected components in the final map because of various occlusion by trees, buildings and shadows. We target these challenges in the thesis and develop new and modify existing algorithms to resolve the issues.

1.2 Problem of Interest

We will be focusing on two sub-problems to extract the road network from satellite images, as described below:

1. **Lack of Labeled Data:** Supervised learning techniques to perform pixel wise semantic semantic segmentation requires large amount of annotated data. However, due to lack of labeled data, in this problem we aim to perform road segmentation with the less labeled/annotated data.

2. **Correct Topological Rod Network:** Performing pixel wise road classification (semantic segmentation) is not appropriate to produce the correct topological road network. Here, the goal is to utilize nearest pixel context/information to generate the correct road network.

1.3 Contributions of this Work

In this thesis, we propose to use improved self-supervised learning [67],[90], [20], [63] technique to mitigate the problem of less labeled data required for training the deep convolution neural networks.

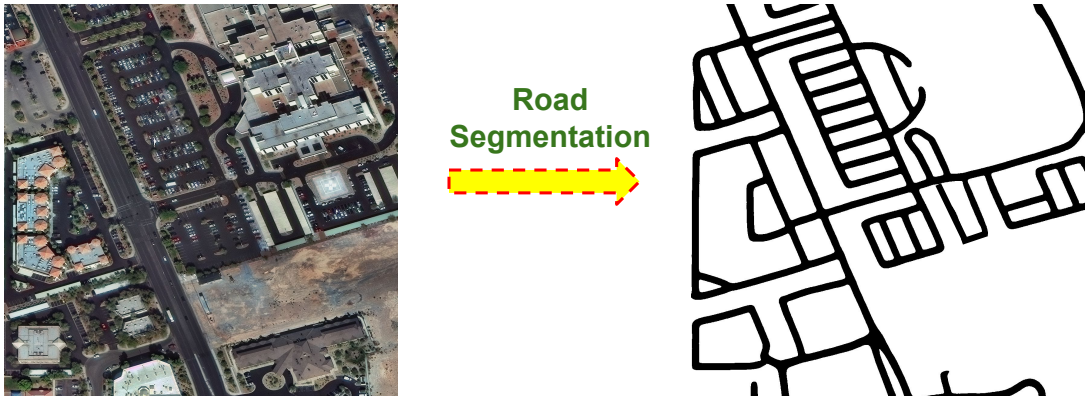

Figure 1.2: An example of road extraction from the satellite imagery. Black pixels corresponds to the extracted road network.

Then we move to relatively harder problem of inferring the connected road topological network using the multi-task learning [7]. The major contribution of the thesis are listed below:

Lack of Labeled Data

1. We propose to initialize both encoder as well as decoder network with pretrained model obtained via self-supervised technique.

2. We propose to use adversarial training scheme for self-supervised learning by increasing the pretext task difficulty gradually and show that it leads to better performance of the target task.

3. We explore the generalization of the proposed approach on different scene understanding tasks e.g. Land classification and Scene Parsing, and show that it improve the performance over training from random initialization of model.

Correct Topological Rod Network

1. We design a novel task of orientation learning to improve the connectivity of road network. We also explore the generalization and significance of the proposed task with other related tasks and different architectures.

2. We use multi-task learning technique to share the knowledge between related orientation learning and road segmentation task. This technique also has an advantage of handling less labeled data.

3. To further improve the connectivity, we use refinement approach using another CNN. We also propose a stacked multi-branch network, which inherently has the capability of refining and connecting the broken road segments.

1.4 Thesis Outline

The structure of thesis is organized in three following parts:

1. We start with prior work on the road extraction and explain knowledge sharing techniques to improve the performance of supervised learning models.

2. In chapter 3, we describe our approach of transfer learning technique, utilizing adversarial approach to increased difficulty task in pre-training stage. Finally, pre-trained network is fine tune to perform semantic segmentation of road network.

3. In chapter 4, we describe our novel task which constrain the convolution neural network to generate connected road segments. To further improve the road topology, an iterative framework is introduced with a multi-branch module.

4. Finally, chapter 5 provides the summary of our work with some future extensions in this field.

Chapter 2

Prior Work and Dataset

In this chapter, we review the prior work studied by the research community in the field of road extraction from satellite images. We briefly describe prior work in two categories: (i) traditional techniques using image processing and simple classification approaches for rad segmentation, and (ii) modern deep learning techniques using the state-of-art *Convolution Neural Networks*. Later, we describe the required background of *Knowledge Sharing* techniques which we adopted and modify appropriately to improve the performance of data hungry deep learning algorithms. Finally, we discuss the public road dataset statistics and show few diverse examples of them.

2.1 Road Extraction Techniques

2.1.1 Traditional Methods

Historically, road extraction is approached by semi-automatic techniques where the human provides a seed point to initiate the extraction algorithms for example road tracking [84] and active contours or snakes [27], [79]. Tracking algorithms follow the roads from a set of seeding points and operate in iterative fashion by choosing the next position and validating it. Vosselman *et al.* [84] describe the initial point with geometrical properties of road such as shape and position and predict the next point using template matching. Finally, predict the parameters (shape and position) of new point using Kalman filter [39] and continue tracking the road. Another semi-automatic approach to extract road from image is active contour model or snakes. Snake is an energy minimizing technique to extract object outline from a possibly noisy 2D image. It is influenced by external and internal forces originated from an image which pulls the snake towards features of interest such as roads. The internal and external energies are obtained from the geometrical (curvature and continuity of homogeneous road structures) and photometric (gradient and intensity values) properties of the roads in the image. [27], [46] perform road extraction using snakes approach.

Barzohar *et al.* [4] proposed the first automatic road extraction method based on the assumption of geometrical and radiometric characteristics of the road. Barzohar *et al.* use dynamic programming to compute the MAP estimate under the Gibbs distribution. Hinz *et al.* [36] incorporate local (building

extraction, vehicle detection etc.) and global (urban, rural and forest classification) contextual information along with geometrical constraints into the road model. Finally, from region of interest road substructures such as markings and lanes are used to extract road segments and further linked them into a global road network by an iterative grouping.

Song *et al.* [76] proposes classification based approach for road extraction using *Support Vector Machine* classifier followed by region growing segmentation. Song *et al.* train *SVM* with pixel spectral features to create two groups of `road` and `non-roads` and obtain masked road image. Finally, partitioned the masked road image into objects using the geometry of roads such as shape. Later Mokhtarzade *et al.* [58] propose the utilization of artificial neural network as better classifier based on RGB values of 3×3 surrounding pixels. Mnih *et al.* [55], [57] use the similar approach but propose to use different architecture named as *Convolution Neural Network* in multiple stages to remove the false road predictions. The network was used in a patch-based framework where pixels in a 16×16 window was classified based on the surrounding 64×64 pixels. Readers are suggested to read [85] for detailed review of traditional techniques for road extraction from satellite images.

2.1.2 Deep Learning Methods

Krizhevsky *et al.* [44] demonstrated the success of deep learning field by training a deep convolution neural network for image classification. With the success of deep learning techniques in various fields (object classification, object recognition and scene parsing) and large availability of very high resolution satellite imagery, the research community focused towards automating geospatial applications like road extraction, land classification and scene parsing from satellite images.

Marcu *et al.* [50] proposed Local-Global CNN architecture for semantic segmentation of buildings and roads from aerial imagery. They used two independent pathways using adjusted VGG-Net [74] for local (64×64 patch) and AlexNet [44] for global (256×256 patch) context interpretation from the image. Marcu *et al.* also demonstrate the significance of the larger spatial context in aerial imagery in contrast to local patch based techniques [55]. Costea *et al.* [16] use the similar architecture for the detection of road and intersection of roads. Similarly, Cheng *et al.* [13] proposed *CasNet* (Cascaded Network), consisting two convolution neural networks for road detection and road center line extraction. Cheng *et al.* utilizes a smaller variant of VGG-16 [74] for *CasNet* due to unavailability of large road dataset. Most of these techniques focused on the road detection instead of generating the connected road topology.

Mattyus *et al.* [54] tackle the connectivity challenge by introducing an algorithm based on the graph theory after the road segmentation. Mattyus *et al.* [54] utilized an encoder-decoder structure model and pose it as multi-class (`roads`, `building` and `background`) segmentation problem. The network performs well in the segmentation to extract the road segments, but lacks to connect them due to occlusions. Then Mattyus *et al.* propose post-processing algorithm by generating a graph from road segmentation. To generate grpah they apply thinning algorithm of skeletonization followed by RamerDouglasPeucker algorithm [70], [21] for smoothing the graph. Once the graph is created for

the road segments, author generate connections between the leaf nodes using shortest path algorithms. Finally, use a binary decision classifier to predict the correctness of connections and make the valid connections to improve the connected road topology. In parallel, Mosinska *et al.* [60] propose to use perceptual loss along with pixel based loss to learn higher order statistics, which authors claim to be important for learning structured output from satellite images for road network. Mosinska *et al.* use U-Net [73] architecture, followed by the refinement process.

Most of architectures used for semantic segmentation of thin and curvilinear road structures in the remote sensing and computer vision community is based on encoder-decoder network. Few prominent architectures are U-Net [73] and LinkNet [10], as shown in Figure 2.1. At the time of writing, these architectures perform significantly better for road detection in Spacenet [80] and DeepGlobe [18] challenges.

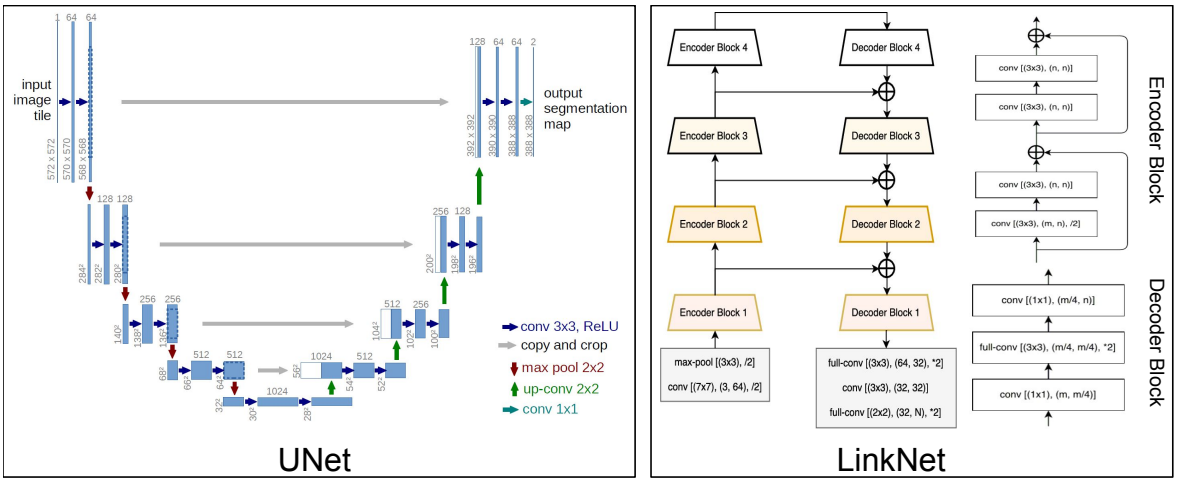

Figure 2.1: Salient Encoder-Decoder Architectures for Semantic Segmentation of thin and curvilinear road structures.

2.2 Knowledge Sharing Techniques

Human beings attempts to find the consistent patterns among our experiences and make certain hypothesis about their features and causes. In real life scenarios humans do not receive tasks in isolation, but instead receive sequence of related tasks over the time. And, humans transfer the knowledge from one scenario to another by utilizing the hypothesis from prior experiences. For example, when humans learns to ride a bicycle, they will learn traffic rules, balancing the vehicle, and how and when to apply brakes. With the experience of ridding a bicycle, humans can learn to ride motor-bike faster by transferring their prior knowledge/experience. In general this ability of knowledge sharing helps the humans to learn complex concepts by first learning simple concepts. In artificial system, the supervised learning techniques learn from a single task in isolation. Machine learning algorithms can take advantage

from human behaviour of knowledge sharing to improve the performance of complex task by learning sequence of smaller tasks.

Supervised machine learning algorithms not able to generalize well to unseen data, if they are trained with insufficient labeled examples for training. In real world applications it is common scenario due to intensive and laborious work to create the labeled data. For example, creating the labeled data for roads or building from satellite images is very tedious task. To address this challenge, another related task is trained in supervised fashion and knowledge is transferred to the target task. Pratt *et al.* [69], [68] pioneered the transfer learning technique to adjust the decision hyperplane by transferring the pretrained neural network weights from a source to target task. Mathematically, in transfer learning source and target task has different labels i.e. $Y^s \neq Y^t$, but have the similar marginal input distribution $P(X)^s \sim P(X)^t$. It can be categorized into (i) inductive transfer learning, (ii) transductive transfer learning, and (iii) unsupervised learning based on the different settings of data availability in source and target domain. We describe two knowledge sharing techniques below and readers are suggested to refer [65], [78] for detailed techniques.

2.2.1 Self-Supervised Learning

Self-Supervised learning is considered as special of case of unsupervised learning. In this technique model is trained in supervised fashion by extracting the relevant labels, called *pseudo labels*, from the embedded metadata of input data. For example, training a convolution neural network to inpaint the erased image regions or colorizing the gray scale images. Self-Supervised techniques [20], [25], [63], [67], [90] learn the strong visual representations of input data by predicting pseudo labels. Features captured by the pre-training stage in the model are transferred to target task, in our case road segmentation. Finally, the model is fine tuned to perform the road segmentation and leads to superior performance.

2.2.2 Multi-Task Learning

Multi-task Learning (MTL) [7] is a learning mechanism, inspired from human beings to acquire knowledge of complex tasks by performing different shared sub-tasks simultaneously It can be considered as an approach to inductive knowledge transfer which improves generalization by sharing the domain information between the related tasks. MTL has been applied successfully in the various domains such as speech recognition, natural language processing [15] and computer vision [41]. Readers are suggested to read survey [91] on multi-task learning.

2.3 Dataset Details

We perform our experiments on the two challenging road datasets Spacenet [80] and DeepGlobe[18] to assess the significance of orientation learning task. In the current scope we utilize only 3-Band RGB image of both datasets to detect the curvilinear road structures.

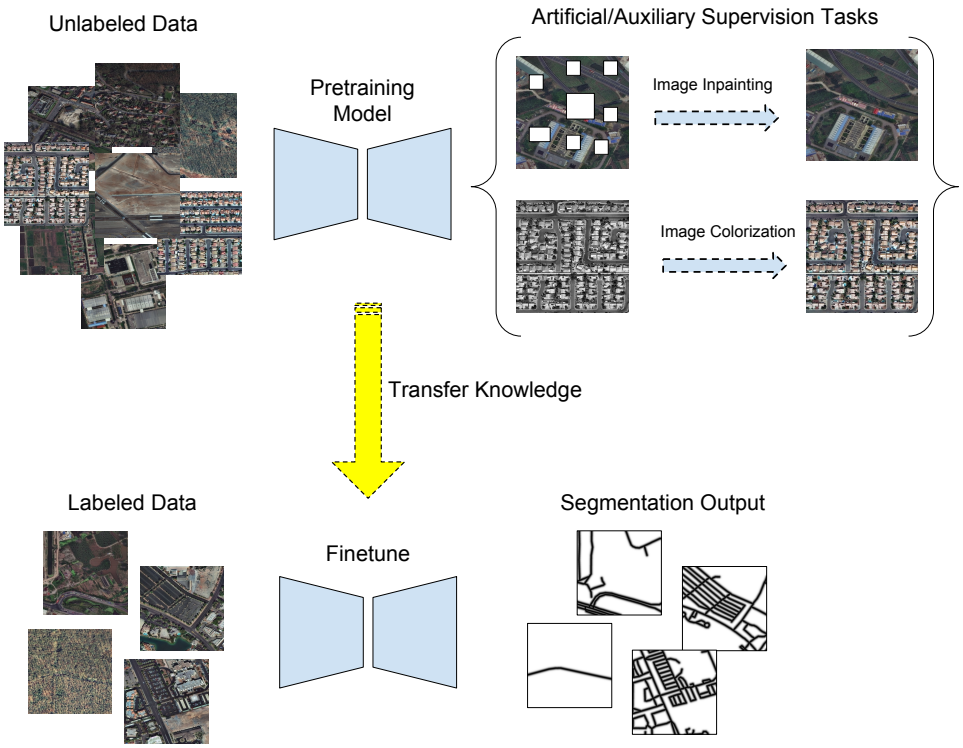

Figure 2.2: *Self Supervised Learning*. This approach use unlabeled data to learn visual representations of the dataset with an artificial/auxiliary supervision. Finally, the model is fine tuned with labeled dataset for target task, in our case target task is road segmentation.

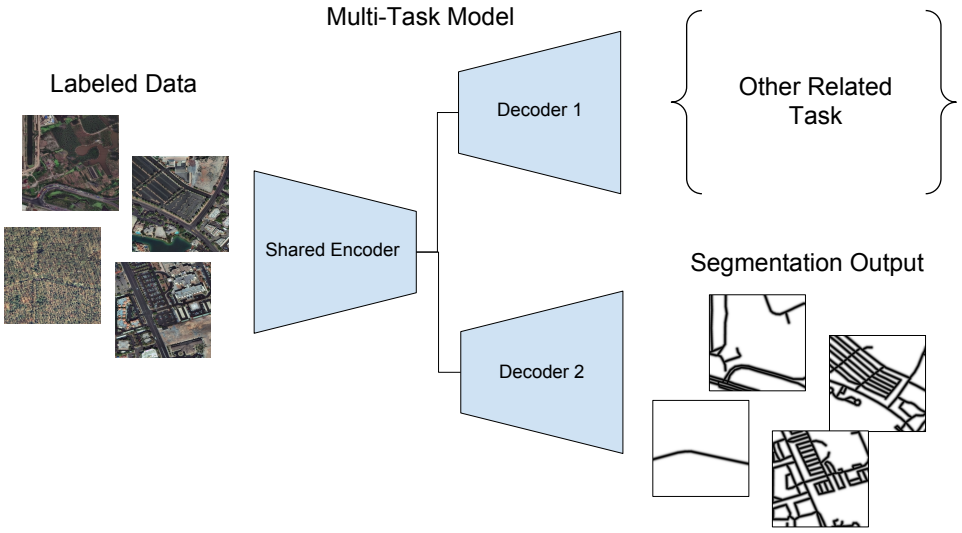

Figure 2.3: *Multi Task Learning*. This approach use labeled data in a joint learning fashion with other related task to learn common visual representations, which leads to improvement in the performance of target task.

Figure 2.4: Sample images from *Spacenet* [80] Road Dataset with road center lines.

Spacenet [80]: This dataset provides imagery from four different cities: `Paris`, `Las Vegas`, `Shanghai`, and `Khartoum`.The imagery is available at ground resolution of 30cm/pixel and pixel resolution of 1300×1300 in GeoTiff format (16-bit). Dataset consists of different road types (`Motorway`, `Primary`, `Secondary`, `Tertiary`, `Residential`, `Unclassified`, `Cart Tracks`) from the four cities, having diverse road widths and visual appearance. Road annotation is provided in the form of line-string, representing the centerline of roads. Each image may have multiple line-strings and each line-string consists of pixel coordinates $\{\text{X Y}\}$ depicting road centerline points in the 2D image plane, assuming top-left corner as the origin.

The public dataset consists of 2567 images with road vector data as labels. We split the dataset into 2000 images for training and 567 for testing. When we split the imagery, each city equally contributes to train (80% per city) and validation set (20% per city). To augment the training dataset we create crops of 650×650 with overlapping region of 215 pixels, thus providing $\sim$32K images. For validation we use the crops of same size without overlap.

DeepGlobe [18]: It includes GeoTiff imagery from three different areas: `Thailand`, `Indonesia`, and `India`. The ground resolution of 3-Band RGB image is 50cm/pixel and pixel resolution is 1024×1024. In this dataset the labels generated are pixel based, where all the pixels belonging to road are annotated, instead of labeling only the center line of road. As shown, the urban morphology, the illumination conditions, the road density, and the structure of the street networks are significantly diverse among the samples. The public dataset contains 6226 images, spanning a total land area of $1632km^2$. We create our own splits of 4696 images for training phase and 1530 for validation.

Figure 2.5: Sample images from *DeepGlobe* [18] Road Dataset with road center lines. Center lines are obtained using skeletonization of ground truth segmentation mask followed by vectorization.

The dataset consists of different types of road surfaces (unpaved, paved, dirt roads), rural and urban areas, etc. We augment it by creating crops of size 512×512 with overlapping region of 256 pixels, yielding $\sim$42K images for training phase.

Chapter 3

Self-Supervised Feature Learning for Semantic Segmentation

Overhead imageries play a crucial role in many applications such as urban planning, crop yield forecasting, mapping, and policy making. Semantic segmentation could enable automatic, efficient, and large-scale understanding of overhead imageries for these applications. However, semantic segmentation of overhead imageries is a challenging task, primarily due to the large domain gap from existing research in ground imageries, unavailability of large-scale dataset with pixel-level annotations, and inherent complexity in the task. Readily available vast amount of unlabeled overhead imageries share more common structures and patterns compared to the ground imageries, therefore, its large-scale analysis could benefit from unsupervised feature learning techniques.

In this work, we study various self-supervised feature learning techniques for semantic segmentation of overhead imageries. We choose image semantic inpainting as a self-supervised task [67] for our experiments due to its proximity to the semantic segmentation task. We (i) show that existing approaches are inefficient for semantic segmentation, (ii) propose architectural changes towards self-supervised learning for semantic segmentation, (iii) propose an adversarial training scheme for self-supervised learning by increasing the pretext task's difficulty gradually and show that it leads to learning better features, and (iv) propose a unified approach for overhead scene parsing, road network extraction, and land cover estimation. Our approach improves over training from scratch by more than 10% and ImageNet pre-trained network by more than 5% mIOU.

3.1 Introduction

Overhead imageries are images captured by imaging satellites, aeroplanes, drones, etc. They can be updated easily as well as frequently [53]. In contrast to ground imageries which are often captured with digital, portable, or surveillance cameras, overhead imageries present a unique and occlusion-free view of a large geographical area (see Figure 3.1 (a)). Due to this, they are extensively used for land cover classification [33, 45], scene parsing [40, 64, 1], road network extraction [55, 57, 87, 52, 53, 60, 6, 5], etc. However, the focus has been towards a specific narrow area of application. In this work, we present a unified approach, based on semantic segmentation, towards a variety of overhead imagery tasks —

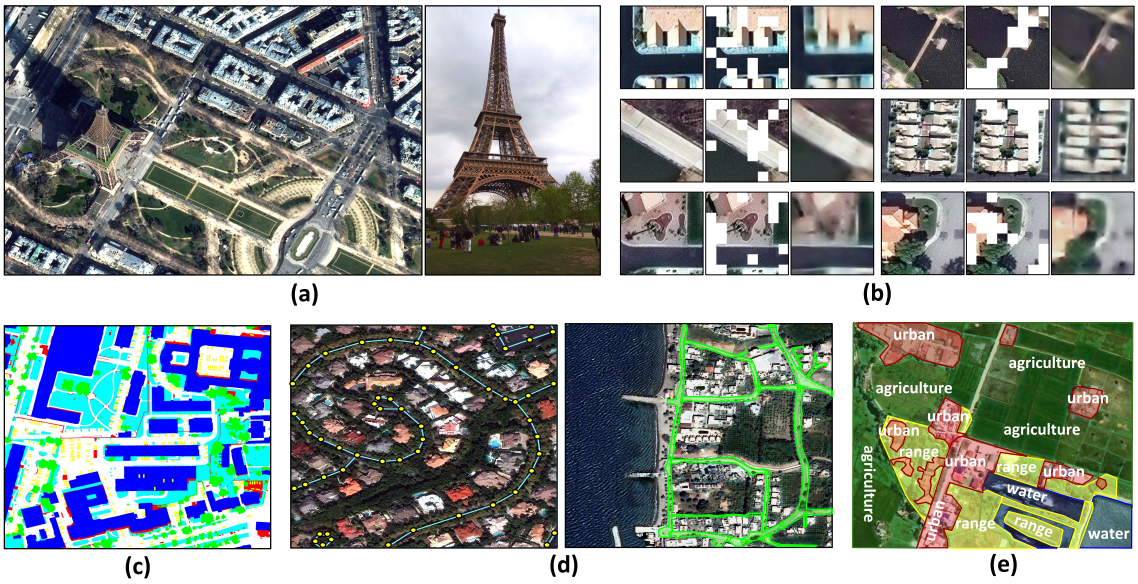

Figure 3.1: (a) There exist a large domain gap between ground and overhead imageries due to the perspective. (b) Overhead imageries exhibit rich context information making it suitable for unsupervised feature learning with inpainting. Semantic segmentation of overhead imagery enables a variety of tasks: (c) scene parsing of a city, (d) road network extraction, and (e) land cover estimation.

(i) scene parsing of a city, (ii) road network extraction in urban and remote areas, and (iii) land cover estimation in diverse geographical terrains.

Overhead imagery captures a vast geographical area with diverse landscapes, objects and extreme variations in their count, size, and aspect-ratio. Moreover, significant diversity arises due to illumination, region's geography, weather conditions, etc. Undoubtedly, there is a large domain gap from existing research in ground imageries primarily due to the overhead perspective. Transfer learning between these domains is, therefore, unsuitable. Overhead view-point of objects and scenes are highly ambiguous and their annotations require domain expertise due to the uncommon appearance. Unavailability of large-scale dataset with pixel-level annotations for different overhead imagery tasks further limits the utility of current semantic segmentation techniques [48, 47, 62, 2].

In this work, we exploit the strong context information and spatial relationship present in overhead imageries to learn useful features at the pre-training stage with self-supervised technique. We employ semantic inpainting as the self-supervised task [67] (Figure 3.1 (b)) due to its proximity to the semantic segmentation task. We propose architectural changes (3.3.1) enabling the pre-training of encoder network which preserves the spatial context of features as well as the decoder network which learns to upsample the features with respect to the semantic boundary of entities, an essential ingredient for semantic segmentation.

Semantic inpainting as self-supervised task leads to learning useful features only when the region filling task is non-trivial. Curated object-centric datasets (ImageNet [19], Pascal VOC [22], etc.) are inherently diverse and objects occupy a significant portion of the image. Erasing fixed or random re-

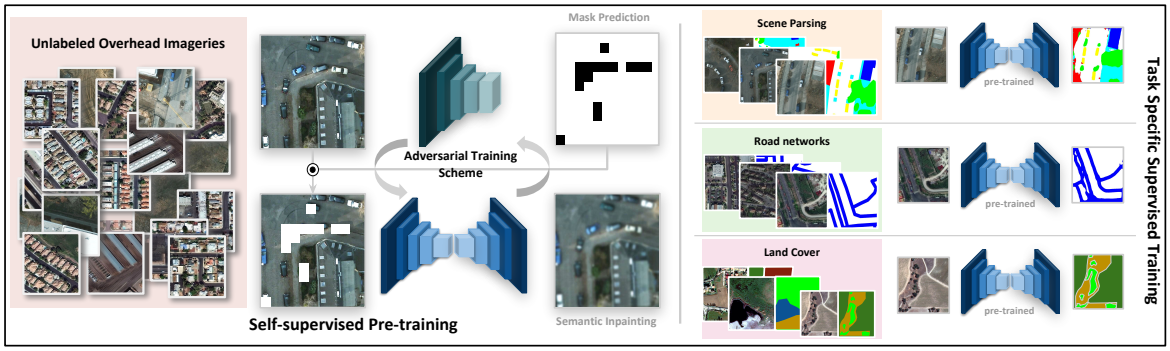

Figure 3.2: An overview of our approach: self-supervised pre-training (left) and task specific supervised training (right). We use semantic inpainting as self-supervised pre-training [67] to exploit the freely available large amount of unlabeled overhead imageries. To ensure the pretext task's difficulty we train the inpainting network with an adversarial mask prediction scheme. The pre-trained encoder-decoder inpainting network is then fine-tuned for a variety of overhead imagery tasks: scene parsing, road network extraction, and land cover estimation.

gions from object centric images, therefore, is adequately difficult. On the contrary, overhead imageries with much wider world-view lacks specific subject in the images. To ensure the pretext task's difficulty, instead of inpainting random regions [67] (Figure 3.1 (b) left), we propose to inpaint difficult and semantically meaningful regions (Figure 3.1 (b) right) with an adversarial training scheme consisting of coach and inpainting networks (3.3.2). The coach network see an entire image and predicts an increasingly difficult mask which is used to erase the corresponding regions of the image. The inpainting network then tries to fill-in the erased regions with the help of available contexts. At the end of the pre-training stage, the inpainting network learn to efficiently encode the available contexts and upsample the activation maps for overhead imageries. The pre-trained model is further used as initialization for different overhead imagery tasks. Figure 3.2 shows the overview of our approach.

Contributions

1. We show that existing self-supervised techniques focusing on the encoder network alone are inefficient for semantic segmentation. We propose architectural changes towards self-supervised pre-training of encoder as well as decoder networks.

2. We propose an adversarial training scheme for self-supervised learning by increasing the pretext task difficulty gradually and show that it leads to superior performance.

3. We also propose a unified segmentation based approach for scene parsing, road network extraction, and land cover estimation in overhead imageries. Our technique improves over training from scratch by more than 10% and ImageNet pre-trained network by more than 5% mIOU.

3.2 Related Work

Overhead Imagery Understanding The overhead imagery community, in the past, has mostly focused on specific task and application individually. The prominent tasks in this domain are land cover classification [33, 45], scene parsing [40, 64, 1], and road network extraction [55, 57, 87, 52, 53, 60, 6, 5]. Readers are suggested to see [94] for a comprehensive survey on recent developments in overhead imagery analysis. Unsupervised input reconstruction [55, 57], supervised pre-training on natural images [1], and data augmentations with balancing class population [40] have been explored to overcome the data scarcity. In contrast to these works, we perform pre-training with self-supervision from the same domain and show its efficacy in a unified semantic segmentation approach for scene parsing, road network estimation, and land cover estimation.

Unsupervised and Self-Supervised Feature Learning Deep learning models require a large amount of annotated data to train from scratch. RBMs [34], Autoencoders [35], and its variants [82, 71, 83] have been popular choice for unsupervised pre-training where labeled data is scarce [55, 57]. Recently, self-supervised learning techniques [20, 67, 63, 86, 25, 90] using freely available *pseudo* labels have emerged as a superior technique due to stronger self-supervision. Doersch *et al.* [20] proposed to learn representations by predicting relative position of two patches in an image. Noroozi *et al.* [63] extended this idea further to train the network for solving jigsaw puzzles. Zhang *et al.* [90] proposed Split-Brain Autoencoders, two disjoint sub-networks each trained to predict the missing image channel(s). Pathak *et al.* [67] proposed Context Encoders to predict the contents of missing regions in the image using the available contexts. Note that, [20, 67, 63, 25, 90] focus on pre-training the encoder networks alone, and therefore, are inefficient for semantic segmentation. Furthermore, difficulty level of the hand-crafted self-supervised tasks are fixed across examples depending on the nature of the task itself. In this work, we propose an adversarial training scheme capable of generating increasingly difficult examples for pre-training based on content of the image.

Semantic Segmentation Recent semantic segmentation [48, 2, 62, 47] techniques rely on backbone model pre-trained on related task such as supervised image classification. Self-supervised pre-training have also shown promising results on popular benchmarks [67, 90]. In both supervised [48, 2, 62, 47] and self-supervised pre-training [67, 90], the decoder network is trained from scratch for semantic segmentation. In contrast to this, we propose pre-training of the encoder as well as the decoder networks with semantic inpainting task.

3.3 Method

3.3.1 Semantic Inpainting as Self-supervision

Image semantic inpainting refers to predicting the actual image x from its corrupted version $\hat{x}$ using a convolution neural network. In this pretext task the training pair consists of corrupted image $\hat{x}$, which is

obtained by randomly erasing regions from the image and pseudo label is the full image x. To generate realistic content in the erased image, the model needs learn structure and contextual information from the surrounding areas of erased region. Thus the model learn visual features through the image generation process.

Pathak *et al.* [67] proposed semantic inpainting as self-supervision to learn visual representation of the image. In [67], a random binary mask M is generated for each image such that the pixels with corresponding mask value 0 are erased from the image, and 1 are kept intact.

$$\hat{x} = M \odot x \tag{3.1}$$

where $\odot$ is the element-wise product operation. The inpainting model F learns to inpaint images by minimizing the masked L_2 distance as reconstruction loss, $\mathcal{L}_{rec}$. We add additional loss term for context regions, $\mathcal{L}_{con}$, to allow the network to reconstruct the entire image and learn to upsample activation maps effectively.

$$L_{rec}(\hat{x}) = \frac{1}{\sum(1-M)} ||(1-M) \odot (x - F(M \odot x))||_2^2 \tag{3.2}$$

$$L_{con}(\hat{x}) = \frac{1}{\sum M} \qquad ||M \odot (x - F((1-M) \odot x))||_2^2 \tag{3.3}$$

The final loss $\mathcal{L}_{inpainting}$ is the weighted sum of reconstruction and context losses.

$$L_{inpainting} = \mathbf{w}_{rec}\mathcal{L}_{rec} + \mathbf{w}_{con}\mathcal{L}_{con} \tag{3.4}$$

Architectural Improvements We propose architectural changes to the semantic inpainting encoder-decoder architecture used in [67]. We use **(a)** a more powerful ResNet-18 [29] as the backbone encoder network, **(b)** do away with the channel-wise fully-connected bottleneck layer, and **(c)** exploit the pre-trained encoder as well as decoder for the segmentation task. ResNet [29], compared to AlexNet [44] equivalent used in [67], have the potential to learn better representations, is more efficient as well as easier and faster to train [31]. Furthermore, `BatchNorm` [37] helps in reducing the domain gap between semantic inpainting of corrupted images and semantic segmentation of natural images since the input to convolutional layers follow the same distribution during both stages.

Fully-connected bottleneck layer in an encoder-decoder network connects all spatial locations together, however, also results in losing the vital spatial context. Deep CNNs' (AlexNet [44], VGG [74], ResNet [29]) convolutional filters possess large enough field-of-view (FOV) to *see* the spatial extent of 195×195 pixels (or more) of input [49]. We use the input size 128×128 for inpainting which is well within reach of the encoder network's FOV. By **not** employing the fully-connected bottleneck layer in our architecture, the resulting network is fully convolutional, able to preserve the spatial context, and has fewer parameters.

Lastly, while learning to inpaint, the decoder network tries to push the low resolution feature up to the semantic boundary of the entities at input resolution. The decoder network learns non-linear weighted upsampling of the low resolution feature maps which we show is useful for the target segmentation task.

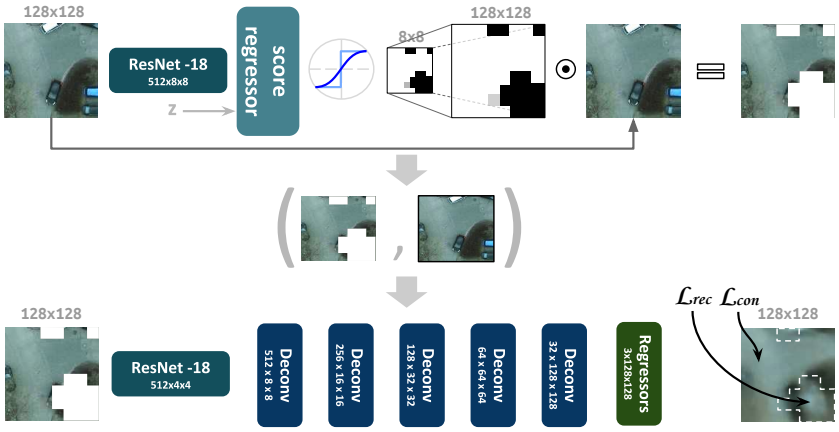

Figure 3.3: The coach network (top) take the image as input and outputs a semantically meaningful binary mask. The mask is used to erase parts of the image which then is used as input to the inpainting network (bottom). The inpainting network learns to encode visual representations as well as upsampling by trying to fill-in the image regions erased by the coach. The coach is trained with loss adversary to the inpainting network making it capable of generating increasingly difficult examples (see 3.3.2).

To the best of our knowledge, ours is the first architecture that re-uses the encoder as well as the decoder network for the target task.

3.3.2 Coach Network

Pathak *et al.* [67] inpaints the image erased by a randomly generated binary mask. The mask dictates the surrounding area of corrupted regions to learn the context and inpaint regions with meaningful content based on the context. For example, in Figure.3.4a the model needs to fill the sea color in the missing region without much learning from the context. However in Figure.3.4b the model needs to learn more contextual information to generate the image, hence lead to learn good quality features in convolution layers.

Overhead imageries with much wider world-view lacks specific subject in the images, therefore to learn useful representations, its inpainting task need masks that can erase semantically meaningful and difficult regions. Identifying meaningful regions or difficult examples without labeled data is extremely difficult. Similar ideas have recently been proposed by Gao *et al.* [24] and Wei *et al.* [89] to identify and use masks with different difficult levels for training, however, with a focus on handling arbitrary levels of corruption in semantic inpainting and weakly-supervised semantic segmentation with a pre-trained model, respectively. In contrast to using random mask from pre-defined distributions in [24], the coach network learns to score the regions based on difficulty in its inpainting. The coach is trained with loss adversarial to the reconstruction loss. We hypothesize that when trained on multiple examples the coach learns to identify the patches having poor reconstruction by predicting the reconstruction difficulty score. In this way, the coach learns to create increasingly difficult examples for the inpainting network. We propose *coach network* that learns a *semantically meaningful* mask M for the given image

(a) (b)

Figure 3.4: Image in-painting on natural images with corrupted image, ground truth and reconstructed image. (a) easy image in-painting task, as the network to generate image with sea color to fill the erased region, (b) difficult task, as the network need to learn context from the image to generate the missing image.

x (see Figure 3.3). The coach model C learns to assign meaningful score to the regions in image x by **maximizing** the reconstruction loss.

$$L_{coach}(x) = 1 - \mathcal{L}_{rec}(x \odot M) = 1 - \mathcal{L}_{rec}(x \odot C(x)) \tag{3.5}$$

However, applying this loss naïvely would result in the masks having 0 value at all regions because then no context information is present for the inpainting model and maximum reconstruction loss is achieved. Therefore, we apply constraints on outputs of the coach model to ensure a constant fraction of the images is always available as context for inpainting.

$$\hat{B}(x) = B(x) - \texttt{SORT}(B(x))^{k|B(x)|} \tag{3.6}$$

$$M = C(x) = \sigma(\alpha \hat{B}(x)) \tag{3.7}$$

The backbone network B of the coach model C has the same architecture as the encoder network of inpainting model. This gives the coach approximately similar representation power as the encoder network. $\texttt{SORT}(B(x))$ represents the sorting operation in descending order over all values in the activation map. $|B(x)|$ denotes the spatial size of activation map, k represents the k^{th} element in the sorted list of scores and controls the fraction of image to be erased. $\hat{B}(x)$ gives the relative difficulty score for each region with respect to the k^{th} element. The regions with score lesser than the k^{th} element are erased from the image while the other regions are kept intact. For example, $k = 0.75$ would erase $\frac{1}{4}$ area of the image. We scale the scores to the range [0, 1] using point-wise sigmoid function $\sigma(\alpha \mathbf{x})$, where α is a scalar that controls the steepness of σ. High α value results in discrete masks value $\{0, 1\}$ (for inpainting mask), whereas low α results in continuous mask values [0,1] (for training coach model). We use $\alpha = 1$ while training the coach network, and step-function ($\alpha \to \infty$) while training the inpainting network.

3.3.3 Training

We train coach and inpainting networks in an alternate fashion creating a competition between the models. The coach model learns to create increasingly difficult examples for the inpainting model while

Figure 3.5: Coach model predicts an increasingly difficult masks for semantic inpainting. For each row, from left to right: Input image (512×512) for the coach network, masks predicted at iterations 0, 1, 6, and 8 with corresponding inpainting output. Note that at iteration 0 the coach predicts random masks.

the inpainting model learns superior feature with more difficult examples. The overall training objective (ignoring $\mathcal{L}_{con}$ for simplicity) is given by

$$L(x) = \min_{\theta_F} \max_{\theta_C} ||x - F(x \odot C(x, \theta_C), \theta_F)||_2^2 \tag{3.8}$$

where θ_F and θ_C are the parameters of the inpainting network and coach network, respectively. To introduce diversity and stochasticity in mask prediction, we inject noise sampled from a standard normal distribution to the coach's penultimate activation maps with the help of reparameterization [43]. In the first iteration of training inpainting model, we fill the mask (output of the coach network) with values drawn from a uniform distribution, $B(x)^{j,k} \sim U[0, 1]$. We use this random mask as a starting point, instead of random patch mask as used in [67], to keep the nature of corruption same across iterations as semantic inpainting tends to overfit to the type of corruption it has been trained for [24]. Figure 3.5 shows few examples of meaningful and increasingly difficult masks predicted by the coach network.

Dataset	Resolution	Ground Resolution	Train	Validation	Crop Size	Stride	Task
Potsdam [38]	6000×6000	5 cm	20	4	600×600	200×200	Scene Parsing
SpaceNet Road [72]	1300×1300	30 cm	2000	567	650×650	250×250	Road network
DeepGlobe Lands [18]	2448×2448	50 cm	803	171	612×612	228×228	Land cover
DeepGlobe Roads [18]	1024×1024	50 cm	6226	1243	512×512	256×256	Road network
fMoW [14]	variable	50 cm	100000	2000	512×512	non-overlapping	Pre-training

Table 3.1: Statistics and other details for the datasets used in our experiments. We use non-overlapping crops for validation images for all datasets.

3.4 Dataset

We validate our ideas by performing experiments on four disparate datasets of overhead imageries with variations in the task, dataset size, and image ground resolutions. Note that, we use only 3 band RGB images in our experiments. The statistics of these datasets are given in Table 3.1.

Potsdam [38] This dataset is used for scene parsing of the Potsdam city. Pixel-level annotations are provided for 6 classes: `impervious surface`, `building`, `tree`, `low vegetation`, `car`, and `background`. We create crops of size 600×600 with a stride of 200×200 for 20 training images and non-overlapping stride for 4 validation images. Post-processing artifacts are present in significant number of images.

SpaceNet Road [72] This dataset is used for road network estimation in four cities: `Paris`, `Las Vegas`, `Shanghai`, and `Khartoum`. The annotations are provided in the form of line-strings corresponding to the mid-line of a road. We obtain the binary masks by computing distance transform with respect to the mid-line, apply a Gaussian kernel of standard deviation of 15 on the distance transform outputs to obtain a heatmap, and then threshold the heatmap at 0.4. This results in foreground road masks of roughly 12 meters. We create crops of 650×650 with a stride of 250×250 for 2000 training images and non-overlapping stride for 567 validation images. The images are very diverse and consists of motorway, primary highway, secondary highway, tertiary highway, residential road, cart track, and unclassified roads. A significant number of roads have not been labeled.

DeepGlobe Lands [18] This dataset is used for land cover estimation. Pixel-level annotations are provided for 7 classes: `urban`, `agriculture`, `range land`, `forest land`, `water`, `barren land`, and `unknown` (not used for evaluation). We create crops of 612×612 with stride of 228×228 for training images and non-overlapping stride for validation images.

DeepGlobe Roads [18] This dataset is used for road network estimation. Pixel-level annotations are provided for `road` and `background` classes. We create crops of 512×512 with stride of 256×256 for training images and non-overlapping stride for validation images.

Functional Map of the World [14] We use only the images from this dataset to study the feature quality learned with respect to the number of unlabeled examples. We use only the `train` split in our experiments. The images are taken from all over the world and significant diversity in terms of geography, terrain, weather condition, illumination exist in the images. We resize the images preserving the original aspect-ratio such that the minimum image dimension becomes 1024 pixels. We then create crops of 512×512 with non-overlapping stride for training as well as validation images.

3.5 Experiments and Results

3.5.1 Implementation Details

AlexNet architecute for Baselines For pre-training with self-supervised methods Context Prediction [20], Context Encoder [67], and Split-Brain Autoencoder, we use a common AlexNet [44] architecture as an encoder till Conv5 (see Table 3.2) and add `BatchNorm` layer after each convolution layer (similar to [90]). Following Doersch *et al.* [20], we train the Context Prediction network with 2 patches of size 96×96 from spatial grid configuration. The patches are spatially separated by 16 pixels and the locations are further randomly jitter by 7 pixels. We replace `pool5` layer with `AvgPool` layer. This is then followed by `fc6 Linear` layer. Features of both patches are concatenated and fed to `fc7` layer which predicts the relative spatial location.

We use the similar architecture for Context Encoder described by [67] i.e., AlexNet [44] till `pool5` as encoder followed by channel wise fully connected and five `Deconvolution` layers. We randomly erase 16 patches of size 16×16 from input image (227×227) and the network tries to reconstruct the image. For the Split-Brain Autoencoder, we follow the training procedure proposed by Zhang *et al.* [90] and train 2 disjoint networks halved along the channel dimensions. Table 3.2 describes the network details for one half i.e. predicting `ab` channel from `L` channel. We use the same architecture for the second network and reverse the input and output channels. Mean squared loss is calculated with respect to heavily downscaled ground truth of size 12×12.

We adapt and fine-tune the pre-trained networks for semantic segmentation task using FCN [48]. We use `SGD` optimizer to train the network for 100 epochs and step learning rate starting at 0.001, step size 0.1, momentum 0.9 and weight decay of 0.0005.

Semantic Inpainting We use input size of 128×128, batch size of 128 and employ random crops, mirroring, resizing, horizontal flip, and rotations for data augmentation. We empirically set $\mathbf{w}_{rec} = 0.99$ and $\mathbf{w}_{con} = 0.01$ in all experiments and find it to be a good balance between inpainting and learned feature quality. We use MSE loss clipped at `2` and observe that it allows the network to converge faster, predict pixel intensities far from the mean of the distribution. We use SGD optimizer with 0.9 momentum and 0.0005 weight decay to train the inpainting network for 100 epochs and step LR starting at 0.1 with step size 0.1.

Layer	Context Encoder						SplitBrain Autoencoder					
	X	C	K	S	P	D	X	C	K	S	P	D
data	227	3	–	–	–	–	180	1	–	–	–	–
conv1	56	96	11	4	2	1	45	48	11	4	5	1
pool1	27	96	3	2	0	1	23	48	3	2	1	1
conv2	27	256	5	1	2	1	23	128	5	1	2	1
pool2	13	256	3	2	0	1	12	128	3	2	1	1
conv3	13	384	3	1	1	1	12	192	3	1	1	1
conv4	13	384	3	1	1	1	12	192	3	1	1	1
conv5	13	256	3	1	1	1	12	128	3	1	1	1
pool5	6	256	3	2	0	1	12	128	3	1	1	1

Table 3.2: AlexNet architecture used as encoder network in all baseline experiments. X : input spatial resolution for the layer, C : number of channels/filters in the layer, K : convolution or pooling kernel size, S : stride, P : padding, and D : kernel dilation.

Coach Networks Inputs, data augmentation, and batch size for this network is kept same as inpainting network. We remove the `maxpool` layer from ResNet-18 to predict the mask at a resolution of 8×8 and then apply $16\times$ nearest neighbor upsampling to scale the mask to 128×128. We erase 25% of the patches or 16 patches ($k = 0.75$) based on the predicted difficulty score. For Context Encoders [67], we remove 16 random patches of size 16×16 from the image. We train the coach network with Adam optimizer [42] at a fixed learning rate of 10^{-5} for 30 epochs at a time. This is followed by training of inpainting network for 30 epochs at a fixed learning rate of 10^{-5}. We repeat this procedure for 10 iterations.

Semantic Segmentation We adapt the inpainting network for semantic segmentation by removing the pixel-wise regressors. For the variant of inpainting network with bottleneck, following Long *et al.* [48], we apply a pixel-wise classifier at 3 scales: $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$. For the variant of inpainting network without the bottleneck, we apply a pixel-wise classifier at all 5 scales. We train all segmentation networks for 100 epochs and step learning rate starting at 0.001 with step size 0.1. We use input size of 256×256, batch size of 64 and employ the same data augmentation used for the training inpainting network. We observe that training segmentation network using small amount of data with cross-entropy loss leads to variations in segmentation results between re-runs. We train the segmentation network with soft-IOU loss [53] which leads to more stable and reproducible results.

3.5.2 Results

We initialize all parameters with the technique proposed by He *et al.* [28]. We use mean Intersection-Over-Union (mIOU) as the metric at different amount of labeled and unlabeled data used for training. We do **not** apply weights to loss with respect to class population in any experiment and found that pre-training helps in alleviating the effect of class imbalance which is a prominent issue in overhead imagery

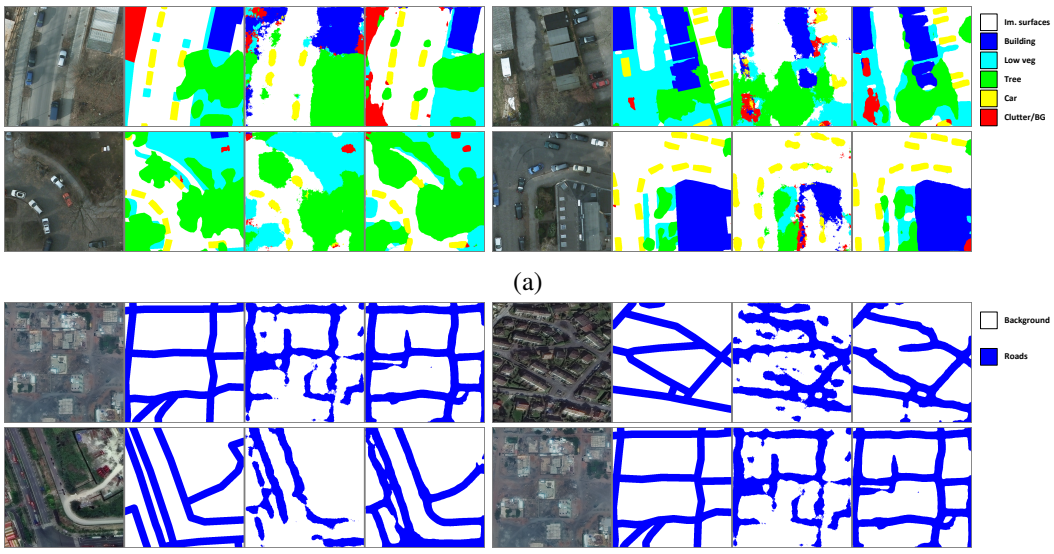

Figure 3.6: Qualitative semantic segmentation results for Potsdam (a) and SpaceNet Road (b), from left to right: input image, ground truth, prediction with model trained from scratch, and prediction with model pre-trained using our approach. 10% of labeled data is used for fine-tuning in all cases.

tasks. Table 3.3 shows the performance of the baselines and our method while using all training images for the self-supervised pipeline and 10% labeled images of respective training set for fine-tuning the segmentation network. Figure 3.6 shows qualitative segmentation results of prediction from a model trained from scratch and a model pre-trained with our method.

Self-supervised Baselines We compare our results with three competitive self-supervised feature learning techniques: (a) Context Prediction [20], (b) Context Encoders [67], and (c) Splitbrain Autoencoders [90]. To evaluate their relative performance, we keep the AlexNet [44] architecture same for all the methods. Context Prediction and Context Encoders both tasks try to learn the structural information in the image, however, Context Encoders perform better in all cases, confirming semantic inpainting task being relatively closer to semantic segmentation task. Splitbrain AE outperforms Context Prediction and Context Encoders, confirming the findings of [90].

Architectural improvement ResNet-18 pre-trained on ImageNet performs better than training from scratch (see Table 3.3). This can be explained with the fact that the weights of earlier layers are generic and rarely change across domains. However, pre-training on ImageNet performs worse than simple autoencoder pre-training suggesting the large gaps between ground and overhead imageries. Table 3.3 also shows that having no bottleneck and re-using the pre-trained decoder network along with the encoder significantly improves the results, specially for road network extraction.

Interestingly, for DG Lands, pre-training on ImageNet performs better than unsupervised and self-supervised pre-training. We hypothesize that this is because image reconstruction and inpainting of

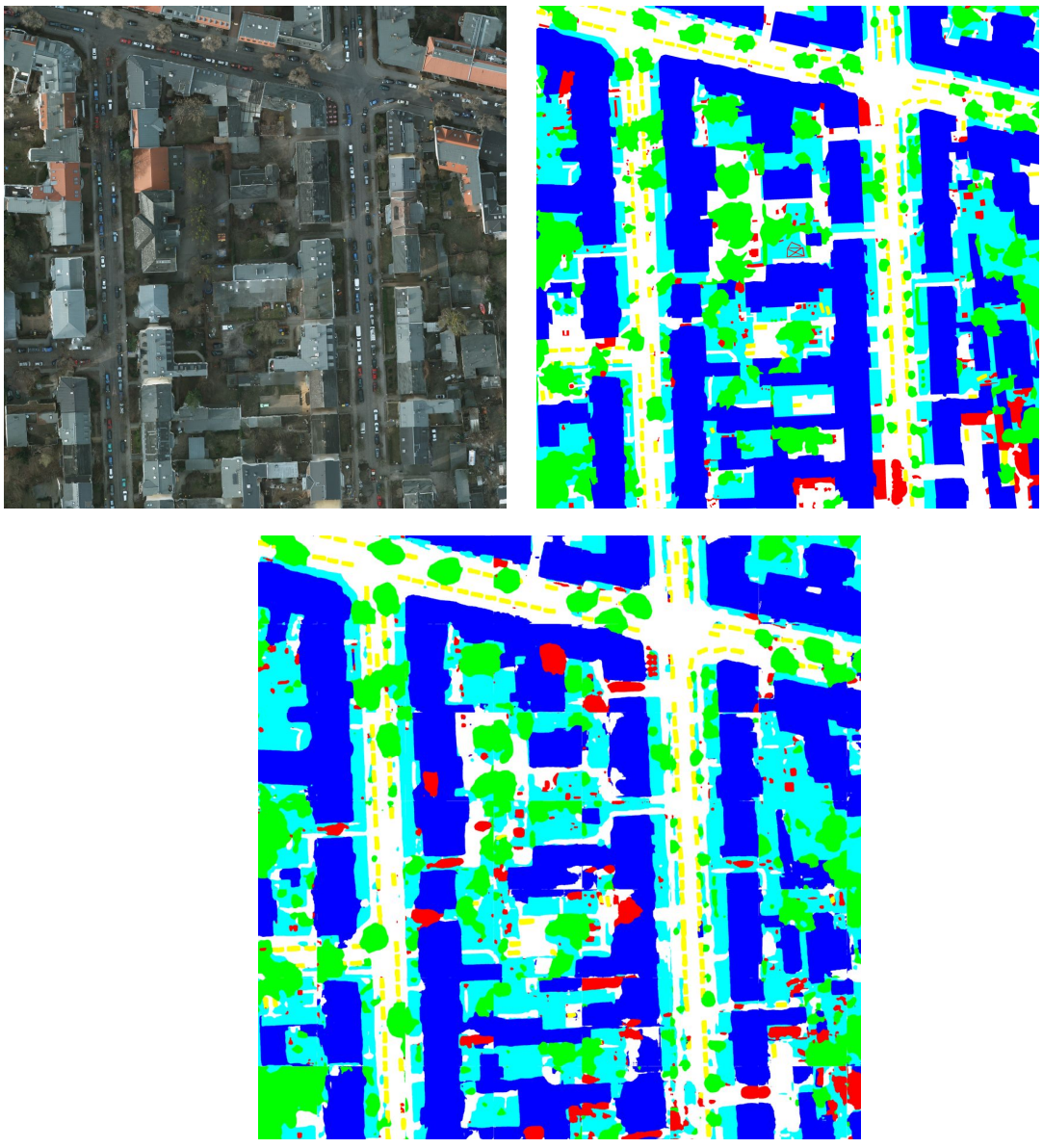

Figure 3.7: Parsing of an unseen region of Potsdam city. Input image (top left), ground truth segmentation map (top right), and predicted segmentation with coach training and 10% labeled data used for fine-tuning (bottom).

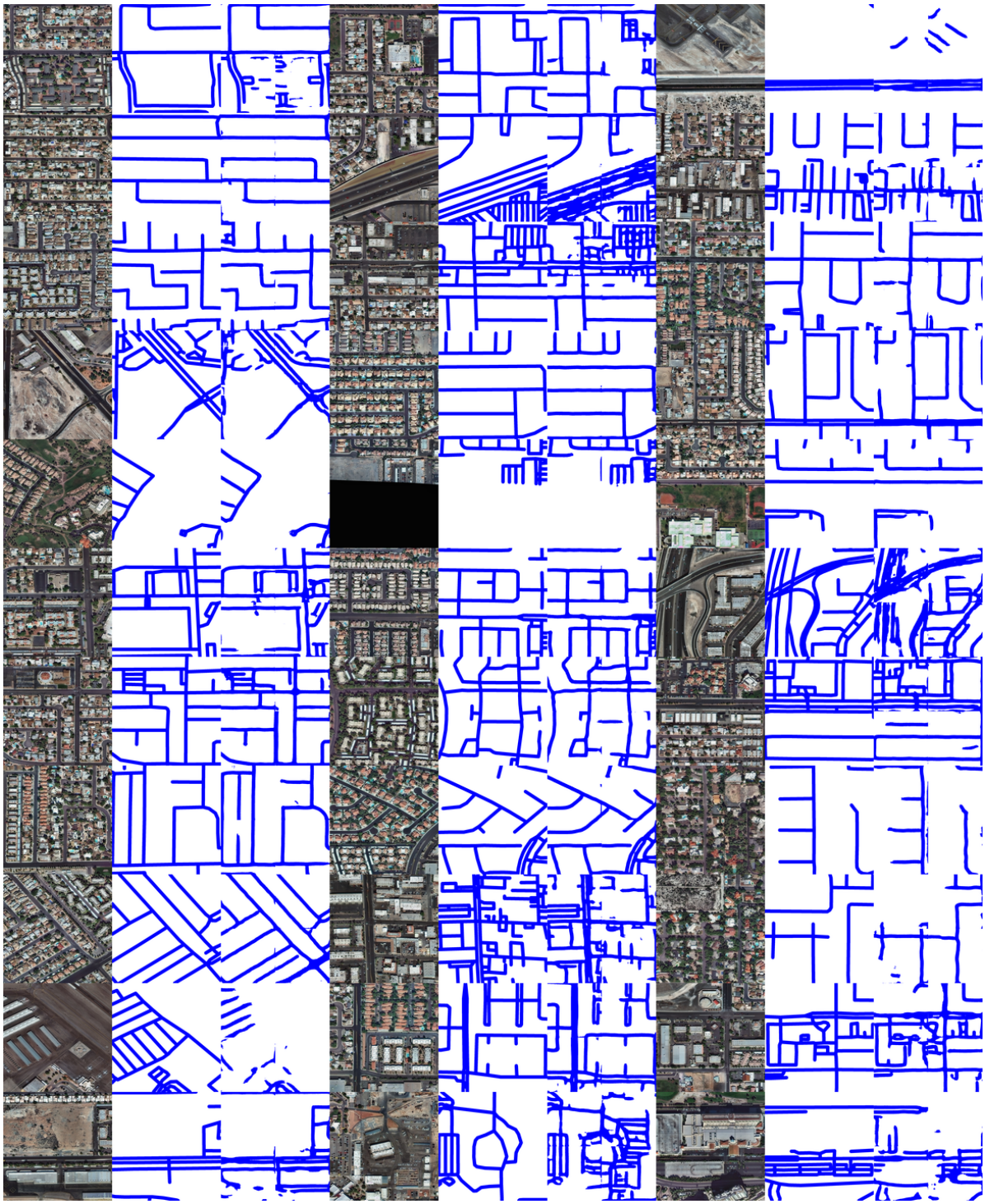

Figure 3.8: Road network extraction. Input image (left), ground truth segmentation map (middle), and predicted segmentation with coach training and 10% labeled data used for fine-tuning (right).

Method	Encoder	Bottleneck	Decoder	**Results**			
				Potsdam	SpaceNet	DG Roads	DG Lands
Context Prediction [20]	AlexNet	✗	✗	0.273	0.593	0.478	0.257
Context Encoders [67]		✓	✗	0.298	0.610	0.478	0.339
Splitbrain AE [90]		✓	✗	0.265	0.641	0.482	0.411
ImageNet	ResNet-18	✗	✗	0.493	0.701	0.669	**0.575**
Scratch	ResNet-18	✗	✗	0.414	0.657	0.643	0.495
Scratch		✗	✓	0.418	0.661	0.607	0.507
Autoencoder		✓	✓	0.502	0.748	0.749	0.515
Autoencoder		✗	✓	0.499	0.742	0.742	0.499
Context Encoders (Ours)	ResNet-18	✓	✗	0.540	0.730	0.478	0.501
		✗	✓	0.562	0.762	0.759	0.503
Coach Mask (Ours)	ResNet-18	✗	✓	**0.568**	**0.770**	**0.768**	0.529

Table 3.3: Semantic segmentation results (mIOU) while using full training set for the self-supervised pipeline and 10% of labeled images of respective datasets for training the segmentation network.

the images used for land cover classification is inherently equivalent to texture completion leading to inferior self-supervision. Superior results with Splitbrain Autoencoders [90] cross channel prediction, among the baseline methods, further confirms that color and texture plays a major role in this task.

Learned masks Adversarial inpainting with increasingly difficult masks outperforms the baselines in all the tasks simultaneously (see Table 4.6 and 3.4). These improvements against the strong baselines, although seems small, is significant primarily because performance gain over sophisticated data augmentation is difficult. Note that, the domain gap in inputs between inpainting and segmentation is similar in cases of random and adversarial masks since an equal amount of region is erased from the input image.

Dataset	**Method**	**(a) Labeled**				**(b) Unlabeled**					
		10%	**25%**	**50%**	**100%**	**1K**	**2K**	**5K**	**10K**	**50K**	**100K**
Potsdam	Scratch	0.418	0.502	0.544	0.582	NA	NA	NA	NA	NA	NA
	Context Encoders (Ours)	0.562	0.628	0.668	0.698	0.432	0.453	0.537	0.561	0.548	0.562
	Coach Mask (Ours)	**0.568**	**0.637**	**0.674**	**0.705**	0.446	0.469	0.541	0.563	0.566	0.565
SpaceNet	Scratch	0.661	0.720	0.748	0.766	NA	NA	NA	NA	NA	NA
	Context Encoders (Ours)	0.762	0.781	0.795	0.804	0.696	0.731	0.754	0.759	0.763	0.765
	Coach Mask (Ours)	**0.770**	**0.786**	**0.797**	**0.806**	0.709	0.731	0.757	0.770	0.774	0.774

Table 3.4: Segmentation performance (mIOU) using the proposed architecture (ResNet-18 encoder, no bottleneck, and decoder) with respect to the (a) fraction of labeled images used for fine-tuning and (b) number of unlabeled images used for self-supervised training with 10% labeled data for fine-tuning.

Number of labeled and unlabeled samples used As expected, there is a consistent improvement for all methods when the number of labeled images is increased (see Table 3.4). Our adversarial training strategy consistently outperforms others with respect to different amounts of labeled images used for fine-tuning. Surprisingly, the performance of self-supervised pre-training remains mostly the same despite a significant increase in number of unlabeled images used for pre-training (see Table 3.4). This behavior is most likely due to domain gap between semantic inpainting and semantic segmentation task. Furthermore, the random mask based inpainting technique suffer more than our proposed technique when the number of unlabeled images used for pre-training is drastically reduced. These results also conclude that our adversarial training have similar advantages and disadvantages when compared to the Context Encoders [67], however, it performs better in all scenarios we tested.

3.6 Summary

In this work, we propose a unified semantic segmentation approach towards a variety of overhead imagery tasks. We employ self-supervised techniques for pre-training due to scarcity of labeled data and availability of a large number of unlabeled data. Experiments show that existing self-supervised techniques, focusing primarily on classification, are inefficient for semantic segmentation. Our proposed architectural changes (3.3.1) leads to significant improvements in various diverse overhead imagery tasks. This is largely due to the use of high capacity ResNet-18 [29] as the backbone network and the re-use of pre-trained decoder networks. Additional improvements over strong baselines are observed on training the inpainting network with an adversarial coach network (3.3.2). The coach model is able to predict an increasingly difficult mask leading to a more difficult self-supervised task. However, existing self-supervised techniques as well as our proposed method do not exploit the availability of large unlabeled images. These insights motivate us to further probe self-supervised learning techniques to unlock the true potential of self-supervision in our future works.

Chapter 4

Improved Road Connectivity by Multi-Task Learning

Road network extraction from satellite images often produce fragmented road segments leading to road maps unfit for real applications. Pixel-wise classification fails to predict topologically correct and connected road masks due to the absence of connectivity supervision and difficulty in enforcing topological constraints. In this paper, we propose a connectivity task called Orientation Learning, motivated by the human behavior of annotating roads by tracing it at a specific orientation. We also develop a stacked multi-branch convolutional module to effectively utilize the mutual information between orientation learning and segmentation tasks. These contributions ensure that the model predicts topologically correct and connected road masks. We also propose Connectivity Refinement approach to further enhance the estimated road networks. The refinement model is pre-trained to connect and refine the corrupted ground-truth masks and later fine-tuned to enhance the predicted road masks. We demonstrate the advantages of our approach on two diverse road extraction datasets SpaceNet [80] and DeepGlobe [18]. Our approach improves over the state-of-the-art techniques by 9% and 7.5% in road topology metric on SpaceNet and DeepGlobe, respectively.

4.1 Introduction

A mapped road network provides routing information to find the traversable paths, which are important for planning in various applications such as navigation and disaster management. Example of a connected road network is shown in Figure 4.1a. Manual mapping of a complex road network is time consuming and requires intensive human effort. Automatic extraction of road networks from satellite imagery has been proposed [46, 4, 9, 77, 88], where recently, deep learning based techniques have shown high quality mapping results in diverse scenarios [56, 57, 54, 92, 5, 75, 17, 13, 59, 81]. However, the extracted road networks often produce fragmented road segments, and therefore, are unfit for real applications (Figure 4.1b). Satellite images pose difficulties in the extraction of roads due to (a) shadows of clouds and trees, (b) diverse appearance and illumination condition due to terrain, weather, geography, etc., and, (c) similarity of road texture with other materials. Label scarcity [75] as well as omission and registration noise in road ground-truths [57] also inhibit the accurate estimation of road maps.

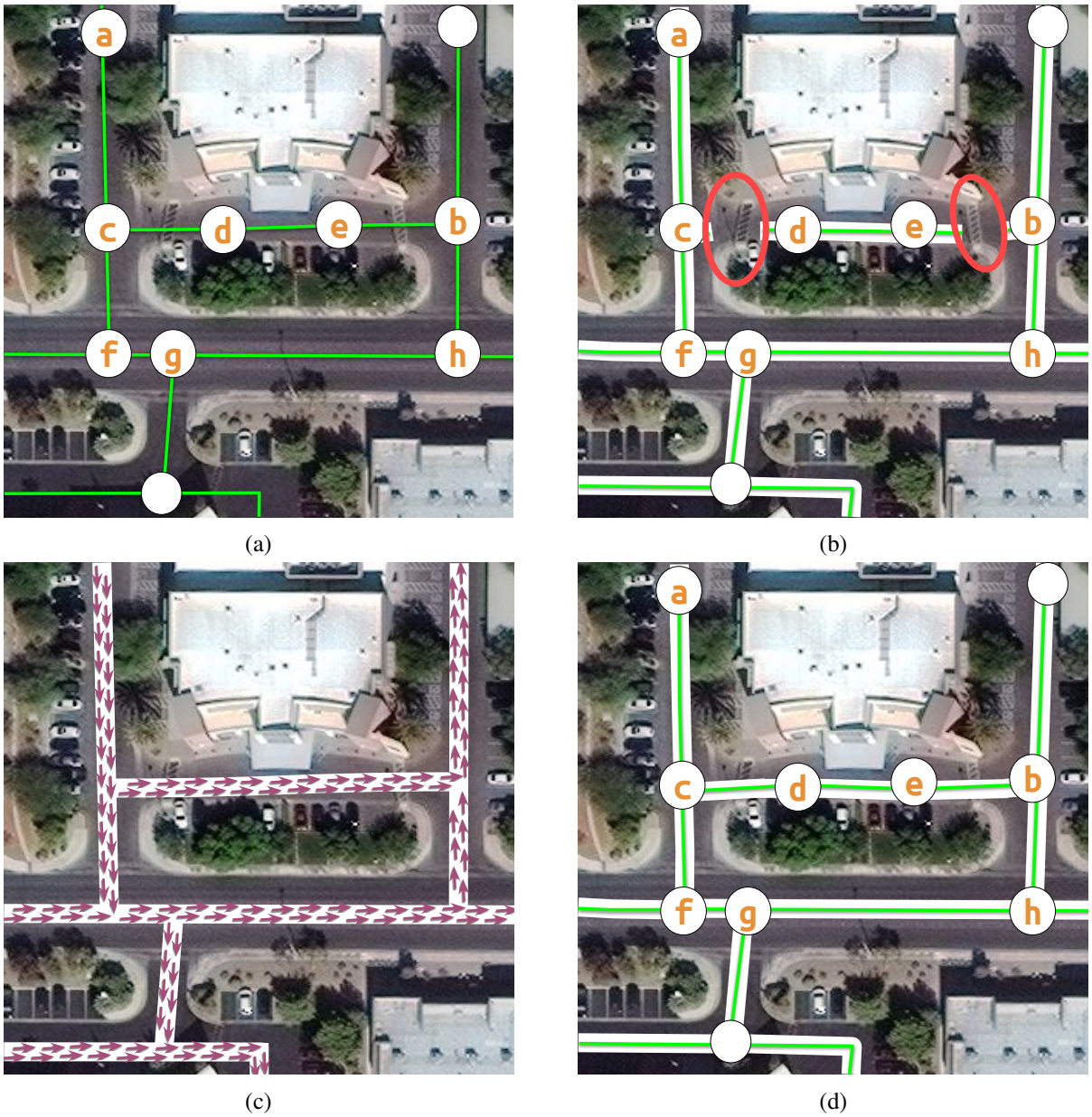

Figure 4.1: Road network extraction formulated as binary segmentation fails to produce topologically correct road map due to change in road appearance. (a) Annotators trace lines (highlighted nodes) along the center of roads with a traversable shortest path (a, c, d, e, b) for $a \rightarrow b$. (b) Fragmented road network estimated using segmentation resulting in path (a, c, f, g, h, b) for $a \rightarrow b$. (c) Tracing roads with orientation to achieve connectivity. (d) We extract connected and topologically correct road networks using segmentation and orientation.

Road network extraction is explored in [54, 56, 17, 13, 57], where the problem is posed as segmentation followed by post-processing steps to refine and couple the missing connections. The pixel-wise classification supervision does not constrain the model to learn representations for connected road segments [59], leading to poor estimation of road topology. Predicting masks with accurate topology is a challenging task due to difficulty in enforcing topological constraints via a loss function [59, 51] or during post-processing [54]. To measure deviations in topology, Mosinska *et al.* [59] rely on higher-level abstract features of ground-truth and predicted road masks whereas Máttyus *et al.* [51] employ an adversarial matching paradigm. To improve road connectivity, Máttyus *et al.* [54] proposed post-processing steps to reason for missing connection hypotheses while Bastani *et al.* [5] and Ventura *et al.* [81] iteratively connect road segments in the neighbouring image patches.

Our focus is on improving connectivity in road network extraction from binary segmentation of overhead imagery. Characterizing connectivity supervision in the way human annotates road maps requires topological and structural information of roads. We build our approach on the intuition that to annotate road maps human trace lines along the road orientation to connect the fragmented road segments. Consider Figure 4.1b, tracing lines $c \rightarrow b$ via d and e can connect the broken roads. This motivates us to design a connectivity task using available road labels to predict road orientation angle along with the road segmentation (Figure 4.1c).

In this paper, we propose to learn a road orientations jointly with per-pixel road segmentation in multi-branch CNN model (Figure 4.2). We also propose connectivity refinement which connect small gaps and reduces false positives in the prediction. The connectivity refinement model is pre-trained to restore the corrupted road ground-truth masks (Figure 4.2 and 4.4). This allows the model to effectively correct diverse failure scenarios. Similar to Mosinska *et al.* [59], our connectivity refinement model can be employed in an iterative manner, however, our refinement approach focuses on improving connectivity with the help of pre-training in addition to segmentation improvement. Lastly, we design a joint learning module by stacking multi-branch encoder-decoder structure (Figure 4.5 and 4.6). This module is a variant of stacked hourglass network [61], however our motivation is different i.e., flow of information between the related tasks to improve the performance of individual task in a multi-task learning framework. In contrast to [54, 57, 75, 5], our segmentation model inherently captures the information of connected road segments in the intermediate representation, leading to an accurate topology in road network estimation (Figure 4.1d).

Contributions:

1. We design an orientation learning task and demonstrate that the joint learning of orientation and segmentation improves the connectivity of road network.
2. We propose a connectivity refinement approach pre-trained with corrupted road ground-truth masks and fine-tuned with segmentation outputs to iteratively enhance the topology of the estimated road networks.

3. We design a stacked multi-branch module to effectively utilize the dual supervision. We show that the proposed module enables the flow of information between the tasks and helps in boosting the connectivity.

4.2 Related Work

Road Segmentation: Numerous techniques have been developed in literature to extract the road topology from satellite images. Most of the traditional methods [46], [88], [77], [9], [4] impose connectivity by incorporating contextual priors such as road geometry, higher order CRF formulation, marked point processes and solving integer programming on road graphs. These methods utilized hand designed features and optimizing complex objectives. In recent deep learning methods road extraction is formulated as binary segmentation problem [54], [59] using encoder-decoder structured models, which are able to capture more spatial context. Different from segmentation based approaches, Bastani *et al.* [5] introduced graph based methodology to generate road line strings. In the current scope, we focus on segmentation based approaches. Mnih *et al.* [56] learns road classification by CNN model in multiple stages (to reduce false negative rate due to label noise), operating on the image patches. Mattyus *et al.* [54] propose encoder-decoder structure model and pose it as multi-class (`roads`, `building` and `background`) segmentation. The network performs well in the segmentation, but lacks to connect the roads. Thus, to improve the connectivity, authors add post processing on extracted graph from segmentation output and generate connections between the leaf nodes of the extracted graph using shortest path algorithms. Finally, authors use a binary decision classifier to predict the correctness of connections. We found that it face difficulty in correctly classifying the generated connections due to large road density, ambiguous visual appearance of roads, occlusions and complex road topology in our datasets.

The other well admired variants of encoder-decoder structure in the remote sensing and computer vision community, to learn thin curvilinear road structures are U-Net [73] and LinkNet [10]. Their variants are proposed to learn the road segmentation in [16], [13]. Recently, LinkNet34 [10] is primarily utilized to segment the roads in DeepGlobe challenge [18]. Nevertheless, connectivity is achieved with more heuristic based post-processing in most of the methods. In contrast, we propose joint learning of connectivity task and road segmentation with a stacked encoder-decoder structure. The most recent work of Mosinska *et al.* [59] add perceptual loss to learn road topology in addition to the pixel based loss in U-Net [73]. Further more, authors add recursive refinement to fill the small gaps in road segments. The introduced loss term favors the road like structures but inefficient to connect the road segments. In this thesis, we use connectivity refinement approach using an another pre-trained network to learn connectivity pattern.

Multi-Task Learning (MTL): It is a learning mechanism [7], inspired from human beings to acquire knowledge of complex tasks by performing different shared sub-tasks simultaneously. Such as, a person trying to learn squash and tennis together, often improves the ability to predict trajectory, to swing and to throw, thus leading to better performance in both games. From machine learning perspective, multi-task

learning improves the performance by inducing mutual source of information of each task in the model. MTL has been applied successfully in the various domains such as speech recognition, natural language processing [15] and computer vision [41]. Readers are suggested to read survey [91] on multi-task learning.

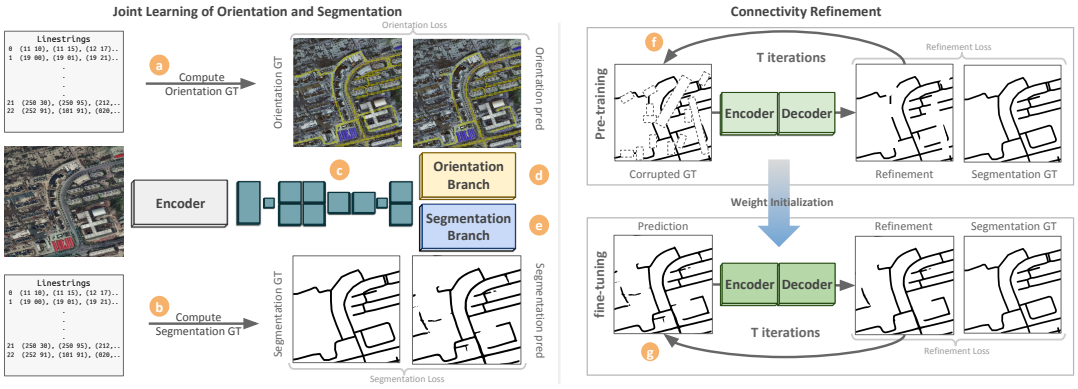

Figure 4.2: Overview of our approach for extracting connected road topology from satellite images. Annotations in the form of line strings, are converted to (a) orientation ground truth and (b) road masks ground truth. We use encoder-decoder structure with (c) *stacked multi-branch module* to jointly learn (d) orientation and (e) segmentation, providing dual supervision to the model. The orientation task is developed to improve the road connectivity. Finally, a connectivity refinement network, (f) pre-trained with corrupted ground truth to remove false roads and further improve the road connectivity, is (g) fine-tuned with segmentation output. (Images are best visualized in color.)

Humans perform two related tasks while annotating the roads i.e. identify the road pixel and angle to trace the line. In our work, motivation to include multi-task learning is to incorporate this behavior of annotating the roads as two tasks i.e. while labeling the satellite images, humans recognize roads and connect them by tracing a line, inherently identifying the orientation. We show that these related tasks improve the connectivity with better encoded representation in the encoder.

4.3 Method

Road extraction from overhead images via segmentation based methods produce disconnected road segments. To address this, we develop an orientation task from the road line strings (Section 4.3.1) and use it as an auxiliary loss along with pixel-wise segmentation loss. The motivation of orientation loss is to capture the relational information between the neighboring pixels through explicit learning of orientations between them. We formulate the problem as a two stage process: (a) joint learning of road orientation and segmentation in multi-task fashion, and (b) a connectivity refinement using a pre-trained CNN model (Section 4.3.2). We first present our novel inductive task followed by a connectivity refinement technique. Finally, we outline the proposed end-to-end joint learning pipeline with two stacks of multi-branch encoder-decoder which can flow the information across the tasks (Section 4.3.3).

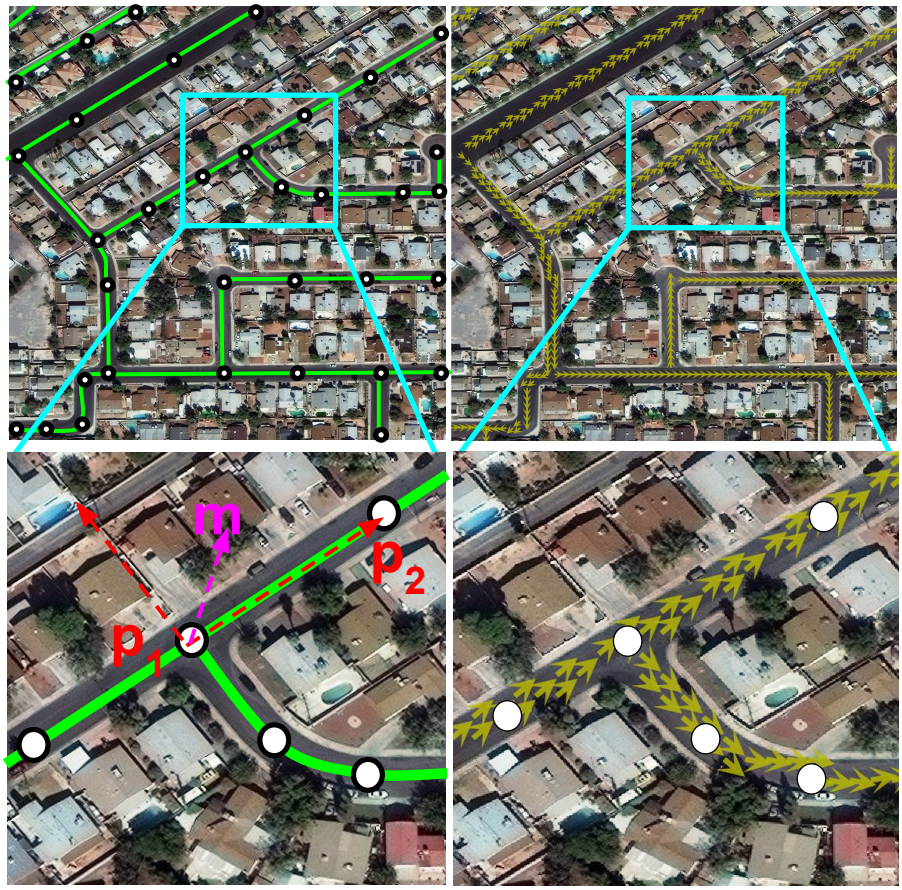

Figure 4.3: Road Orientations. Top left: road line strings annotations. Bottom left: two consecutive points to compute the orientation angle. Top right: Ground-truth road orientation vectors. Bottom right: Road orientation ground-truth in an image patch.

4.3.1 Orientation Learning

The pixel level annotation of roads is a computationally costly and time consuming task. To reduce the human effort, roads are preferably annotated with line strings connecting 2D points. We visualize each road line string as a directional vector between two consecutive points in 2D image plane (see Figure 4.3). The directional vector provides the orientation (tracing angle) of each road segment.

The orientation learning task is partly inspired from Part Affinity Fields [93] and bears resemblance with the deep watershed technique for instance segmentation [3]. Intuitively, representations learned for instance (road segments) segmentation would lead to improved connectivity in the estimated road network. However, road segments, unlike object instances or human body parts, do not have defined boundary between them and are rather interconnected. Therefore, instead of predicting orientation from the object boundary towards its centroid [93], we encode and predict the unit vector pointing towards the next pixel in the same or the connected adjacent road segment. Learning orientation with a pixel

based cross-entropy loss poses a connectivity constraint in the encoded representation as learning of road orientations favors the connected road segments and joint learning of related tasks often leads to more generalizable features [41, 7]. Orientation learning can be extended to applications like automatic segmentation along the object boundary [5, 8], connect the occluded lanes in lane detection, connect broken alphabets in OCR, etc.

We now describe the process to generate the orientation ground-truths from line strings. Consider an image shown in Figure 4.3 with road line strings $\{l_1, l_2, \ldots, l_m\}$ and each line string l_k consists of 2D points $\{p_1, p_2, \ldots, p_n\}$. We assume undirected road network, ignoring driving direction of the roads. We sort the coordinates of the points of each line string such that most of the directional vectors point from left to right and top to bottom, which we find to be appropriate for the neural network to learn and focus on connected road representation. We compute a unit directional vector $|\vec{v}(x, y)| \in [-1, 1]$ between two consecutive point pairs $\{(p_1, p_2), \ldots, (p_{n-1}, p_n)\}$ of l_k using (4.1) and convert it into polar domain to obtain orientation angle o_r using (4.2). For each point pair (p_i, p_j) using (4.3), the pixels lying within the threshold width λ_{orient} along the perpendicular direction of l_k, are assigned the same orientation value; for all other pixels non-road orientation angle o_b is assigned.

$$\vec{v}_{ij}(x, y) = \frac{p_i(x, y) - p_j(x, y)}{||p_i(x, y) - p_j(x, y)||_2^2} \tag{4.1}$$

$$\vec{v}_{ij}(x, y) \equiv \langle 1 \quad \measuredangle o_r \rangle \tag{4.2}$$

$$o_{l_k}(m) = \begin{cases} o_r & \text{if } |\vec{v}_\perp \cdot \overrightarrow{(m - p_1)}| < \lambda_{orient} \\ o_b & \text{otherwise.} \end{cases} \tag{4.3}$$

where $||p_i - p_j||_2^2$ is the total length between the consecutive points, $v_\perp$ is a vector perpendicular to unit directional vector, (x, y) are the coordinates of points and o is ground truth for orientations. We ignore non-road orientation angle during plotting of the vectors in Figure 4.2 and 4.3.

4.3.2 Connectivity Refinement

The orientation supervision improves the connectivity in the estimated road network. However, complex and dense road topology such as bridges and parking lots leads to failure in orientation prediction. The model also hallucinates roads in regions with similar textures e.g. road like patterns in farms. To further improve the prediction topology and suppress false positives, we employ the connectivity refinement (see Figure 4.4). Motivated by the success of restoring the images from corruption [67, 75], we interpret missing and spurious road segments as corrupted road ground-truth mask. We first pre-train the refinement network to restore the corrupted masks allowing the model to learn connectivity pattern as well as remove false roads. Note that, we opt for weight initialization and do not train the connectivity refinement using segmentation outputs and corrupted GT simultaneously to avoid overfitting to a single distribution of corruptions [23]. In pre-training stage, we concatenate satellite image X, corrupted ground-truth y' along with previous road prediction $\bar{y}_{t-1}$ (where $\bar{y}_0 = y'$) and feed it as input to the

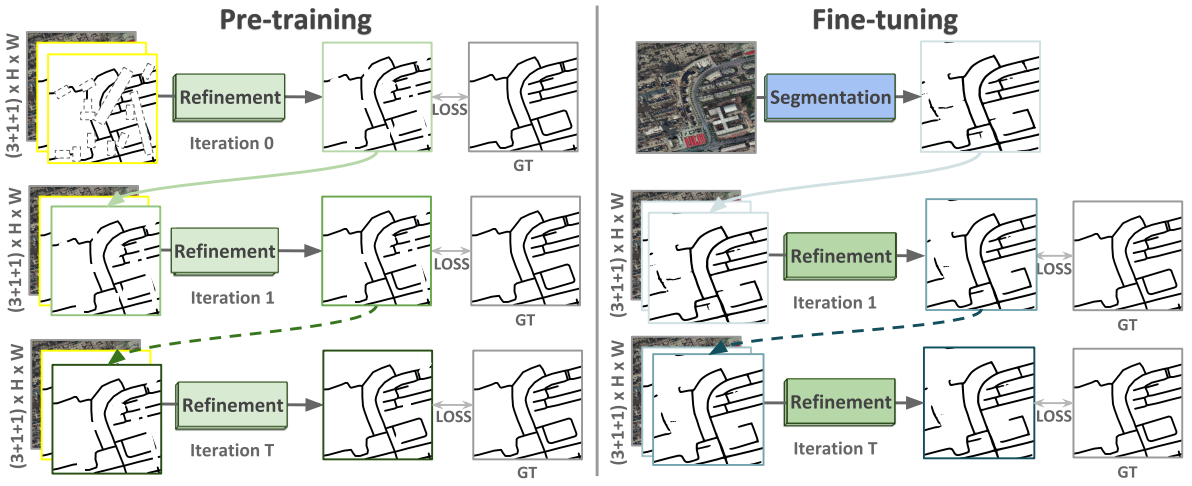

Figure 4.4: Connectivity Refinement. We pre-train the encoder-decoder CNN to remove false roads with pre-text task of correcting the corrupted road ground-truth masks. The model is later fine-tuned to refine the road segmentation outputs.

refinement model $g(\cdot)$.

$$\bar{y}_t = g\Big([X, y', \bar{y}_{t-1}]\Big) \quad t = 1, \ldots, T \tag{4.4}$$

At the end of pre-training stage the neural network learns to effectively encode the available contexts and fills the missing road segments. The pre-trained model is further fine-tuned to improve the road segmentation. In fine tuning stage, we replace the manually corrupted ground truth mask with the output of segmentation network.

$$\hat{y}_t = g\Big([X, \hat{y}, \hat{y}_{t-1}]\Big) \quad t = 1, \ldots, T \tag{4.5}$$

where $\hat{y} = f_{seg}(X)$, $\hat{y}_0 = \hat{y}$, and $[\cdot]$ denotes concatenation along channel axis. We use $T = 3$ and identical encoder-decoder architectures for $g(\cdot)$ and $f_{seg}(\cdot)$.

4.3.3 Stacked Multi-branch Module

The stacked multi-branch module as shown, in Figure 4.5 is composed of three blocks: (a) shared encoder, (b) iterative fusion with multi-branch, and (c) prediction branches for orientation and segmentation. The proposed CNN model performs the following tasks simultaneously: (a) learn a robust common representation for connected road segments in the shared encoder, (b) predicts road orientation and road segmentation, and (c) allows the information flow between the tasks to encourage road connectivity.

The shared encoder takes the input image X and learns a mapping function E, which projects the input to an encoded representation for both tasks. The encoding $z = E(X)$ is fed to the stacked multi-branch module to learn the coarse predictions. The motivation for n-stack multi-branch module is three

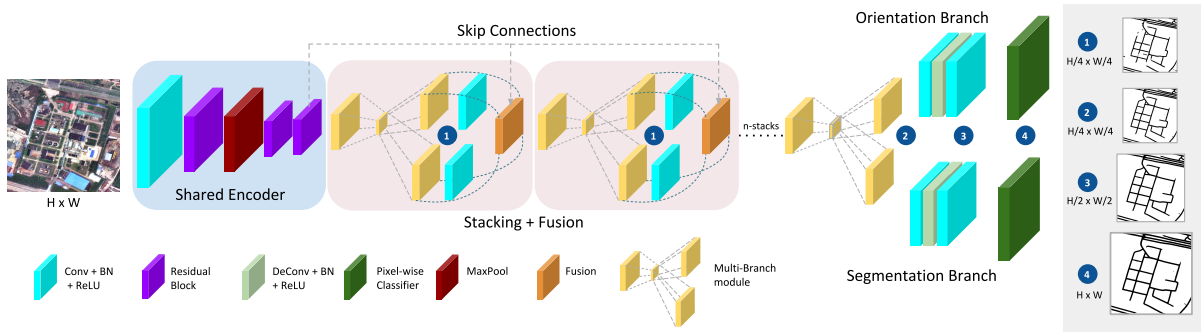

Figure 4.5: Architecture of *n-stacked multi-branch* CNN to learn road orientation and segmentation simultaneously. The stacked module is capable to calculate losses L_{seg} & L_{orient} at different scales $\big(\{\frac{1}{4}, \frac{1}{4} \ldots \text{n times}\}, \frac{1}{2}$ and $1\big)$ to optimize the CNN. We use two stacks of *multi-branch module* (Figure 4.6) with features fusion in first stack only. Refer to supplementary material for additional architectural details.

fold: (a) large receptive field to capture the spatial context, (b) mini encoder-decoder structure learns to re-calibrate features and coarse predictions in a repetitive fashion, and (c) it allows the information to flow from previous stack to the subsequent stack and refine the coarse predictions. We denote the stacking with a function H_n, where n is number of stacked multi-branch modules and coarse predictions with $\bar{o}$ for orientation and $\bar{y}$ for roads in (4.6).

To learn refine predictions $\hat{o}$ and $\hat{y}$ from the coarse predictions $\bar{o}_n$ and $\bar{y}_n$, we create two symmetric branches for each task. Each branch learns to up-sample the predictions using decoder networks consisting of two transposed convolutions followed by a pixel-wise convolutional classifier.

$$\bar{o}_n, \bar{y}_n = \begin{cases} H_n(\bar{o}_{n-1} + \bar{y}_{n-1} + z) & \text{if } n > 1 \\ H(z) & \text{if } n = 1 \end{cases} \tag{4.6}$$

Loss Function: The proposed network is capable of yielding the intermediate outputs at different scales, `n` outputs from each stack of multi-branch module at $\frac{1}{4}$ scale and two from successive transposed convolution at $\frac{1}{2}$ and 1. Hence, this allows to use multi-scale loss to guide the network while training. Let (X, y, o) be a given labeled sample from the dataset and $f(\cdot)$ denotes the prediction function using our model. We optimize the following loss functions:

$$L_{seg}(\hat{y}, y) = -\texttt{SoftIoU}\Big(f_{seg}(X), y\Big) \tag{4.7}$$

$$L_{orient}(\hat{o}, o) = -\sum_{c=0}^{o_l} o_c \log\Big(f_{orient}(X)\Big) \tag{4.8}$$

$$Loss = \sum_{s} \big(L^s_{seg} + L^s_{orient}\big) \tag{4.9}$$

where `SoftIoU` is differentiable IoU loss function [54], o_l is the number of bins in the quantized orientation, and s is scale having values $\{\frac{1}{4}, \frac{1}{4}, \ldots \text{n times}\}, \frac{1}{2}$ and 1.

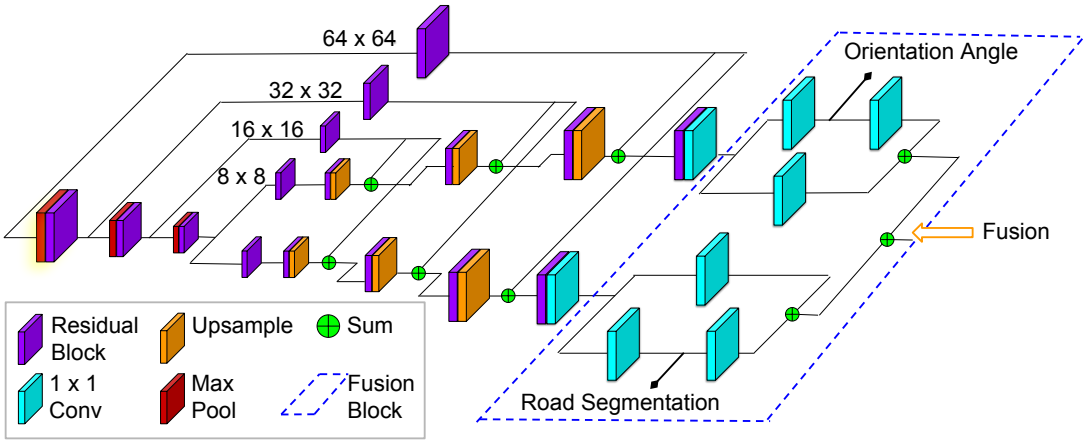

Figure 4.6: A multi-branch module. The intermediate output is extracted from each branch using 1×1 convolution and are merged using a fusion block.

4.4 Evaluation Metric

Pixel Based Metrics: In road segmentation, classifier's performance is measured using standard bench marking metrics like intersection over union (IoU) and F1-score. As we raster the ground truth using vector map in [80], which has constant mask for varying road widths. These variations adversely effect the pixel based metrics. Mnih *et al.* [57] suggest to use relaxed metric by adding buffer pixels. While evaluating precision, a predicted road pixel is considered as true-positive, if there is a road pixel in ground truth within a buffer. Similarly, in recall ground truth road pixel is considered predicted correctly, if there is a road pixel in output within the buffer. We use the relaxed metrics, suggested by Mnih *et al.* [57] with buffer of 4 pixels in our evaluations of road segmentation.

Graph Based Metric: The effective road network is capable of identifying the shortest route between the geographical locations. Hence to measure the such capability, we evaluate the connectedness of road network using *Average Path Length Similarity (APLS)* [80] as evaluation metric on the ground truth graph $G(V, E)$ and predicted graph $\hat{G}(\hat{V}, \hat{E})$. The metric finds the shortest path between each node of road network and penalizes for longer routes or no route. It ($S_{P \to T}$) (eq.4.10) measures the sum of difference of shortest path between each nodes in G and $\hat{G}$ for each image X. To penalize the false positives, symmetric term is added to APLS metric which considers predicted graph as ground truth and true graph as prediction.

$$S_{P \to T} = 1 - \frac{1}{|V|} \sum min\left(1, \frac{|L(a,b) - L(\hat{a},\hat{b})|}{L(a,b)}\right) \tag{4.10}$$

$$APLS = \frac{1}{N} \sum_{\forall X} \left(\frac{1}{\frac{1}{S_{P \to T}(G,\hat{G})} + \frac{1}{S_{T \to P}(\hat{G},G)}} \right) \tag{4.11}$$

where $|V|$ = total number of vertices in ground truth graph, $L(a, b)$ is path length between nodes a and b in G, $L(\hat{a}, \hat{b})$ is path length between nodes $\hat{a}$ and $\hat{b}$ in $\hat{G}$, and N is total number of images.

Algorithm 1: APLS metric

Result: Score $S_{P \to T}$
Data: ground truth $G(V, E)$, proposed $G'(V', E')$
begin
  $diff = 0$
  $count = 0$
  foreach $a \in V$ **do**
    $a' \leftarrow$ MatchingNode(a,V')
    if $a' \notin V'$ **then**
      $diff \leftarrow diff + 1$
      $count \leftarrow count + 1$
    else
      foreach $b \in V$ **do**
        $b' \leftarrow$ MatchingNode(b,V')
        $count \leftarrow count + 1$
        if $b' \notin V'$ **then**
          $diff \leftarrow diff + 1$
        else if $edge(a', b') \notin E'$ **then**
          $diff \leftarrow diff + 1$
        else
          $diff \leftarrow diff + min\big(1, \frac{|(L(a,b) - L(a',b')|}{L(a,b)}\big)$
        end
  end
  $S_{P \to T} = 1 - \big(\frac{diff}{count}\big)$
end

4.5 Experiments and Results

4.5.1 Dataset

We perform our experiments on SpaceNet [80] and DeepGlobe [18] datasets using only 3-band RGB images. We follow the experimental protocols and dataset splits of [75]. We evaluate and report the road connectivity metrics on full resolution images at inference time for each dataset.

Spacenet [80]: This dataset provides imagery from four different cities. The imagery is available at ground resolution of 30cm/pixel and pixel resolution of 1300×1300. The public dataset consists of 2567 images with road vector data as labels. We split the dataset into 2000 images for training and 567 for testing. To augment the training dataset we create crops of 650×650 with overlapping region of 215 pixels, thus providing $\sim$32K images. For validation we use the crops of same size without overlap.

DeepGlobe [18]: It includes imagery from three different areas with pixel level annotation for roads. The ground resolution of RGB image is 50cm/pixel and pixel resolution is 1024×1024. We use 4696 images for training and 1530 for validation. We augment it by creating crops of size 512×512 with overlapping region of 256 pixels, yielding $\sim$42K images for training phase.

4.5.2 Implementation Details

Dataset Preprocessing: We generate road heatmaps for Spacene [80] using distance transform along the center line of road with Gaussian kernel (σ = 15) and create binary mask while training with threshold of 0.76. For calculating the ground truth road orientation vectors in DeepGlobe [18], we obtain graph after skeletonization and smooth it using Ramer-Douglas-Peucker algorithm [70], [21].

Training Details: An identical training procedure for all the experiments is implemented in PyTorch [66] framework and the parameters of each model are initialized with He *et al.* [30]. We use random crops of size 256×256 from the image followed by mean subtraction. To improve the generalization of network, random horizontal flip, mirroring and rotation is employed as data augmentation. We train the joint network with batch size of 32 for 120 epochs. We use `SGD` optimizer with momentum = 0.9, weight decay = 0.0005 and initial learning rate of 10^{-2} with step scheduler having drop factor of 10 at epochs $\{60, 90, 110\}$. We use $\lambda_{orient} = 12$ pixels in equation 4.3 as orientation width along the roads. Segmentation outputs from the joint learning module is converted into network graph with each linear road segment representing an edge. We perform simple graph processing to remove small hanging road segments and graph smoothing. The proposed graph is converted into line strings and used it to evaluate the connectivity metric APLS [80] of road network. We use bin size of 10° for orientation angles.

Architecture Details We use encoder-decoder structured model with two intermediate stacks of multi-branch module to predict orientation and segmentation at different scales. We perform downsampling in the model using 2×2 `max pool` layer. We use three basic Resnet [30] blocks in the shared encoder and multi-branch module. We add `BatchNorm` layer after each convolution layer. The shared encoder reduces the input resolution to $H/4, W/4$ using the strided convolution and max-pool layer, which is fed to the multi-branch module. The shared encoder and final decoder has 64, 64, 64, 128, 64, 32 channels and each multi-branch module has 128. The final decoder block uses bottleneck deconvolution similar to LinkNet [10]. There are 29.02 Million parameters in the joint model.

4.5.3 Results

Road Width of Spacenet Mask: We convert the road line strings of Spacenet dataset [80] using distance transform with Gaussian kernel along the center line of roads. This provide a choice to choose threshold corresponding to different road widths. We perform experiments with different thresholds (as shown in Figure.4.7) using LinkNet34 [10] model and choose the threshold of 0.76 in all the experiments. The threshold of 0.76 correspond to road width of 6-7 meters.

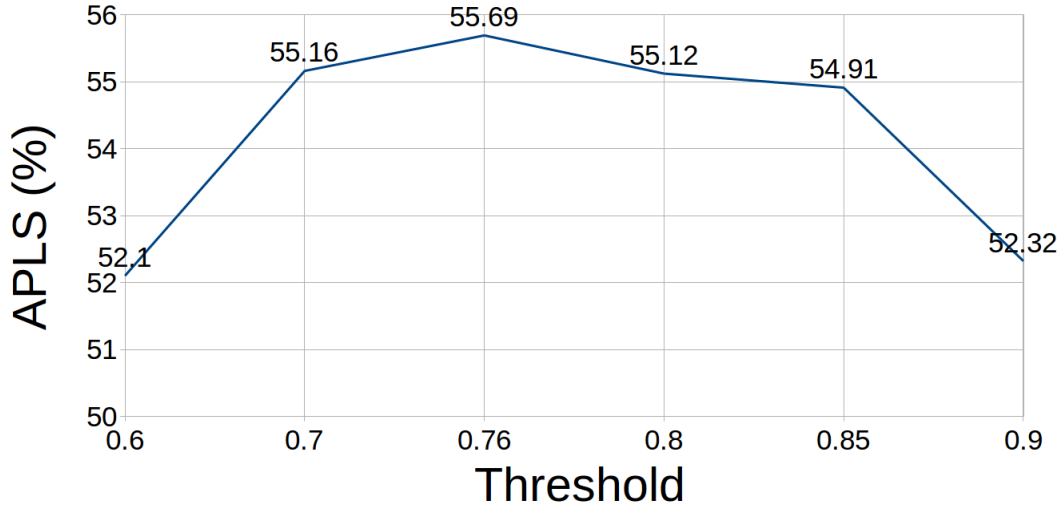

Figure 4.7: Effect of different road widths for Spacenet [80] Road masks using LinkNet34 [10] model on the connectivity metric APLS.

Quantization Size	**Spacenet**		**DeepGlobe**	
	road IoU^a	APLS	road IoU^a	APLS
5°	63.35	63.12	66.62	72.39
10°	63.75	**63.65**	**67.21**	**73.12**
20°	**63.80**	63.01	67.02	72.75

Table 4.1: Effect of different quantizations on orientation angles in the proposed stacked multi-branch module. **road IoU**^{a}: accurate pixel based intersection over union. **APLS**: average path length similarity on the extracted graph from road segmentation.

Orientation Learning: We choose two architectures Resnet-18 [30] and LinkNet34 [10] to study the performance of orientation learning. We modify both architectures with dual and identical decoders having shared encoder. The results in Table. 4.2 shows that our proposed task for road connectivity generalizes to different architectures. Incorporating the orientation learning as auxiliary loss in both architectures with multi-branches improves the APLS by 6.41% and 5.08% for Spacenet [80]. This suggests that multi-task learning of two related task improves the intermediate representation, leading to better generalization. To study the significance of orientation task in road connectivity as an auxiliary loss, we formulate another shared task of predicting junctions in multi-task learning framework. The results in Table. 4.2 shows that the connectivity metric APLS improve significantly with orientation task and not due to the multi-task learning. This validates the effectiveness of the orientation task in predicting the connected road topology.

Method	Spacenet		DeepGlobe	
	road IoU^a	APLS	road IoU^a	APLS
Resnet18	59.04	52.65	62.12	63.31
Resnet18 + Orientation	61.90	59.06	**64.77**	**68.93**
Resnet18 + Junctions	58.41	52.76	63.54	66.20
LinkNet34	60.33	55.69	62.75	65.33
LinkNet34 + Orientation	**62.45**	**60.77**	64.72	68.71
LinkNet34 + Junctions	60.72	55.91	63.79	67.42

Table 4.2: Comparison of two auxiliary tasks of orientation and junction learning for road connectivity. It shows that improvement in the road connectivity is due to orientation learning in contrast to multi-task learning. **road IoU**^{a}: accurate pixel based intersection over union. **APLS**: average path length similarity on the extracted graph from road segmentation.

Method	Spacenet		DeepGlobe	
	road IoU^a	APLS	road IoU^a	APLS
Resnet18 [30] + Orientation	61.90	59.06	64.77	68.93
LinkNet34 [10] + Orientation	62.45	60.77	64.72	68.71
Unet [73] + Orientation	60.12	58.59	65.21	67.81
Multi-branch(1 Stack) + Orientation	63.26	60.92	65.60	70.23
Multi-branch(2 Stack) + Orientation	**63.75**	**63.65**	**67.21**	**73.21**
Multi-branch(3 Stack) + Orientation	63.73	62.89	66.61	72.48

Table 4.3: Comparison of **joint learning modules employed for road segmentation and orientation learning**. It shows that our stacked multi-branch module increase the APLS by 2.7%.

Quantization of Orientation Angles: We perform ablation study on the different quantization levels for orientation angles and report the results in Table.4.1. The results shows that quantization of 10° is good choice for better road connectivity and we use it in all the comparison methods.

Connectivity Refinement: In contrast to [59] we use connectivity refinement with pre-trained model. We analyze different manipulations like randomly erasing the road masks with random blocks or linear lines of random lengths. Also, we insert few linear artifacts in the road mask. We found that random erasing with linear structures appear similar to the real segmentation outputs, as it intermix with linear structures thus, we report results for only such manipulations. Figure 4.8 shows the improvement with connectivity refinement and marginal improvement in road IoU. This shows that proposed refinement is able to connect the road segments with gaps and remove the false predicted roads, rather than enhancing road width.

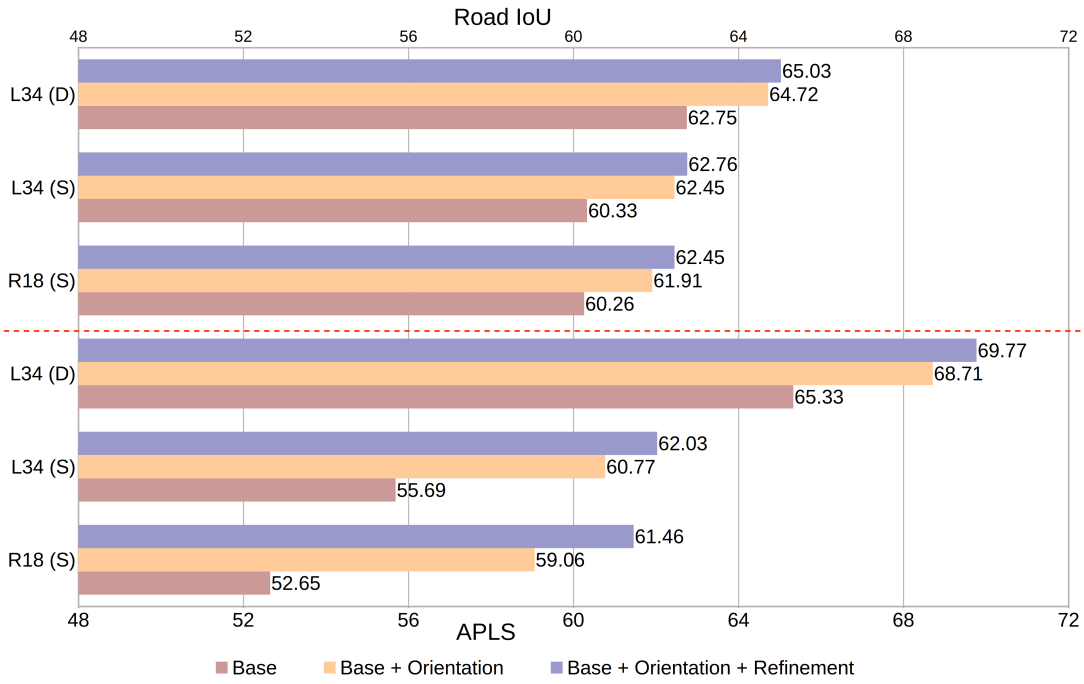

Figure 4.8: Quantitative Improvement with Orientation task and Connectivity Refinement. **R18**, **L34**: Resnet-18 and LinkNet34 based encoder-decoder as joint learning model. **S** and **D**: denote the Spacenet and DeepGlobe dataset. Bars upper and lower to red line shows road IoU and APLS improvement.

Stacked multi-branch module: We perform experiments to compare our proposed stacked multi-branch module with the state-of-art CNN models as joint learning module, which are commonly used to perform segmentation of thin structures. We did additional comparisons by stacking three multi-branch modules and found that performance stabilizes with two modules. This can be explained with the fact that additional modules increase the number of learnable parameters, which degrades the segmentation. Hence, we restrict our pipeline to consist two stacks of multi-branch modules. The results in Table. 4.3 shows that stacking of multi-branch modules improve the road connectivity over the single encoder-decoder modules by $\sim 2.5\%$.

We study the incremental improvement of individual proposed contribution for stacked multi-branch module and show the results in Table. 4.4. Initially, we hypothesize that knowledge of orientation angles helps in tracing a line to connect the broken road segments, which we achieve by cross information flow between both the tasks in stacked multi-branch module. We discover that adding the orientation features with segmentation performs better. This also confirms that the neural network utilize the orientation information to connect the broken road segments (see Table. 4.4) and improve APLS by 1.87% and 1.18% on respective data sets.

Performing the connectivity refinement on the road segmentation of stacked multi-branch model improves the road connectivity marginally. Our intuition for this behaviour is due to the fusion of task specific and the shared encoder features before feeding it to the second stack of multi-branch module, as it modulates the shared features (shown in Figure.4.9). And the second multi-branch module function

Multi-Scale	Orientation Learning	Feature Fusion	Connectivity Refine	Spacenet		DeepGlobe	
				IoU^a	APLS	IoU^a	APLS
				61.51	58.70	64.23	67.98
✓				61.80	58.49	64.44	67.92
✓	✓			63.44	61.78	66.81	72.03
✓	✓	✓		63.75	63.65	**67.21**	**73.21**
✓	✓	✓	✓	**63.76**	**63.79**	67.02	73.20

Table 4.4: Effect of step-wise improvement with multi-scale loss, orientation learning and cross task information flow by feature fusion of both outputs. Further adding the connectivity refinement improves the APLS marginally, which shows that second stack of multi-branch module function as refinement network.

Fusion	Spacenet		DeepGlobe	
	IoU^a	**APLS**	**IoU^a**	**APLS**
No Fusion	63.44	61.78	66.81	72.03
Sum	**63.75**	**63.65**	**67.21**	**73.21**
Concatenate	63.53	63.01	66.59	72.23

Table 4.5: Effect of fusion type in our proposed module to cross the information flow between the orientation learning and segmentation in first stack.

as refinement network on the modulated feature space. In the end, joint learning and fusion of both tasks improve the road IoU by $\sim 2.5\%$ and APLS by $\sim 5\%$ on both datasets over the classification supervision in stacked multi-branch network.

Effect of fusion: We perform ablation study on multiple fusion techniques for the information flow and report the results in Table.4.5. We discover that adding the orientation features with segmentation performs better. It shows that the simple feature addition modulates the shared feature with bias and improve the APLS by 1.87% and 1.18% over the no fusion on both datasets.

Comparisons with state-of-the-art results: We compare the effectiveness of the proposed methods with state-of-art segmentation based methods [54], [51] and [59] (see Table 4.6 and Figure 4.10). Máttyus *et al.* [54] hypothesize the connections with shortest path algorithms between the nodes of road graph and validates the connection with a classifier. We found that the classifier is unable to detect the false connections in cases with densely connected roads which leads to a decrease in $APLS$ after post-processing. Mosinka *et al.* [59] introduce the topology loss term with recursive refinement. However, it also face challenges in predicting the roads in densely connected areas, and unpaved roads.In-spite of large diversity in both datasets, our approach significantly improves the connectivity in the extracted

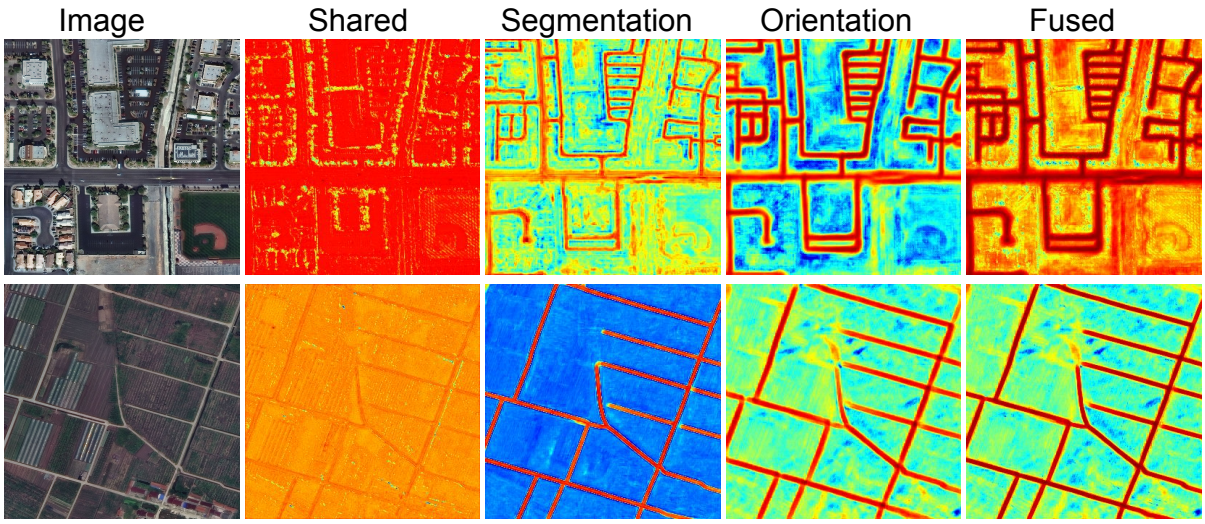

Figure 4.9: Feature maps for different stages in proposed model. **Image**: a satellite image, **Shared**: feature map after the shared encoder and before the first stack, **Segmentation/Orientation**: feature map of segmentation/orientation in the first stack before fusion, **Fused**: additive fusion of all feature maps and fed to second stack of the proposed model.

Method	**Spacenet**						**Deepglobe**					
	Precision	Recall	F1	IoU^r	IoU^a	APLS	Precision	Recall	F1	IoU^r	IoU^a	APLS
DeepRoadMapper* [54]	60.61	60.80	60.71	43.58	59.99	54.25	79.82	80.31	80.07	66.76	62.58	65.56
DeepRoadMapper** [54]	57.57	58.29	57.93	40.77	N/A	50.59	77.15	77.48	77.32	63.02	N/A	61.66
Topology Loss† [59]	50.35	50.32	50.34	33.63	56.29	49.00	76.69	75.76	76.22	61.58	64.95	56.91
Topology Loss‡ [59]	52.94	52.86	52.90	35.96	57.69	51.99	79.63	79.88	79.75	66.32	64.94	65.96
L34 [10]	61.30	61.45	61.39	44.27	60.33	55.69	78.34	78.85	78.59	64.73	62.75	65.33
L34 [10] + Orient (Ours)	63.82	63.96	63.89	46.94	62.45	60.76	81.24	81.73	81.48	68.75	64.71	68.71
MatAN [51]	49.84	50.16	50.01	33.34	52.86	46.44	57.59	56.96	57.28	40.13	46.88	47.15
RoadCNN (segmentation) [5]	62.82	63.09	62.95	45.94	62.34	58.41	82.85	83.73	83.29	71.36	**67.61**	69.65
Ours (full)	**64.65**	**64.77**	**64.71**	**47.83**	**63.75**	**63.65**	**83.79**	**84.14**	**83.97**	**72.37**	67.21	**73.12**

Table 4.6: Comparison of our technique with the state-of-the-art road network extraction techniques. IoU^r and IoU^a refers to relaxed and accurate road IoU. Ours (full) include the proposed stacked multi-branch module with orientation learning. We use implementation from [5] for DeepRoadMapper [54] and our own implementation for [59].

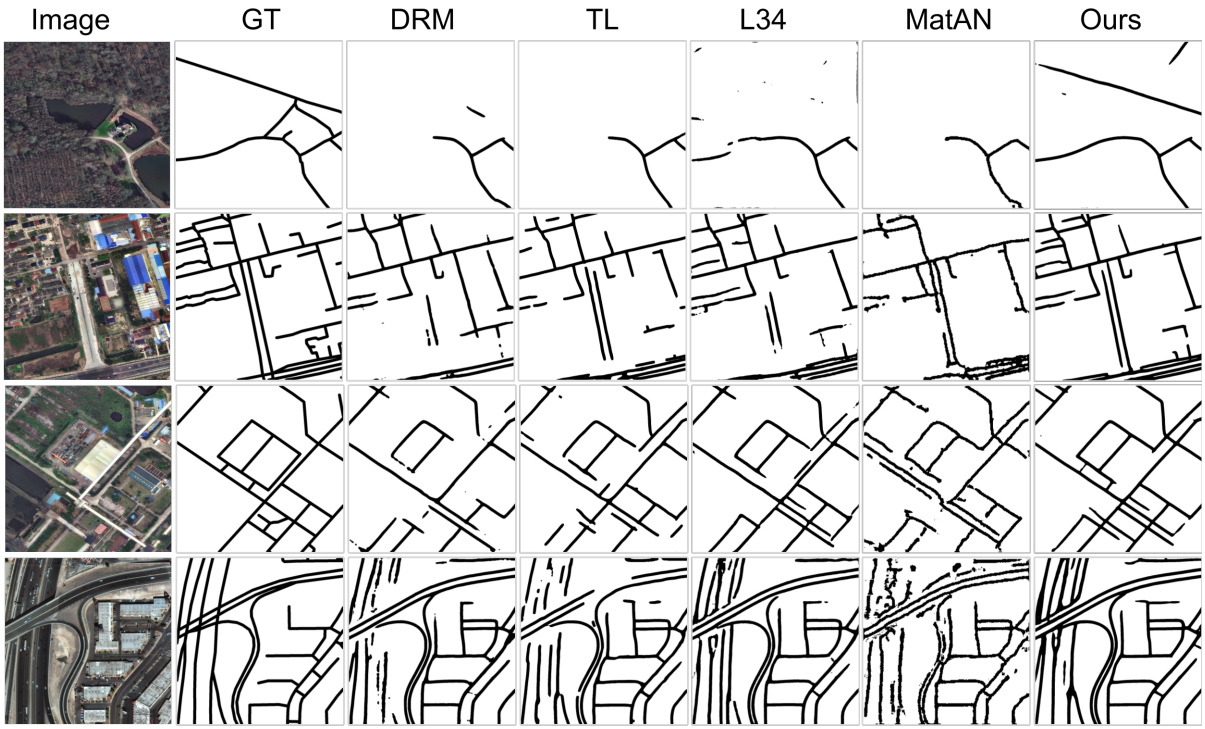

Figure 4.10: Qualitative Comparisons with state-of-the-art methods — DRM [54], TL [59], L34 [10], and MatAN [51].

road graph against the baselines. However, the proposed technique faces challenges to accurately connect roads under the bridges as well as in the presence of large occlusion (see row #4 in Figure 4.10).We also observe the false road detection in farm outlines due to it's visual similarity with unpaved roads and parking lots on top of buildings due to the absence of relative depth cues.

4.6 Conclusions

In this work, we propose a novel task of orientation learning that constrain the model to produce connected and topologically accurate road networks. We show that pixel-wise classification supervision leads to road networks with fragmented road segments and poor connectivity. Our experiments show that the joint learning of orientation and segmentation followed by connectivity refinement leads to a significant improvement in the road connectivity. We also show the effectiveness of the stacked encoder-decoder structure model as a joint learning module, which can efficiently utilize the information from related tasks.

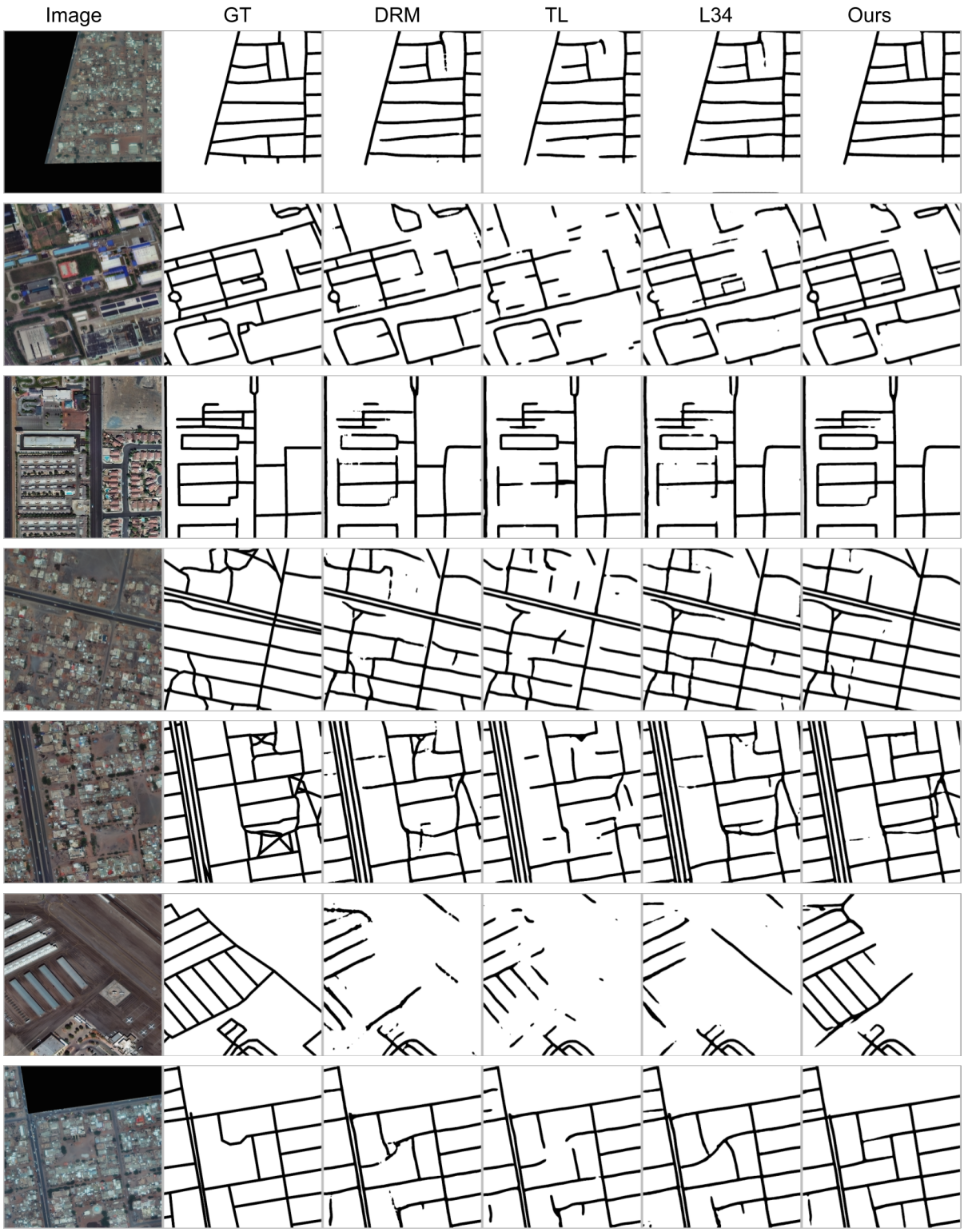

Figure 4.11: Qualitative Comparisons with state-of-the-art methods - **DRM**: DeepRoad Mapper[54], **TL**: Topology Loss [59], **L34**: LinkNet34 [10]

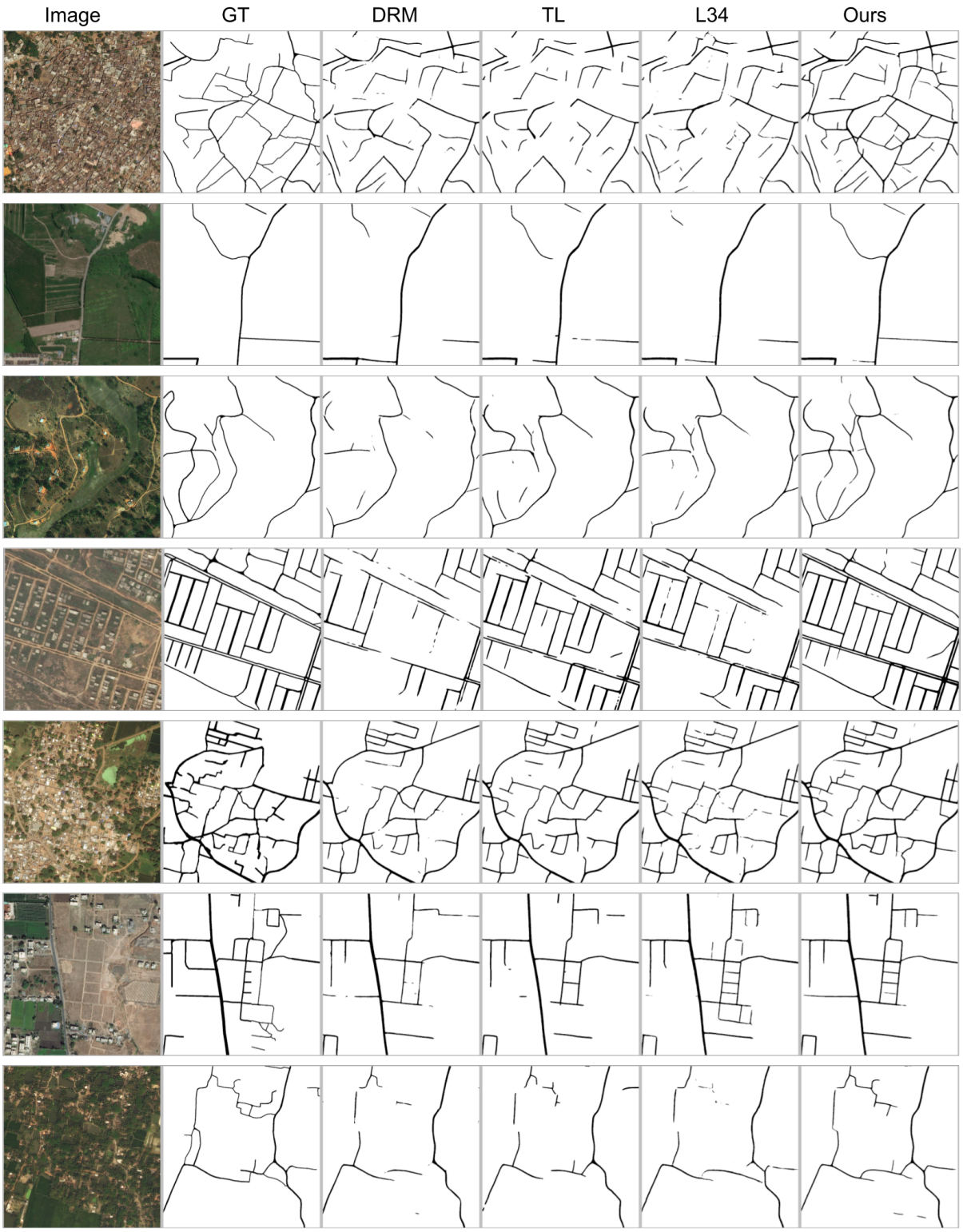

Figure 4.12: Qualitative Comparisons with state-of-the-art methods - **DRM**: DeepRoad Mapper[54], **TL**: Topology Loss [59], **L34**: LinkNet34 [10]

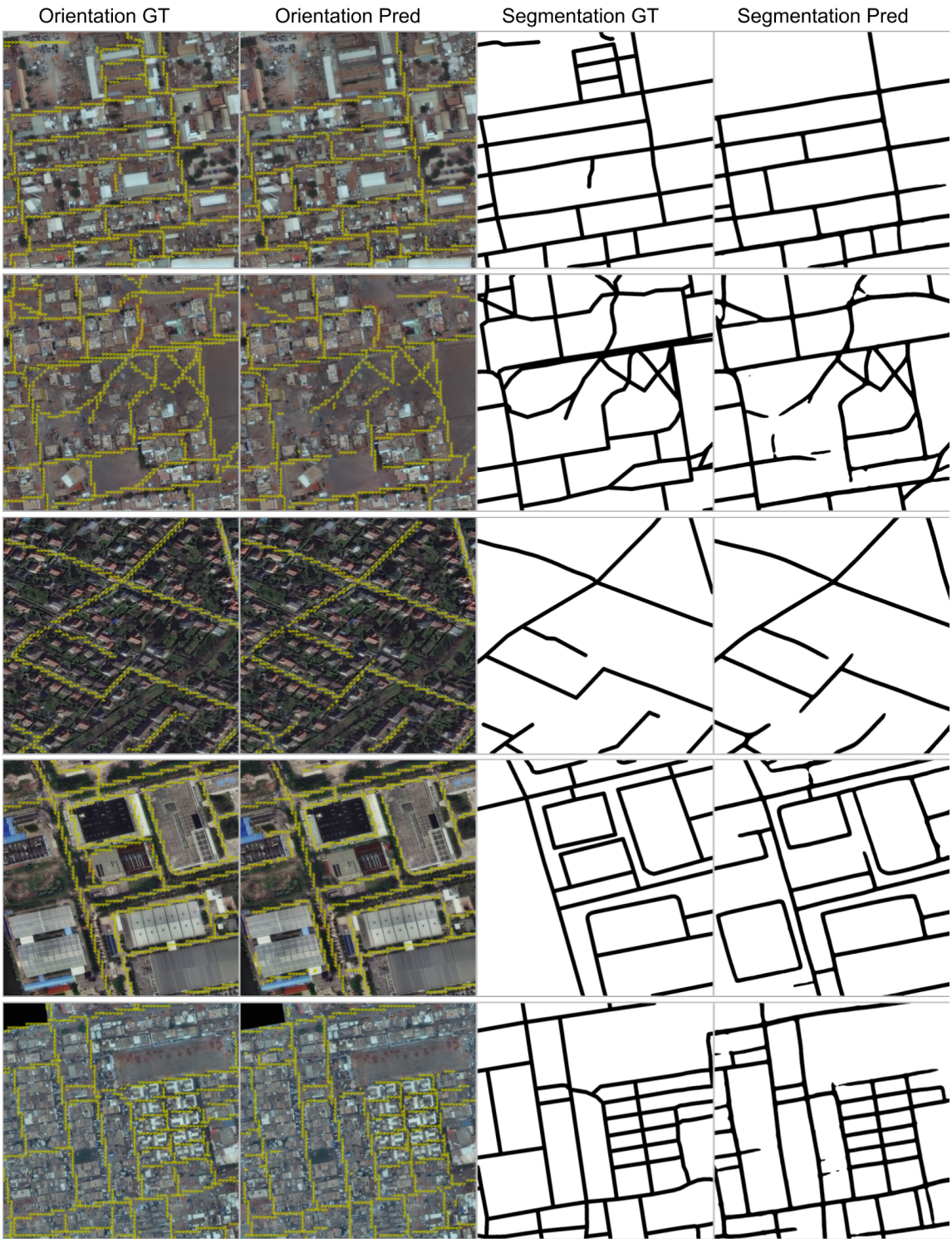

Figure 4.13: Qualitative results of Orientation and Segmentation prediction of **Ours** method. Orientation GT and Prediction are visualized as overlay on the image.

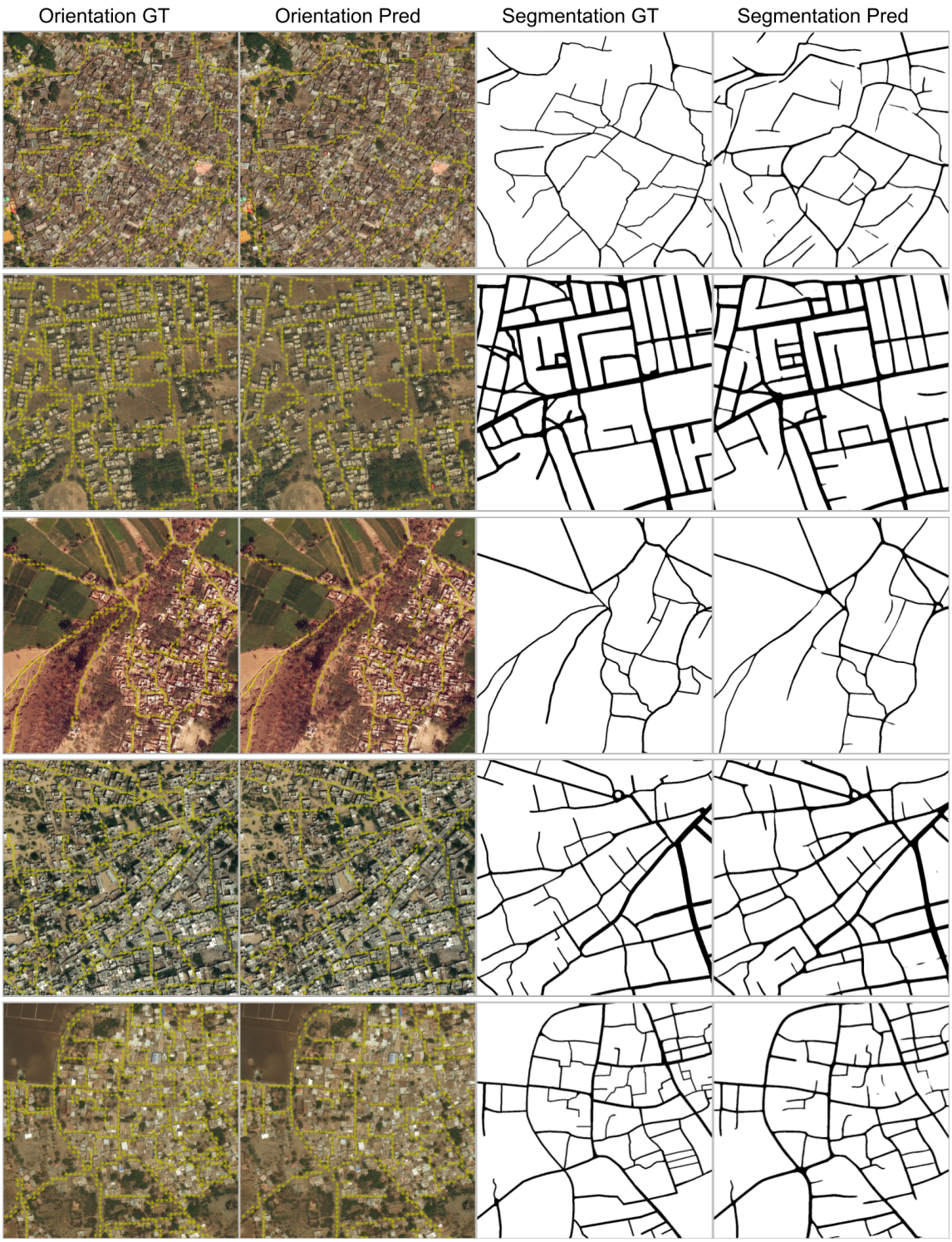

Figure 4.14: Qualitative results of Orientation and Segmentation prediction of **Ours** method. Orientation GT and Prediction are visualized as overlay on the image.

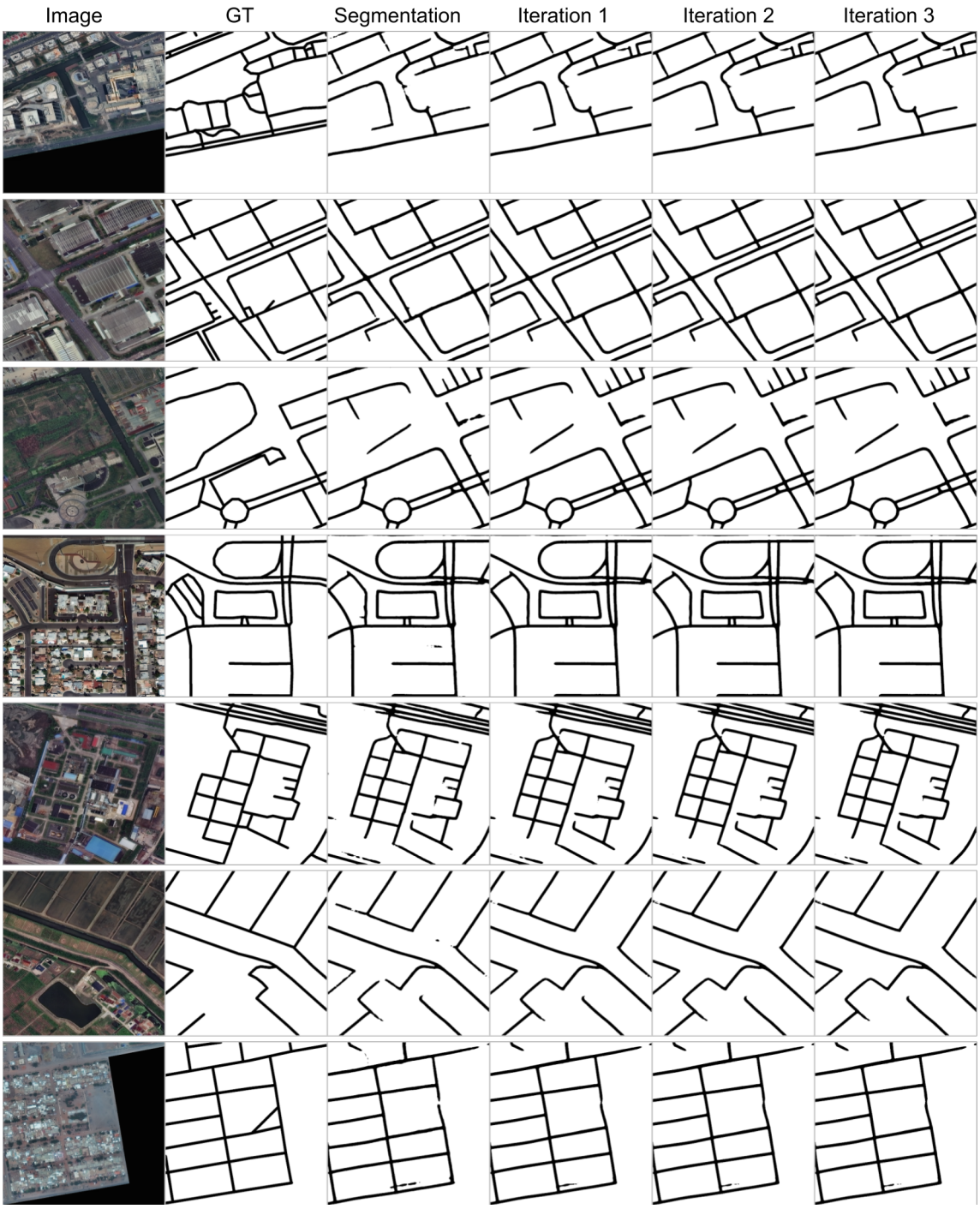

Figure 4.15: Qualitative results of Connectivity refinement over segmentation and orientation learning with LinkNet34 [10] as joint learning module.

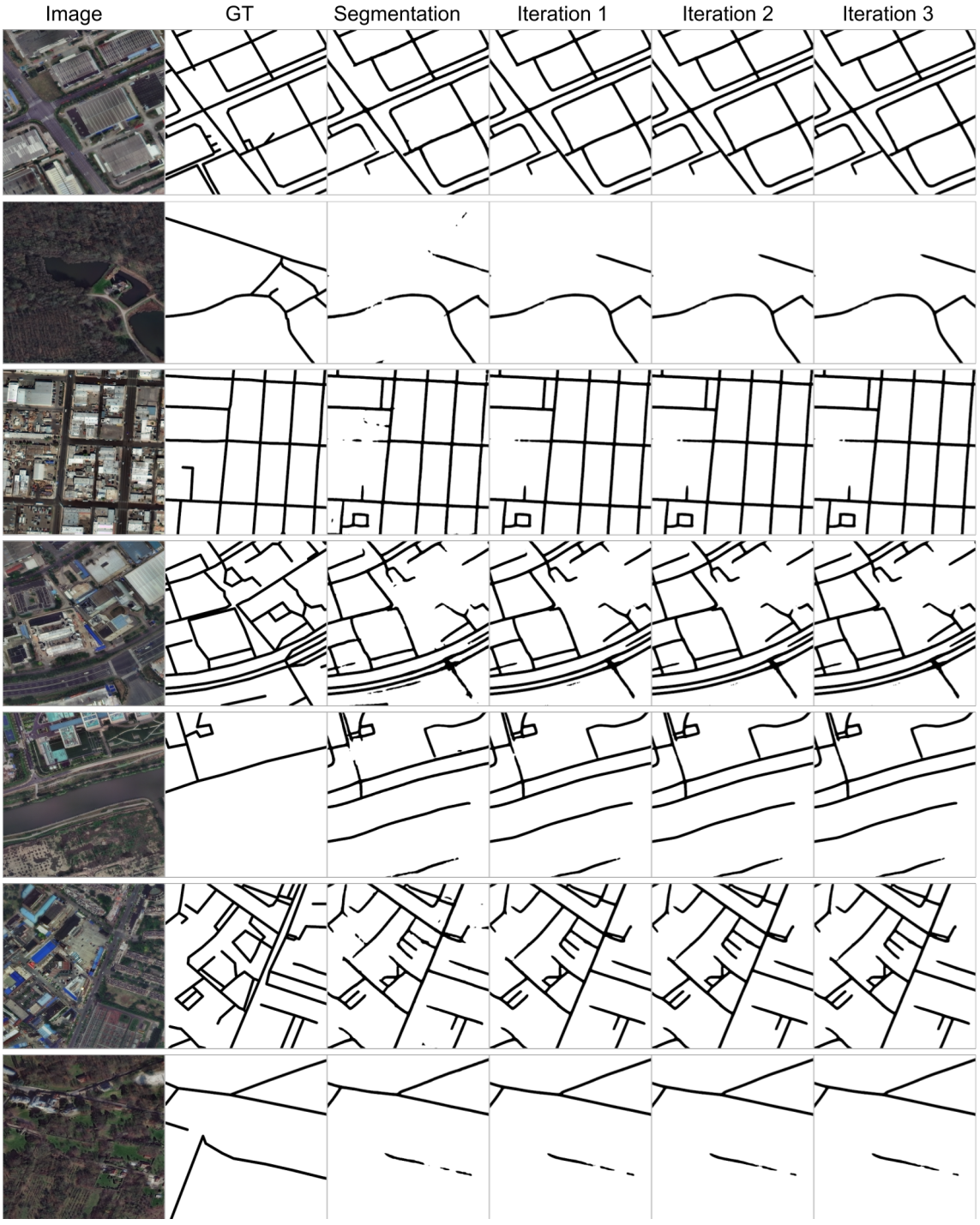

Figure 4.16: Qualitative results of Connectivity refinement over segmentation and orientation learning with LinkNet34 [10] as joint learning module.

Chapter 5

Summary and Future Directions

The research community has witness the fast development of computer vision methods using the natural images in last few years. This progress energize the Geo-spatial field to apply wide variety of techniques to satellite imagery. The goal of this thesis was to apply such algorithms to satellite imagery.

Particularly, in this thesis we propose a unified framework that enables efficient and large scale understanding of satellite imagery on various tasks such as land classification, road estimation and scene parsing of a city. We employ self-supervised techniques for pre-training due to scarcity of labeled data and availability of a large number of unlabeled data. Specifically we use semantic in-painting as an auxiliary task by randomly erasing certain areas of the image, then fine tune to improve the semantic segmentation over the training from scratch. We also propose a *Coach Network* which works in an alternate fashion with in-painting network. In contrast to using random mask, this training strategy learns a semantically meaningful mask for the given image and increase the representation capability of the neural network. A strong representation network, when employed for semantic segmentation further improves the performance of target task.

We also observe that image colorization induce strong semantic representation using the Split-Brain architecture [90]. Motivated by this insight, we plan to further probe the proposed approach with image colorization to increase the performance of semantic segmentation. Alternatively, the researcher in this field can develop a stronger and difficult auxiliary tasks such as extracting the artifacts in the image.

Lastly, in chapter 4, we develop a novel task of *Orientation Learning*, to capture the relational information in road pixels by explicitly learning angles between them. We demonstrate the significance of the task by contrasting with another related task of *Junction Learning*. And experiments show that road connectivity has been improved due the proposed task instead of the multi-task learning framework. We also propose to perform connectivity refinement model by first learning to rectify the linear road corruptions, followed by fine tuning of road segmentation outputs. Finally, we design an integrated stacked multi-branch convolution neural network, which is capable to learn strong representation of connected road segments in the shared encoder. Stacking of multi-branch module allows to build relation among the *Orientation Learning* and *Road Segmentation* task with a fusion operation. With the proposed techniques, we are able to improve the road connectivity metric on Spacenet [80] and DeepGlobe [18] by $\sim 9\%$ and $\sim 7.5\%$ over the state-of-art techniques.

In the proposed approach we utilize only simple post processing on the skeleton graph obtained from road segmentation which includes removal of small segments and duplicate edges. One method to improve the road connectivity is utilizing the learned angle information to trace along the broken road segments and finally connect them. Also, to further enhance the spatial context of roads alternative related task of building footprints extraction is a natural extension. In contrast of using simple fusion in stacked multi-branch model, learning based fusion technique is a good option to improve the mutual information among the related task. Alternative solution to the problem would be to increase the performance of related tasks such as orientation learning by exploring the different architectures split. Another possible future direction is to utilize GAN [26] framework to learn higher statistical contextual relationship and as an alternative to unsupervised learning.

Related Publications

Conference:

1. **Anil Batra**, Suriya Singh, Guan Pang, Saikat Basu, C.V. Jawahar and Manohar Paluri. **Improved Road Connectivity by Joint Learning of Orientation and Segmentation.** *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2019*

2. Suriya Singh, **Anil Batra**, Guan Pang, Lorenzo Torresani, Saikat Basu, C.V. Jawahar and Manohar Paluri. **Self-Supervised Feature Learning for Semantic Segmentation of Overhead Imagery.** *British Machine Vision Conference (BMVC) 2018*

Bibliography

[1] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In *ACCV*, 2016. 13, 16

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *PAMI*, 2017. 14, 16

[3] Min Bai and Raquel Urtasun. Deep watershed transform for instance segmentation. In *CVPR*, 2017. 34

[4] Meir Barzohar and David B Cooper. Automatic finding of main roads in aerial images by using geometric-stochastic models and estimation. *PAMI*, 1996. 3, 6, 29, 32

[5] Favyen Bastani, Songtao He, Sofiane Abbar, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, Sam Madden, and David DeWitt. Roadtracer: Automatic extraction of road networks from aerial images. In *CVPR*, 2018. xiii, 3, 13, 16, 29, 31, 32, 35, 45

[6] Alexander Buslaev, Selim Seferbekov, Vladimir Iglovikov, and Alexey Shvets. Fully convolutional network for automatic road extraction from satellite imagery. In *CVPRW*, 2018. 13, 16

[7] R Caruna. Multitask learning: A knowledge-based source of inductive bias. In *ICML*, 1993. 4, 9, 32, 35

[8] Lluıs Castrejón, Kaustav Kundu, Raquel Urtasun, and Sanja Fidler. Annotating object instances with a polygon-rnn. In *CVPR*, 2017. 35

[9] Dengfeng Chai, Wolfgang Forstner, and Florent Lafarge. Recovering line-networks in images by junction-point processes. In *CVPR*, 2013. 29, 32

[10] Abhishek Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *VCIP*, 2017. xii, 8, 32, 40, 41, 42, 45, 46, 47, 48, 51, 52

[11] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018. 3

[12] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv preprint arXiv:1802.02611*, 2018. 3

[13] Guangliang Cheng, Ying Wang, Shibiao Xu, Hongzhen Wang, Shiming Xiang, and Chunhong Pan. Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 55(6):3322–3337, 2017. 7, 29, 31, 32

[14] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *CoRR abs/1711.07846*, 2017. 21, 22

[15] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML*, 2008. 9, 33

[16] Dragos Costea and Marius Leordeanu. Aerial image geolocalization from recognition and matching of roads and intersections. *CoRR*, abs/1605.08323, 2016. 7, 32

[17] Dragos Costea and Marius Leordeanu. Aerial image geolocalization from recognition and matching of roads and intersections. *arXiv preprint arXiv:1605.08323*, 2016. 29, 31

[18] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *CVPRW*, 2018. x, 8, 9, 11, 12, 21, 29, 32, 39, 40, 53

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 14

[20] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015. 3, 9, 16, 22, 24, 27

[21] David H Douglas and Thomas K Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 10(2):112–122, 1973. 7, 40

[22] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 14

[23] Ruohan Gao and Kristen Grauman. On-demand learning for deep image restoration. In *CVPR*, 2017. 35

[24] Ruohan Gao and Kristen Grauman. On-demand learning for deep image restoration. In *CVPR*, 2017. 18, 20

[25] Lluis Gomez, Yash Patel, Marçal Rusiñol, Dimosthenis Karatzas, and CV Jawahar. Self-supervised learning of visual features through embedding images into text topic spaces. In *CVPR*, 2017. 9, 16

[26] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. 2014. 54

[27] Armin Gruen and Haihong Li. Semi-automatic linear feature extraction by dynamic programming and lsb-snakes. *Photogrammetric engineering and remote sensing*, 63(8):985–994, 1997. 3, 6

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015. 23

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 17, 28

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 40, 41, 42

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, 2016. 17

[32] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017. 3

[33] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. In *IGARSS*, 2018. 13, 16

[34] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 2006. 16

[35] Geoffrey E Hinton and Richard S Zemel. Autoencoders, minimum description length and helmholtz free energy. In *NIPS*, 1994. 16

[36] Stefan Hinz and Albert Baumgartner. Automatic extraction of urban road networks from multi-view aerial imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 58(1-2):83–98, 2003. 6

[37] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 17

[38] ISPR. Potsdam 2d semantic labeling contest. `http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html`. 21

[39] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960. 3, 6

[40] Michael Kampffmeyer, Arnt-Børre Salberg, and Robert Jenssen. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In *CVPRW*, 2016. 13, 16

[41] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018. 9, 33, 35

[42] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 23

[43] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 20

[44] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 7, 17, 22, 24

[45] Nataliia Kussul, Mykola Lavreniuk, Sergii Skakun, and Andrii Shelestov. Deep learning classification of land cover and crop types using remote sensing data. *GRSL*, 2017. 13, 16

[46] Ivan Laptev, Helmut Mayer, Tony Lindeberg, Wolfgang Eckstein, Carsten Steger, and Albert Baumgartner. Automatic extraction of roads from aerial images based on scale space and snakes. *MVA*, 2000. 6, 29, 32

[47] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 2017. 14, 16

[48] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 3, 14, 16, 22, 23

[49] Jonathan L Long, Ning Zhang, and Trevor Darrell. Do convnets learn correspondence? In *NIPS*, 2014. 17

[50] Alina Elena Marcu. A local-global approach to semantic segmentation in aerial images. *arXiv preprint arXiv:1607.05620*, 2016. 7

[51] Gellért Máttyus and Raquel Urtasun. Matching adversarial networks. In *CVPR*, 2018. xii, 31, 44, 45, 46

[52] Gellert Mattyus, Shenlong Wang, Sanja Fidler, and Raquel Urtasun. Enhancing road maps by parsing aerial images around the world. In *ICCV*, 2015. 13, 16

[53] Gellert Mattyus, Wenjie Luo, and Raquel Urtasun. Deeproadmapper: Extracting road topology from aerial images. In *ICCV*, 2017. 13, 16, 23

[54] Gellért Máttyus, Wenjie Luo, and Raquel Urtasun. Deeproadmapper: Extracting road topology from aerial images. In *ICCV*, 2017. xii, xiii, 3, 7, 29, 31, 32, 37, 44, 45, 46, 47, 48

[55] Volodymyr Mnih and Geoffrey E Hinton. Learning to detect roads in high-resolution aerial images. In *ECCV*, 2010. 3, 7, 13, 16

[56] Volodymyr Mnih and Geoffrey E Hinton. Learning to detect roads in high-resolution aerial images. In *ECCV*, 2010. 29, 31, 32

[57] Volodymyr Mnih and Geoffrey E Hinton. Learning to label aerial images from noisy data. In *ICML*, 2012. 7, 13, 16, 29, 31, 38

[58] Mehdi Mokhtarzade and MJ Valadan Zoej. Road detection from high-resolution satellite images using artificial neural networks. *International journal of applied earth observation and geoinformation*, 9(1):32–40, 2007. 7

[59] Agata Mosinska, Pablo Márquez-Neila, Mateusz Kozinski, and Pascal Fua. Beyond the pixel-wise loss for topology-aware delineation. In *CVPR*, 2018. xii, xiii, 29, 31, 32, 42, 44, 45, 46, 47, 48

[60] Agata Justyna Mosinska, Pablo Marquez Neila, Mateusz Kozinski, and Pascal Fua. Beyond the pixel-wise loss for topology-aware delineation. In *CVPR*, 2018. 8, 13, 16

[61] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 31

[62] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *CVPR*, 2015. 14, 16

[63] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016. 3, 9, 16

[64] Sakrapee Paisitkriangkrai, Jamie Sherrah, Pranam Janney, and Anton van den Hengel. Semantic labeling of aerial and satellite imagery. *J-STARS*, 2016. 13, 16

[65] Sinno Jialin Pan, Qiang Yang, et al. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010. 9

[66] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPSW*, 2017. 40

[67] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. x, 3, 9, 13, 14, 15, 16, 17, 18, 20, 22, 23, 24, 27, 28, 35

[68] Lorien Y Pratt. Discriminability-based transfer between neural networks. In *Advances in neural information processing systems*, pages 204–211, 1993. 9

[69] Lorien Y Pratt, Jack Mostow, Candace A Kamm, and Ace A Kamm. Direct transfer of learned information among neural networks. In *AAAI*, volume 91, pages 584–589, 1991. 9

[70] Urs Ramer. An iterative procedure for the polygonal approximation of plane curves. *Computer graphics and image processing*, 1(3):244–256, 1972. 7, 40

[71] Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *ICML*, 2011. 16

[72] SpaceNet Road. Spacenet on amazon web services (aws). `https://spacenetchallenge.github.io/datasets/datasetHomePage.html`, 2017. 21

[73] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 8, 32, 42

[74] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 7, 17

[75] Suriya Singh, Anil Batra, Guan Pang, Lorenzo Torresani, Saikat Basu, Manohar Paluri, and C. V. Jawahar. Self-supervised feature learning for semantic segmentation of overhead imagery. In *BMVC*, 2018. 29, 31, 35, 39

[76] Mingjun Song and Daniel Civco. Road extraction using svm and image segmentation. *Photogrammetric Engineering & Remote Sensing*, 70(12):1365–1371, 2004. 7

[77] Radu Stoica, Xavier Descombes, and Josiane Zerubia. A gibbs point process for road extraction from remotely sensed images. *IJCV*, 2004. 29, 32

[78] Tatiana Tommasi. Learning to learn by exploiting prior knowledge. 2013. 9

[79] John C Trinder and Haihong Li. Semi-automatic feature extraction by snakes. In *Automatic Extraction of Man-Made Objects from Aerial and Space Images*, pages 95–104. Springer, 1995. 6

[80] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018. vii, x, xii, 8, 9, 11, 29, 38, 39, 40, 41, 53

[81] Carles Ventura, Jordi Pont-Tuset, Sergi Caelles, Kevis-Kokitsi Maninis, and Luc Van Gool. Iterative deep learning for road topology extraction. In *BMVC*, 2018. 29, 31

[82] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, 2008. 16

[83] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *JMLR*, 2010. 16

[84] George Vosselman and Jurrien De Knecht. Road tracing by profile matching and kaiman filtering. In *Automatic Extraction of Man-Made Objects from Aerial and Space Images*, pages 265–274. Springer, 1995. 3, 6

[85] Weixing Wang, Nan Yang, Yi Zhang, Fengping Wang, Ting Cao, and Patrik Eklund. A review of road extraction from remote sensing images. *Journal of traffic and transportation engineering (english edition)*, 3(3):271–282, 2016. 7

[86] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *CVPR*, 2015. 16

[87] Jan D Wegner, Javier A Montoya-Zegarra, and Konrad Schindler. A higher-order crf model for road network extraction. In *CVPR*, 2013. 13, 16

[88] Jan D Wegner, Javier A Montoya-Zegarra, and Konrad Schindler. A higher-order crf model for road network extraction. In *CVPR*, 2013. 29, 32

[89] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *CVPR*, 2017. 18

[90] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *CVPR*, 2017. 3, 9, 16, 22, 24, 27, 53

[91] Yu Zhang and Qiang Yang. A survey on multi-task learning. *CoRR*, abs/1707.08114, 2017. URL `http://arxiv.org/abs/1707.08114`. 9, 33

[92] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IGRSL*, 2018. 29

[93] Cao Zhe, Simon Tomas, Wei Shih-En, and Sheikh Yaser. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 34

[94] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *GRS Magazine*, 2017. 16