PROJECT DOCUMENTATION: ROADS SEGMENTATION USING U-NET, DEEPLABV3+, AND DINKNET.

Salem Al-Nsour
sal0213371@ju.edu.jo

Mohammed Batayneh
Mhm0216340@ju.edu.jo

Anas Mousa
ans0212567@ju.edu.jo

Khaled Al-Shrouf
kal0215106@ju.edu.jo

Dunia Alatoom
dny0205996@ju.edu.jo

ABSTRACT

This project addresses the critical task of extracting road networks from satellite and aerial imagery, which is essential for understanding and analyzing urban infrastructure. The objective is to develop a robust deep learning-based model that classifies each pixel as either road or background, facilitating automated and accurate analysis of road networks. Leveraging state-of-the-art architectures, the proposed model is trained on a dataset of satellite and aerial images sourced from Google Maps. Through systematic experimentation with model architectures, loss functions, and training strategies, the project achieves competitive performance metrics, with F1 scores reaching up to 0.925 and 0.917 (our result and the competition result respectively). The results highlight the potential of advanced deep learning techniques to automate road network extraction effectively, offering significant implications for urban planning, disaster management, and autonomous navigation systems.

.

CONTENTS

1 INTRODUCTION

The development of high-resolution satellite imaging technology has revolutionized the interpretation of geographical data (1), driving advancements in diverse fields such as agriculture (2), urban mapping (3), road segmentation (4) and even poverty analysis (5)

One of the most featured applications of this technology lies in the extraction and classification of road networks from satellite images. Roads are vital components of the infrastructure playing a critical role in accessibility (6), transportation management (7) and urban planning (8). when road networks are accurately mapped, satellite imaging can help in optimizing traffic flow (9), to potentially identify areas in need of maintenance, and supporting disaster response efforts such as flood risks (10).

In this study we will be focusing on developing a robust deep learning model to accurately segment roads from a dataset of satellite and aerial images sourced from Google Maps. By labeling each pixel as either a road (1) or background (0), we seek to create a precise road extraction model.

The approach is proposed based on state-of-the-art deep learning methods, for high accuracy with the segmentation of roads against diverse and complex backgrounds.

2 THEORETICAL BACKGROUND

In this work, we employed a series of deep learning models and techniques for semantic segmentation, including Smile U-Net, DeepLabV3+, and DinkNet. These models leverage convolutional neural networks (CNNs) for pixel-wise classification, enabling the differentiation between roads and background in satellite imagery.

2.1 U-NET

The U-Net architecture, which Smile U-Net is based on, follows a symmetric encoder-decoder structure. The encoder captures spatial features, while the decoder reconstructs the spatial resolution for segmentation. The model minimizes a loss function $\mathcal{L}$ to optimize its performance, specifically the **Binary Cross-Entropy (BCE) loss**, which is defined as:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right],$$

where y_i is the ground truth label ($y_i \in \{0,1\}$), p_i is the predicted probability for pixel i, and N is the total number of pixels in the dataset.

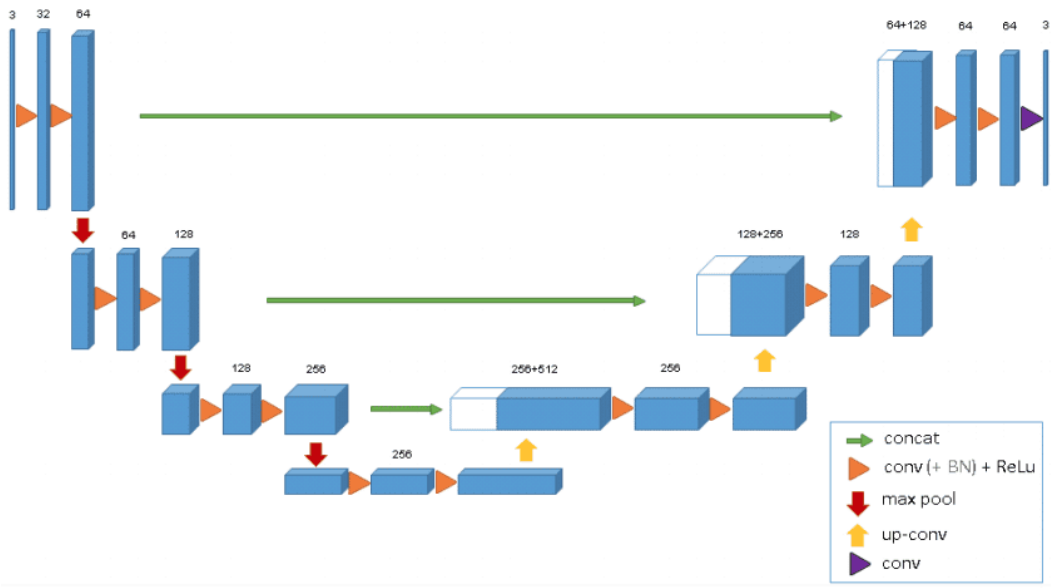

Figure 1: The U-Net architecture: A symmetric encoder-decoder structure with skip connections.

2.2 DeepLabV3+

DeepLabV3+ is a semantic segmentation model that enhances performance by integrating **atrous (dilated) convolutions** and the **Atrous Spatial Pyramid Pooling (ASPP)** module. It is designed to capture rich contextual information at multiple scales while preserving spatial resolution, making it particularly effective for segmenting complex features such as roads.

Atrous (Dilated) Convolutions

Atrous convolutions introduce a **dilation rate** r, which controls the spacing between kernel elements. This increases the receptive field without increasing the number of parameters or reducing the spatial resolution. The output of an atrous convolution is defined as:

$$y[m,n] = \sum_{i,j} x[m + r \cdot i, n + r \cdot j] \cdot w[i,j],$$

where:

- x: Input feature map.
- w: Convolutional kernel.
- $y[m,n]$: Output feature map.
- r: Dilation rate.

Atrous Spatial Pyramid Pooling (ASPP)

The ASPP module is a key component of DeepLabV3+, enabling the model to extract features at multiple scales. It uses parallel atrous convolutions with different dilation rates to capture both fine and global contextual information. The ASPP output is defined as:

$$\text{ASPP}(x) = \sum_{k=1}^{K} f_k(x),$$

where $f_k(x)$ represents feature extraction at scale k.

The ASPP module includes:

- **1x1 Convolution:** Captures fine-grained details.
- **3x3 Atrous Convolutions:** Operates at varying dilation rates (e.g., $r = 6, 12, 18$).
- **Image Pooling:** Encodes global contextual information.

Decoder Design

The decoder refines segmentation results by upsampling and incorporating low-level features from the encoder via skip connections. This design ensures accurate boundary prediction while maintaining computational efficiency.

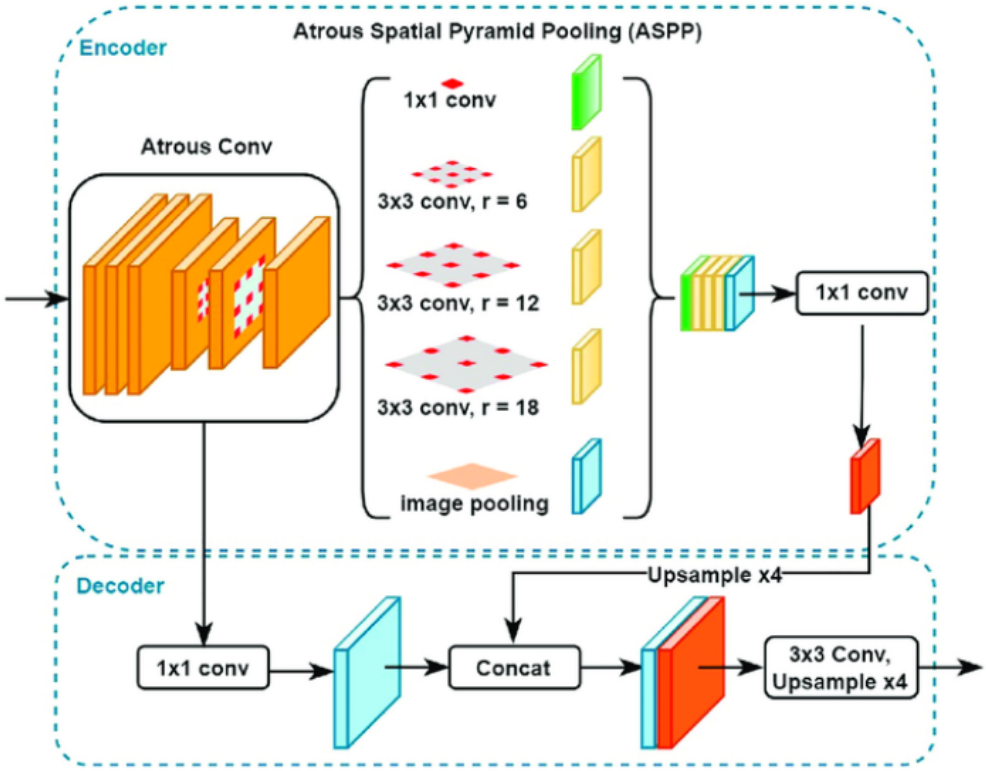

Figure 2: DeepLabV3+ architecture: Atrous convolutions and the ASPP module enable multi-scale feature extraction, while the decoder refines segmentation predictions using skip connections.

2.3 DinkNet

DinkNet is a segmentation model that combines a **ResNet backbone** for feature extraction, **dilated convolutions**, and **attention mechanisms** to improve the segmentation accuracy of road networks. Its architecture is designed to balance global context and local detail, making it highly effective for semantic segmentation tasks.

Key Components

- **Dilated Convolutions:** Increase the receptive field without reducing spatial resolution or increasing the number of parameters. This enables the model to capture multi-scale contextual information, which is critical for segmenting structures like roads that vary in size and shape.
- **Residual Blocks:** ResNet's residual connections prevent vanishing gradients and enable deeper feature extraction.
- **Decoder Design:** The decoder reconstructs the segmentation map using transpose convolutions and combines multi-level features via skip connections, as shown in **Figure 3**.

Loss Function

While the standard loss function for semantic segmentation is the **Binary Cross-Entropy (BCE) Loss**, DinkNet in this study utilized a **hybrid loss function** combining BCE and **Dice Loss**:

- **BCE Loss:** Ensures pixel-wise classification accuracy.
- **Dice Loss:** Optimizes the overlap between predicted and ground truth segmentation masks, particularly useful for imbalanced datasets.

The hybrid loss function stabilizes training and ensures robust performance across different threshold settings, addressing issues observed when using Dice Loss alone.

The hybrid loss function is defined as:

$$L_{\text{Hybrid}} = \alpha L_{\text{BCE}} + (1 - \alpha) L_{\text{Dice}},$$

where:

- L_{BCE} is the binary cross-entropy loss.
- L_{Dice} is the Dice loss:

$$L_{\text{Dice}} = 1 - \frac{2 \sum_{i=1}^{N} p_i y_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} y_i^2}.$$

- α is a weight parameter balancing the contributions of BCE and Dice Loss.

DinkNet Architecture

Below is a visual representation of the DinkNet architecture, illustrating its encoder-decoder structure with dilated convolutions and skip connections.

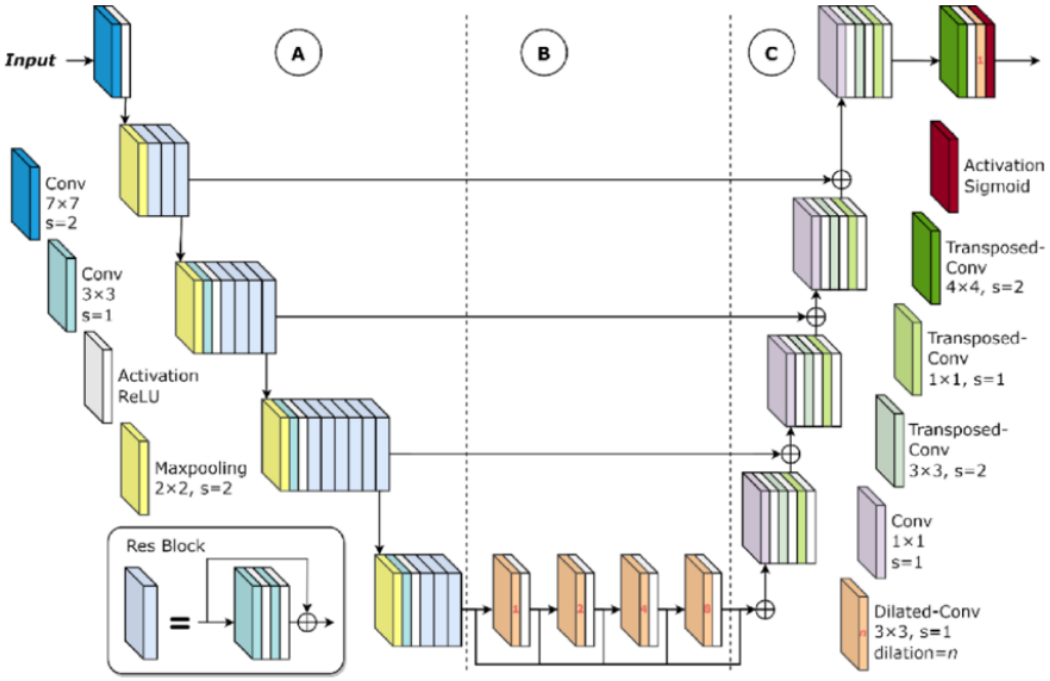

Figure 3: DinkNet architecture: An encoder-decoder structure utilizing ResNet blocks, dilated convolutions, and attention mechanisms.

IMPORTANCE OF THE HYBRID LOSS FUNCTION

- The hybrid loss function ensures stability during training by balancing pixel-wise accuracy with mask overlap.
- **BCE Loss** focuses on classifying each pixel correctly, while **Dice Loss** emphasizes the overall shape and structure of segmented objects.
- Combining these losses leads to improved performance, particularly in scenarios with **class imbalance**, where road pixels are less frequent than background pixels.

This methodology enabled DinkNet to achieve high segmentation accuracy in this study, leveraging the strengths of both loss functions while mitigating their individual weaknesses.

2.4 SYSTEMATIC AUGMENTATION AND TRAINING TECHNIQUES

We developed a novel systematic data augmentation strategy, targeting batches with low average F1 scores to improve model performance. Additionally, our training loop incorporated learning rate reduction after three epochs without improvement and early stopping upon reaching a predefined threshold.

.

3 DATASET DESCRIPTION

The dataset is divided into two main subsets: training and testing.

3.1 TRAINING DATASET

The training dataset has two components. The first component is satellite/aerial images acquired from Google Maps, which include 100 images of urban or suburban areas. These images show roads, buildings, parking, factories, and cars. Figure 4 shows an example of a satellite image.

Figure 4: satellite image

The second component is the ground truth masks where each pixel is labeled either as road or background. Each satellite image is paired with a corresponding binary ground truth mask. Figure 5 shows an example of a ground truth image.

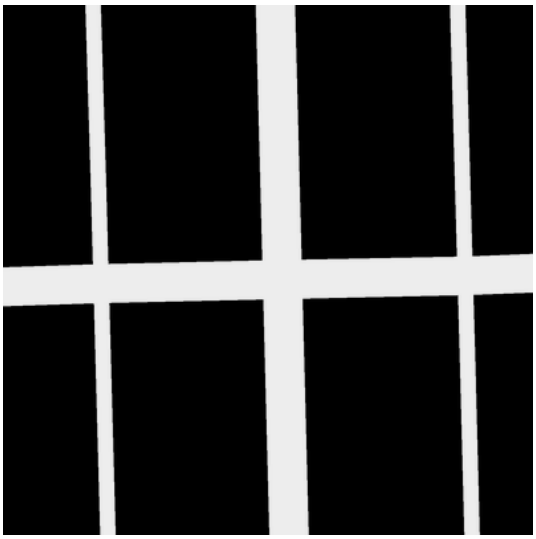

Figure 5: ground truth mask

3.2 Testing Dataset

The testing dataset consists of 50 satellite images that are distinct from the training dataset. This data is used to evaluate the model's performance and determine how well it generalizes to unseen road configurations.

4 Methodology

The following flowchart demonstrates the step-by-step process of how data is processed for road segmentation using deep learning techniques.

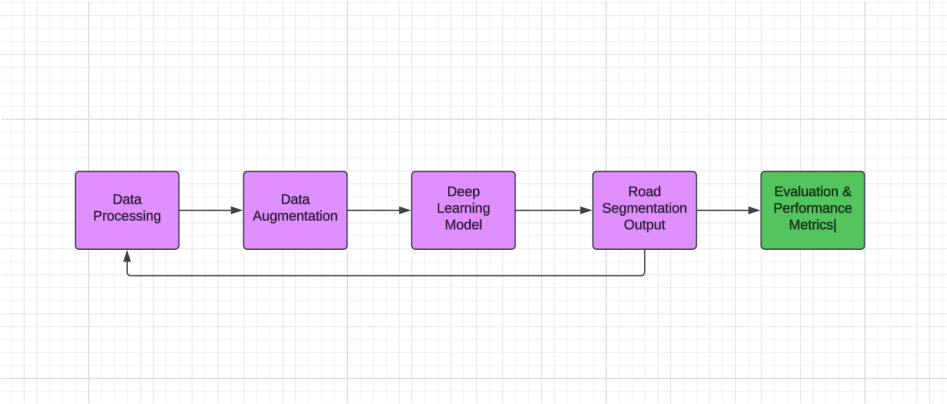

Figure 6: step-by-step process

4.1 Data Processing

The data was preprocessed to ensure consistency and suitability for deep learning models. This involved:

- **Resizing:** All images were resized to a uniform resolution.
- **Normalization:** Pixel values were normalized to improve convergence.
- **Cleaning:** Noisy or inconsistent data were removed.

4.2 Data Augmentation

Two augmentation strategies were implemented:

1. **Random Augmentation:** Flips, rotations, and color adjustments.
2. **Systematic Augmentation:** Targeting batches with low F1 scores.

Table 1: Performance Comparison of Augmentation Strategies

Augmentation Strategy	Model	F1 Score
Random Augmentation	DeepLabV3+	0.75
Systematic Augmentation	DeepLabV3+	0.79

4.3 Training Strategies

- **Learning Rate Scheduling:** Decreased learning rate after 3 epochs without improvement.
- **Early Stopping:** Stopped training once a predefined threshold was met.

4.4 Deep Learning

Table 2: Model Development Results

Model	Backbone	Loss Function	F1 Score
Smile U-Net	-	BCE	0.71
Pretrained Model	-	BCE	0.77
DeepLabV3+	ResNet-50	Hybrid (Dice + BCE)	0.79
DinkNet	ResNet-152	Hybrid (Dice + BCE)	0.916

5 Model Development and Results

The model development process involved iterative improvements through architectural enhancements, loss function modifications, and training optimizations. Starting with a Smile U-Net model trained from scratch, we achieved an initial F1 score of 0.71. To improve performance, pretrained models were introduced, resulting in a notable increase to an F1 score of 0.77.

Subsequent experiments with different loss functions revealed that using Dice loss provided a slight accuracy improvement, reaching 92% compared to 91% with the standard binary cross-entropy loss. To further enhance stability, we adopted a hybrid loss function that combined Dice loss with binary cross-entropy, effectively mitigating the threshold sensitivity associated with Dice loss.

A transition to a more advanced DeepLabV3+ model, using a pretrained ResNet-50 backbone, yielded an F1 score of 0.75. To address performance consistency, we developed a novel systematic data augmentation approach. By targeting batches with low average F1 scores, we improved the performance of DeepLabV3+ to 0.79. Additionally, a robust training loop was implemented, incorporating adaptive learning rate reduction after three epochs without improvement and early stopping upon reaching predefined thresholds.

The introduction of the DinkNet model with a pretrained ResNet-152 backbone marked a significant performance boost, achieving an F1 score of 0.88 and 0.914 on the competition leaderboard, securing 18th place at that stage. Applying the systematic augmentation approach with DinkNet further improved the F1 score to 0.916, elevating our position to 12th place. Through iterative tuning and optimization, we achieved a final F1 score of 0.925 as our result and 0.917 on the competition leaderboard, along an accuracy of 0.958. This effort secured us 10th place on the final day of the competition. Figures 7 and 8 show an example of the test data using the final model.**??**

Figure 7: Input image

Figure 8: Predicted mask

5.1 Summary of Results

Table 3: Summary of Model Performance and Rankings

Iteration	**F1 Score**	**Rank**
Initial Smile U-Net	0.71	-
Pretrained Model	0.77	-
DeepLabV3+ + Systematic Augmentation	0.79	-
DinkNet + ResNet-152	0.914	18th
DinkNet + Systematic Augmentation	0.916	12th
DinkNet + ResNet-152 + Systematic Augmentation	0.917	10th

5.2 Key Insights

- Systematic augmentation consistently improved performance by focusing on low-scoring batches.
- The hybrid loss function (Dice + BCE) stabilized training.
- DinkNet with ResNet-152 achieved the best result with an F1 score of 0.917.

6 Future Work

While the current study demonstrates the effectiveness of advanced segmentation models like DinkNet and DeepLabV3+ for road extraction, there are several avenues for further improvement:

- **Incorporation of Transformer-Based Architectures:** Explore the integration of vision transformers (ViTs) or hybrid transformer-CNN models to better capture global and local dependencies in road segmentation.
- **Improved Augmentation Techniques:** Investigate more sophisticated data augmentation strategies, such as GAN-based synthetic data generation, to enhance model robustness in diverse environments.

- **Domain Adaptation:** Address the challenges of domain shift by incorporating unsupervised or semi-supervised domain adaptation techniques to improve performance across different geographical regions.
- **Lightweight Models for Real-Time Applications:** Develop lightweight architectures optimized for deployment on edge devices, enabling real-time road segmentation for navigation and autonomous driving.
- **Multi-Modal Data Fusion:** Integrate additional data modalities, such as LiDAR or aerial imagery with different spectral bands, to improve segmentation accuracy in challenging scenarios.

These directions aim to further enhance the accuracy, efficiency, and adaptability of road segmentation models for practical applications in transportation and urban planning.

7 Conclusion

In this study, we explored state-of-the-art deep learning techniques for road segmentation from satellite and aerial imagery, focusing on models such as Smile U-Net, DeepLabV3+, and DinkNet. Through systematic experimentation, including the adoption of hybrid loss functions and advanced data augmentation strategies, significant improvements in segmentation performance were achieved.

The integration of atrous convolutions and multi-scale feature extraction in DeepLabV3+ demonstrated its capability to capture complex contextual information, while the DinkNet architecture, with its ResNet backbone and attention mechanisms, achieved superior segmentation accuracy. Additionally, the use of systematic data augmentation proved to be an effective method for addressing imbalances in the dataset and enhancing model robustness.

Despite these advancements, challenges remain, particularly in improving domain generalization and achieving real-time performance for practical applications. Future work should focus on leveraging emerging technologies, such as transformer-based architectures and multi-modal data fusion, to further enhance segmentation accuracy and adaptability.

This work underscores the potential of advanced deep learning methodologies for road segmentation, contributing to applications in transportation management, urban planning, and disaster response. By continuing to refine these models, we can pave the way for more efficient and scalable solutions in remote sensing and geographic information systems.

Supplementary Material

The supplementary material for this paper is available on our GitHub repository: GitHub Repository.

References

[1] K. Jacobsen, et al., High resolution satellite imaging systems-an overview, Photogrammetrie Fernerkundung Geoinformation 2005 (6) (2005) 487.

[2] C. Zhang, A. Marzougui, S. Sankaran, High-resolution satellite imagery applications in crop phenotyping: An overview, Computers and Electronics in Agriculture 175 (2020) 105584.

[3] P. M. Dare, Shadow analysis in high-resolution satellite imagery of urban areas, Photogrammetric Engineering & Remote Sensing 71 (2) (2005) 169–177.

[4] C. Henry, S. M. Azimi, N. Merkle, Road segmentation in sar satellite images with deep fully convolutional neural networks, IEEE Geoscience and Remote Sensing Letters 15 (12) (2018) 1867–1871.

[5] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, S. Ermon, Combining satellite imagery and machine learning to predict poverty, Science 353 (6301) (2016) 790–794.

[6] T. Man, R. Rusu, C. Moldovan, M. Ionescu-Heroiu, N.-S. Moldovan, I. HĂRĂNGUŞ, Spatial impact of the road infrastructure development in romania. an accessibility approach., Romanian Review of Regional Studies 11 (1) (2015).

[7] S. Gargoum, K. El-Basyouny, Automated extraction of road features using lidar data: A review of lidar applications in transportation, in: 2017 4th International Conference on Transportation Information and Safety (ICTIS), IEEE, 2017, pp. 563–574.

[8] N. J. Yuan, Y. Zheng, X. Xie, Segmentation of urban areas using road networks, Microsoft, Albuquerque, NM, USA, Tech. Rep. MSR-TR-2012-65 (2012).

[9] J. Qiu, L. Du, D. Zhang, S. Su, Z. Tian, Nei-tte: Intelligent traffic time estimation based on fine-grained time derivation of road segments for smart city, IEEE Transactions on Industrial Informatics 16 (4) (2019) 2659–2666.

[10] C. Van der Sande, S. De Jong, A. De Roo, A segmentation and classification approach of ikonos-2 imagery for land cover mapping to assist flood risk and flood damage assessment, International Journal of applied earth observation and geoinformation 4 (3) (2003) 217–229.