T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

FUSING THE RGB IMAGE AND LIDAR DATA FOR ROAD DETECTION

ARDA TAHA CANDAN
A THESIS SUBMITTED FOR THE DEGREE OF
MASTER OF SCIENCE
DEPARTMENT OF COMPUTER ENGINEERING

GEBZE
2023

T.R.

GEBZE TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

FUSING THE RGB IMAGE AND LIDAR DATA FOR ROAD DETECTION

ARDA TAHA CANDAN

A THESIS SUBMITTED FOR THE DEGREE OF

MASTER OF SCIENCE

DEPARTMENT OF COMPUTER ENGINEERING

THESIS SUPERVISOR

ASSOC. DR. HABİL KALKAN

GEBZE

2023

T.C.

GEBZE TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

YOL TESPİTİ İÇİN RGB KAMERA VE LİDAR FÜZYONU

ARDA TAHA CANDAN

YÜKSEK LİSANS TEZİ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DANIŞMANI

DOÇ. DR. HABİL KALKAN

GEBZE

2023

GEBZE TEKNİK ÜNİVERSİTESİ	**YÜKSEK LİSANS JÜRİ ONAY FORMU**

GTÜ Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 18/05/2023 tarih ve 2023/29 sayılı kararıyla oluşturulan jüri tarafından 06/06/2023 tarihinde tez savunma sınavı yapılan Arda Taha Candan'ın tez çalışması Bilgisayar Mühendisliği Anabilim Dalında YÜKSEK LİSANS tezi olarak kabul edilmiştir.

JÜRİ

ÜYE (TEZ DANIŞMANI)	: Doç. Dr. Habil KALKAN
ÜYE	: Dr. Öğr. Üyesi Yakup GENÇ
ÜYE	: Dr. Öğr. Üyesi Savaş ÖZTÜRK

ONAY

Gebze Teknik Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun

..…/……/…… tarih ve ……/…..... sayılı kararı.

SUMMARY

Drivable road detection is one of the most fundamental problems for autonomous vehicles. The most commonly used sensors for this problem are RGB cameras and LiDARs. While RGB camera data contains a wealth of visual information, such as colour, LiDAR data provides precise position information without being affected by ambient light. Several studies show that using these two sensors together for road detection produces more robust results. However, the fact that the data produced by the sensors are in different forms and spaces makes the sensor fusion method critical to utilise the obtained features efficiently. In this thesis, first, LiDAR data was processed to make them suitable for fusion. Afterwards, three different U-Net-based novel image segmentation architectures were developed: early fusion, late fusion, and cross fusion. Then, the models based on these architectures were trained and evaluated on the KITTI road detection dataset. The early fusion and cross fusion models using the RGB image and LiDAR altitude difference image achieved the highest MaxF score among the evaluated models. The models were also compared with other state-of-the-art models in the literature, and the results were at a competitive level.

Key Words: Autonomous vehicles, Sensor fusion, Computer vision, Deep learning.

ÖZET

Sürülebilir yol tespiti, otonom araçlar için en temel sorunlardan biridir. Bu problem için en sık kullanılan sensörler RGB kameralar ve LiDAR'lardır. RGB kamera verileri renk gibi çok sayıda görsel bilgi içerirken, LiDAR verileri ortam ışığından etkilenmeden hassas konum bilgisi sağlamaktadır. Çeşitli çalışmalar, yol tespiti için bu iki sensörün birlikte kullanılmasının daha kararlı sonuçlar ürettiğini göstermiştir. Ancak sensörlerin ürettiği verilerin farklı biçimlerde ve boyutlarda olması, elde edilen özniteliklerin verimli bir şekilde kullanılması için sensör füzyon yöntemini kritik hale getirmektedir. Bu tezde, ilk olarak LiDAR verileri işlenerek füzyona uygun hale getirilmiştir. Daha sonra, erken füzyon, geç füzyon ve çapraz füzyon olmak üzere üç farklı U-Net tabanlı özgün görüntü bölütleme mimarisi geliştirilmiştir.Ardından, bu mimariler ile KITTI yol tespiti veri seti üzerinde modeller eğitilmiş ve değerlendirilmiştir. RGB görüntüsünü ve LiDAR yükseklik fark görüntüsünü kullanan erken füzyon ve çapraz füzyon modelleri, değerlendirilen modeller arasında en yüksek MaxF skorlarını elde etmişlerdir. Modeller ayrıca literatürdeki diğer başarılı modeller ile de karşılaştırılmış ve rekabetçi sonuçlar elde edilmiştir.

Anahtar Kelimeler: Otonom araçlar, Sensör füzyonu, Bilgisayarlı görü, Derin öğrenme.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor Assoc. Prof. Habil Kalkan for his support and guidance at every stage of this thesis. Apart from his invaluable experience and expertise, his constant motivation and encouragement made it possible for me to complete this thesis and all the related publications.

I am grateful to my beloved parents, Fatma and Arif, and my brother Tolga for their endless efforts and unwavering faith in me. I would also like to express my special thanks to my dear wife, Gizem Nur, who has shared my burden and shown me her great love and constant support throughout the process. Without their encouragement and support, this thesis would not have been possible.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS AND ACRONYMS

Abbreviations and Acronyms		Explanations
ADC	:	Analog-to-Digital Converter
ADI	:	Altitude Difference Image
ALT	:	Altitude Image
ASPP	:	Atrous Spatial Pyramid Pooling
CNN	:	Convolutional Neural Network
CRF	:	Conditional Random Fields
CV	:	Coefficient of variation
DCNN	:	Deep Convolutional Neural Network
FCN	:	Fully Convolutional Neural Networks
JPU	:	Joint Pyramid Upsampling
LiDAR	:	Light Detection and Ranging
PRE	:	Precision
REC	:	Recall
ReLU	:	Rectified Linear Unit
RGB	:	Red-Green-Blue

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Autonomous vehicles are revolutionising the transportation sector by becoming safer, more efficient and more accessible. According to a report by McKinsey [1], %37 of the vehicles sold in 2035 are expected to support at least level 3 autonomy [2]. The report also predicts that driving assistance systems and autonomous driving capabilities could generate between $300 billion and $400 billion in revenue in the passenger car market. However, the successful deployment of autonomous vehicles on public roads mostly relies on developing robust and accurate perception systems that can interpret and understand the surrounding environment.

1.1. Problem Definition

Road detection is a critical task for autonomous vehicles, as it involves detecting and segmenting the driveable road using the data obtained from onboard sensors. It provides a basis for many essential applications for autonomous vehicles, such as lane detection, object detection and path planning. In more detail, lane detection is used to position the vehicle inside the lane and detect possible deviations. Object detection is used to detect objects that may threaten the vehicle in the environment, such as other vehicles and pedestrians on the road. Path planning, on the other hand, is used to determine how the vehicle will reach the target point on the detected road. For all these applications, the area that can be used as a road by the vehicle must be well segmented.

The most commonly used sensors in autonomous vehicles are cameras. Cameras are often sufficient in many autonomous vehicle applications, mainly thanks to the rich colour and texture information they provide. However, their performance depends on the ambient light due to being passive sensors. For example, in a low-light environment, cameras cannot capture sufficient information, or on a sunny day, the shadows on the road may distort road detection [3]. Furthermore, adverse weather like snow, rain, and fog can can seriously affect camera performance [4]. Another frequently used sensors in autonomous vehicles are LiDARs. LiDARs are active sensors not affected by ambient

light and provide accurate distance measurements of the surroundings. However, LiDARs do not provide colour information like cameras and require preprocessing to make them suitable for segmentation. A more robust and accurate representation of the environment can be obtained by fusing these two sensors. However, the fusion process is challenging since these sensors generate data at different resolutions and spaces. Therefore, an efficient fusion approach should be designed to achieve robust segmentation.

In conclusion, an accurate and robust road detection system is critical for developing reliable autonomous vehicles. For accurate road detection, the boundaries of the road must be distinguished from the environment and the other object in a high-resolution manner. The results must be unaffected by ambient light and weather conditions for robust road detection. By combining LiDAR and camera sensors using an efficient fusion approach, a road segmentation system can be realised to ensure these criteria.

1.2. Aim and Scope

This thesis aims to adapt the deep learning architectures to sensor fusion and investigate the effectiveness of camera and LiDAR fusion for road segmentation. Specifically, the aim is to answer the following research questions:

- How can LiDAR data be used for image segmentation?
- How can deep learning architectures be adapted to fusion approaches for the road segmentation task?
- Which fusion approach is more effective for camera and LiDAR fusion regarding road detection?
- Can camera and LiDAR fusion improve the accuracy and robustness of road segmentation?

This thesis focuses on developing a method that effectively fuses camera and LiDAR for road segmentation. To that end, various deep learning architectures for segmentation were realised for different fusion approaches. The proposed architectures were evaluated on a public dataset and compared with each other and the other state-of-the-art methods.

1.3. Outline

The remainder of this thesis is organised as follows. Section 2 provides a background on the camera and LiDAR sensors. It also includes related work and discusses the limitations of the existing methods. Section 3 describes the proposed methodology in detail, including the dataset, data preprocessing, deep learning architectures, and fusion approaches. Section 4 presents the results obtained from the experiments conducted on the dataset. These results are then compared with the state-of-the-art methods to show the effectiveness of the proposed approaches. Finally, Section 5 concludes the thesis by summarising the main contributions of this study. It also highlights the limitations of the current approach and suggests potential directions for future research.

2. BACKGROUND AND RELATED WORK

Autonomous vehicles are becoming increasingly popular due to their potential to revolutionise transportation. However, a robust perception system that can accurately interpret and understand the surrounding environment is required for autonomous vehicles to operate safely and effectively on public roads. This section provides an overview of the perception systems used in autonomous vehicles, including the fundamentals of camera and LiDAR sensors. Then, the existing methods related to road detection using cameras and LiDAR sensors are reviewed by highlighting their strengths and limitations.

2.1. Perception for Autonomous Driving

The perception system of an autonomous vehicle is responsible for feeding the decision mechanism of the vehicle by interpreting the data collected by the environment. Such a perception system consists of various sensors, such as RGB cameras and LiDARs. These sensors enable the capturing of unique features of the environment. For example, an RGB camera captures high-resolution and colour images of the scene and is convenient for object detection, lane detection and traffic sign detection. A LiDAR sensor, on the other hand, provides a 3D mapping of the environment with high precision and is suitable for object detection and localisation.

Different methods are used to interpret the data collected from the sensors, such as computer vision and machine learning algorithms. Then, the outputs are given to the autonomous vehicle's decision mechanism to decide how to navigate the environment.

2.1.1. RGB Camera Sensors

Most cars on the roads today have already installed at least one camera. In fact, as of 2018, it has become mandatory for all newly manufactured vehicles in the U.S. to have a backup camera [5]. Besides being used in driving assistance systems, cameras are also the most common perception sensor for autonomous vehicles.

The working principle of the cameras is dropping the light absorbed from surrounding objects onto a photosensitive lens. The light-sensitive surface converts the light into electrons when the light falls on the lens. These electrons are then converted into pixels by being converted to voltage and passed through the Analog-Digital Converter (ADC). Thus, a 2-dimensional image of the environment is captured. The resulting image can be expressed as a $(height \times width)$ matrix depending on the camera specifications. Each pixel value in the matrix is defined as $[R, G, B]$ to represent colour information.

Cameras have many advantages that make them popular for autonomous vehicles. They provide high-resolution data along with rich colour information. These features allow problems such as object detection and classification to be easily solved using image processing or deep learning algorithms on camera images. In addition, since the cameras are relatively small and inexpensive sensors, they can be installed at multiple positions in the vehicle at a low cost.

The downside of cameras is mainly caused by that they are passive sensors. Passive sensors need input from the physical environment, such as light. Therefore, they are directly dependent on ambient light. The performance of the cameras, similar to the human eye, degrades drastically when the light is low, for example, at night [3]. In addition, adverse weather like snow, rain, and fog that causes poor visibility also affects camera performance [4]. In particular to the image segmentation problem, shadows or glares on the image also affect the performance [6]. Considering that autonomous vehicles are expected to operate at any time of the day and in all weather conditions, these disadvantages are safety-critical problems.

2.1.2. LiDAR Sensors

LiDAR, which stands for Light Detection and Ranging, are remote sensing sensors that measure the distances of the surrounding objects and model the environment using laser pulses. The working principle of the LiDAR sensors is emitting laser pulses and measuring the time it takes for the laser to reflect off objects and return to the sensor.

Precise environmental sensing can then be done by analysing the time-of-flight and intensity of the reflected lasers.

2D and 3D LiDAR sensors are available, but 3D LiDARs are commonly preferred for autonomous vehicles. The data obtained from the LiDAR sensor is called the point cloud. A 3D point cloud collected from an autonomous vehicle is shown in Figure 2.1 [7]. Section 3.2 gives detailed information about the point cloud data type.

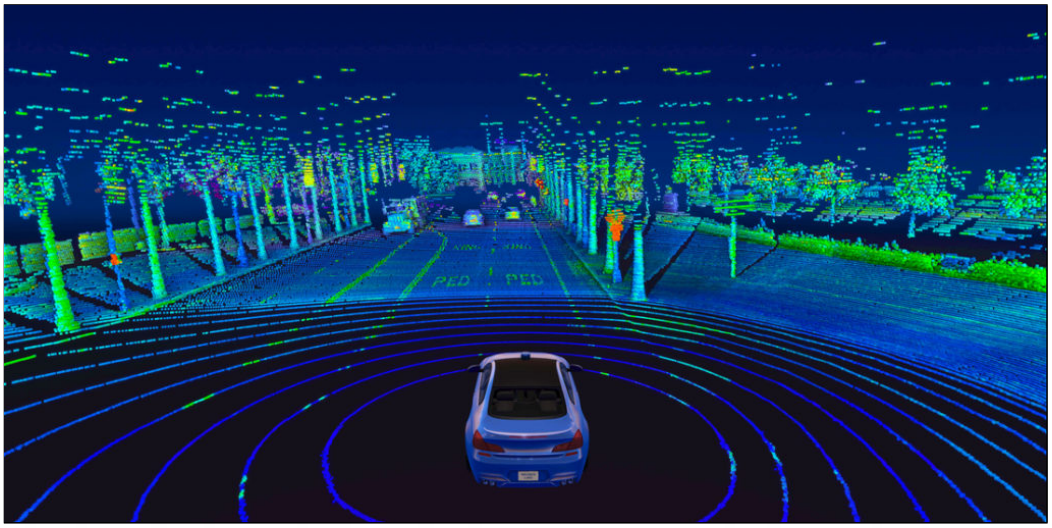

Figure 2.1: Point cloud data obtained from an onboard 3D LiDAR.

LiDAR sensors are classified as active sensors since they emit laser pulses into the environment instead of using light in the environment. Therefore, they are not dependent on ambient light. They are also more robust to the adverse weather such as snow, rain and fog than cameras [8]. Another advantage of the LiDARs is that they generate 3D data. Thus, LiDARs are commonly used for mapping and localisation [9], [10]. These qualities make LiDARs advantageous to use on autonomous vehicles.

LiDAR sensors also have several weaknesses. Since LiDARs cannot detect colours like cameras, they cannot provide visual features. Thus, they cannot be used for problems such as lane detection or traffic sign recognition. Also, although they can measure distance with high precision, they cannot provide high-resolution images like cameras because the resolution of the data depends on the quantity of reflected laser pulses. Consequently, LiDAR data is in discrete form and using methods such

as clustering or interpolation may be required. Finally, while the cameras capture the visual line of sight, LiDARs can only operate at a specific range due to their working principles.

2.2. Related Work

Road detection can be realised with image processing methods in ideal environments by utilising high-resolution and rich colour information provided by the camera images [11]. However, for these methods to be reliable, the lane markings on the road must be clear and distinct, as on highways, and all external conditions, such as light and weather, must be ideal. Nevertheless, in the real world, there are many situations where lane markings are not distinct or nonexistent [12]. Also, the sidewalks in urban areas and various external conditions can cause such methods non-functional. Developing a robust method for real-world problems is possible by approaching road detection as a segmentation problem. Besides, the successful results of the deep learning segmentation models published recently accelerated the developments in this field. Popular segmentation models in the literature are discussed in detail in Section 4.

Various methods for road detection have been proposed in the past years. Most of these methods focus on using camera images [13], [14]. Although good results are obtained overall, the methods' performance is seriously affected when the challenging cases are examined, which consist of poor ambient light and weather conditions [6]. There are also methods that focus on using LiDAR data [15], [16]. Albeit these methods achieved better results in the challenging cases, the overall performance was not competitive enough. The results of the proposed models for camera and LiDAR lead to the development of new methods focused on camera and LiDAR fusion that fully benefits the high resolution and rich colour features of cameras and is also robust against challenging cases by utilising LiDAR data.

Sensor fusion involves combining features from multiple sensors to acquire a more accurate and complete understanding of the environment. There are two main

approaches for sensor fusion: early fusion and late fusion [17]. In early fusion, raw data from multiple sensors are combined before any processing. In contrast, in late fusion, data from each sensor is combined after being processed independently. These fusion approaches are described in detail in Section 3.5, including their benefits and drawbacks. In addition to these approaches, many different sensor fusion approaches for the camera and LiDAR have been proposed in the literature.

Caltagirone et al. used fully convolutional neural networks (FCN) for the camera and LiDAR fusion and proposed the cross fusion approach [6]. In this approach, sensor data is fused at specific rates at each model layer, different from early fusion and late fusion approaches. Cross fusion approach ensures that the distinctive features of the sensor data are preserved during the fusion. A comparison of the proposed architectures, early fusion, late fusion, and cross fusion, is shown in Figure 2.2. In the result of the study, the performance of the cross fusion approach surpassed the early fusion and late fusion approaches, especially in challenging cases. In Section 4, the performance of the cross fusion method is compared with the U-Net-based method proposed in this thesis.

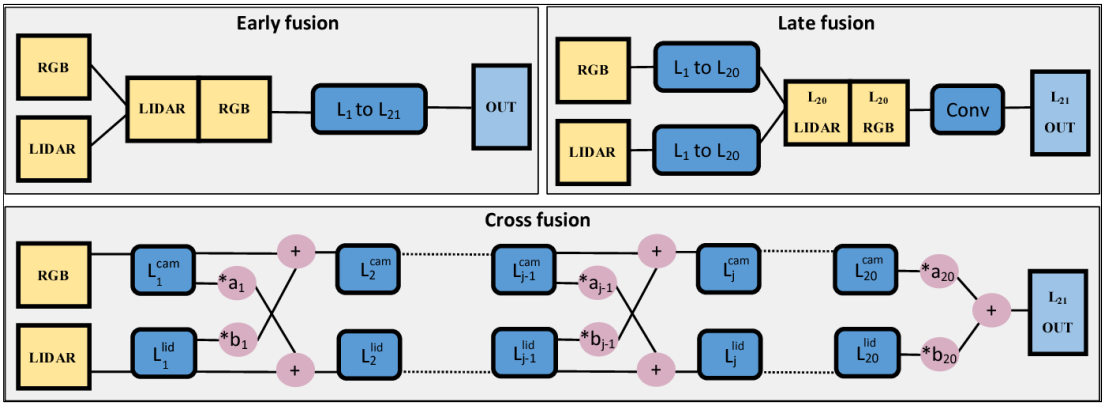

Figure 2.2: Early fusion, late fusion, and cross fusion architectures of the LidCamNet.

Chen et al. proposed the Progressive LiDAR Adaptation for Road Detection (PLARD) method [20]. The PLARD method consists of two sub-modules. In the first module, Data Space Adaptation, altitude difference transformation is applied to the LiDAR data to align the LiDAR data with the camera images and make them suitable for segmentation. Altitude difference transformation is also used in this thesis; all details

are explained in Section 3.2.2. In the second module, Feature Space Adaptation, LiDAR features and camera features are concatenated and passed through a transformation network consisting of three (1×1) convolutional layers. The output of this network is the parameters that determine the linear transformation required to fuse LiDAR features with camera features. This module aims to improve performance by eliminating feature space inconsistency. Figure 2.3 shows the architecture of the PLARD method. According to the data pipeline, ResNet-101 [18] based DCNNs are deployed for ADI images obtained from LiDAR data as an output of Data Space Adaptation and RGB images, respectively. After each network layer, the features on the ADI side are fused to the features on the RGB image side by passing through the Feature Space Adaptation Module. Thus, the fusion method, which is quite similar to the cross fusion approach, is used. When writing this thesis, the PLARD method was the most successful in the KITTI Road Detection leaderboard [19] among the models utilising LiDAR. In Section 4, the performance of the PLARD method is compared with the U-Net-based method proposed in this thesis.

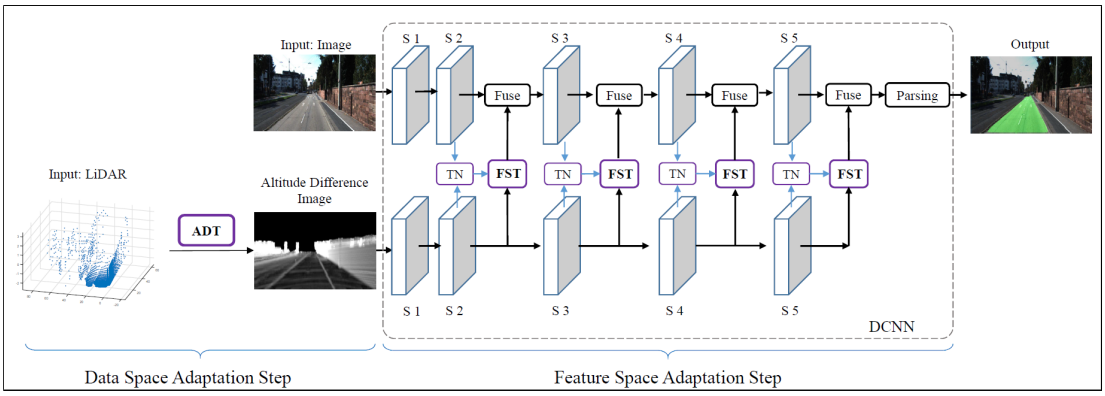

Figure 2.3: The architecture of the PLARD.

Unlike the other studies, Gu et al. used the Conditional Random Fields (CRF) method for sensor fusion [21]. First, they convert the LiDAR data to a camera image. By applying Flat Region Extraction on it, regions where the height difference is less than a specific threshold value are highlighted. Afterwards, in the Horizontal-Vertical Scanning step, the rows and columns of the LiDAR images are scanned, and local

plains are detected. Finally, images are interpolated using the Delaunay Triangulation [22] method, and road detection results are obtained. The DeepLab model [23] is also trained for camera images in parallel, and results are obtained. The results from the LiDAR and camera are then fused using CRF to get the final result. The steps of the proposed method are shown in Figure 2.4. In Section 4, the performance of the LC-CRF method is compared with the U-Net-based method proposed in this thesis.

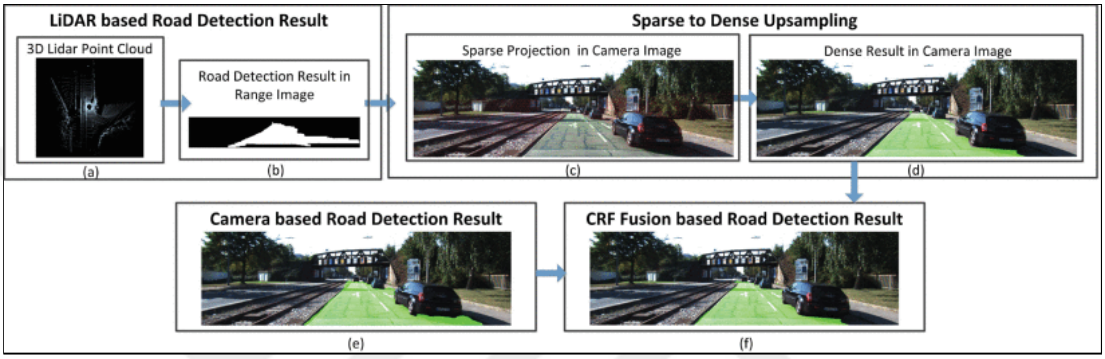

Figure 2.4: The flowchart of the LC-CRF.

3. METHODOLOGY

In this section, the preprocessing and implementation stages of the models proposed in this thesis are explained in detail. Section 3.1 describes the dataset used for training and evaluating the models and its characteristics. Section 3.2 explains the processing steps of the LiDAR data to create images suitable for image segmentation in detail. Section 3.3 gives information about the data augmentation applied to the dataset. Section 3.4 summarises popular models in the literature used for image segmentation and explains the selected model and its architecture in detail. Finally, Section 3.5 describes the proposed fusion model architectures.

3.1. Dataset

KITTI road benchmark dataset [24], created by Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago in collaboration with Honda Research Institute Europe GmbH, is selected for training and evaluating the models proposed in this thesis. Being that the dataset is commonly used in the literature, it is an excellent benchmark for comparing new results with the literature. It also includes both types of sensors needed for this thesis. The data was collected with a Volkswagen Passat B6 vehicle equipped with a Velodyne HDL-64E laser scanner and two 1.4 Megapixels Point Grey Flea 2 colour cameras with four 4-8 mm Edmund Optics NT59-917 varifocal lenses attached. The setup of the vehicle is shown in Figure 3.1.

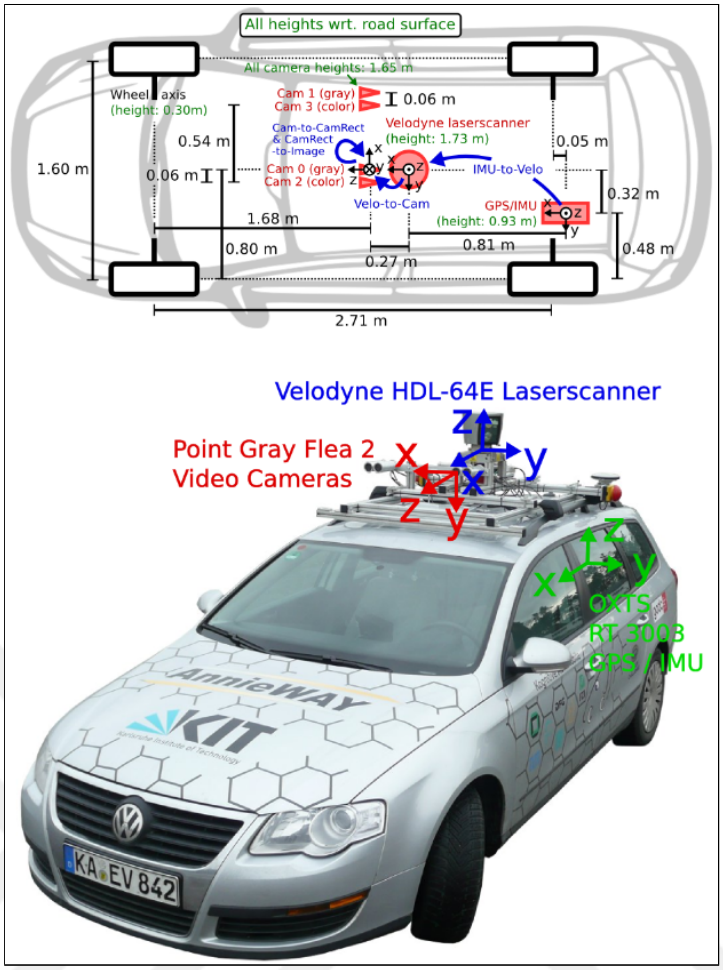

Figure 3.1: Sensor setup of the KITTI vehicle.

The dataset contains RGB images and 3D LiDAR data collected from various types of urban roads, such as unmarked, marked, and multiple marked lanes. It is organised into two categories as training and test by default. The number of samples for each category and road type is shown in Table 3.1.

Table 3.1: The number of samples for each category and road type in the dataset.

Road category	**Training**	**Test**
Urban unmarked	98	100
Urban marked	95	96
Urban multiple marked lanes	96	94
Total	**289**	**290**

KITTI provides an online evaluation leaderboard for the researchers using the dataset. All results are publicly available on the official website [19]. Researchers submit their results, as bird-eye-view, to the website, and all the performance metrics are calculated after the evaluation by the official team. To keep the evaluation fair, ground truth images of the test dataset are not published to the public. Therefore, in this thesis, only the training dataset is used by applying various data augmentation operations, which are explained in detail in Section 3.3. The dataset also contains projection and transformation matrices for transition between sensor coordinates, allowing the processing camera and LiDAR data in the same coordinate space. The processing steps of the LiDAR data is given in detail in Section 3.2. An RGB camera image and its ground truth from the training dataset are shown in Figure 3.2.

Figure 3.2: RGB image and ground truth sample from the dataset.

3.2. Data Processing

In order to realise the fusion, two sensor data need to be in the same format. In the dataset, RGB camera images are in 2D shape as $(height, width)$, and the pixel

values represent RGB colours as $[R,G,B]$. On the other side, 3D LiDAR data is a list of arrays where each array represents the reflected laser point parameters. A laser point is represented in an array as in Equation 3.1.

$$P = [x,y,z,i] \tag{3.1}$$

P represents the LiDAR point, x and y represent the coordinates of the point in the X and Y axes, respectively, with respect to the sensor, z represents the altitude of the point with respect to the sensor, and i represents the intensity of the reflected laser pulse. All reflected points together construct a point cloud in each sensor cycle, which is shown in Equation 3.2. In each cycle, a point cloud contains tens of thousands of points.

$$Point\ cloud = [P_1, P_2, ..., P_n] \tag{3.2}$$

n represents the number of points in the point cloud. It can be seen that the data of the camera and LiDAR are in different shapes. Moreover, due to the placement of the sensors on the vehicle, their data are in different coordinate spaces. To overcome this problem, calibration files provided with the dataset are used. The calibration files contain the required projection and transformation matrices to realise the transition between sensor coordinates. Using these matrices, all LiDAR points are reflected into the camera plane using Equation 3.3.

$$x_i = PR_{rect} * RO_{rect} * TR^{-1}_{lidar2cam} * P_i \tag{3.3}$$

PR_{rect} represents (3×4) projection matrix for the left colour camera in rectified coordinates, RO_{rect} represents (3×3) rotation matrix for non-rectified to rectified camera coordinates, $TR_{lidar2cam}$ represents (3×4) (non-rectified) rigid transformation

matrix for LiDAR to camera coordinates, P_i represents the LiDAR point and x_i represents the pixel. RO_{rect} and $TR_{cam2road}$ are extended to (4×4) matrices by adding a fourth row and a fourth column for RO_{rect} only.

3.2.1. Altitude Image

Once all LiDAR points in the point cloud are reflected into the camera plane, 2D grayscale LiDAR image is constructed, where each pixel value represents the normalised altitude value of the LiDAR point. Two examples are shown in Figure 3.3, that LiDAR points are reflected on the RGB image (LiDAR pixels were coloured in red to yellow scale for better visualisation).

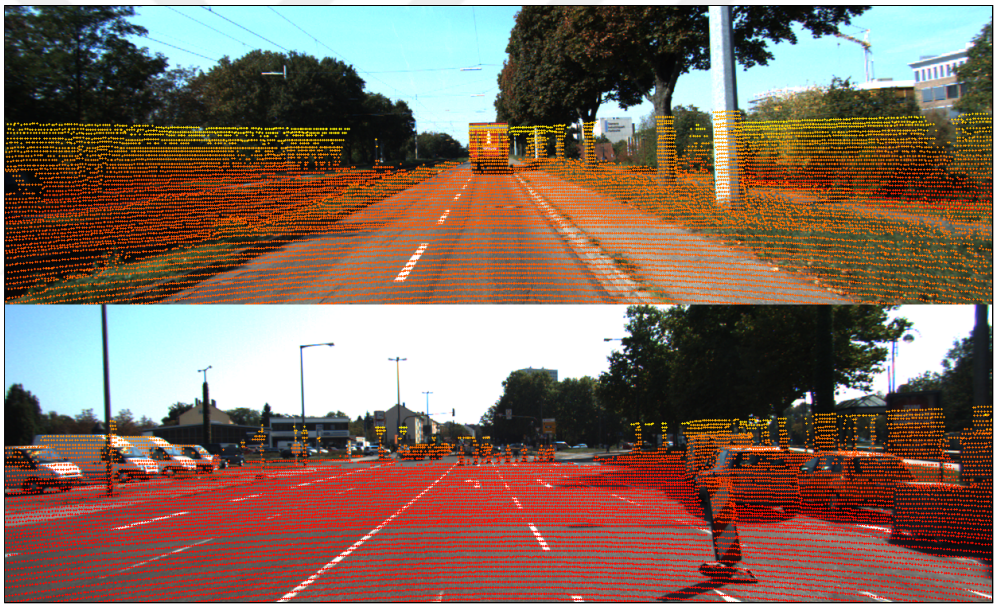

Figure 3.3: Two examples of point cloud reflected on the RGB image.

Even though LiDAR data is in the same shape and coordinate space after the projection, there are empty pixels on the LiDAR image since the LiDAR data is sparse. Consequently, obtained LiDAR image is not suitable for image segmentation. To solve this, linear interpolation is applied to the LiDAR image using the Python SciPy library [25]. After the interpolation, a grayscale image is obtained from LiDAR data, where each pixel corresponds to the altitude value of the point. This image is referred to as the ALT image throughout the thesis. An example of an ALT image is shown in Figure 3.4.

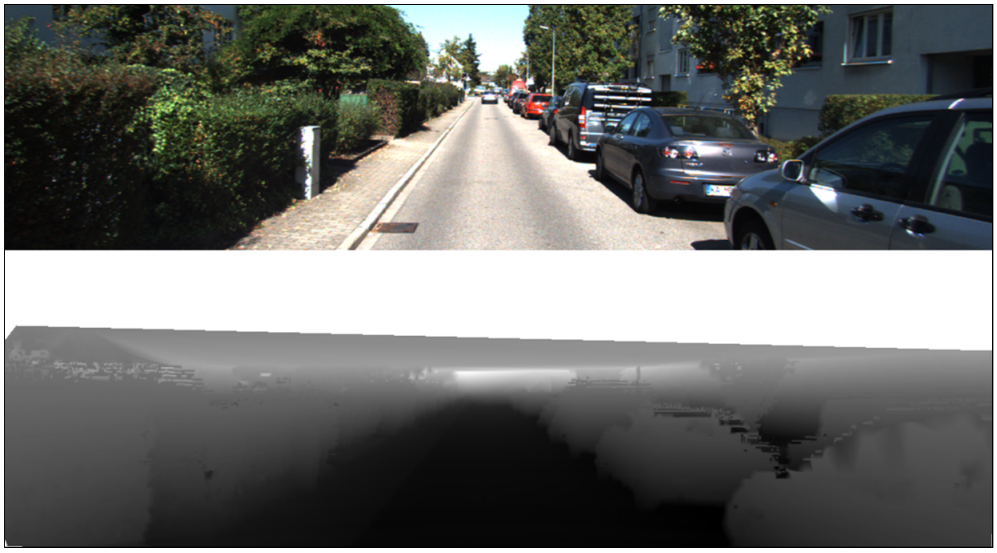

Figure 3.4: Comparison of an RGB image and the corresponding ALT image.

3.2.2. Altitude Difference Image

In Progressive LiDAR Adaptation for Road Detection [20], Chen et al. proposed the Data Space Adaptation for obtaining Altitude Difference Image (ADI) from LiDAR data. In this approach, altitude differences are calculated and used to obtain LiDAR images that highlight flat parts on a scene by utilising the smoothness of the road surface in terms of altitude compared to the other components in the surroundings, such as cars and pavements. In order to accomplish this, new pixel values are obtained by calculating the altitude differences of each pixel in the ALT image with its neighbours in the designated frame. The performed operation is shown in Equation 3.4.

$$V_{x,y} = \frac{1}{M} \sum_{N_x,N_y} \frac{\|Z_{x,y} - Z_{N_x,N_y}\|}{\sqrt{(N_x - x)^2 + (N_y - y)^2}} \tag{3.4}$$

$V_{(x,y)}$ represents the altitude difference value at location (x,y), $Z_{(x,y)}$ represents the altitude values at that location; N_x and N_y represents the locations of the neighbouring pixels; and M indicates the total number of neighbouring pixels. An example of an ADI image obtained after the operation is shown in Figure 3.5.

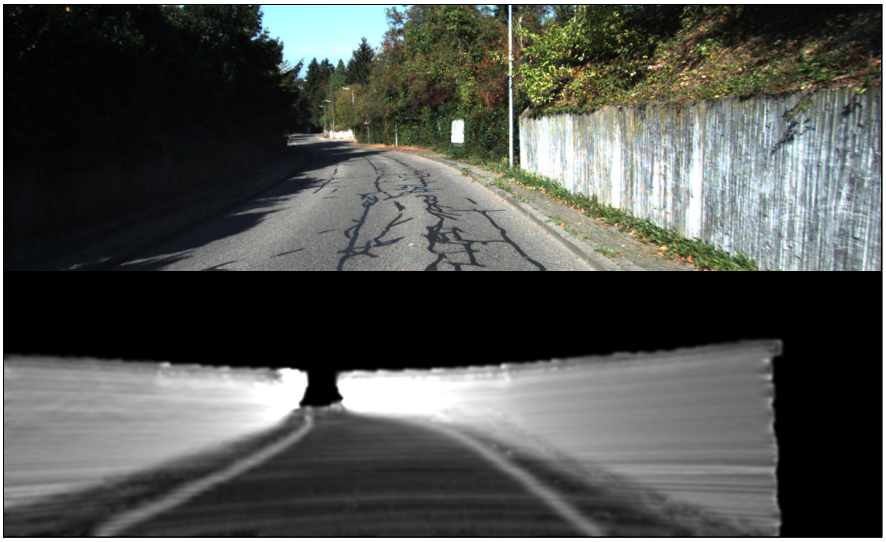

Figure 3.5: Comparison of an RGB image and the corresponding ADI image.

As shown in Figure 3.5, the ADI image is doing well in terms of highlighting both flat and un-levelled planes. Also, it is seen that the left part is in shadows in the RGB image, and it is difficult to detect the road boundaries, whereas it is evident in the ADI image. This emphasises the advantage of the LiDAR sensor.

In the ADI approach, while calculating altitude differences, the altitude values are lost. Nevertheless, one can expect the road to be the lowest plane on a scene; thus, altitude is a valuable feature. Therefore, both ALT and ADI images are used in this thesis for the fusion, and obtained results are compared in Section 4.

3.3. Data Augmentation

The KITTI road benchmark test dataset does not contain ground truth images, as explained in Section 3.1. The training dataset only consists of 289 images, which is insufficient to train a deep model. Therefore, data augmentation techniques were applied to the training images to artifically increase the amount of data. Python Albumentations library was used for realising the augmentation [26].

First, 89 images from 289 training images were separated as test data. Then remaining 200 images were mirrored horizontally to double the dataset size. Next, the elastic transform, a sophisticated version of the affine transform based on [27],

was applied to the images with the OpenCV border replicate method as the pixel extrapolation method. Images generated by the elastic transform augmentation look like they are taken from different angles of the scene in the original image. Finally, the brightness and contrast of the generated images were randomly changed to vary the lighting conditions of the dataset. As a result, all applied augmentations increased the training dataset by a factor of 10, leading to 2000 images in total. The new dataset was separated into 1400 training and 600 validation images. Since test images were separated before the augmentations, no augmented versions of the training or validation images exist in the test dataset. Figure 3.6 shows some artificially produced images.

Figure 3.6: Artificially produced images using data augmentation techniques.

3.4. Base Segmentation Model

In recent years there have been many studies on image segmentation, mainly built on top of FCN architecture [28]. Chen et al. proposed DeepLab [23], a CNN-based segmentation model. The DeepLab model introduced three novel approaches for semantic segmentation. First, they introduced Atrous convolution, or Dilated convolution, which adds gaps to the convolutional filter (see Figure 3.7 for illustration [29]).

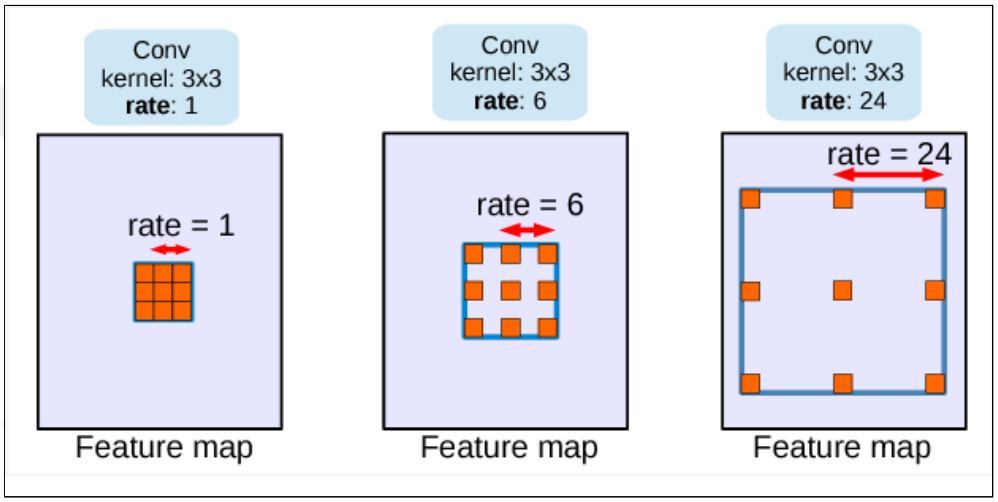

Figure 3.7: DeepLab Atrous (Dilated) convolution.

This approach helps to reduce the parameter space and enlarge the filter view to obtain more features. Second, they introduced atrous spatial pyramid pooling (ASPP). This approach improves the model to detect classes that appear in different scales in the images by applying multiple sampling rates to the convolutional layers. Third, they used CRF to capture fine details and reinforce the edges. DeepLab advanced the results on the popular datasets in the field, such as CityScapes [30]. They also developed augmented versions of the model [29]. Zhao et al. proposed PSPNet [31], a semantic segmentation model for scene parsing. PSPNet exploits the global context information using the pyramid pooling approach. This approach pools input at different sizes before passing through convolutional layers. Then, they are upsampled and concatenated with the original input, which fuses with the local and global contexts of the input. PSPNet

also advanced the results on the popular segmentation datasets, including Cityscapes. Wu et al. proposed FastFCN [32], which resolves the speed bottleneck caused by dilated convolutions of the DeepLab model by proposing the Joint Pyramid Upsampling (JPU) approach. In JPU, the last three feature maps are taken as inputs and create a fused feature map that contains rich contextual information at different scales (see Figure 3.8). The FastFCN model achieved complexity reduction by more than three times comparing the DeepLab without performance loss.

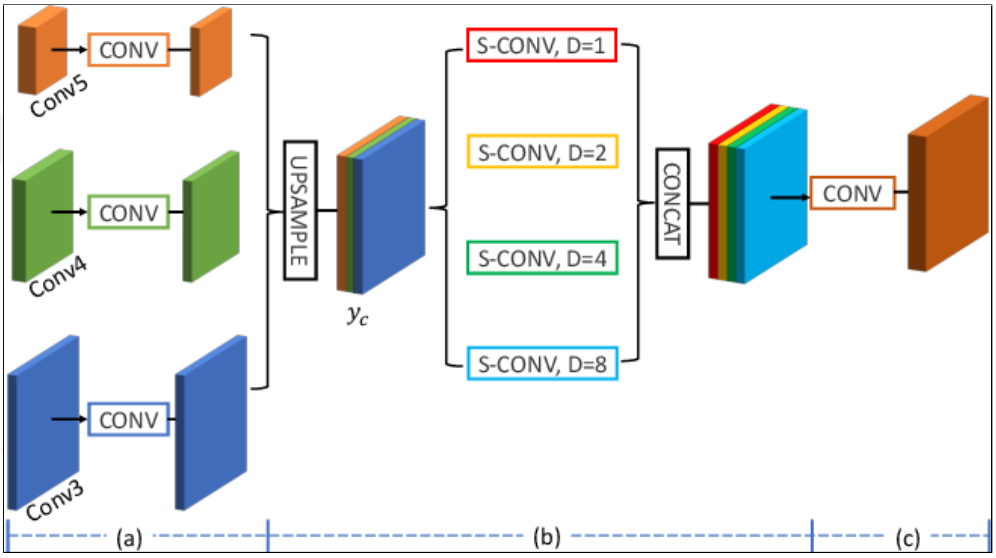

Figure 3.8: FastFCN JPU design.

While all these models can be modified and used for road detection, the U-Net model is preferred as the base model in this thesis. U-Net was proposed by Ronneberger et al. in 2015 to find tumours in the lungs and brain medical images [33]. U-Net is a U-shaped encoder-decoder architecture based on FCN. The architecture is shown in Figure 3.9.

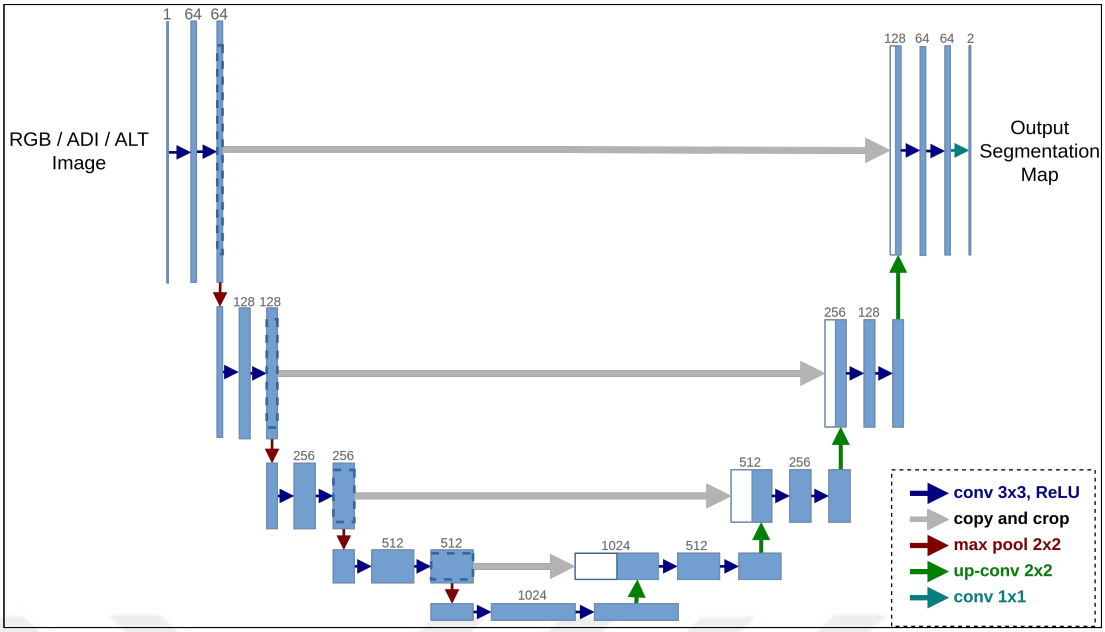

Figure 3.9: Network architecture of the U-Net model.

The U-Net architecture composes of five components: contracting path, bottleneck, skip connections, expansive path and final layer. The contracting path (left part of the architecture) consists of a sequence of four encoder blocks. Each encoder block consists of two (3×3) convolutions, each followed by a rectified linear unit (ReLU) activation function and a (2×2) max pooling operation with stride 2 for downsampling. At each downsampling step, the spatial dimensions (height and width) of the feature maps are reduced by half, and the number of feature channels is doubled. The output of the contracting path is given to the bottleneck layer (bottom part of the architecture), which connects the contracting path and the expansive path. It consists of two (3×3) convolutions, each followed by a ReLU activation function. The spatial dimensions of the feature maps are reduced during the contracting path, which can cause losing high-resolution features. To mitigate this issue, U-Net uses skip connections. Skip connections convey the feature maps from the encoder block to the corresponding decoder block. This allows the network to fuse low-level and high-level features, which can improve the localisation accuracy of the segmentation. The expansive path (right part of the architecture) starts with a (2×2) transposed convolution. Then the output is concatenated with the feature map carried by the corresponding skip connection. The

decoder block is finalised with two (3×3) convolutions, each followed by a rectified linear unit (ReLU) activation function. A (1×1) convolution with sigmoid activation is used to obtain an output segmentation map representing the pixel-wise classification at the final layer.

The U-Net model is selected as the base model to implement the proposed fusion models in this thesis because it provides multiple advantages in this manner. First, U-Net's flexible architecture can easily adapt to handle different input types. Second, the skip connections of the U-Net are great tools to realise cross fusion since they allow the network to combine features at multiple scales. Third, the model is designed to be able to learn even with small datasets, which is also the case when using the KITTI road detection dataset. Finally, the model has already achieved state-of-the-art performance on various image segmentation tasks.

3.5. Fusion Models

This thesis proposes three U-Net-based models for road detection using RGB camera and LiDAR fusion: early fusion, late fusion and cross fusion. All the models were implemented and evaluated for comparison. Their results are shown in Section 4.

3.5.1. Early Fusion

In the early fusion model, the RGB camera and LiDAR images are fused before feeding into the model. To do this, images are concatenated in the dept dimension, and the obtained image is given as input to the model. The number of channels of the fused image is the sum of the number of channels of the images used for the fusion. For example, if the RGB image and ADI image are fused, the obtained image has four channels as $[R,G,B,z]$. This model aims to extract features from both images jointly from the very beginning. This approach can improve the results by utilising complementary information provided by the images throughout all layers. The early fusion network architecture is shown in Figure 3.10.

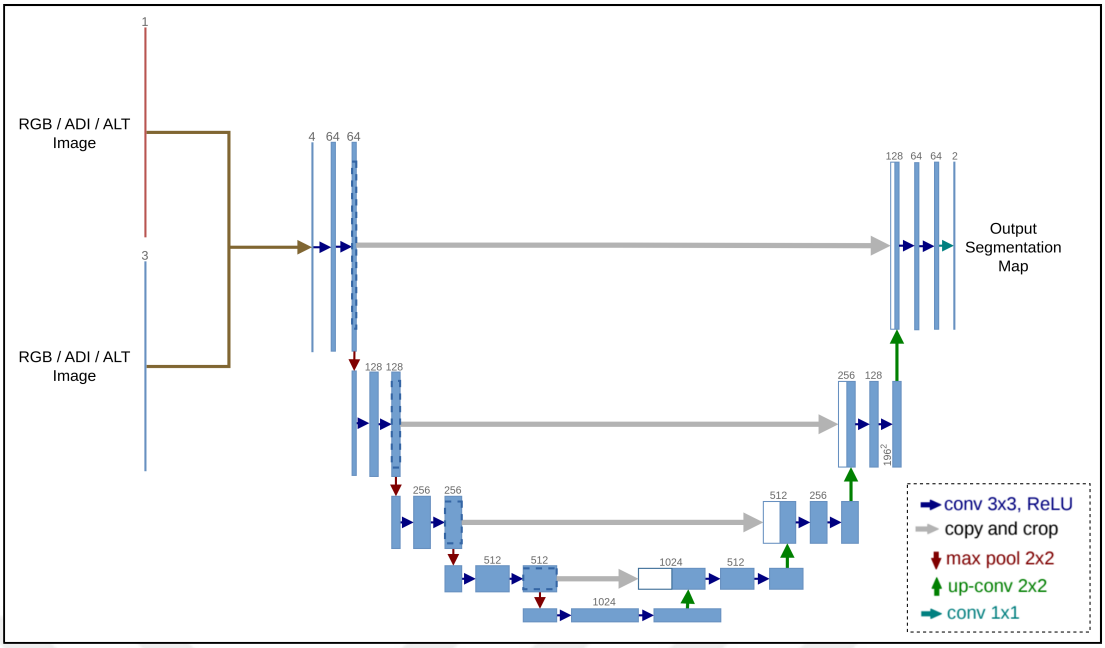

Figure 3.10: Network architecture of the early fusion model.

3.5.2. Late Fusion

In the late fusion model, separate models are trained in parallel for each image type until the final layer. When the models reach the final layer, the feature maps of the models are concatenated in the depth dimension. Then, the output segmentation map is obtained by passing through the final layer. The late fusion approach aims to allow models to specialise in unique features of different types of images. This approach is expected to be more efficient, especially if the features of the images to be fused are highly distinct. In addition, training the models is computationally more efficient as it can be parallelised. The late fusion network architecture is shown in Figure 3.11.

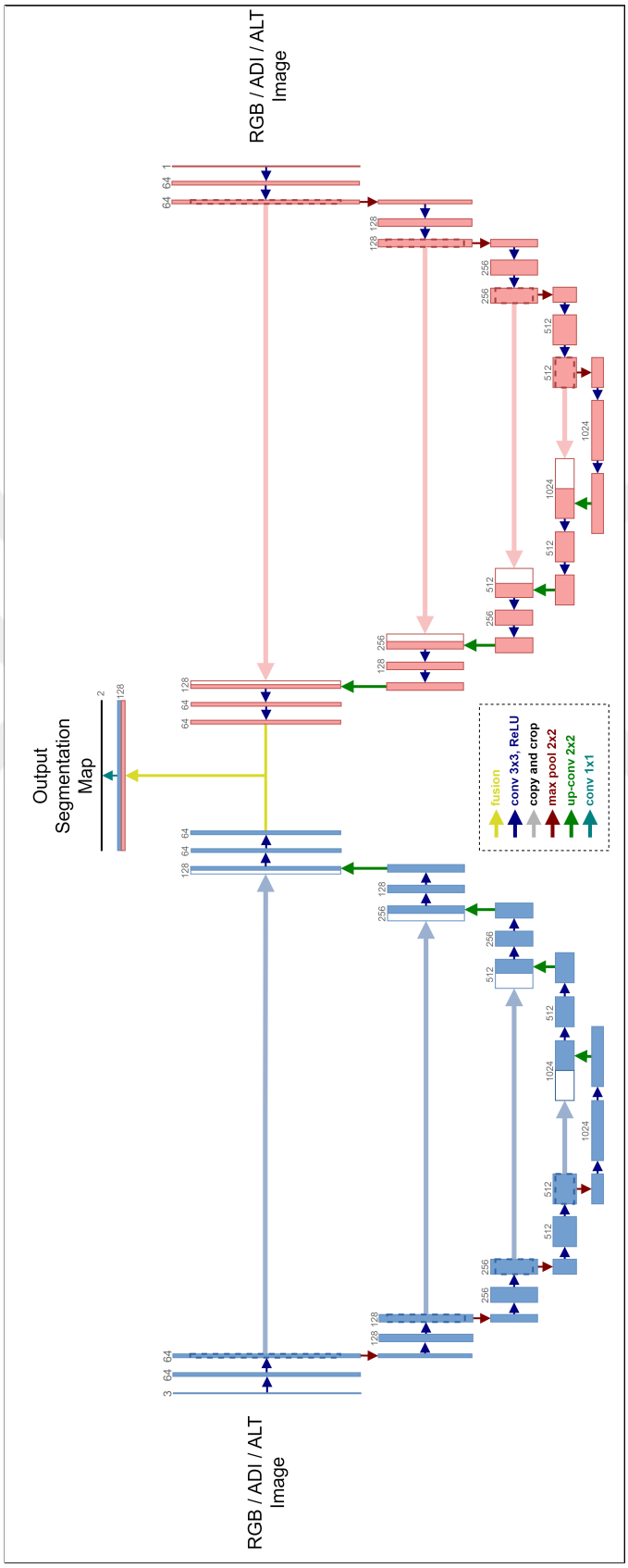

Figure 3.11: Network architecture of the late fusion model.

3.5.3. Cross Fusion

In the cross fusion model, separate models are trained in parallel for each image type to be fused, similar to the late fusion model. However, this time fusion is realised in more than just the final layer; instead, it is realised at the beginning of every decoder block by utilising skip connections. First, two models are processed until the end of the bottleneck part, using the two image types. Then, at the beginning of each decoder block, skip connections convey the feature map from the encoder of the parallel model instead of the same model. That is, the feature map on the decoder block of the RGB image model is concatenated with the feature map from the skip connections of the LiDAR image model. In contrast, the feature map on the decoder block of the LiDAR image model is concatenated with the feature map from the skip connections of the RGB image model. Thus, cross fusion is achieved. Finally, similar to the late fusion model, the outputs of the two models are concatenated before the final layer and passed through the final layer to obtain an output segmentation map. The cross fusion approach aims to capture complex and non-linear relationships between the image types that cannot be captured by simple concatenation, such as in early fusion and late fusion. However, it is computationally expensive since such parallelism between the models is tough to implement. The cross fusion network architecture is shown in Figure 3.12.

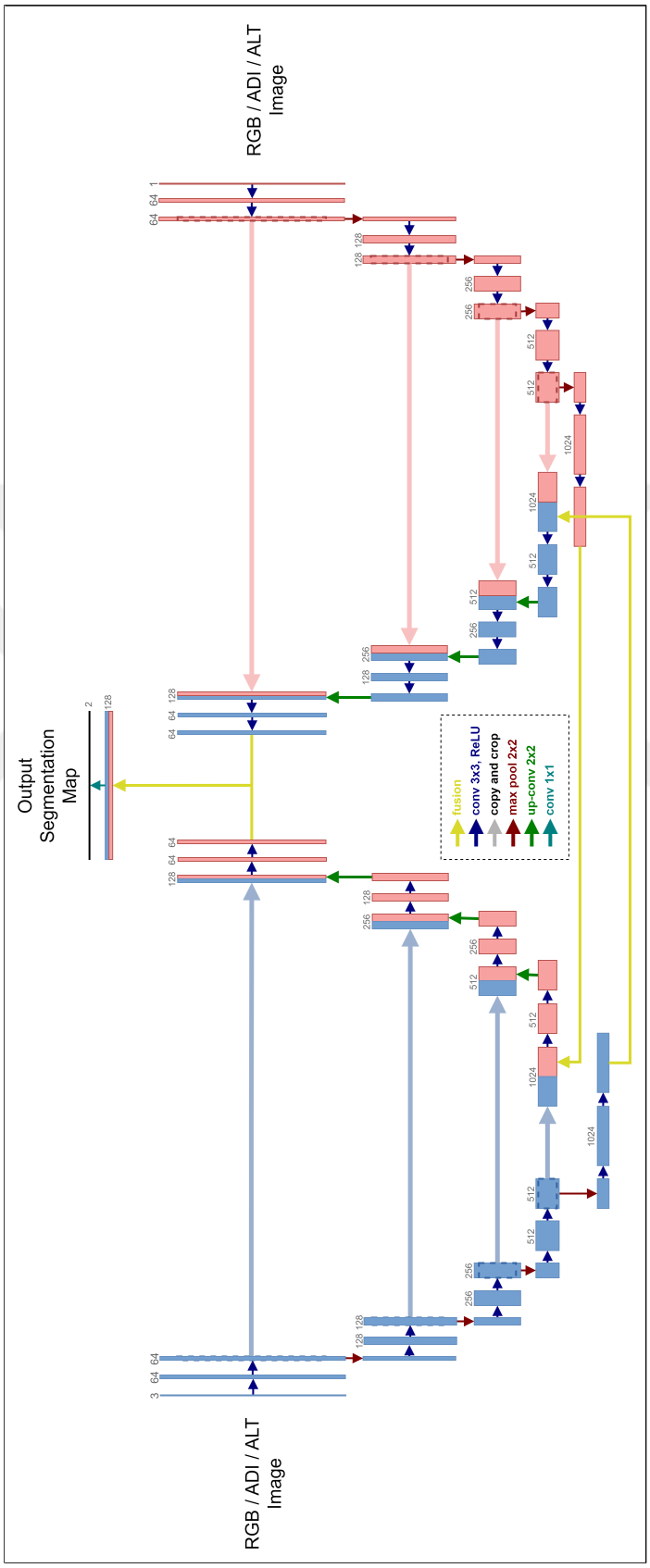

Figure 3.12: Network architecture of the cross fusion model.

4. RESULTS

In this thesis, three different sensor fusion techniques (early fusion, late fusion, and cross fusion see Section 3.5) were used for road detection using three different types of images (RGB, ADI, and ALT, see Section 3.2) collected from two sensors, RGB camera, and 3D LiDAR. In order to make better comparisons, an original U-Net model was trained for each image type. Then, to show and compare the contribution of the fusion results, all combinations of the image types with the fusion models were trained.

All hyper-parameters, except the input layer and the dataset, were fixed during the experiments. Adam Optimization Algorithm [34] was used with 10^{-4} initial learning rate and the batch size 4. Images were resized to (125×414) during training and tests. The models were trained for 60 epochs on NVIDIA GeForce RTX 2060. For evaluation, the leave-p-out cross-validation method was used. The training dataset was randomly divided into 5 folds and training was carried out by excluding one fold in each iteration, with a total of 5 iterations for each model. The results of the models were then obtained by taking the average of the MaxF scores of each iteration of the model on the test dataset.

All results were compared in terms of coefficient of variation (CV) between each iteration of the model, precision, recall, and MaxF scores [24]. The calculation formula for the coefficient of variation is given in Equation 4.1.

$$CV = \frac{s}{\overline{x}} * 100 \tag{4.1}$$

s represents the standard deviation and $\overline{x}$ represents the mean of the iterations. The calculation formulas for the F1 score and the MaxF score are given in Equation 4.2 and Equation 4.3.

$$F1\,score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4.2}$$

$$MaxF\,score = \arg\max_{\tau} F1\,score \tag{4.3}$$

τ represents the classification threshold chosen to maximise the F1 score. The results obtained from all models are shown in Table 4.1 as percentages based on the coefficient of variation, precision, recall, and MaxF scores. The best models for each category are highlighted in bold.

Table 4.1: Results of all proposed models on the test dataset. The best models for each category are highlighted in bold.

Fusion Method	Model	CV (%)	PRE (%)	REC (%)	MaxF (%)
Single source	RGB	0.20	96.34	96.74	96.50
	ADI	0.30	96.23	96.54	**96.51**
	ALT	0.43	95.41	95.48	95.33
Early fusion	RGB + ADI	0.09	96.87	97.20	**97.02**
	RGB + ALT	0.26	96.27	96.66	96.41
	ADI + ALT	0.12	97.00	96.88	96.93
	RGB + ADI + ALT	0.31	96.88	96.62	96.72
Late fusion	RGB + ADI	0.06	97.03	96.69	**96.83**
	RGB + ALT	0.30	96.77	96.82	96.77
	ADI + ALT	0.25	96.85	96.48	96.64
	RGB + ADI + ALT	0.21	97.13	96.57	96.82
Cross fusion	RGB + ADI	0.05	97.22	96.84	**96.99**
	RGB + ALT	1.12	95.96	96.95	96.38
	ADI + ALT	0.09	96.51	96.61	96.52

4.1. Single Source

In the single source models, the model trained with RGB images achieved a MaxF score of 96.50%, while the model trained with ADI images achieved 96.51%. Although the two results are similar, when the CV values are examined, it can be deduced that the model trained using RGB image is more robust since it has a lower CV. This shows that RGB cameras can produce better results in road detection than LiDAR. The other single source LiDAR-based model, the ALT model, achieved a 95.33% MaxF score. Looking at the results, it is seen that the ALT model lags behind other models. When the ALT images were examined, it was seen that some of them were distorted because some components in the scene have lower altitude values than the road. For example, while the vehicle crosses a bridge, LiDAR rays reflected from the area under the bridge have much lower altitude values than the road itself. These unexpected altitude differences distort the images and are thought to lead to a poorer model for segmentation. An example of the altitude difference distortions is shown in Figure 4.1.

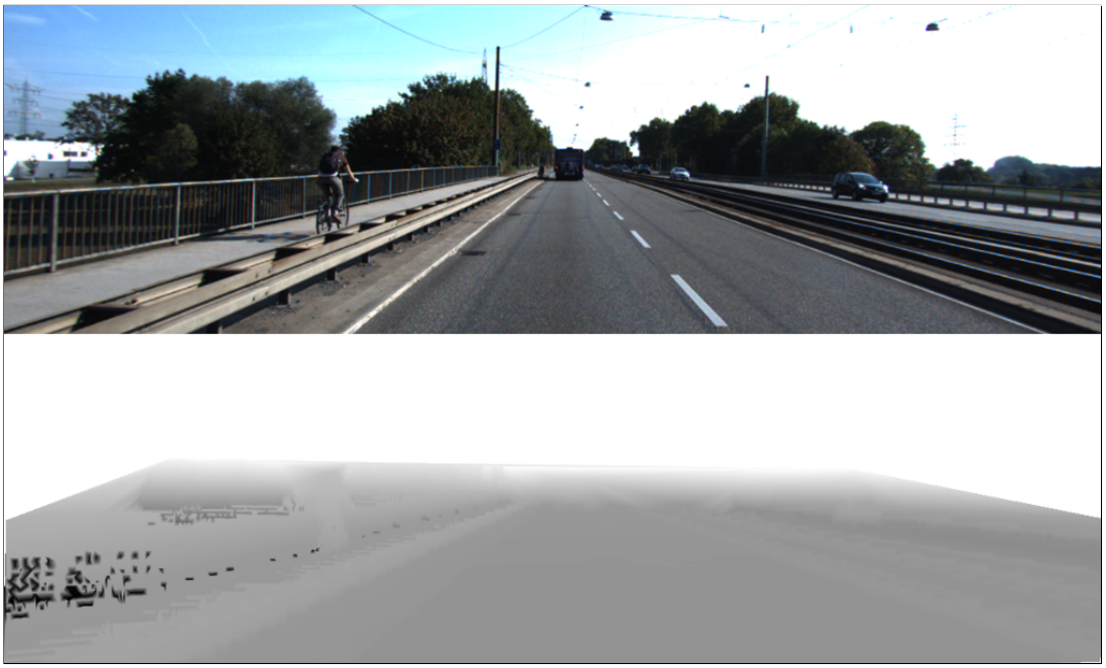

Figure 4.1: Outlier ALT image. Top: RGB image. Bottom: ALT image. There are black pixels on the left of the ALT image due to the reflected lasers from the terrain under the bridge.

4.2. Early Fusion

The models using the early fusion approach show that the ADI + ALT model is more successful with a MaxF score of 96.93% than the single source ADI model and single source ALT model. This shows that the ALT image contains features not in the ADI image and contributes to the fusion. This inference is also confirmed by the other fusion methods. The highest MaxF score, both in the early fusion method and among all models, is the RGB + ADI model with a MaxF score of 97.02%. The RGB + ADI + ALT model, combined with three images, achieved a MaxF score of 96.72% and fell behind the ADI + ALT model. The RGB + ALT model, on the other hand, achieved a MaxF score of 96.41%, which is lower than the single source RGB model.

4.3. Late Fusion

In the late fusion approach, the RGB + ADI achieved a MaxF score of 96.93% and became the best model for this approach, while the RGB + ADI + ALT model also achieved very close success with a MaxF score of 96.82% but with a higher CV. The RGB + ALT model achieved 96.77%, and the ADI + ALT model that is only trained with LiDAR-based images achieved 96.64% MaxF scores.

4.4. Cross Fusion

The most successful model in the cross fusion approach is the RGB + ADI model with a MaxF score of 96.99%, very close to the highest MaxF score of 97.02% of the early fusion RGB + ADI model. This model also achieved the lowest CV value among all the models. While the ADI + ALT model achieved a MaxF score of 96.52%, the RGB + ALT model became the most unsuccessful model in this approach, with 96.38%. The RGB + ADI + ALT model could not be trained with the cross fusion approach because only two images can be fused simultaneously due to the architecture.

4.5. Overall

When the results are examined, it is seen that all fusion models are more successful than all single source models, except the RGB + ALT models. The most successful models overall were the models using RGB and ADI images. Furthermore, among the U-Net-based fusion models, the most successful fusion method overall was the early fusion method, while the most unsuccessful method was the late fusion method. The results of the RGB + ADI model trained with the early fusion model, which achieved the highest MaxF score, are shown in Figure 4.2. Although the model with the highest MaxF score in the study was achieved by the early fusion RGB + ADI model, the cross fusion RGB + ADI model achieved the lowest CV and very close MaxF score to the early fusion model, showing the success of the cross fusion method.

Figure 4.2: RGB + ADI early fusion model results. Green: True positive, Blue: False positive, Red: False negative.

In order to show the contribution of LiDAR in road detection distinctly, the results of the single source RGB model and the early fusion RGB + ADI model on a sample containing intense shadows are compared in Figure 4.3. Although there is a 0.52% difference between the results of the two models on the test dataset, this difference can be seen more clearly in images with intense shadows.

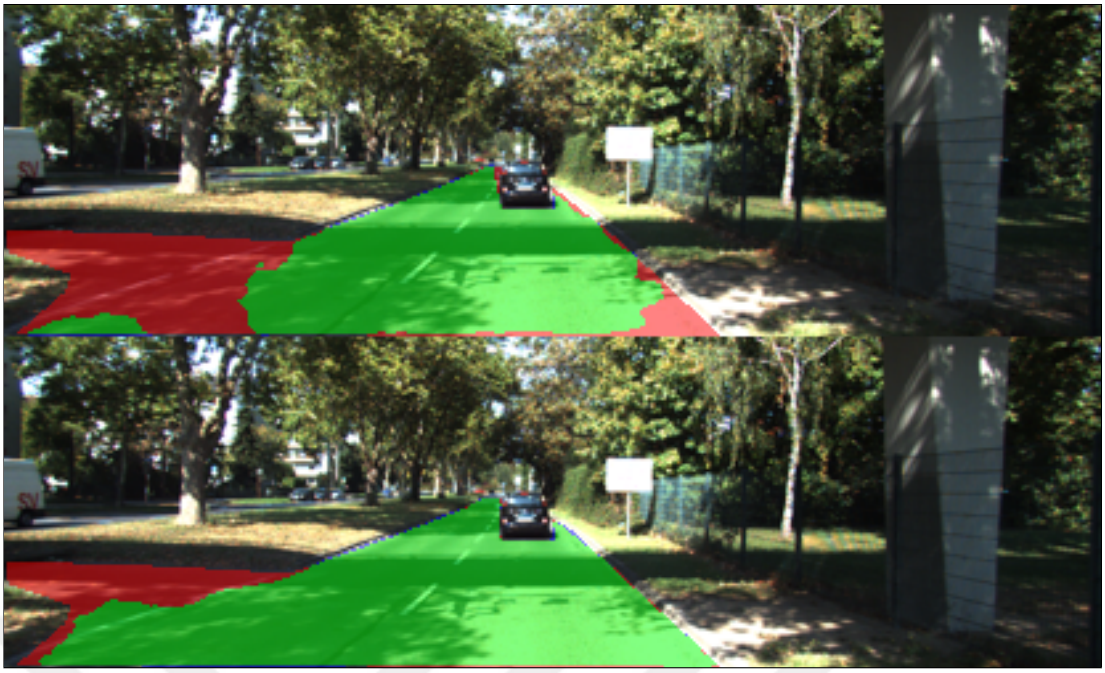

Figure 4.3: Comparison of single source RGB model and early fusion RGB + ADI models on an intense shadow environment. Top: Single source RGB model result, Bottom: Early fusion RGB + ADI model result. Green: True positive, Blue: False positive, Red: False negative.

The successful models for each fusion method are compared with the PLARD [20] model, which is the most successful model using camera and LiDAR fusion in the official KITTI leaderboard, described in Section 2.2. A trained PLARD model and implementation were used for this comparison [35]. All models were compared on the test dataset described in Section 3.1. The results are shown in Table 4.2. When the results are examined, it is seen that the trained models lagged behind the PLARD model. Improvements that can be made for achieving better results are discussed in Section 5. This comparison is mainly used as a reference to compare the implemented models with other models in the literature.

Table 4.2: Comparison of the proposed models with the PLARD.

Model	PRE (%)	REC (%)	MaxF (%)
PLARD [35]	95.72	98.89	97.28
Early fusion RGB + ADI	96.87	97.20	97.02
Late fusion RGB + ADI	97.03	96.69	96.83
Cross fusion RGB + ADI	97.22	96.84	96.99

The successful models implemented in this thesis are also compared with other successful road detection models in the literature that use the same dataset for training and fuse RGB camera and LiDAR (Table 4.3). Since new submissions are restricted due to KITTI's policies, the results of the implemented models are obtained on the dataset explained in Section 3.1, whereas literature results are obtained from the official KITTI road detection leaderboard. Table 4.2 can be used as a reference to compare the proposed models and literature models. When the results are examined, it is seen that the proposed models, especially the early fusion and the cross fusion RGB + ADI models, achieved competitive results in the literature.

Table 4.3: Comparison of the proposed models with the literature. Implemented models are evaluated on the test dataset, which is explained in Section 3.1, whereas literature results are obtained from the official KITTI road detection leaderboard.

Model	PRE (%)	REC (%)	MaxF (%)
PLARD [20]	97.19	96.88	97.03
CLCFNet [36]	96.38	96.39	96.38
LidCamNet [6]	96.23	95.83	96.03
LC-CRF [21]	93.62	97.83	95.68
Early fusion RGB + ADI	96.87	97.20	97.02
Late fusion RGB + ADI	97.03	96.69	96.83
Cross fusion RGB + ADI	97.22	96.84	96.99

5. CONCLUSION AND FUTURE WORK

In this thesis, multiple deep learning models based on the U-Net architecture were developed and compared for road segmentation using RGB camera and 3D LiDAR fusion. The KITTI road detection dataset was used for training and evaluating the models since it is a good benchmark popular in the literature and provides all the required information for the transformations between the sensor frames. The implementation steps of the models are explained as follows.

First, the LiDAR data was processed to be suitable for segmentation and fusion with the RGB images. To achieve this, the LiDAR data was projected into the camera plane using the transformation matrices provided by the dataset. Then, interpolation was applied to the LiDAR images to obtain segmentable images. The obtained images are called ALT and were used in the trained models. The ADI image was also constructed as an alternative and complementary to the ALT image. The ADI image was obtained by calculating the weighted average of the altitude differences of the pixels in the ALT image with their neighbours in a designated frame. Then, data augmentation was applied to enlarge the dataset. The training and validation dataset sizes were increased by a factor of 10 by applying mirror transform, elastic transform, brightness and contrast. Next, the U-Net segmentation model was selected as the base model of this thesis since the U-Net model provides a flexible architecture to implement fusion approaches and also achieves good results even with a small dataset. Finally, three fusion approaches were implemented and evaluated: early fusion, late fusion, and cross fusion.

In the early fusion approach, the camera and LiDAR images were fused before being given to the model to extract features from both images. The original U-Net architecture was directly used for the approach. In the late fusion approach, the U-Net architecture was mirrored to train separately for camera and LiDAR images. The approach aimed to allow the model to specialise in the unique features of different types of images. Finally, in the cross fusion approach, the original architecture was mirrored similarly to the late fusion except for the camera and LiDAR fusion realised by

utilising the skip connections. The approach aimed to capture complex and non-linear relationships between the image types that simple concatenation cannot capture.

According to the results, the early fusion approach achieved the highest MaxF score, whereas cross fusion achieved a similar MaxF score with a lower coefficient of variation. In terms of the image types used, the models fusing the RGB and ADI images were more successful. Also, it is important that all models trained using the RGB camera and LiDAR fusion outperformed those trained using only one of them. When the results were compared with the state-of-the-art models, the early fusion and cross fusion models were at a competitive level.

Several areas of future research can be identified based on the results and analysis presented in this thesis. First, a large dataset is required for training. This issue can also be solved by using methods such as transfer learning. Another issue may be using more efficient methods (e.g., improving the ALT image method by dealing with outliers) when converting LiDAR data into segmentable images. Last but not least issue open to improvement is the development of more efficient fusion approaches and architectures. Modules such as the Feature Space Adaptation presented in the PLARD [20] can be implemented to achieve better results.

In conclusion, this thesis shows that camera and LiDAR fusion gave better results than the single source models. Then, it shows that different fusion methods severely affect the results, and therefore the fusion approach is critical. Finally, a novel road segmentation model using an RGB camera and LiDAR fusion was developed with the early fusion approach, and state-of-the-art results were obtained.

REFERENCES

[1] Deichmann J., Ebel E., Heineke K., Heuss R., Kellner M., Steiner F., (2023), "Autonomous driving's future: Convenient and connected", McKinsey Center for Future Mobility.

[2] Web 1, (2021), https://www.sae.org/blog/sae-j3016-update, (Date of Access: 09/05/2023).

[3] F. Reway, W. Huber, E. P. Ribeiro, (2018), "Test Methodology for Vision-Based ADAS Algorithms with an Automotive Camera-in-the-Loop", 2018 IEEE International Conference on Vehicular Electronics and Safety (ICVES), 1-7, Madrid, Spain, 12-14 September.

[4] Zhang Y., Carballo A., Yang H., Takeda K., (2023),"Perception and sensing for autonomous vehicles under adverse weather conditions: A survey", ISPRS Journal of Photogrammetry and Remote Sensing, 196, 146–177.

[5] Web 2, (2018), https://www.cnbc.com/2018/05/02/backup-cameras-now-required-in-new-cars-in-the-us.html, (Date of Access: 09/05/2023).

[6] Caltagirone L., Bellone M., Svensson L., Wahde M., (2019), "Lidar–camera fusion for road detection using fully convolutional neural networks", Robotics and Autonomous Systems, 111, 125–131.

[7] Web 3, (2020), https://velodynelidar.com/press-release/velodyne-lidar-announces-three-year-sales-agreement-with-baidu, (Date of Access: 09/05/2023).

[8] Zywanowski K., Banaszczyk A., Nowicki M. R., (2020), "Comparison of camera-based and 3D LiDAR-based place recognition across weather conditions", 2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV), 886–891, Shenzhen, China, 13-15 December.

[9] Hess W., Kohler D., Rapp H., Andor D., (2016), "Real-time loop closure in 2D LIDAR SLAM," 2016 IEEE International Conference on Robotics and Automation (ICRA), 1271-1278, Stockholm, Sweden, 16-21 May.

[10] Stefano F. D., Chiappini S., Gorreja A., Balestra M., Pierdicca R., (2021), "Mobile 3D scan LiDAR: a literature review", Geomatics, Natural Hazards and

Risk, 12 (1), 2387–2429.

[11] Assidiq A., Khalifa O. O., Islam M. R., Khan S., (2008), “Real time lane detection for autonomous vehicles”, 2008 International Conference on Computer and Communication Engineering, 82–88, Kuala Lumpur, Malaysia, 13-15 May.

[12] Bar Hillel A., Lerner R., Levi D., Raz G., (2014), “Recent progress in road and lane detection: A survey”, Machine Vision and Applications, 25, 727–745.

[13] Muñoz-Bulnes J., Fernandez C., Parra I., Fernández-Llorca D., Sotelo M. A., (2017), ”Deep fully convolutional networks with random data augmentation for enhanced generalization in road detection”, 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), 366-371, Yokohama, Japan, 16-19 October.

[14] Wang H., Fan R., Cai P., Liu M., (2021), ”SNE-RoadSeg+: Rethinking Depth-Normal Translation and Deep Supervision for Freespace Detection”, 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 1140-1145, Prague, Czech Republic, 27 September - 01 October.

[15] Caltagirone L., Scheidegger S., Svensson L., Wahde M., (2017), ”Fast LIDAR-based road detection using fully convolutional neural networks”, 2017 IEEE Intelligent Vehicles Symposium (IV), 1019-1024, Los Angeles, CA, USA, 11-14 June.

[16] Lyu Y., Bai L., Huang X., (2019), ”ChipNet: Real-Time LiDAR Processing for Drivable Region Segmentation on an FPGA”, IEEE Transactions on Circuits and Systems I: Regular Papers, 66 (5), 1769-1779.

[17] Khaleghi B., Khamis A., Karray F. O., Razavi S. N., (2013), ”Multisensor data fusion: A review of the state-of-the-art”, Information Fusion, 14 (1), 28-44.

[18] He K., Zhang X., Ren S., Sun J., (2016), ”Deep Residual Learning for Image Recognition”, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778, Las Vegas, NV, USA, 27-30 June.

[19] Web 4, (2013), https://www.cvlibs.net/datasets/kitti/eval_road.php, (Date of Access: 09/05/2023).

[20] Chen Z., Zhang J., Tao D., (2019), ”Progressive LiDAR adaptation for road detection”, IEEE/CAA Journal of Automatica Sinica, 6 (3), 693-702.

[21] Gu S., Zhang Y., Tang J., Yang J., Kong H., (2019), ”Road Detection through CRF based LiDAR-Camera Fusion”, 2019 International Conference on Robotics and Automation (ICRA), 3832-3838, Montreal, QC, Canada, 20-24 May.

[22] Lee D. T., Schachter B. J., (1980), ”Two algorithms for constructing a Delaunay triangulation”, International Journal of Computer and Information Sciences, 9 (3), 219-242.

[23] Chen L. -C., Papandreou G., Kokkinos I., Murphy K., Yuille A. L., (2018), ”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.

[24] Fritsch J., Kuehnl T., Geiger A., (2013), ”A new performance measure and evaluation benchmark for road detection algorithms”, 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013), 1693-1700, The Hague, Netherlands, 06-09 October.

[25] Virtanen P., Gommers R., Oliphant T. E., Haberland M., Reddy T., Cournapeau D., Burovski E., Peterson P., Weckesser W., Bright J., van der Walt S. J., Brett M., Wilson J., Millman K. J., Mayorov N., Nelson A. R. J., Jones E., Kern R., Larson E., Carey C. J., Polat ˙I., Feng Y., Moore E. W., VanderPlas J., Laxalde D., Perktold J., Cimrman R., Henriksen I., Quintero E. A., Harris C. R., Archibald A. M., Ribeiro A. H., Pedregosa F., van Mulbregt P., SciPy 1.0 Contributors, (2020), ”SciPy 1.0: fundamental algorithms for scientific computing in Python”, Nature Methods, 17 (3), 261-272.

[26] Buslaev A., Iglovikov V. I., Khvedchenya E., Parinov A., Druzhinin M., Kalinin A. A., (2020), “Albumentations: Fast and Flexible Image Augmentations”, Information, 11 (2), 125.

[27] Simard P., Steinkraus D., Platt J., (2003), ”Best practices for convolutional neural networks applied to visual document analysis”, Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings, 958-963, Edinburgh, UK, 06 August.

[28] Long J., Shelhamer E., Darrell T., (2015), ”Fully Convolutional Networks for Semantic Segmentation”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3431-3440, Boston, MA, USA, 07-12 June.

[29] Chen L., Papandreou G., Schroff F., Adam H., (2017), ”Rethinking Atrous Convolution for Semantic Image Segmentation”, arXiv e-prints, arXiv:1706.05587.

[30] Cordts M., Omran M., Ramos S., Rehfeld T., Enzweiler M., Benenson R., Franke U., Roth S., Schiele B., (2016), ”The Cityscapes Dataset for Semantic Urban Scene Understanding”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3213-3223, Las Vegas, NV, USA, June 27-30.

[31] Zhao H., Shi J., Qi X., Wang X., Jia J., (2017), ”Pyramid Scene Parsing Network”, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6230-6239, Honolulu, HI, USA, 21-26 July.

[32] Wu H., Zhang J., Huang K., Liang K., Yizhou Y., (2019), ”FastFCN: Rethinking Dilated Convolution in the Backbone for Semantic Segmentation”, arXiv preprint, arXiv:1903.11816.

[33] Ronneberger O., Fischer P., Brox T., (2015), ”U-Net: Convolutional Networks for Biomedical Image Segmentation”, Medical Image Computing and Computer-Assisted Intervention (MICCAI), 234-41, Munich, Germany, 05-09 October.

[34] Kingma D. Ba J., (2014), ”Adam: A Method for Stochastic Optimization”, International Conference on Learning Representations, arXiv:1412.6980.

[35] Web 5, (2022), https://github.com/zhechen/PLARD, (Date of Access: 09/05/2023).

[36] Gu S., Yang J., Kong H., (2021), ”A Cascaded LiDAR-Camera Fusion Network for Road Detection”, 2021 IEEE International Conference on Robotics and Automation (ICRA), 13308-13314, Xi’an, China, 30 May - 05 June.

BIOGRAPHY

Arda Taha Candan is a graduate of the Department of Computer Engineering at Yıldız Technical University, Faculty of Electrical and Electronics. He has been working as a researcher at TÜBİTAK BİLGEM since 2019, where he specialises in robotics and autonomous technologies. During his master's, he worked on sensor fusion for autonomous vehicles. His current research interests focus on developing autonomous cooperative, connected and automated mobility applications.

APPENDICES

Appendix A: Publications within the Scope of the Thesis Study

Candan A. T., Kalkan H., (2022), ”RGB Camera and LiDAR Fusion for Road Detection”, 2022 Innovations in Intelligent Systems and Applications Conference (ASYU), 1-5, Antalya, Turkey, 07-09 September.

Candan A. T., Kalkan H., (2023), ”U-Net-based RGB and LiDAR image fusion for road segmentation”, Signal, Image and Video Processing (SIViP).