Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19175-y

Semantic road segmentation using encoder-decoder architectures

Burhanuddin Latsaheb [1] · **Sanjeev Sharma** [1] · **Sanskar Hasija** [1]

Received: 2 October 2022 / Revised: 7 February 2024 / Accepted: 2 April 2024

Abstract
Road detection is a fundamental task in autonomous driving, making accurate and efficient road area segmentation essential for the safe and precise navigation of autonomous vehicles. This paper proposes various models for road segmentation, employing an encoder-decoder architecture for fully automatic segmentation of road areas. As part of the encoder, this work explores different models, such as ResNet50V2, DenseNet121, DenseNet169, and DenseNet201, and utilizes them in one of the few dedicated methods for road area segmentation. Here, the dataset, derived from the Mapillary Vistas Dataset, has been meticulously pre-processed to convert it into a binary segmentation problem for road detection, comprising 8041 training images and 919 validation images with their respective masks. The models were trained on our dataset, achieving the highest Dice coefficient value of 99.61% on the training dataset and 93.85% on the validation dataset using the DenseNet169 encoder model. This research contributes to advancing the state-of-the-art in road segmentation for autonomous driving applications.

Keywords Road segmentation · Deep learning · Encoder-decoder architecture · DenseNet · ResNet

1 Introduction

Roads are vital to various aspects of life, including economic development, social benefits, and safety. The automated detection of roads from road images is a crucial goal, providing essential information for applications like autonomous driving [30]. The challenge of road detection is not merely about identifying drivable surfaces but also encompasses understand-

✉ Burhanuddin Latsaheb
burhan.latsaheb@gmail.com

Sanjeev Sharma
sanjeevsharma@iiitp.ac.in

Sanskar Hasija
sanskarhasija19@cse.iiitp.ac.in

[1] Indian Institute of Information Technology, Pune, India

Published online: 13 April 2024

ing various road conditions, patterns, and anomalies. It requires an intricate understanding of diverse features such as lane markings, surface texture, and road boundaries. Algorithms for path planning and motion control can significantly benefit from accurate drivable region or road area recognition. Several public datasets are available for road detection, including the KITTI dataset [14], Mapillary Vistas Dataset [28], CULane [42], and Citiscapes [11].

Traditional methods have often fallen short in adapting to the dynamic and complex nature of roads, especially under varying weather and lighting conditions. Deep learning models, while promising, still face challenges in generalization, scalability, and real-time application. These challenges highlight the need for innovative solutions that can effectively address the multifaceted nature of road detection, particularly in the context of autonomous driving.

Due to the success of deep learning, the computer vision field has made remarkable advances over the past few decades. New algorithms have emerged for tasks such as image segmentation [29], object detection [44], and instance segmentation.

While some papers focus on lane detection [6, 8, 26], others address road detection by classifying roads as snowy, dry, wet, etc. The limited availability of images for snowy and wet roads can lead to overfitting, as discussed in [9, 22]. Research has also focused on road damage detection [12, 26, 31]. However, several papers targeting road area segmentation primarily focus on satellite images and remote sensing [15, 25, 39].

In light of these challenges, this research aims to push the boundaries of current road detection technologies by developing a robust and efficient encoder-decoder-based model. By focusing on semantically segmenting road areas and classifying various road features, this approach seeks to create a more comprehensive understanding of the road environment. The creation of a new dataset tailored to the specific needs of this study further adds to the novelty of our work. This allows for a more targeted evaluation, bridging the gap between existing datasets and real-world applications. The contributions of this work are expected to have significant implications for autonomous driving systems, providing a foundation for safer and more intelligent navigation.

This work creates a custom two-class dataset that is meticulously curated to include various road scenarios under different weather conditions and lighting situations. The dataset serves as a test bed to evaluate the efficacy and efficiency of each of the five architectures. Our primary objective is to identify the architecture that offers the best balance between accuracy and computational efficiency for automated road detection, particularly in the context of autonomous driving applications. This work proposes a comprehensive study involving five distinct encoder-decoder-based models for semantically segmenting road areas from images. These architectures are designed to classify the rest of the image regions as the background. The five architectures are uniquely distinguished by their encoder components: a traditional Unet model, ResNet50V2, DenseNet121, DenseNet169, and DenseNet201. These encoder architectures were selected to offer a broad spectrum of feature extraction capabilities, from the straightforward yet effective structure of Unet to the densely connected nature of the DenseNet series. The decoders are unified structures tailored to synergize with each encoder's specific characteristics, comprising layers for upsampling, convolution, and activation, with additional batch normalization steps for optimization. This rigorous approach provides a comprehensive understanding of how different encoder-decoder combinations can influence the task of road segmentation, thereby setting a new benchmark in the field and paving the way for future research and innovations.

Novel contributions: This research introduces a novel approach to road area segmentation using encoder-decoder architectures, which are traditionally successful in image segmentation tasks but have not been fully explored in the specific context of road segmentation in autonomous vehicles. Unlike previous models that primarily focus on satellite or aerial

imagery, this study leverages these architectures for ground-level image processing, offering a new perspective in the field. This approach is particularly novel in its application of different encoder models such as ResNet50V2, DenseNet121, DenseNet169, and DenseNet201, each offering unique strengths in feature extraction and representation, tailored for road detection in diverse environments. Additionally, the use of the Mapillary Vistas Dataset, processed into a binary segmentation problem, presents a unique take on dataset customization, facilitating a more focused and effective training process for autonomous driving scenarios.

Due to the success of deep learning, the computer vision field has made remarkable advances over the past few decades. New algorithms have emerged for tasks such as image segmentation [29], object detection [44], and instance segmentation. In our work, we leverage these advancements to address the specific challenges in road segmentation, combining traditional and innovative approaches to develop a robust and efficient encoder-decoder-based model. This model is not only designed for high accuracy but also for adaptability to various road conditions and environments, a crucial aspect for real-world applications in autonomous driving.

In summary, this research pushes the boundaries of current road detection technologies, introducing a comprehensive and adaptable approach to semantically segment road areas, effectively bridging the gap between existing methods and the practical requirements of autonomous vehicle systems. The creation of a new dataset, along with the novel application of encoder-decoder architectures, adds significant value to our work, setting it apart from existing literature in the field.

The primary objectives and contributions of this work are:

- **Objective:** Develop an efficient algorithm to detect roads automatically from images of roads for autonomous driving.
- **Contribution 1:** Introducing a new road segmentation method based on deep learning that offers higher accuracy and efficiency.
- **Contribution 2:** The Mapillary dataset was modified and a new dataset was created to tailor it to the specific needs of the study. This allowed for a more targeted evaluation of the proposed method.
- **Contribution 3:** Providing open-source implementation and dataset, fostering future research and development in the field.
- **Contribution 4:** An encoder-decoder architecture model is proposed for the segmentation of road areas. To extract features from an image, this work experiment with different encoders like Resnet and DenseNet and, after the experiment, suggest the best architecture.

The paper is organized into several sections that provide a comprehensive overview of the research. Section 2 contains a literature review that covers the research problem. Section 3 discusses the dataset and methodology used for the research, while Section 4 provides detailed information on the experiments and results. Finally, the conclusion of the work is presented in Section 5 of the paper.

2 Literature review

Road detection and lane detection have been the subject of research for quite some time. Deep learning techniques such as Convolutional Neural Networks (CNNs), pre-trained frameworks using transfer learning, and segmentation have made road detection and lane detection easier. This topic has gained tremendous momentum in recent years.

The purpose of this section is to highlight some noteworthy studies. Some of the most popular open-source image segmentation datasets include the Mapillary Vistas dataset [28], KITTI dataset [14], Cityscapes dataset [11], and others. These datasets have been the subject of many studies.

In the study by Danping Liu et al. [23], a novel approach to semantic segmentation for autonomous driving scenes is proposed, using a Multi-Scale Adaptive Mechanism (MSAAM). This method effectively tackles challenges prevalent in complex driving environments, such as large-scale variations, occlusions, and varied object appearances. The MSAAM integrates features across multiple scales and utilizes a novel attention module that combines spatial, channel-wise, and scale-wise attention mechanisms to enhance feature discrimination. Demonstrating superior performance on the Cityscapes dataset, with key metrics like ClassAvg: 81.13, mIoU: 71.46, and computational efficiency, the method outperforms comparative models, underlining its effectiveness in autonomous driving scene understanding

Kölle, M., et al. [34] here the authors address the need for efficient semantic segmentation in off-road environments for autonomous vehicles. Most existing models prioritize accuracy over inference speed and are tailored for urban settings. This research introduces SwiftNet, a deep learning model optimized for both high inference speed and accuracy, particularly for large images typical in off-road scenarios. SwiftNet, pre-trained on ImageNet and further trained on the Rellis-3D dataset, achieved an impressive average inference speed of 24 FPS and an mIoU score of 73.8% on a Titan RTX GPU, marking a significant advancement for autonomous vehicle systems in diverse terrains.

In the research conducted by Gagliardi et al [13], a novel approach using Deep Neural Networks (DNN) for the detection and assessment of asphalt pavement distress is presented. The study employs advanced object detection and semantic segmentation techniques, notably YOLO v7 and U-Net, to accurately identify various types of pavement distress. This methodology, validated on a road in Rome, Italy, highlights the efficiency of DNNs in automating the process of identifying, localizing, and assessing pavement damages, offering significant improvements for Pavement Management Systems.

In the study by Kölle, M., et al [20] , the focus shifts to a data-centric approach in semantic segmentation of 3D geospatial point clouds. The paper highlights the significance of creating high-quality datasets for machine learning models. Employing Active Learning, the research demonstrates that labeling under 1% of training points can yield accuracies close to full-scale benchmarks. This method's application on ISPRS point cloud datasets and a large-scale National Mapping Agency point cloud showcases its efficiency and cost-effectiveness

Kangcheng Liu et al. [25] proposed an approach that combines confidence-level-based contrastive learning with a multi-stage fusion network architecture. The aim was to enhance instance discrimination and alignment of low-confidence features with high-confidence counterparts. This framework proved effective in real industrial crack segmentation and road components extraction, contributing to evolving methods for segmentation in autonomous driving applications, achieving an mIoU score of 0.561.

Heyang Thomas Li et al. [21] introduced a pipeline for accurate segmentation and extraction of rural road surface objects in 3D lidar point-cloud. The method involved transforming the point-clouds into a 2D image space and utilizing the Mask R-CNN algorithm for object localization, segmentation, and classification. This approach allowed for geometric parameter estimation such as road width and proved effective in improving the segmentation of needle-type objects.

Sun et al. [24] introduced the Hybrid Multi-resolution and Transformer semantic extraction Network (HMRT). This approach addresses existing method limitations by providing a global

receptive field for high-resolution images and retaining detailed information during downsampling. The method outperformed existing techniques, achieving up to 91.29% recall, 90.41% F1 score, 91.32% OA, and 84.00% MIoU.

Ghandorh et al. [15] proposed a novel road detection technique that combines semantic segmentation and edge detection. This approach overcomes challenges in high-resolution satellite images where objects are small, and not all information is crucial. Using attention blocks in the encoder and a combination of weighted cross-entropy loss and focal Tversky loss, the method produces sharp-pixel segmentation maps and fine edges. Experiments conducted in Saudi Arabia and Massachusetts demonstrate that the method effectively predicts sharp segmentation masks against complex backgrounds, enhancing road detection accuracy.

Tashnim Chowdhury et al. [10] introduced RescueNet, a high-resolution post-disaster dataset collected using Unmanned Aerial Vehicles (UAVs) for semantic segmentation to assess damages after natural disasters. RescueNet provides detailed pixel-level annotation of affected areas, including buildings, roads, pools, trees, debris, and more. The authors demonstrated the effectiveness of the dataset by implementing state-of-the-art segmentation models, achieving a Mean IoU score of 93.98%. This unique dataset contributes comprehensive annotations compared to existing datasets and marks a significant step towards precise damage assessment using deep learning techniques in the wake of natural disasters.

Volpi et al. [39] proposed an online learning protocol that requires continuous, frame-by-frame adaptation from sequences of temporally correlated images. Accompanied by various baselines and extensive analysis of their behaviors, this work serves as a starting point for research into adaptive learning systems that can adapt to real-world situations without supervision.

Muhammad et al. [27] highlighted the significant role of visual sensory data but noted the current insufficiency in achieving level-5 autonomy. They provide an in-depth analysis of deep learning models, state-of-the-art performance, and existing challenges, emphasizing that substantial improvements are still needed in this critical domain.

Wang et al. [40], the authors tackled the issue of image segmentation for autonomous driving in low-light scenarios. They proposed SFNET-N, a unique framework that employs synthetic data collection and a light enhancement network to recognize objects in dark environments. The approach was validated through extensive testing, achieving mIoU scores of 56.9% and 57.4% on specific datasets, demonstrating promising applicability for nighttime autonomous driving.

Han et al. [16] introduced Yolopv2, a multi-task learning network designed for panoptic driving perception, focusing on traffic object detection, road area segmentation, and lane detection. The proposed method achieved an mIoU of 93.2% for drivable area segmentation on the BDD100K dataset. With its superior accuracy and significantly reduced inference time, Yolopv2 contributes to the feasibility of real-time autonomous driving applications.

Bogdoll et al. [5] conducted an extensive survey on anomaly detection techniques in autonomous driving, covering camera, lidar, radar, and multimodal data. The authors systematically categorized methods, considering various attributes such as detection approach and corner case level, and provided insights into current state-of-the-art techniques. The survey highlights the strides made in the field while emphasizing existing challenges and research gaps.

Vojir et al. [38] introduced a method for road anomaly detection through partial image reconstruction and segmentation coupling. Using various backbones, the proposed JSR-Net achieved mIoU scores of 61.2% with Mobilenet v2, 50.3% with Xception, 51.6% with Resnet-101 checkp1, and 66.1% with Resnet-101 checkp2. The approach demonstrated state-

of-the-art performance in detecting unknown objects in autonomous driving, significantly reducing false positives and highlighting robust precision.

Nima Khairdoost et al. [19] the use of a CNN-based regression method was investigated for detecting ego lane boundaries. After the lane detection stage, they classified it using the ResNet101 network to verify the detected lanes or possible road boundary, achieving an accuracy of 94.52% in the lane classification stage.

Qin Zou et al. [45] proposed hybrid deep architecture was by combining a CNN with a recurrent neural network (RNN). Rural and highway images were both accurately classified by the model named UnetConvLSTM at 98.4% and 98.00%, respectively.

Yong Chen et al. [6] developed a gradient direction feature and a lane boundary projection model. By using images captured with vehicle-mounted monocular cameras, it was able to detect lanes accurately despite broken and worn land markings, glare from the sun, and curved lanes.

Abdulhakam.AM.Assidiq et al. [3] developed a vision-based lane detection approach that is robust to lightning changes and shadows while allowing for real-time operation. On smooth roads, the accuracy was more than 95%, while on congested roads, it was more than 90%.

Hsu-Yung Cheng et al. [8] developed a robust algorithm that could detect the left and right boundaries of lanes regardless of lighting conditions.

Yang Xing et al. [41] proposed vision-based lane detection and evaluation system was proposed, based on fusion of camera and Lidar sensors. Based on the cross-fusion network, the maximum average accuracy for road detection on urban streets was 85.54%, and the correct rate for lane detection was 96.34%.

Li Yong et al. [22] proposed an algorithm that was 40 times faster than existing algorithms at that time for road detection. By using dark channel-based image segmentation, they were able to distinguish rough road regions from complex background noise with 91.7% accuracy.

Wansik Choi et al. [9] proposed a method to balance an imbalanced dataset using Cycle-GAN to improve the performance of various road surface detection algorithms. Using generated snowy and wet images to avoid overfitting, they achieved an accuracy of 84% and an Intersection Over Union (IOU) score of 0.81 on the augmented dataset.

Carles Ventura et al. [36] created a CNN that predicted local connectivity between the central pixel of an input patch and its borders, achieving a precision-recall of 83.5%.

Keval Doshi and Yasin Yilmaz [12] proposed four ensemble models for object detection containing 5, 15, 25, and 30 models. These were trained on various types of road damages from Czech, Japan, and India, achieving an F1 score of 0.628 on the test 1 dataset and 0.6358 on the test 2 dataset.

Deeksha Arya et al. [2] proposed a large-scale heterogeneous road damage dataset comprising 26,620 images collected from multiple countries using smartphones. They also proposed generalized models capable of detecting and classifying road damage in more than one country.

[31, 35], and [1] proposed similar R-CNN and Faster R-CNN architectures by varying the region proposal network (RPN). These architectures were trained on different datasets, such as images captured from smartphones and the Global Road Damage Detection Challenge 2020, achieving F1 scores of 0.51 and 0.528 and an accuracy of 98.88%, respectively.

A deep learning-based crack detection method was proposed by Zhang et al. [43]. Using a supervised deep convolutional neural network, they classified each image patch in the collected images, achieving an F1 score of 89.65% and a Recall score of 92.51%.

Johan Vertens et al. [37] published a dataset with 20,000 RGB - thermal image pairs and proposed an architecture, HeatNet, that takes RGB and thermal images to train and predict

segmentation masks in the daytime and nighttime domains. The mIoU scores achieved for the daytime domain were 70.8, and for the nighttime domain, it was 64.9.

Bowen Cheng et al. [7] proposed MaskFormer, a simple mask classification transformer that predicts a set of binary masks, each associated with a single global class prediction, achieving a mIoU score of 54.1.

Rudra P K Poudel et al. [32] present ContextNet, a new architecture that utilizes factorized convolution, network compression, and pyramid representation. It produces competitive semantic segmentation in real-time with low memory requirements, achieving 66.1% accuracy.

After extensive research on road area and lane detection, road damage detection and classification, and semantic segmentation, it was found that deep learning models can achieve significant results in detecting and classifying roads and lanes from images, videos, and in real-time. This paper proposes a semantic segmentation model to detect road areas from images using a dataset created by selecting images and their labels from the Mapillary Vistas dataset [28] and processing them to meet our needs.

3 Materials and methods

This section provides an overview of the proposed work. Firstly, the dataset collection process will be described. Then, the collected dataset will be transformed into a road segmentation problem dataset. In this dataset, images will be categorized into two groups: one representing the road, and the other representing the background. Following this, models will be designed for road segmentation using the Encoder-decoder architecture. Finally, the models will be trained, hyperparameters will be adjusted, and evaluation metrics will be calculated. Figure 1 shows the flow graph of the overall work.

3.1 Original dataset

The original dataset used is the Mapillary vistas dataset [28]. It consists of a total of 25,000 images, out of which 18,000 are training images, 2,000 are validation images, and 5,000 are testing images, with variable sizes. There are two versions of the dataset, v1.2 and v2.0. In v1.2, there are 66 labels for semantic segmentation, and in v2.0, there are 124 labels. Additionally, there are labels for panoptic segmentation and instance segmentation. To address the specific problem, we convert the segmented labels from multi-class segmentation to binary segmentation, creating a custom dataset using the Mapillary Vistas dataset.

3.2 Dataset for road segmentation

The dataset was prepared from the Mapillary vistas dataset [28]. Some images and their corresponding labels were taken from the Mapillary vistas dataset. Then the images were passed through a function that resized them from their original size to 224 x 224. The corresponding labels were passed through another function that resized the labels from their original size to 224 x 224 and converted them to grayscale images with two classes: ”Road” and ”Background.” The colors representing these classes are white (255,255,255) for ”Road” and black (0,0,0) for ”Background.” Figure 2 shows the images of the sample data.

Unique to our approach is the method of transforming the Mapillary Vistas Dataset into a format specifically tailored for road segmentation. We have innovatively processed the dataset

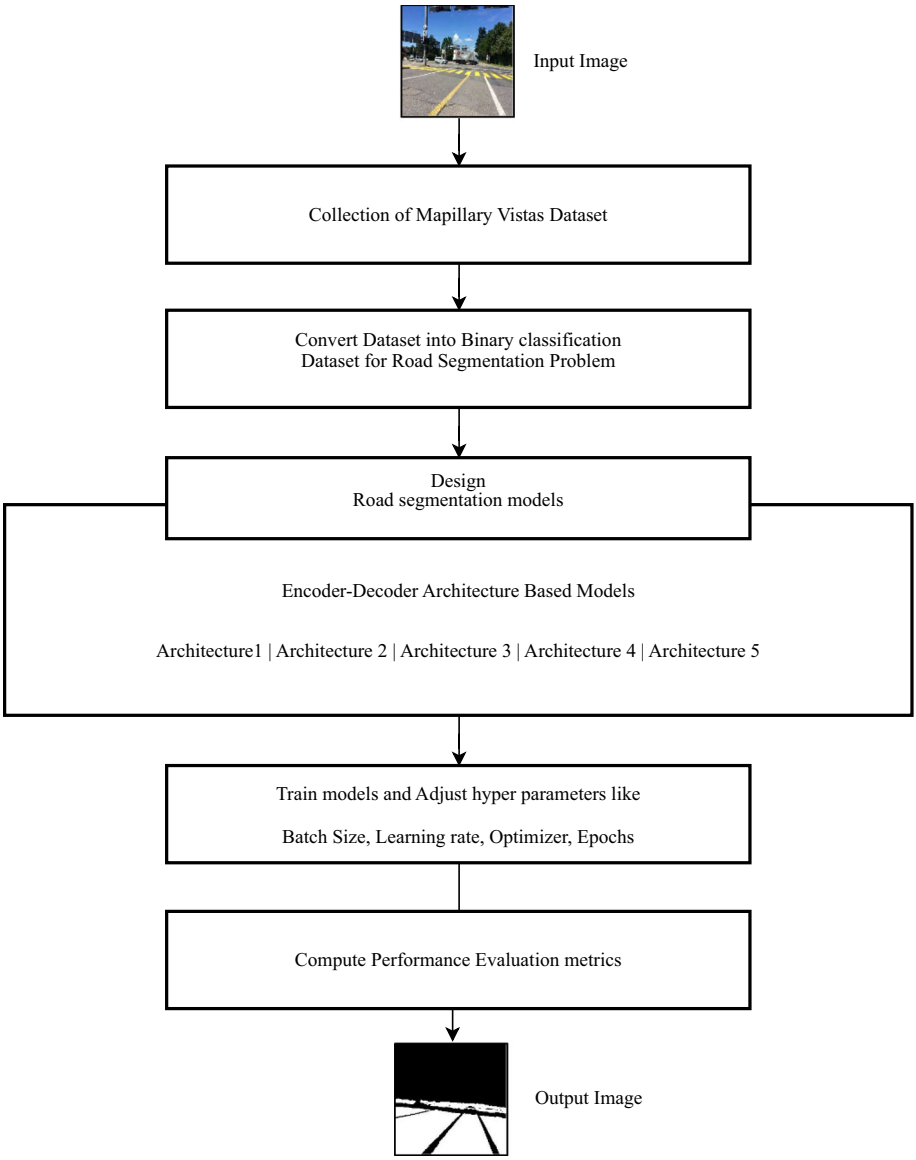

Fig. 1 Overall Flow diagram

into binary segmentation, distinguishing 'Road' from 'Background,' which is a significant departure from the dataset's original multi-class segmentation format. This customization enables our models to focus intensely on the critical task of road detection, a key aspect in autonomous driving applications. Moreover, standardizing image sizes to 224 x 224 pixels ensures uniformity, aiding in more efficient and effective model training. Such specific dataset processing techniques are pivotal in achieving the high accuracy and generalizability of our models.

Fig. 2 Road Net data Generation from the original dataset

3.3 Methodology

To segment roads, this work used an encoder-decoder architecture. There are two parts to it: the encoder, or contractive part, similar to a classification network, and the decoder, which enlarges the representation into a high-resolution segmentation.

3.3.1 Encoder architecture

The encoder consists of four convolutional blocks, each followed by a bottleneck layer. Each convolutional block includes two convolutional layers and a pooling layer for spatial dimension reduction. The pooling layer captures hierarchical information in a contracted form, and the bottleneck layer acts as a bridge between the contracting and expansive parts.

3.3.2 Decoder architecture

The decoder comprises four decoder blocks, each handling upsampling and feature refinement. Each decoder block includes a `Conv2DTranspose` layer for upsampling, a concatenation layer for merging upsampled features with corresponding features from the contracting part (skip connection), and two convolutional layers for further feature processing. The number of filters is halved after each decoder block, gradually reducing complexity. After the last decoder block, a `Conv2D` layer with a 1×1 kernel is applied for the final segmentation output.

3.3.3 Choice of encoders

Different variants of ResNet (`ResNet50V2`) and DenseNet (`DenseNet121`, `DenseNet169`, `DenseNet201`) serve as encoders. The number in ResNet and DenseNet denotes the layers in the model, indicating a deep architecture with skip connections for effective feature learning. Pre-trained models like `ResNet50V2` and DenseNet variants leverage learned features from large datasets, enhancing the model's ability to generalize to the segmentation task.

Figure 3 shows the general architecture of the models. A summary of the architectures is given in Table 1.

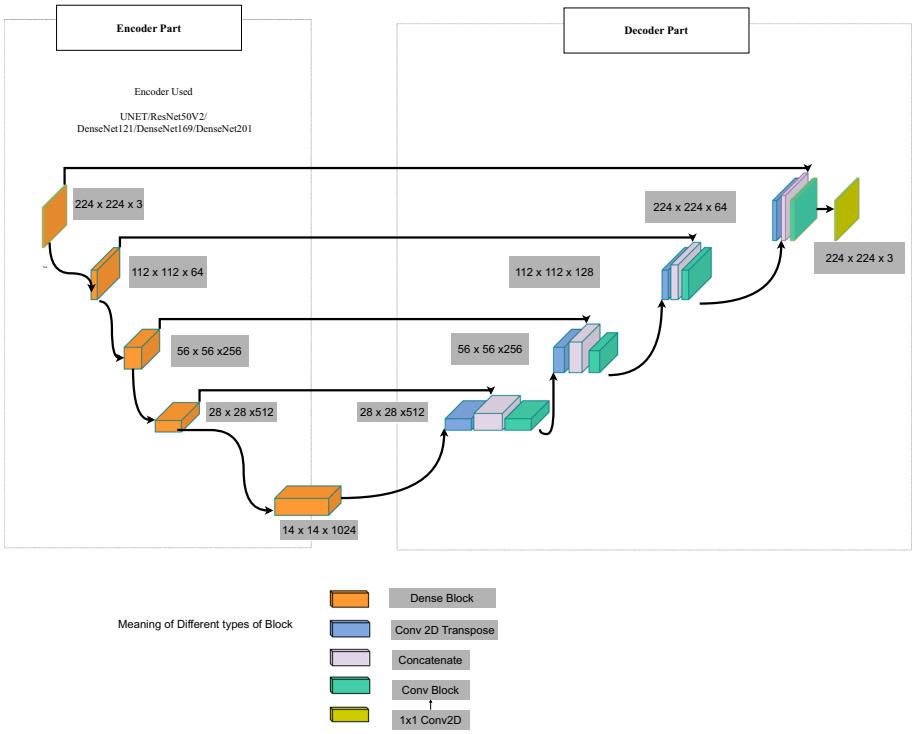

Fig. 3 Architecture

3.3.4 The role of convolutions in the model

The encoder-decoder architectures contain specific convolutional operations to perform contextual information fusion across scales.

Residual networks (ResNet): Residual connections are an essential part of ResNet architectures. They are mathematically represented as:

$$\text{output} = \mathcal{F}(\text{input}, W) + \text{input} \tag{1}$$

Table 1 Summary of Models

Architecture	Encoder	Decoder
Architecture 1	Unet encoder	2x2 convolution (“up-convolution”), two 3x3 convolution layers, 1x1 convolution
Architecture 2	ResNet50V2	2x2 convolution("up-convolution"), 3x3 convolution layer, Batch normalization layer, Activation Layer, 1x1 convolution layer
Architecture 3	DenseNet121	2x2 convolution("up-convolution"), 3x3 convolution layer, Batch normalization layer, Activation Layer, 1x1 convolution layer
Architecture 4	DenseNet169	2x2 convolution("up-convolution"), 3x3 convolution layer, Batch normalization layer, Activation Layer, 1x1 convolution layer
Architecture 5	DenseNet201	2x2 convolution("up-convolution"), 3x3 convolution layer, Batch normalization layer, Activation Layer, 1x1 convolution layer

Densely connected networks (DenseNet): Dense connections are the main concept behind DenseNet architectures. They are described as:

$$\text{output}_i = \mathcal{H}_i([\text{output}_{i-1}, \text{output}_{i-2}, \ldots, \text{output}_0]) \quad (2)$$

3.3.5 Impact of DenseNet and residual networks on segmentation

The choice of DenseNet and ResNet for the encoder part had a significant impact on the road segmentation task. These architectures were selected because they allow for a more efficient training process and achieve better performance in terms of accuracy. The detailed impact of using these networks is reflected in the experimental results and discussions in the following sections.

3.3.6 Architecture 1

"The architecture used is identical to the Unet model that was proposed " [33]. There are two parts to this system: the encoder and the decoder. For the encoder part, two 3x3 convolutions are applied repeatedly with the same padding parameters, Rectified Linear Activation (ReLU) as the activation parameter, and a 2x2 max pooling operation with stride 2. Each downsampling step doubles the number of feature channels. The decoder part consists of an upsampling of the feature map followed by a 2x2 convolution ("up-convolution") that halves the number of feature channels, a concatenation with the correspondingly cropped feature map from the contracting path, and two 3x3 convolutions with the activation parameter set to Rectified Linear Activation (ReLU). Finally, there is a 1x1 convolution with the activation parameter set as a sigmoid activation function, with the number of filters equal to 1.

3.3.7 Architecture 2 (ResNet50V2 as Encoder)

It is called a Residual Convolutional Network because it makes use of residual connections between layers via Residual Blocks. In this model, we use transfer learning for the encoder part, utilizing the pre-trained ResNet50V2 architecture [17]. The layers from the pre-trained model have been taken as the layers with their corresponding shapes (224,224,3), (112,112,64), (56,56,128), (28,28,256), (14,14,512). The decoder block consists of an upsampling of the feature map followed by a 2x2 convolution ("up-convolution" or "Conv2DTranspose"), a concatenation with the correspondingly cropped feature map from the encoder path, followed by a 3x3 convolution with the padding parameter set to the same. Each convolution layer is followed by a Batch Normalization layer and an Activation layer set to Rectified Linear Activation (ReLU).

3.3.8 Architecture 3 (DenseNet121 as Encoder)

It is called a Densely Connected Convolutional Network because DenseNet121 uses dense connections between layers through Dense Blocks [18]. In the DenseNet121 Unet model, traditional convolutional layers in the encoder blocks are replaced with Dense Blocks of DenseNet121. We use transfer learning for the encoder part, utilizing the pre-trained DenseNet121 architecture. The layers from the pre-trained model have been taken as the layers with their corresponding shapes (224,224,3), (112,112,64), (56,56,128), (28,28,256), (14,14,512). The decoder block consists of an upsampling of the feature map followed by a

2x2 convolution (”up-convolution” or ”Conv2DTranspose”), a concatenation with the correspondingly cropped feature map from the encoder path, followed by a 3x3 convolution with the padding parameter set to the same. Each convolution layer is followed by a Batch Normalization layer and an Activation layer set to Rectified Linear Activation (ReLU).

3.3.9 Architecture 4 (DenseNet169 as Encoder)

In the DenseNet169 Unet model, traditional convolutional layers in the encoder blocks are replaced with Dense Blocks of DenseNet169. We use transfer learning for the encoder part, utilizing the pre-trained DenseNet169 architecture. The layers from the pre-trained model have been taken as the layers with their corresponding shapes (224,224,3), (112,112,64), (56,56,128), (28,28,256), (14,14,512). The decoder block consists of an upsampling of the feature map followed by a 2x2 convolution (”up-convolution” or ”Conv2DTranspose”), a concatenation with the correspondingly cropped feature map from the encoder path, followed by a 3x3 convolution with the padding parameter set to the same. Each convolution layer is followed by a Batch Normalization layer and an Activation layer set to Rectified Linear Activation (ReLU).

3.3.10 Architecture 5 (DenseNet201 as Encoder)

This model uses transfer learning for the encoder part, utilizing the pre-trained DenseNet201 architecture. The layers from the pre-trained model have been taken as the layers with their corresponding shapes (224,224,3), (112,112,64), (56,56,128), (28,28,256), (14,14,512). The decoder block consists of an upsampling of the feature map followed by a 2x2 convolution (”up-convolution” or ”Conv2DTranspose”), a concatenation with the correspondingly cropped feature map from the encoder path, followed by a 3x3 convolution with the padding parameter set to the same. Each convolution layer is followed by a Batch Normalization layer and an Activation layer set to Rectified Linear Activation (ReLU).

3.4 Mathematical representation of model operations

3.4.1 Convolutional layer (Conv2D)

The convolutional layers are pivotal in feature extraction, applying filters to the input data. Each convolution operation is mathematically represented as:

$$Y_{ij} = \sum_{m} \sum_{n} X_{i+m,j+n} \cdot W_{mn} + b \tag{3}$$

where Y_{ij} is the output, X is the input matrix, W_{mn} are the weights of the convolutional filter, and b is the bias term. The indices m, n traverse the kernel dimensions.

3.4.2 Batch normalization

Batch normalization follows convolutional layers and standardizes the outputs, facilitating a more stable and efficient learning process:

$$Y' = \gamma \left(\frac{Y - \mu}{\sqrt{\sigma^2 + \epsilon}} \right) + \beta \tag{4}$$

Here, Y denotes the layer's input, μ and σ^2 are the mean and variance, respectively, γ and β are trainable parameters, and ϵ ensures numerical stability.

3.4.3 Activation function (ReLU)

ReLU, a non-linear activation function, is essential for the model to capture complex relationships in the data:

$$f(Y') = \max(0, Y') \tag{5}$$

where f denotes the ReLU function, and Y' is the output from the batch normalization layer.

3.4.4 Transposed convolution (Conv2DTranspose)

Transposed convolution layers in the decoder part perform upsampling, increasing the spatial dimensions of the feature maps.

3.4.5 Concatenation in decoder

In the decoder, concatenation plays a crucial role in merging the upsampled features with the corresponding features from the encoder, enhancing the detail in the segmentation output:

$$D = [U; E] \tag{6}$$

where U represents the upsampled feature map from the decoder, E is the corresponding feature map from the encoder, and D is the combined feature map after concatenation.

3.4.6 Final output layer (Conv2D with Sigmoid Activation)

The model's final output layer uses a 1x1 convolution followed by a sigmoid activation function to generate the segmentation map:

$$O = \sigma(\sum_m \sum_n D_{mn} \cdot W''_{mn} + b'') \tag{7}$$

In this formula, O is the final segmentation output, D is the input from the last layer of the decoder, W''_{mn} are the weights of the convolutional layer, b'' is the bias term, and σ represents the sigmoid activation function.

3.5 Customized encoder-decoder architectures

In our study, the novelty extends to the careful selection and integration of various encoder models, each bringing unique strengths to the task of road segmentation. We have adapted ResNet and DenseNet models, renowned for their efficacy in feature extraction, and ingeniously incorporated them into encoder-decoder architectures. This innovative adaptation is critical for extracting complex features relevant to road scenarios from diverse environmental conditions. Additionally, our custom-designed decoder complements each encoder's strengths, efficiently reconstructing the road segmentation from the encoded features. This level of customization in both the encoder and decoder aspects of our models is a significant contributor to their effectiveness in accurately segmenting road areas, thereby advancing the state-of-the-art in road detection technologies.

4 Experiments and results

This section outlines the various experiments conducted on the dataset to assess the efficiency of the proposed deep learning models. Comprehensive experimental analysis was performed on different images to semantically segment the road from the corresponding images. Each model was prepared and trained separately, following the procedures described in the methodology section.

4.1 Hardware and software specifications

The experiments were implemented using TensorFlow and the cv2 library of Python and conducted on the Kaggle platform, utilizing an Nvidia Tesla P100 16 GB graphics card. The Weights and Biases platform was employed for monitoring and tracking the hyperparameter tuning of the different models. [4]

4.2 Performance evaluation metrics

Performance evaluation metrics play a crucial role in assessing the effectiveness of deep learning segmentation models. The performance of each model on the validation dataset was evaluated and compared using the Dice coefficient and Mean Intersection over Union (IoU) metrics.

Dice coefficient: This statistical tool measures the similarity between two sets of data, A and B, and is expressed as:

$$\textbf{Dice coefficient} = \frac{2 \cdot (A \cap B)}{A \cup B} \tag{8}$$

The area of convergence is calculated by overlaying the predicted segmentation on the true label of the image and identifying overlapping road pixels.

Mean intersection over union (Mean IoU): This essential metric calculates the average intersection over the union of sets A and B, where A represents the predicted pixels and B is the ground truth pixels. The Mean IoU is computed as:

$$\textbf{Mean IoU} = \frac{1}{n} \sum_{i=1}^{n} \text{IoU}_i \tag{9}$$

where n is the number of classes.

4.3 Hyperparameter optimization

The main goal of this study is to design a segmentation model capable of semantically segmenting road areas from an image. This requires identifying the optimal hyperparameter configuration to achieve the best Dice coefficient. Table 2 presents the hyperparameters used to train the five models discussed in the methodology section.

4.4 Performance of models

This subsection details the results obtained from the different models, each trained using the dataset. Table 3 lists the evaluation metrics for different models, and Figs. 4, 5, 6, and 7

Table 2 Hyperparameters Setting

Parameters	Value
Input shape	224x224
Image channels	3
Mask shape	224x224
Mask channels	1
Optimizer	Adam
Batch Size	32
Learning Rate	0.0001
Epochs	200
Activation	ReLU & Sigmoid (For the last layer)

display the graphs for training Dice coefficient, validation Dice coefficient, train loss, and validation loss, respectively. Sample outputs for each model are presented in Figs. 8, 9, 10, 11 and 12.

Each model's performance, as shown in Table 3, offers a comprehensive understanding of the strengths and weaknesses of different architectures. For instance, the DenseNet169 encoder achieved a Dice coefficient of 99.69% on the training set, indicating its superior feature extraction capabilities for road segmentation tasks. This detailed analysis helps in identifying the most promising models for further optimization and application.

4.5 Novel insights and implications

The results of our study contribute novel insights into the field of semantic road segmentation. The high performance of models like DenseNet169 in accurately segmenting road areas underlines their potential in autonomous driving applications. These findings not only advance our understanding of encoder-decoder architectures but also pave the way for future research aimed at enhancing the reliability and efficiency of autonomous navigation systems.

4.6 Discussion

The results of our study, summarized in Tables 2 and 3, demonstrate the comparative performance of various encoder-decoder architectures. The study not only highlights the benefits

Table 3 Calculation of Parameters for different models

Name of the Architecture	Train Dice Cofficient	Train Loss	Validation Dice Cofficient	Validation Loss	Total No of Parameters
Architecture 1	98.51	0.0115	92.24	.2086	16,474,785
Architecture 2	99.43	0.0043	92.66	0.3147	12,883,841
Architecture 3	99.65	0.0026	93.83	0.2641	13,272,001
Architecture 4	99.69	0.0023	93.85	0.2632	15,295,937
Architecture 5	99.72	0.0022	93.8	0.2577	20,154,817

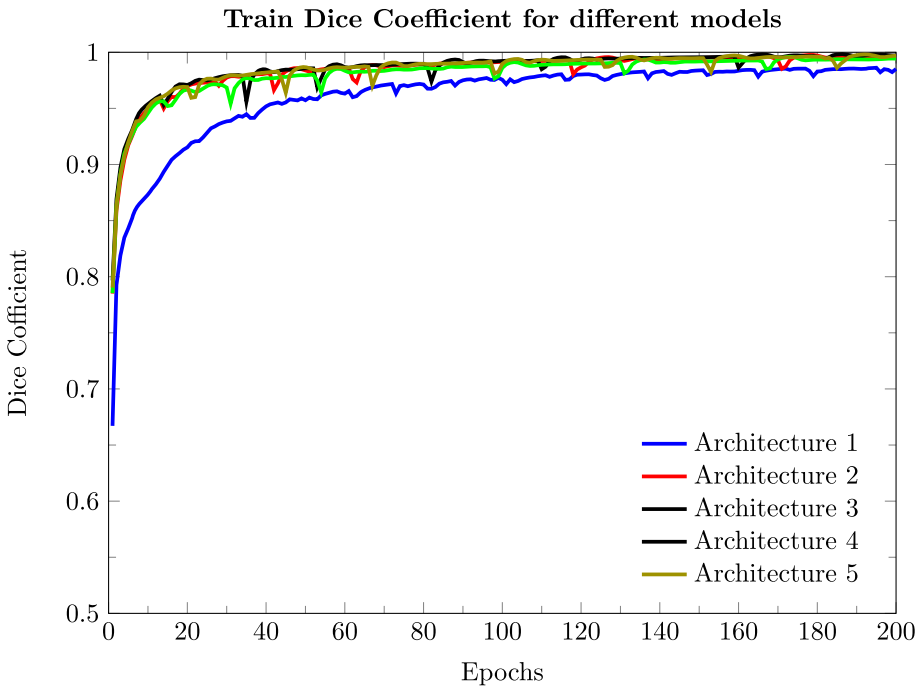

Fig. 4 Graph for Train dice coefficient for different models

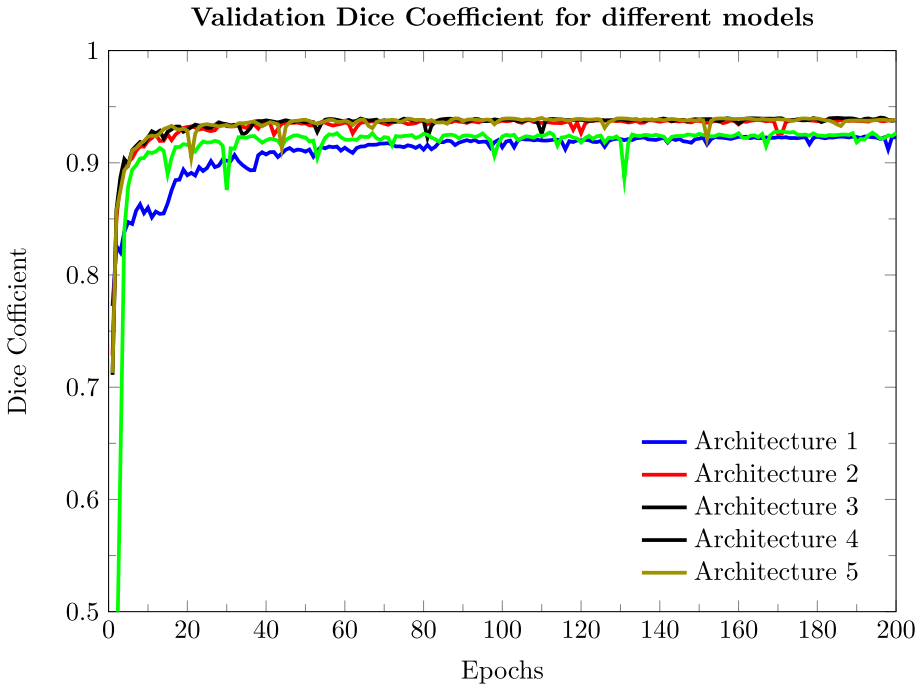

Fig. 5 Graph for Validation dice coefficient for different models

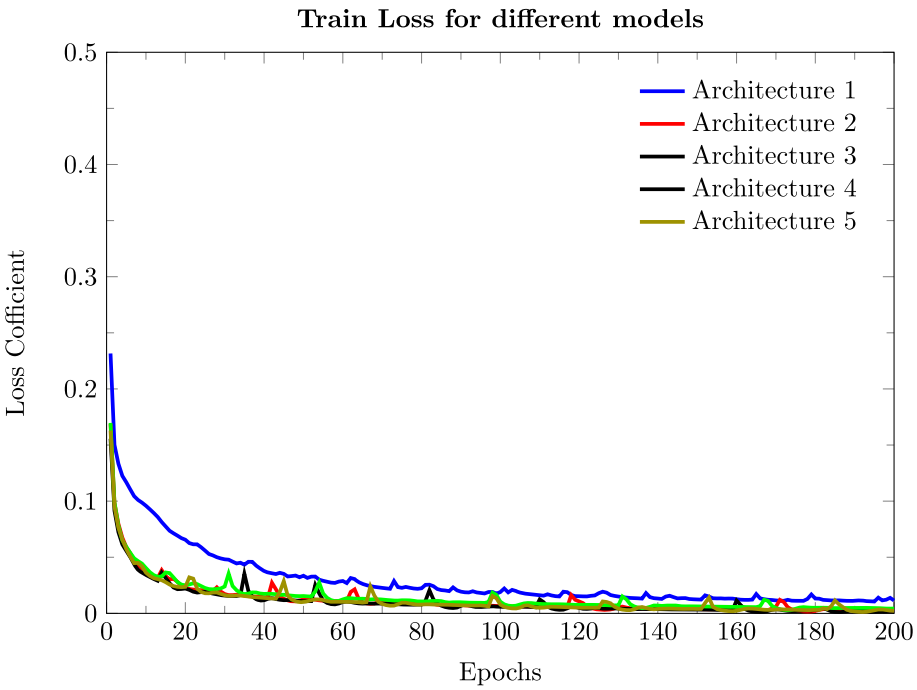

Fig. 6 Graph for Train loss coefficient for different models

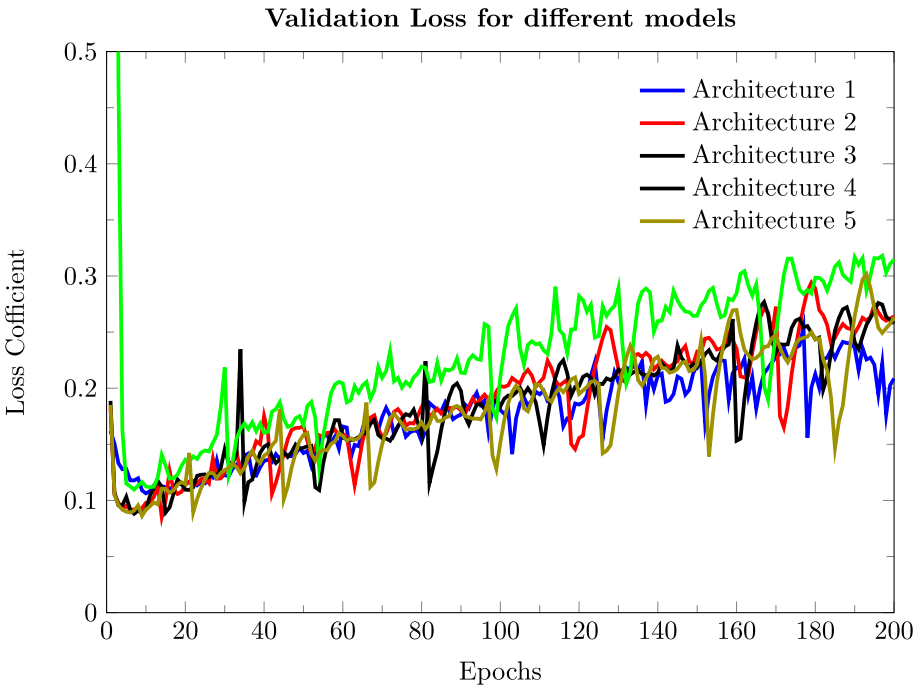

Fig. 7 Graph for Validation loss coefficient for different models

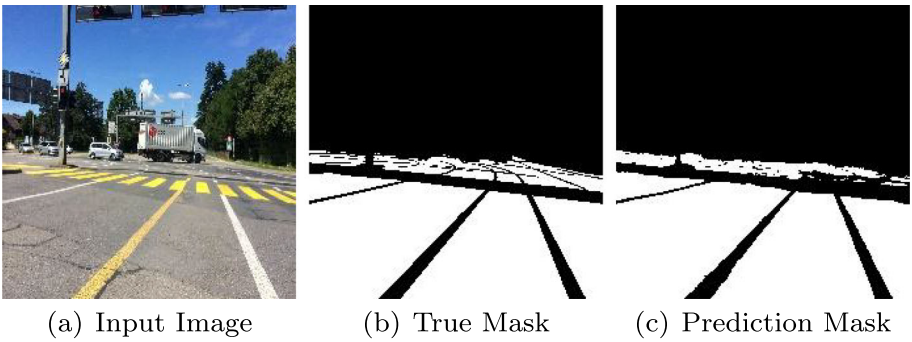

Fig. 8 Predictions for Architecture 1 on the validation dataset

of using residual and dense blocks as part of the encoder but also uncovers some potential shortcomings of previously reported methods, including DenseNet and UNet models.

ResNet models: Residual Networks (ResNet) have emerged as a powerful architecture with certain distinctive features:

- **Residual Connections:** By introducing shortcut connections that bypass one or more layers, ResNet alleviates the vanishing gradient problem, allowing for deeper networks.
- **Improved Training Stability:** The residual connections promote smoother gradient flow, enhancing the stability of training.
- **Potential Complexity:** While offering strong feature learning capabilities, ResNet's architecture might lead to increased complexity and computational demands.
- **Difference with DenseNet:** Our experiments indicated a more pronounced difference in performance between ResNet and DenseNet encoders, highlighting a potential trade-off between complexity and generalizability.

DenseNet models: While DenseNet architectures offer advantages in terms of feature reuse and reduction in vanishing gradient problems, they may suffer from certain demerits:

- **Computational Complexity:** Dense connections lead to an increase in the computational burden, making the model more resource-intensive.
- **Potential Overfitting:** The dense connectivity may sometimes result in overfitting, especially with limited training data.

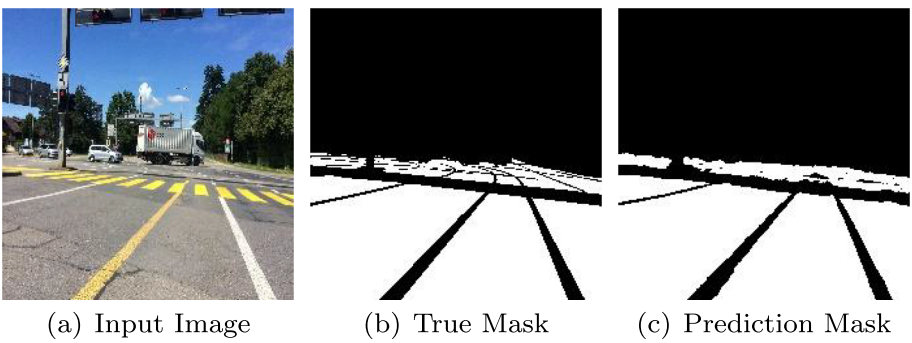

Fig. 9 Predictions for Architecture 2 on the validation dataset

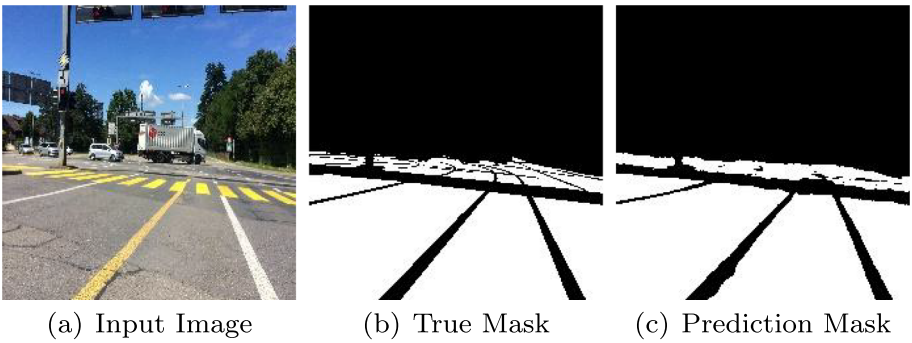

Fig. 10 Predictions for Architecture 3 on the validation dataset

– **Consistency:** The experiments showed minimal differences in training and validation Dice coefficient values for the DenseNet encoders, indicating good generalization without overfitting. Their dense connectivity pattern encourages more effective gradient flow during training.

UNet models: The UNet model is a widely used architecture for semantic segmentation but has its limitations:

– **Lack of Residual Connections:** Unlike ResNet-based models, UNet does not incorporate residual connections, which may hinder the flow of gradients during training.
– **Limited Generalization:** UNet may struggle with generalizing across diverse and complex scenarios, especially when there is a significant domain shift between training and testing data.
– **Trade-offs with ResNet:** The greater difference in the ResNet-based model may indicate a potential trade-off between complexity and generalizability, with residual connections offering powerful feature learning but possibly leading to less stable performance.

Through our experiments, we observed a significant increase in the value of the Dice coefficient when using residual and dense blocks in the encoder. These findings shed light on the potential trade-offs associated with different encoder architectures, emphasizing the role of architectural design in achieving high performance.

Implications and future directions: The success of DenseNet encoders in this study points to the potential benefits of densely connected networks in segmentation tasks. It uncov-

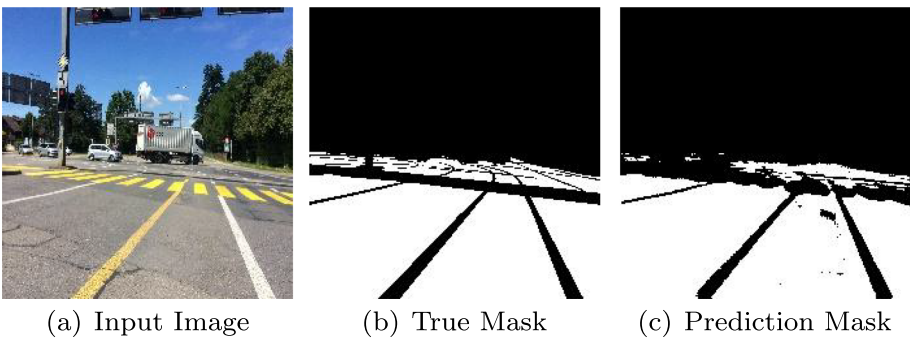

Fig. 11 Predictions for Architecture 4 on the validation dataset

Fig. 12 Predictions for Architecture 5 on the validation dataset

ers specific challenges associated with commonly used architectures and offers guidance for future explorations, including further variations, combinations, and optimization strategies in the field of semantic segmentation.

In summary, the nuanced analysis provided in this discussion contributes valuable insights into the design and selection of encoder-decoder models for road detection tasks, with a balanced view of the merits and demerits of different architectural choices.

5 Conclusion and future scope

This paper conducted an in-depth comparative study of various encoder-decoder architectures, specifically focusing on road detection in semantic segmentation. Through extensive experimentation and analysis, the study demonstrated the effectiveness of different models, revealing the nuances of their performance and generalization capabilities.

5.1 Conclusion

The key findings of this study include:

- **Effectiveness of Residual and Dense Blocks:** The integration of residual and dense blocks within the encoder significantly enhanced the performance of the models, as evidenced by the higher Dice coefficient values.
- **Comparison of Encoder Types:** The models employing Densenet encoders exhibited consistent performance across training and validation sets, while Resnet-based models showed slightly more variability.
- **Quality of Predictions:** The predicted segmentation masks were found to be highly accurate and closely aligned with the true masks, indicating the robustness and precision of the models.

These insights underscore the potential of carefully designed encoder-decoder architectures in semantic segmentation tasks and contribute valuable knowledge to the field of road detection.

5.2 Future scope

The field of road area detection continues to present exciting opportunities for further research and innovation. Some promising directions include:

- **Exploration of New Models:** Future work can explore novel architectures and techniques for semantic segmentation, potentially leading to even more accurate and efficient road detection models.
- **Enhancements to Decoder Design:** Modifying the decoder part of existing models may unlock additional performance gains, enabling more nuanced feature reconstruction and improved segmentation quality.
- **Optimization Strategies:** Investigating different optimization algorithms and hyperparameter tuning strategies can lead to more refined models with better generalization and fewer parameters.
- **Panoptic Segmentation:** An emerging and unexplored area within road detection is panoptic segmentation, which combines instance and semantic segmentation. This offers a new avenue for understanding and representing roads within complex scenes.

In conclusion, this study has shed light on the capabilities and limitations of current models for road area detection and has paved the way for further advancements in this vital area of research. The continued exploration and development of innovative techniques hold great promise for enhancing the state-of-the-art in road detection and contributing to safer and more intelligent transportation systems.

Funding No funding was received to assist with the preparation of this manuscript.

Data Availability The dataset was created from Mapillary Vistas Dataset [28] it is publicly available on kaggle https://www.kaggle.com/datasets/burhanuddinlatsaheb/roadnet-dataset

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

References

1. Arman M, Hasan M, Sadia F, Shakir AK, Sarker K, Himu FA et al (2020) Detection and classification of road damage using r-cnn and faster r-cnn: a deep learning approach. In: International conference on cyber security and computer science, Springer, pp 730–741
2. Arya D, Maeda H, Ghosh SK, Toshniwal D, Mraz A, Kashiyama T, Sekimoto Y (2020) Transfer learning-based road damage detection for multiple countries. arXiv:2008.13101
3. Assidiq AA, Khalifa OO, Islam MR, Khan S (2008) Real time lane detection for autonomous vehicles. In: 2008 international conference on computer and communication engineering, IEEE, pp 82–88
4. Biewald L (2020) Experiment tracking with weights and biases. URL https://www.wandb.com/. Software available from wandb.com
5. Bogdoll D, Nitsche M, Zöllner JM (2022) Anomaly detection in autonomous driving: A survey. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4488–4499
6. Chen Y, He M, Zhang Y (2011) Robust lane detection based on gradient direction. In: 2011 6th IEEE conference on industrial electronics and applications, IEEE, pp 1547–1552
7. Cheng B, Schwing A, Kirillov A (2021) Per-pixel classification is not all you need for semantic segmentation. Adv Neural Inf Process Syst 34
8. Cheng HY, Jeng BS, Tseng PT, Fan KC (2006) Lane detection with moving vehicles in the traffic scenes. IEEE Trans Intell Trans Syst 7(4):571–582
9. Choi W, Heo J, Ahn C (2021) Development of road surface detection algorithm using cyclegan-augmented dataset. Sensors 21(22):7769
10. Chowdhury T, Murphy R, Rahnemoonfar M (2022) Rescuenet: A high resolution uav semantic segmentation benchmark dataset for natural disaster damage assessment. arXiv:2202.12361
11. Cordts M, Omran M, Ramos S, Scharwächter T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B (2015) The cityscapes dataset. In: CVPR Workshop on the future of datasets in vision, vol. 2. sn

12. Doshi K, Yilmaz Y (2020) Road damage detection using deep ensemble learning. In: 2020 IEEE International conference on big data (Big Data), IEEE, pp 5540–5544
13. Gagliardi V, Giammorcaro B, Bella F, Sansonetti G (2023) Deep neural networks for asphalt pavement distress detection and condition assessment. Earth Resources and environmental remote sensing/GIS Applications XIV, SPIE 12734:251–262
14. Geiger A, Lenz P, Stiller C, Urtasun R (2013) Vision meets robotics: The kitti dataset. Int J Robotics Res 32(11):1231–1237
15. Ghandorh H, Boulila W, Masood S, Koubaa A, Ahmed F, Ahmad J (2022) Semantic segmentation and edge detection—approach to road detection in very high resolution satellite images. Remote Sens 14(3):613
16. Han C, Zhao Q, Zhang S, Chen Y, Zhang Z, Yuan J (2022) Yolopv2: Better, faster, stronger for panoptic driving perception. arXiv:2208.11434
17. He K, Zhang X, Ren S, Sun J (2015) Deep residual learning for image recognition. arXiv:1512.03385
18. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
19. Khairdoost N, Beauchemin SS, Bauer MA (2021) Road lane detection and classification in urban and suburban areas based on cnns. In: VISIGRAPP (5: VISAPP), pp 450–457
20. Kölle M, Walter V, Schmohl S, Soergel U (2023) Learning on the edge: Benchmarking active learning for the semantic segmentation of als point clouds. ISPRS Ann Photogrammetry, Remote Sens Spatial Inf Sci 10:945–952
21. Li HT, Todd Z, Bielski N, Carroll F (2022) 3d lidar point-cloud projection operator and transfer machine learning for effective road surface features detection and segmentation. Visual Comp 38(5):1759–1774
22. Li Y, Ding W, Zhang X, Ju Z (2016) Road detection algorithm for autonomous navigation systems based on dark channel prior and vanishing point in complex road scenes. Robotics Autonomous Syst 85:1–11
23. Liu D, Zhang D, Wang L, Wang J (2023) Semantic segmentation of autonomous driving scenes based on multi-scale adaptive attention mechanism. Front Neurosci 17
24. Liu K (2022) Multi-resolution transformer network for building and road segmentation of remote sensing image. ISPRS Int J Geo-Inf 11(3):165
25. Liu K (2022) Semi-supervised confidence-level-based contrastive discrimination for class-imbalanced semantic segmentation. In: 2022 12th International conference on CYBER technology in automation, control, and intelligent systems (CYBER), IEEE, pp 1230–1235
26. Maeda H, Sekimoto Y, Seto T, Kashiyama T, Omata H (2018) Road damage detection and classification using deep neural networks with smartphone images. Computer-Aided Civil Infrastructure Eng 33(12):1127–1141
27. Muhammad K, Hussain T, Ullah H, Del Ser J, Rezaei M, Kumar N, Hijji M, Bellavista P, de Albuquerque VHC (2022) Vision-based semantic segmentation in scene understanding for autonomous driving: Recent achievements, challenges, and outlooks. IEEE Trans Intell Transportation Syst
28. Neuhold G, Ollmann T, Rota Bulo S, Kontschieder P (2017) The mapillary vistas dataset for semantic understanding of street scenes. In: Proceedings of the IEEE international conference on computer vision, pp 4990–4999
29. Noble J, Boukerroui D (2006) Ultrasound image segmentation: a survey. IEEE Trans Med Imaging 25(8):987–1010. https://doi.org/10.1109/TMI.2006.877092
30. Okuda R, Kajiwara Y, Terashima K (2014) A survey of technical trend of adas and autonomous driving. In: Proceedings of technical program - 2014 INTERNATIONAL SYMPOSIUM on VLSI technology, systems and application (VLSI-TSA), pp 1–4. https://doi.org/10.1109/VLSI-TSA.2014.6839646
31. Pham V, Pham C, Dang T (2020) Road damage detection and classification with detectron2 and faster r-cnn. In: 2020 IEEE international conference on big data (Big Data), IEEE, pp 5592–5601
32. Poudel RP, Bonde U, Liwicki S, Zach C (2018) Contextnet: Exploring context and detail for semantic segmentation in real-time. arXiv:1805.04554
33. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. https://doi.org/10.48550/ARXIV.1505.04597, arXiv:1505.04597
34. Selee B, Faykus M, Smith M (2023) Semantic segmentation with high inference speed in off-road environments. Tech Rep SAE Technical Paper
35. Singh J, Shekhar S (2018) Road damage detection and classification in smartphone captured images using mask r-cnn. arXiv:1811.04535
36. Ventura C, Pont-Tuset J, Caelles S, Maninis KK, Van Gool L (2018) Iterative deep learning for road topology extraction. arXiv:1808.09814
37. Vertens J, Zürn J, Burgard W (2020) Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images. In: 2020 IEEE/RSJ International conference on intelligent robots and systems (IROS), IEEE, pp 8461–8468

38. Vojir T, Šipka T, Aljundi R, Chumerin N, Reino DO, Matas J (2021) Road anomaly detection by partial image reconstruction with segmentation coupling. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 15651–15660
39. Volpi R, De Jorge P, Larlus D, Csurka G (2022) On the road to online adaptation for semantic image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 19184–19195
40. Wang H, Chen Y, Cai Y, Chen L, Li Y, Sotelo MA, Li Z (2022) Sfnet-n: An improved sfnet algorithm for semantic segmentation of low-light autonomous driving road scenes. IEEE Trans Intell Transportation Syst 23(11):21405–21417
41. Xing Y, Lv C, Cao D (2020) Advanced Driver Intention Inference: Theory and Design. Elsevier
42. Xingang Pan Jianping Shi PLXW, Tang X (2018) Spatial as deep: Spatial cnn for traffic scene understanding. In: AAAI Conference on artificial intelligence (AAAI)
43. Zhang L, Yang F, Zhang YD, Zhu YJ (2016) Road crack detection using deep convolutional neural network. In: 2016 IEEE international conference on image processing (ICIP), IEEE, pp 3708–3712
44. Zhao ZQ, Zheng P, Xu ST, Wu X (2019) Object detection with deep learning: A review. IEEE Trans Neural Netw Learn Syst 30(11):3212–3232. https://doi.org/10.1109/TNNLS.2018.2876865
45. Zou Q, Jiang H, Dai Q, Yue Y, Chen L, Wang Q (2019) Robust lane detection from continuous driving scenes using deep neural networks. IEEE Trans Veh Technol 69(1):41–54