Urban Road Segmentation with Transformers*

Bartas Lisauskas[1], Rytis Maskeliūnas[1]

[1]Kaunas University of Technology, Faculty of Informatics, Studentų St.50, Kaunas, Lithuania

Abstract

This paper introduces a transformer-based computer vision system for segmenting different urban road scenes. Detection and understanding of objects in the environment is a critical task for autonomous vehicles or advanced self-driving robots. The approach integrates a MiT transformer-based backbone network for feature extraction with a decoder that incorporates CNN depthwise separable convolution layers, to efficiently fuse features and reduce computational cost. The system detects and separates different objects and environments into 19 semantic classes, as defined by the Cityscapes dataset. The computer vision model consists of 44.61 million parameters and reaches the mean intersection over union of the 73.95% accuracy metric with the Cityscapes dataset. The results gathered demonstrate the good ability of the model to detect different objects and environments in urban road scenes. The proposed computer vision system approach demonstrates the balance between good segmentation accuracy and efficient network structure for more reliable autonomous solutions in complex urban environments.

Keywords

Computer vision, Deep learning, Image processing, Neural networks, Semantic segmentation

1. Introduction

Computer vision is a field of artificial intelligence that teaches computers to understand the world as humans see it. Using deep learning models and digital image data, systems can accurately identify and classify objects in various road scene environments, and based on that information, autonomous systems can make further decisions. Today, computer vision systems help automate processes in various domains. As with any rapidly evolving field, it is increasingly challenging to keep up with the latest knowledge. Computer vision systems use neural networks to perform image processing tasks. One such task is to extract useful information from digital images. Neural networks are applied to object detection, classification, and segmentation tasks. From an engineering perspective, the goal of computer vision is to develop autonomous systems that can perform tasks that human beings do, and often do so more quickly and efficiently.

Image segmentation is one of the most important digital image processing techniques and has been widely used in the automotive and robotics industries. Road scene segmentation is a critical problem when computer vision systems are deployed for autonomous driving, pedestrian detection, and traffic monitoring. In autonomous vehicles, the quality and reliability of computer vision systems are very important for the safety of the driver and other road users. A precise understanding of traffic participants or obstacles is essential to prevent potential accidents, and accurate object detection in environments with many traffic users is a fundamental requirement to achieve safe, efficient, and reliable autonomous driving. However, developing a system that can reach high precision remains a challenging task.

In recent years, the need for accurate segmentation of the road scene has grown significantly, especially within the automotive and robotics industries. Autonomous vehicles are highly dependent on accurate understanding of the surroundings, pedestrians, and obstacles to ensure safe navigation. According to a 2021 investigation conducted in the US by the National Highway Traffic Safety Administration, the results showed that approximately 94% of all car accidents are due to human error [1].

The automotive and robotics industries have a potential market for different computer vision systems. The automotive industry is rapidly moving towards the realization of autonomous vehicle technologies.

*IVUS 2025: Information Society and University Studies 2025, May 15, Kaunas, Lithuania

bartas.lisauskas@ktu.edu (B. Lisauskas); rytis.maskeliunas@ktu.lt (R. Maskeliūnas)

0009-0005-2732-8502 (B. Lisauskas); 0000-0002-2809-2213 (R. Maskeliūnas)

The integration of computer vision systems in self-driving cars is expected to continue to drive substantial market growth. According to Forbes, in 2022, investments in the automotive industry to achieve autonomous driving technologies exceeded 200 billion dollars. In addition, data from 2021 reveal that more than 80 companies in the United States are actively testing more than 1400 different autonomous vehicles. This information underscores the potential and critical importance of computer vision systems in improving the efficiency and safety of autonomous transportation. Autonomous driving systems not only aim to reduce human driving errors, but can also offer benefits such as more convenient travel, increased mobility, lower operating costs, increased road safety, and reduced ecological footprints [2, 3].

This paper presents a computer vision system that is specifically developed for the segmentation task of different road scene environments. The system performs a detailed image analysis by identifying and classifying various objects and environments on the road and generating segmentation masks for each detected element in the digital view.

2. Semantic Segmentation task and Modern Approaches

Computer vision is a field of artificial intelligence that focuses on the recognition and processing of different objects and environments with digital images. Semantic segmentation is the process of partitioning a digital image into different segments based on shared visual patterns, and it is a fundamental and challenging task in the field of computer vision. During the segmentation process, each pixel in the image is identified, and pixels with similar visual characteristics are grouped into the same classes to differentiate objects from each other and from the background environment. During the past decade, the rapid evolution of deep learning and neural network architectures has substantially improved performance in these tasks, and deep learning models have become indispensable for extracting and processing complex information. For deep learning models, architecture and data set are critical components, and the choice and quality of these directly affect the accuracy of the systems in the execution of computer vision tasks [4].

The path to modern semantic segmentation was introduced with early convolutional neural networks (CNNs). In 1989, French scientist Yann LeCun proposed one of the first CNN architectures, it was called LeNet-5. The computer vision model was built for the handwritten digit recognition task. This work not only demonstrated the effectiveness of using CNN architectures for pattern recognition but also paved the way for their application to more complex tasks, such as image classification, object detection, or image segmentation. During the past decade, different CNN architectures have become the backbone of the neural network of different computer vision applications due to their ability to learn hierarchical representations directly from raw pixel data [5].

Although CNN architectures have been dominating the computer vision field for many years, a proposed transformer architecture approach in 2017 by Ashish Vaswani changed the situation [6]. The Vision Transformers (ViTs) architecture demonstrated a new approach to the important feature extraction process by processing images as sequences of patches and using self-attention mechanisms. This demonstrated approach allowed models to capture long-range dependencies and global context without the constraints that CNN architecture had with fixed-size convolutional filters. Compared with many years used CNN architectures, transformer architectures dynamically focus on the most relevant parts of the image, enhancing robustness to variations in scaling, rotation, or occlusions. In addition, transformer-based models can scale better with larger amounts of data, making them a competitive alternative for semantic segmentation tasks, where global scene understanding is critical [7, 8].

Modern semantic segmentation systems now often use these transformer-based methods with conventional CNN techniques. Earlier demonstrated proposals, such as Fully Convolutional Networks (FCNs), used a different approach by replacing fully connected layers with convolutional layers to enable end-to-end pixel-wise prediction. When implementing systems in FCNs, encoder-decoder architectures such as U-net introduced skip connections that recover fine-grained spatial details that were lost during the downsampling process, while models such as DeepLabV3+ further advanced the field by using dilated convolutions and spatial pyramid pooling to capture multiscale context effectively [9, 10].

In addition, recent published approaches have combined the strengths of CNNs and transformers in hybrid architectures. These models use the efficiency of CNNs for local feature extraction together with the transformer's ability to integrate global contextual information using self-attention mechanisms. Using this type of combination allows for a more accurate segmentation process, particularly in complex scenes such as urban road environments where precise object boundaries and contextual information are essential [11].

In summary, the evolution from the proposed first CNN architecture model LeNet-5 to modern transformer-based architectures today shows significant progress in the field of computer vision, performing different computer vision tasks. Today's state-of-the-art models can achieve high-accuracy results with different image segmentation tasks. Transformer-based architecture models has an advantage with improved global context understanding by using self-attention transformer mechanisms. By integrating of these advantages it is possible to reach good results in the semantic segmentation task, which leads to more accurate and efficient computer vision systems.

3. Methods and Dataset

In this section all the details are provided about the structure of architecture, configuration settings for training and evaluating phases of the computer vision model. In addition, more detailed information is provided about the data set that was used for the training and evaluation steps.

3.1. Dataset

For the experiments which were made with the computer vision model, only the Cityscapes dataset was used. This data set is widely known for benchmarking purposes for many computer vision models in image segmentation tasks. The whole data set consists of 5000 high-resolution images, which were recorded in 50 different cities across Germany. Each pixel in each image is annotated in one of 19 semantic classes such as road, vehicles, pedestrians, traffic signs, or sidewalks. The Cityscapes dataset has three different data splits:

- **Training set:** 2975 images that are used in the training process of a computer vision model.
- **Validation set:** 500 images that are used during the training process to evaluate the model and monitor performance during the training phase.
- **Test set:** 1525 images that are used for the final evaluation process phase, to gather information on the accuracy metrics of the trained model with unseen images.

Figure 1: Sample images from Cityscapes dataset across different Germany cities.

In Figure 1, we can see the images provided from the Cityscapes dataset with different landscape scenes in urban areas. All information on the data set used in the training and evaluation phases of the model and the full list of semantic classes with annotation examples is publicly available at https://www.cityscapes-dataset.com/.

3.2. Configuration Details

The proposed computer vision transformer-based road scene segmentation model is implemented using the mmsegmentation framework codebase. The model consists of an encoder-decoder architecture and uses MiT b3 configuration settings in the encoder module [8]. The system encoder module is pre-trained on ImageNet-1k dataset for better feature extraction task. To improve model accuracy and generalization, in the training process, data augmentation techniques were applied using only a Cityscapes dataset. In the training process, additional functionality was used for random horizontal flipping, cropping, and scaling. The crop size of 512 x 512 pixels was chosen during the training phase. At inference time, the entire image testing strategy was used to generate segmentation predictions. The computer vision model was trained using an AdamW optimization algorithm with an initial learning rate set to 0.00005. Due to GPU resource constraints, a batch size of 1 image was used in the training process. The training schedule was set for 160 000 iterations. The performance of the model was evaluated using the widely used mean intersection-over-union (mIoU) accuracy metric.

3.3. Decoder

The decoder in the system receives four different sets of feature maps with different resolution from the encoder after the feature extraction process. Later these feature maps are projected into a 256 lower dimensional embedding space by using multi-layer perceptron modules. After this projection, different features are resized to the same spatial resolution and concatenated along the channel dimension. The combined feature map is processed by a depth-wise separable convolutional layer. This layer first performs a spatial convolution on each channel independently and then applies a pointwise 1x1 convolution to fuse the information across different channels. Using this approach, an efficient multiscale feature fusion process is possible. Finally, in the decoder part, a dropout layer is applied for better regularization, and a 1x1 convolutional layer produces the final output of the segmentation map.

4. Results

This section provides results of the computer vision road scene segmentation model. In the following, information is provided on the results of the model training process, global accuracy, and finer per-class accuracy metrics. Qualitative visual results are also provided to better understand how the system is capable of detecting different objects and environments with the road scene images.

4.1. Training Process

During the computer vision model training process, the training schedule was set to 160 000 iterations. No further training process was conducted beyond 160 000 iterations line. Model training was performed using AdamW optimizer and CrossEntropyLoss function. In Figure 2, the training phase graph is provided to better understand how the model reduced the loss parameter during the training cycle.

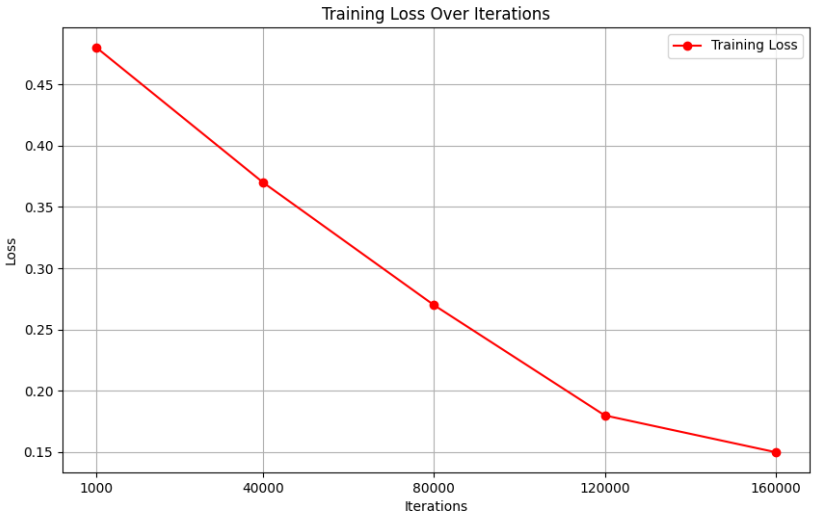

Figure 2: Graph with model training loss values over 160 000 iterations.

The information provided in Figure 2 is gathered from the training logs. It is clearly visible that the model loss parameter was steadily decreasing throughout the training cycle, up to the 160.000 iteration line. From the initial loss of 2.12 to the final phases, where it reached approximately 0.15 at the end of the training cycle. The model best performance was reached at the 160 000 iterations line.

4.2. Per-Class Performance

The computer vision model segments scenes into 19 classes according to the standard of the cityscape data set. In the table below, an intersection over union and accuracy metrics is provided for each class.

Table 1
Per-Class Performance Metrics

Class	IoU (%)	Accuracy (%)
Road	96.80	98.63
Car	93.94	97.23
Sky	92.36	95.07
Vegetation	92.19	96.86
Building	90.84	96.66
Person	80.68	88.73
Bus	77.65	87.00
Sidewalk	76.82	86.02
Traffic Sign	76.28	82.97
Bicycle	75.85	86.91
Truck	74.18	80.17
Traffic Light	67.62	78.79
Train	66.04	79.17
Motorcycle	64.85	76.83
Terrain	60.89	65.92
Pole	60.15	68.87
Rider	58.70	77.80
Wall	50.81	56.66
Fence	48.41	53.78

From the data provided in Table 1, it is clear that the system can detect objects and elements of the

environment, such as roads, cars, the sky, vegetation, and buildings, with a high precision above 90%. Road and car classes reach the highest accuracy metrics. Still, there is room for improvement in the future with other vehicle classes that have medium accuracy results. In addition, classes with lowest accuracy metrics are the most difficult to detect for many computer vision models. It is a challenging task when a distinction is needed between the wall and the fence, but the accuracy can be improved with additional data or the modified architecture with the bigger neural network.

4.3. Global Metrics

To obtain the overall performance accuracy results, the computer vision model was evaluated on the Cityscapes validation set with 500 images to get the global accuracy results. The following global accuracy metrics were collected:

- **Mean IoU (mIoU):** 73.95%
- **Mean Accuracy (mAcc):** 81.79%
- **Overall Accuracy (aAcc):** 95.07%

The metrics provided demonstrate that the computer vision model, which has 44.61 million parameters, is capable of reaching a global 73.95% mean intersection over the union accuracy metric. This metric reflects the average overlap between the predicted segments and the ground-truth values in all classes. It is used primarily to evaluate the performance of many computer vision models. Based on the data provided, we can also see that the computer vision model reached the mean accuracy value (mAcc) of 81.79%. This metric represents the average classification accuracy per pixel for each class. In addition, the overall accuracy metric (aAcc) of 95.07% was reached. This metric reflects the ratio of correctly classified pixels to the total number of pixels in the Cityscapes validation set.

4.4. Qualitative Analysis

Figure 3 demonstrates an example of computer vision system capability to detect different objects and environment in the images of the road scene. In the provided figure, the original images are on the left side, and the results after the segmentation process are on the right side.

Figure 3: Example images with original one on the left, and segmented result on the right.

From this side-by-side images comparison, we can see that the system is capable to reach good accuracy when detecting different objects and environment on the road scene images in close distance. Objects or environments such as cars, roads, or pathways are detected with high accuracy. However, it is worth mentioning that some challenges remain in accurately segmenting distant objects, which leaves a potential direction for future improvements.

4.5. Computational complexity

As shown in Table 2, computer vision model with 44.61 million parameters scales from 41.84 GFLOPs at 512×512 resolution to 238.24 GFLOPs at 1024×1024 resolution, illustrating the trade-off between computational cost and input resolution.

Table 2
Computational Complexity at Different Input Resolutions

Input Resolution	GFLOPs	Parameters (M)
512 × 512	41.84	44.61
768 × 768	110.76	44.61
1024 × 1024	238.24	44.61

5. Conclusion

The experimental results demonstrate that this transformer-based approach for road scene segmentation achieves good accuracy results when detecting different objects and environment details with diverse urban scenes, while keeping the architecture relatively lightweight. With 44.61 million parameters in the network, the computer vision model can reach the accuracy metric of 73.95% mIoU with the Cityscapes validation set. The computer vision model uses the MiT transformer-based encoder as the backbone component for feature extraction, and the CNN-based decoder incorporating depthwise separable convolution layers to efficiently fuse features and reduce computational resources. The visual results provided show that the system can accurately detect major objects and environmental elements in a close distance. However, lower-accuracy classes with small or more distant objects indicate the future area for improvements, suggesting that additional training data or architecture modifications may enhance system performance in these challenging cases. To further improve segmentation accuracy in future work, the computer vision model can be trained on additional road scene datasets such as Berkeley Deep Drive, Mapillary Vistas, or CamVid, which together have tens of thousands of varied urban driving images. The current results were obtained using only the 2975 images from the Cityscapes training set, without any additional data. On the architecture side, it is possible to replace the current CNN-based decoder with a custom attention-based fusion module designed to preserve fine spatial details and capture long-range context. Early experiments indicate that an alternative approach can even use the smaller feature extraction transformer network, reducing overall model size while improving global segmentation accuracy results.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] Zohaib Rafique, Improving efficiency of computer vision for autonomous vehicles (2020). URL: http://rgdoi.net/10.13140/RG.2.2.11761.92009. doi:10.13140/RG.2.2.11761.92009.

[2] Q. Sellat, S. Bisoy, R. Priyadarshini, A. Vidyarthi, Intelligent semantic segmentation for self-driving vehicles using deep learning, 2022. URL: https://www.researchgate.net/publication/357907010_Intelligent_Semantic_Segmentation_for_Self-Driving_Vehicles_Using_Deep_Learning.

[3] Forbes, Autonomous vehicles and their impact on the economy, 2022. URL: https://www.forbes.com/councils/forbestechcouncil/2022/02/14/autonomous-vehicles-and-their-impact-on-the-economy/.

[4] X. Zhou, W. Gong, W. Fu, Application of deep learning in object detection, 2017. URL: https://www.researchgate.net/publication/318035834_Application_of_deep_learning_in_object_detection.

[5] F. Sultana, A. Sufian, P. Dutta, Evolution of image segmentation using deep convolutional neural network: A survey, Knowledge-Based Systems 201–202 (2020) 106062. doi:10.1016/j.knosys.2020.106062.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. arXiv:1706.03762, [arXiv:cs.CL/1706.03762].

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[8] E. Xie, W. Wang, Z. Yu, X. Lei, P. Luo, Segformer: Simple and efficient design for semantic segmentation with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1202–1211.

[9] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, 2015, pp. 234–241.

[10] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, 2018, pp. 801–818.

[11] K. Yuan, S. Guo, Z. Liu, A. Zhou, F. Yu, W. Wu, Incorporating convolution designs into visual transformers, arXiv preprint arXiv:2103.11816 (2021).