machines

MDPI

Article

Efficient Transformer-Based Road Scene Segmentation Approach with Attention-Guided Decoding for Memory-Constrained Systems

Bartas Lisauskas † and Rytis Maskeliunas *,†

Software Engineering Department, Faculty of Informatics, Kaunas University of Technology, 44249 Kaunas, Lithuania; bartas.lisauskas@ktu.edu

* Correspondence: rytis.maskeliunas@ktu.lt

† These authors contributed equally to this work.

Abstract: Accurate object detection and an understanding of the surroundings are key requirements when applying computer vision systems in the automotive or robotics industries, namely with autonomous vehicles or self-driving robots. A precise understanding of road users or obstacles is essential to avoid potential accidents. Due to the presence of many objects and the diversity of the environment, the segmentation of the road scene remains a challenging task. In our approach, a Transformer-based backbone is employed for robust feature extraction in the encoder module. In addition, we have developed a custom decoder module in which we implement attention-based fusion mechanisms to effectively combine features. The decoder modification is specifically designed to maintain fine spatial details and enhance the global context understanding, setting our method apart from conventional approaches that typically use simple projection layers or standard query-based decoders. The implemented model consists of 17.2 million parameters and achieves competitive performance, with a mean intersection over union (mIoU) of 76.41% on the Cityscapes validation set. The results gathered indicate the ability of the model to capture both the global context and fine spatial details that are critical to the accurate segmentation of urban scenes. Furthermore, the lightweight design makes the approach suitable for deployment on memory-limited devices.

Keywords: computer vision; deep learning; image processing; neural networks; semantic segmentation

Academic Editor: Raul D. S. G. Campilho

Received: 28 March 2025
Revised: 16 May 2025
Accepted: 26 May 2025
Published: 28 May 2025

Citation: Lisauskas, B.; Maskeliunas, R. Efficient Transformer-Based Road Scene Segmentation Approach with Attention-Guided Decoding for Memory-Constrained Systems. *Machines* **2025**, *13*, 466. https://doi.org/10.3390/machines13060466

1. Introduction

Neural networks are used to perform image processing tasks in computer vision. One such task is to extract useful information from digital images. These networks are used to perform object detection, classification, and segmentation tasks. From an engineering perspective, the goal of computer vision is to create autonomous systems that can perform tasks that humans do, and, in many cases, to do so faster and more efficiently.

Autonomous driving is a revolutionary technology that is likely to have a very significant impact on people's daily lives in the future. Image segmentation systems provide autonomous cars with a view of the surrounding world and are critically important in achieving safe autonomous vehicle driving. According to a study conducted by the US National Highway Traffic Safety Administration, 94% of all traffic accidents are caused by human error [1]. The realization of autonomous car driving aims to solve this serious problem regarding car accidents. Since autonomous systems are programmed to drive

efficiently and safely, they reduce—and sometimes even eliminate—the need for human driving, thus eliminating the aforementioned human error [2]. The automotive industry is a very promising area for the development of computer vision solutions in autonomous vehicles. The more accurate the image segmentation process and the less the time that it takes, the more accurately autonomous vehicles will understand their surroundings and the more likely they will be to make safer decisions. Providing autonomous vehicles with a view of the world around them using computer vision solutions can offer many benefits, such as increased road safety, lower costs, more comfortable travel, greater mobility, and smaller ecological footprints [3].

With the rapid development of autonomous driving technology, an accurate visual understanding of the surroundings is crucial to ensure road safety and efficient autonomous vehicle navigation. The accurate detection and classification of objects such as pedestrians, vehicles, and traffic signs is essential because it directly impacts the safety and efficiency of autonomous driving systems. Image segmentation is one of the most important processes in digital image processing and has been widely used in the automotive industry and robotics in recent years. The field of computer vision, which is related to artificial intelligence, has seen great progress in the past decade, and today's computer vision systems can recognize visual data more quickly than humans. In the field of computer vision, the semantic segmentation task remains one of the most challenging ones. The assignment of class labels to each pixel, and the classification task at the pixel level, is a key approach that enables vehicles to distinguish between different objects, roads, pathways, pedestrians, and the remaining environmental elements of the road.

As computer vision systems are widely applied in the automotive industry, the same methods can be equally applied to a broader spectrum of autonomous systems in the robotics or aerospace industries. For example, during the NASA Mars 2020 mission, to land the Perseverance rover, a computer vision system was used for hazard and obstacle detection, enabling it to land safely by autonomously selecting the safest landing position [4]. Autonomous robots play an important role in various applications, where accurate perception and effective path planning are key requirements in achieving full autonomy. The perception component is dedicated to understanding the surrounding environment, enabling these robots to make informed decisions [5]. To obtain different fully autonomous vehicles in the future, accurate perception systems are indispensable, ensuring the reliable monitoring and interpretation of complex, dynamic environments [6]. Moreover, the use of semantic segmentation techniques can offer higher precision in detecting urban environments [7]. Semantic data can help to reduce the dependence of a robot on raw sensor input and external signals such as GPS by providing useful environmental information for navigation [8]. In fact, in this work, the demonstrated approach to segmenting road scene images from a driver's perspective can also be used in other applications, such as delivery or taxi robots and various service robots, which need to accurately understand their surroundings when navigating urban environments and avoiding obstacles. Among the most notable applications is parcel delivery, with autonomous delivery robots emerging as key components in the solutions to different delivery challenges [9].

Autonomous robotic platforms operate in complex urban environments and require precise understanding and reliable path planning, while having limited processing resources. Many computer vision models often require extensive computational power, and a Transformer-based system can offer a lightweight alternative, making it suitable for deployment across different autonomous robotic platforms. The ability of the autonomous robot to understand its working environment is the basis for the solution of more complicated problems [10]. The results described in this work show that the proposed approach, which is effective in detecting objects at a close distance, can be particularly advantageous for

use in autonomous self-driving robots. These robots generally operate at lower speeds compared to cars. While a car traveling at 60 km/h covers 100 m in just a few seconds, a robot may take up to 30 s to travel the same distance. Thus, for the robot, distant objects in the environment are not as critical as they are for an autonomous car.

Autonomous systems and their effectiveness are highly dependent on the ability to navigate a complex and unstructured environment [11]. Recent advances have further improved robotic navigation by enabling real-time performance in critical tasks such as environment perception, obstacle detection, obstacle avoidance, path planning, and path tracking [12]. Within path planning, obstacle avoidance is a crucial task in robotics, as autonomous robot operation requires that they reach their destinations without collision [13]. Effective object detection strategies face static obstacles, such as infrastructure or parked vehicles, and dynamic obstacles, including pedestrians and moving vehicles, with each presenting unique challenges to safe navigation. To prevent a potential collision with pedestrians, an accurate detection system enables autonomous robots to intervene early, reducing the risk of accidents [14]. Thus, an accurate computer vision system is essential to ensure the safe and efficient operation of autonomous robots, especially in dynamic urban environments. The main solutions currently focus on understanding the environment through visual information using various computer vision techniques, machine learning, and algorithms [15]. The use of computer vision techniques enables robots to autonomously understand their surroundings, adapt their trajectories, and perform tasks such as maintenance or exploration without human intervention. In addition, for autonomous robots to navigate in urban environments, it is very important to navigate on designated paths, such as footpaths or sidewalks, and avoid areas such as grass to ensure both safety and social conformity. Robots deployed in public environments as autonomous delivery robots operate in spaces in which people live and work [16]. For example, package delivery robots must be able to identify and follow safe and appropriate routes that allow them to navigate autonomously in a manner that is not only efficient but also socially acceptable to the people sharing the environment [17]. Since robots increasingly share space with humans in everyday environments, ensuring safety is paramount [18]. Computer vision applications help to improve the efficiency of transportation systems, increase their levels of intelligence, and improve traffic safety [19]. Moreover, the integration of computer vision with robotics holds significant promise for environmental protection efforts by enabling more efficient resource management and reducing urban environmental impacts [20]. In general, the development of robotic vision is more than just a scientific curiosity or a passing trend. It marks a significant step forward in what machines can do, and this is expected to strongly impact our daily lives [21].

1.1. Evolution of Deep Learning Architectures

In 1989, the French scientist Yann LeCun created one of the first convolutional neural networks, which was called LeNet-5. This neural network was designed for a handwritten digit recognition task. The emergence of the LeNet-5 architecture paved the way for the continued success of convolutional neural networks in performing high-complexity computer vision tasks, and it encouraged researchers to explore the capabilities of convolutional neural networks in performing image segmentation tasks [22]. Among the different deep learning models, convolutional neural networks have achieved excellent performance in different computer vision tasks, such as image classification, object detection, or digital image segmentation. Convolutional neural networks have become one of the most successful and widely used deep learning architectures in computer vision tasks over the past decade.

While, for many years, the convolutional neural network architecture was the state of the art in computer vision tasks, this situation changed in 2017, when the Transformer

architecture was introduced. Recently, Vision Transformers have emerged as a competitive alternative to the long-standing convolutional neural network for computer vision tasks. The Transformer neural network was developed and introduced by the scientist Ashish Vaswani in 2017 [23]. Today, these neural networks compete with state-of-the-art convolutional neural network architectures in terms of efficiency and accuracy.

During research, it was found that it is possible to create accurate computer vision models without using convolutional layers as the main components. One such idea is to use the Vision Transformer neural network architecture for feature extraction; it applies an attention-based mechanism to input images and can achieve competitive efficiency in performing the computer vision semantic segmentation task [24]. The use of Transformer-based architectures over traditional convolutional neural networks as the main component for the feature extraction part is due to the Transformer architecture's improved capacity for global context understanding. The use of self-attention mechanisms in the Transformer architecture allows the system to capture relationships between distant regions of the image. This approach enables the network to integrate information throughout the image. It is especially beneficial for semantic segmentation tasks that require full-scene understanding. Transformers process images as sequences of patches, which gives the opportunity to model both global and local interactions without being constrained by convolutional fixed-size kernels. Thus, the results can be more robust, particularly in situations where the boundaries of the object and contextual cues are crucial [23].

Another important quality of the Transformer architecture is its robustness to variations. Attention mechanisms can dynamically focus on the most relevant parts of the image, making them more resilient to different changes, like rotation, scaling, or occlusion. This adaptability in complex road scenes is beneficial when noise or local variations hinder the performance. In addition, Transformer architectures show improved performance as the data volume increases, making them highly scalable to large datasets [25].

1.2. Modern Approaches to Semantic Segmentation

Semantic segmentation is the core task and remains one of the most challenging in computer vision; it involves the classification of every pixel in a digital image into a corresponding class. It gives the complete context of the scene by incorporating the categories, locations, and shapes of all elements in the scene, including the background [26]. However, it is more challenging and usually more time-consuming than object detection and requires more advanced techniques and more high-quality annotated training data [27]. Over the years, modern approaches have been introduced to perform semantic segmentation tasks—from the early fully convolutional networks that introduced pixel-level predictions to advanced encoder–decoder architectures, which integrate multiscale feature fusion and context-aware processing [23,28]. Recently, the incorporation of attention mechanisms and Transformer-based models have offered new capabilities to capture long-range dependencies and global contextual information, pushing the performance to new heights. The progression of modern techniques is reshaping the state-of-the-art performance in the computer vision field. Table 1 presents a systematic summary of these methods in the field of computer vision image segmentation, including their core architectures, benchmark datasets, and mIoU results.

Fully convolutional networks (FCNs) were among the first breakthroughs in the semantic segmentation task. By replacing the fully connected layers of traditional CNNs with convolutional layers, FCNs enabled end-to-end pixel-wise prediction, which allowed models to generate spatially dense outputs by upsampling the low-resolution feature maps from the convolutional layers. An evaluation of FCNs shows that this method can effectively segment images, establishing a solid baseline for subsequent models [28].

Table 1. Comparison of different architectures.

Method	Architecture	Dataset	Result (mIoU)
FCN	Fully Convolutional Network	PASCAL VOC	67.5
UNet	Encoder–Decoder	Cityscapes	83.6
DeepLabV3+	ASPP + Decoder	Cityscapes	79.6
SETR	Transformer	Cityscapes	76.7
Mask2Former	Hybrid CNN–Transformer	Cityscapes	83.3

Building on the FCN framework, encoder–decoder architectures such as U-Net have become a popular approach to segmentation tasks. U-Net uses a symmetric architecture in which an encoder gradually reduces the spatial dimensions while capturing semantic features, and a decoder progressively upsamples the features to produce a prediction at the pixel level. Skip connections are used between the corresponding encoder and decoder layers to help recover spatial details lost during downsampling. The demonstrated U-Net approach is particularly effective for tasks that require the precise detection of objects, such as the segmentation of the road scene [29].

The DeepLabV3+ model uses a dilated convolution-based approach to better capture multiscale contextual information without decreasing the resolution. Using this type of approach increases the receptive fields of convolutional filters, without additional parameters or a reduction in the spatial resolution. Using spatial pyramid pooling and an encoder–decoder structure, the DeepLabV3+ model can fuse features from multiple scales, which is critical for the complex and varied environments encountered in road scenes [30].

The adaptation of Transformer architectures is another trend. Although originally developed for natural language processing, they were later used in the domain of image segmentation. Models such as SETR or SegFormer use a self-attention mechanism to capture long-range dependencies and global contexts. The images in these models are partitioned into patches and then processed as token sequences, enabling the network to model relationships between distant regions. The global modeling capability is especially advantageous for complex scenes, such as urban road environments, where contextual signals are critical [31,32].

The integration of convolutional neural networks with Transformer modules is a promising direction, combining the strengths of both architectures. Hybrid computer vision models often use convolutional layers for efficient local feature extraction and are combined with additional Transformer layers to capture global context details through self-attention mechanisms. The use of this approach can lead to improved segmentation performance, especially in scenarios where both detailed spatial information and a wider contextual understanding are necessary. One work incorporating the convolution technique into Vision Transformers shows that the combination of these two methodologies can yield competitive performance in segmentation tasks [33].

The main purpose of this study is to develop a Transformer-based semantic segmentation approach, specifically designed for the road scene segmentation task, which takes advantage of the latest advances in Vision Transformers. This paper is organized as follows. Section 2 describes our proposed semantic segmentation approach. Section 3 presents the experimental results, and Sections 4 and 5 conclude the document with a discussion and future directions.

2. Materials and Methods

In this section, we provide the details of the implementation, the structure of the model, and the configuration settings used to train and evaluate our Transformer-based road scene segmentation system. Moreover, we describe the components of other state-of-the-art computer vision models, the components of our approach, and how our approach is different, as well as presenting some advantages over the previously used implementations. In addition, we describe the dataset in more detail and explain how the data are divided into training, validation, and test set splits. In further sections, we also explain the internal structure of the encoder module, as well as the decoder part of the system with attention-based fusion mechanisms for the extraction and combination of multiscale features. Details of each component's structure are provided in further subsections.

The current state-of-the-art computer vision approaches for road scene segmentation have demonstrated significant success by using different types of architectures and achieving competitive results. Methods that use Transformer-based architectures typically rely on standard feature fusion, and they often use simple projection layers to merge different features from different Transformer-based encoder resolutions. Computer vision models have also been introduced, such as RoadFormer, which uses a query-based decoder module to iteratively update query features for mask predictions [34]. Although these approaches work well in many cases, they can struggle to maintain important fine-grained details when working with complex road scene environments.

Our decoder module processes the data with attention-based fusion mechanisms, providing competitive results in the road scene segmentation task in complex urban scenes compared to other lightweight computer vision models that have a similar number of parameters in the network. The mechanism in the decoder part uses specialized attention blocks to progressively fuse low-resolution decoder features with high-resolution features from the encoder part. This can effectively emphasize the most relevant spatial information at multiple scales. Using this approach, it is possible to ensure that the fine-grained features of the encoder can be integrated together with the high-level semantic features of the decoder, giving accurate and competitive segmentation results when compared with other introduced approaches. Moreover, instead of using traditional normalization techniques, the model uses group normalization, which normalizes over the feature channels instead of on the batch dimension. This can give better stability in the training process, particularly in scenarios where the batch size may be different or in cases of high-resolution feature fusion, where traditional methods such as BatchNorm may not be as effective.

It is also worth mentioning that, while many existing approaches rely on standard feature fusion techniques with different scales, this approach can provide better control over which features are fused via the use of attention blocks, which allow it to adapt more effectively to the different complexities of road scene segmentation. Using this approach gives competitive results in terms of segmentation accuracy, particularly in challenging road scene environments with fine details.

2.1. Dataset

All experiments were performed using the widely known Cityscapes dataset, which is used for the benchmarking of computer vision models in semantic segmentation tasks with different urban scene images. The dataset consists of 5000 high-quality, finely annotated, pixel-level, high-resolution images recorded in 50 different cities across Germany. Each pixel in every image in the dataset is annotated to one of 19 semantic categories, representing typical elements of the urban road scene, such as roads, cars, pedestrians, sidewalks, or traffic signs. These 5000 images are split as follows.

- Training set, consisting of 2975 images: This subset of the data is used to learn the parameters of the model during the training process.
- Validation set, consisting of 500 images: This part of the data is used during the training process to configure the model, adjust the hyperparameters, and monitor the performance of the model during the training phase.
- Test set, consisting of 1525 images: The last part of the data is used for the final evaluation phase to evaluate the precision of the model in an unseen dataset.

In Figure 1, we provide some examples from the Cityscapes dataset, demonstrating different environments and urban conditions. All details of the dataset used in the model training process and the final evaluation, including the annotated classes and examples of annotations, are publicly available at https://www.cityscapes-dataset.com/ (accessed on 11 March 2025).

Figure 1. Sample images from the Cityscapes dataset, illustrating the diversity of urban scenes in different cities across Germany.

2.2. Implementation Details

Our Vision Transformer-based road scene segmentation model is implemented using the `mmsegmentation` framework codebase and uses an MiT as a backbone network, with attention-based fusion mechanisms in the decoder module. The system's encoder module is pre-trained on the ImageNet-1k dataset for the extraction of robust visual features, while the system's decoder module is randomly initialized to learn task-specific upsampling. To improve the model's robustness and generalization, in the training process, data augmentation techniques were applied using the Cityscapes dataset. We used additional measures of random horizontal flipping, scaling, and cropping. The crop size of 768 $\times$ 768 pixels was chosen during the training phase. At inference time, a sliding-window strategy was used to generate full-size segmentation predictions.

The computer vision model was trained using the AdamW optimization algorithm with an initial learning rate of 0.00005. In addition, the polynomial decay learning rate schedule with the power parameter set to 1.0 was used, and the linear warm-up phase consisted of up to 1500 iterations at the start of the training process. Due to GPU resource constraints and the high-resolution input images, a batch size of 1 image was used in the training process. The training schedule was set for 160,000 iterations, and it was later extended to 200,000 iterations to explore the possibility of further model accuracy improvements. In the end, it was clear that no significant improvements would be achieved in the training cycle after 160,000 iterations. The best-performing model checkpoint was

taken at 156,000 iterations. Model performance was primarily evaluated using the widely used mean intersection over union (mIoU) metric.

2.3. Encoder

In the system, the encoder backbone network is implemented using the MiT architecture, with the configuration settings named "mit b1" [32]. Using these configuration settings, the input image is processed through a patch embedding module that divides the image into smaller patches and projects them into a feature space. The first stage of the network uses an OverlapPatchEmbed module with a fixed patch size of 7×7 and a stride parameter of 4. The module is responsible for mapping each 7×7 image patch to a 64-dimensional embedding space and generating the initial feature map. The patch embedding can be written as

$$X_0 = \text{Conv}_{7\times7}(I)$$

where I is the input image. The output after this processing stage is then flattened and normalized. The subsequent stages of the encoder network use patch embedding modules with a set patch size of 3×3 and a stride parameter of 2. These network encoder stages progressively reduce the spatial resolution while increasing the number of channels. Using the "mit b1" configuration, the embedding dimensions are preset to [64, 128, 320, 512] for the four network stages. The encoder progressively extracts hierarchical features with increasing channel numbers:

- Stage 1 (c1): 64 channels;
- Stage 2 (c2): 128 channels;
- Stage 3 (c3): 320 channels;
- Stage 4 (c4): 512 channels.

Each of the stages is further processed by a sequence of Transformer blocks. Using the b1 network configuration, the depth of each block is [2, 2, 2, 2] for the four stages. In addition, with each Transformer block, the multi-head self-attention mechanism is used. The attention computation can be written as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

where Q is a query, K is the key, and V is the value received by linear projections of the input features. At each stage, the number of attention heads used is [1, 2, 5, 8], and the MLP within each block expands the dimensions of the feature by a factor of 4.

In each stage, spatial reduction ratios are also used with values [8, 4, 2, 1] to downsample the spatial dimensions during the self-attention processing step. In this way, we reduce the computational load while maintaining an essential global context.

After the processing step is performed through different Transformer blocks, each stage uses a final normalization layer, and the resulting feature maps are reshaped in the [B, C, H, W] format. The final outputs of all four stages are named c1, c2, c3, and c4. These outputs capture a rich hierarchy of features that combine fine spatial details from the early stages of the network with high-level semantic information from the deeper stages of the network. In a further step, all generated feature maps from each encoder network stage are passed to the decoder module for further processing.

2.4. Decoder

The decoder module approach combines the features from the encoder using an attention-based mechanism. In this subsection, we propose an attention-guided decoding

module that fuses multiscale Transformer features via hierarchical attention and residual convolutional fusion. The decoder design allows the rich semantic information from deeper layers to be combined with the fine spatial details of shallower layers; in this way, it can achieve good accuracy and spatial precision. Figure 2 shows the complete model architecture, including the encoder and the attention-guided decoding module.

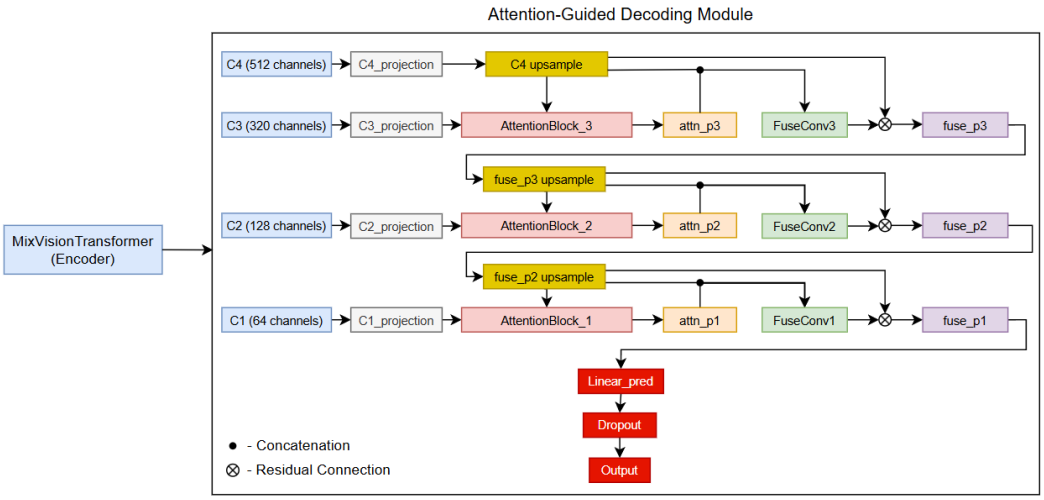

Figure 2. Architecture diagram of the computer vision model with the proposed attention-guided decoding module.

Using the MiT as a backbone feature extraction network, the encoder part produces a set of feature maps at different resolutions, typically named c1, c2, c3, and c4, from the shallowest to the deepest network stage. Each feature map captures information at a different scale. Although deeper features contain more semantic information, shallow features preserve detailed spatial information, which is critical for accurate segmentation in complex road scenes.

The decoder first projects each of the feature maps to a common embedding space with a fixed number of channels. This process is achieved by using a 1×1 pixel convolutional layer, which is later followed by the normalization of the feature maps across channels to help improve the consistency of the feature representations. After this, the ReLU activation function is used to introduce non-linearity, allowing the network to learn more complex patterns. The mentioned process not only standardizes the channel dimensions across different scales but also improves the stability of the whole training process. Mathematically, this projection can be described as follows:

$$p_i = \mathrm{ReLU}\Big(\mathrm{GroupNorm}\big(\mathrm{Conv}_{1\times1}(c_i)\big)\Big), \quad i \in \{1,2,3,4\}.$$

In addition, in the decoder part, we use an attention-based fusion mechanism. When using separate 1×1 convolutional and normalization layers for upsampled decoder features and corresponding encoder features, we generate an attention map that highlights the most relevant spatial regions. This selective weighting approach helps the decoder to focus on important details. As a result, the integration of rich semantic features with fine spatial details helps to achieve a more effective and accurate segmentation process in different road scenes. The attention block in the decoder part is the main component that serves to fuse features from different scales. In this decoder implementation, the fusion process is performed in a top-down approach. In the first stage, the deepest feature map p4 is first upsampled to match the spatial resolution of p3. After this, the attention block receives the upsampled p4 feature map (acting as a gating signal) and the p3 feature map from the encoder backbone. Later, the attention block projects both received inputs using a

1×1 convolutional layer followed by a normalization layer. Specifically, the projections can be formulated as follows:

$$g_1 = \text{GroupNorm}\Big(\text{Conv}_{1\times1}(g)\Big), \quad x_1 = \text{GroupNorm}\Big(\text{Conv}_{1\times1}(x)\Big),$$

where g is the gating signal (upsampled p4) and $\times$ is the encoder feature (p3). The resulting outputs are then summed, activated by the ReLU function, and further processed by another 1×1 convolutional layer with the sigmoid activation function to produce an attention map:

$$\psi = \sigma\Big(\text{GroupNorm}\big(\text{Conv}_{1\times1}\big(\text{ReLU}(g_1 + x_1)\big)\big)\Big).$$

The produced map selectively weights the p3 features, suppressing less relevant regions. The weighted encoder feature is concatenated with the upsampled p4, and a convolutional module refines this fusion. Additionally, a residual connection is added by introducing the upsampled p4.

In addition, the fused features from the previous stage are upsampled in the same way to match the resolution of p2 and fused with p2 using an analogous attention-guided procedure. The same process is repeated once again with the highest-resolution p1 feature map to produce the final fused representation.

In this decoder approach, there are combined residual connections at each fusion stage. These connections add the upsampled features back into the fused output, and, in this way, we ensure that essential spatial details are kept throughout the upsampling process. By using residual connections at each stage, we stabilize the training process by maintaining a smoother gradient flow. In addition, this allows the model to take advantage of the refined fused features and the original upsampled signals, thereby improving the overall robustness of the segmentation process.

After the progressive fusion step, a dropout layer is applied to the refined feature map for regularization. After this step, the final 1×1 convolutional layer is used to project the features to the preset number of segmentation classes, generating the final segmentation output. The final prediction can be written as

$$\text{Output} = \text{Conv}_{1\times1}\Big(\text{Dropout}(\text{Fused Feature})\Big)$$

This decoder approach can be effective for road scene segmentation tasks due to its ability to preserve spatial precision by maintaining high-resolution feature maps and integrating fine-grained details using the attention mechanism, which ensures the accurate distinction of road boundaries and different objects. The attention mechanism used in the decoder module works as a gating function that merges context-rich deep features with detailed shallow features. In doing so, it emphasizes the spatial regions that are most relevant. In addition, the use of residual connections throughout the fusion process helps with training stability and convergence, even when the model incorporates multiple upsampling stages.

Overall, this decoder approach proves robust and efficient in fusing multiscale features, and, when combined with the MiT backbone for feature extraction, it can achieve competitive accuracy in the road scene semantic segmentation task.

3. Results

In this section, we present the experimental results for our Vision Transformer-based road scene segmentation system. First, the training process is described, followed by additional details regarding the performance metrics per class and the global metrics of the

system. At the end of this section, we provide the visualization results for the road scene environment to demonstrate that the system is capable of detecting different objects on the road.

3.1. Training Process

As described in more detail in Section 2.2, the model training schedule was preset to 160,000 iterations. Moreover, to examine whether additional training could lead to better accuracy metrics, the training process later was extended to 200,000 iterations. An analysis of the training log from the `mmsegmentation` framework yielded the following observations.

- Initial Training Phase: The training phase from the beginning showed a high decoder loss parameter of 2.43. During the full 160,000 iterations of training, the model loss parameter decreased as we approached 160,000 iterations. This behavior demonstrates that the computer vision model was continuously learning and improving until the end of the training cycle, while it achieved a loss parameter of approximately 0.11 in the final iterations.
- Iteration 156,000 as the Final Checkpoint: At this point in the training cycle, the model demonstrated the highest precision metrics. From the training log review, the decoder loss parameter was recorded at 0.1161 and we achieved the highest accuracy values. This confirms that the highest performance was reached at this point, leading us to select this checkpoint as our final model.
- Extended Training Experiment: To determine whether the computer vision model could achieve better accuracy, an extended training process of up to 200,000 iterations was carried out. After this, the results of the training log showed that additional training did not give accuracy improvements, indicating that further training beyond the 160,000 point is not effective.

Figure 3 illustrates the training convergence curves of our proposed model compared to traditional segmentation architectures, specifically U-Net+FCN and DeepLabV3+. Our model demonstrates a consistent decrease in loss, reaching approximately 0.11 at 160,000 iterations. In contrast, U-Net, trained up to 160,000 iterations, and DeepLabV3+, trained up to 80,000 iterations, exhibit higher loss values at corresponding iteration points, indicating slower convergence and less efficient learning.

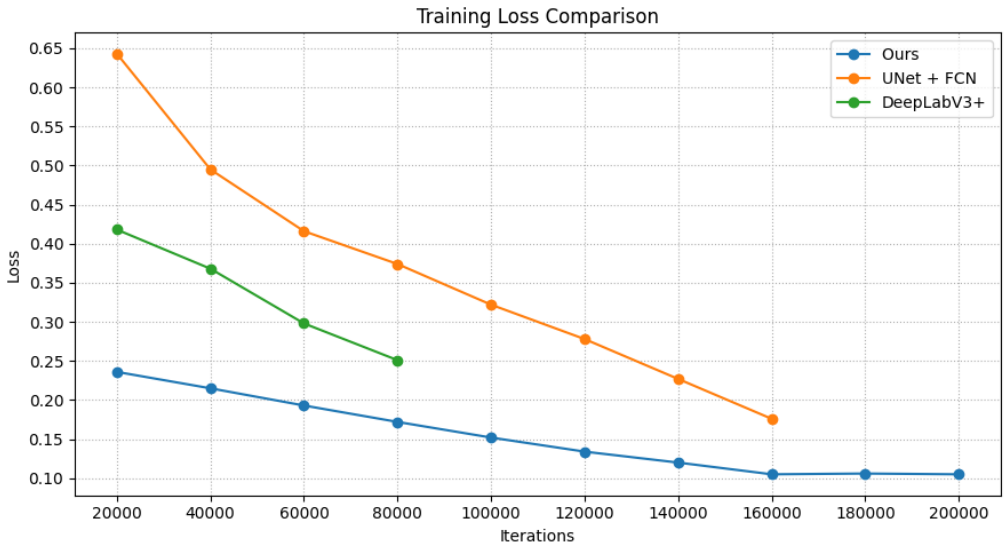

Figure 3. Training loss curves of our model versus U-Net and DeepLabV3+, showing faster convergence and a lower final loss.

The results show the efficiency of our proposed training approach. Although we extended the training to 200,000 iterations, we saw no further improvement beyond 160,000 iterations, so further training would not have been cost-effective.

3.2. Per-Class Performance

Our computer vision model segments urban road scenes into 19 different classes according to the standard of the Cityscapes dataset. It uses the cross-entropy loss function for optimization. For better understanding, we divide the classes into two different groups based on the segmentation accuracy that they achieved, measured by the main intersection over the union value. The first group of classes consists of those with the highest mIoU, which is above $\geq$76%, while the second group of classes consists of those who reached the mean intersection over union far below the 76% line. It is worth mentioning that, despite being trained exclusively on the Cityscapes dataset, without using any additional datasets, the computer vision model is capable of achieving good accuracy in the semantic segmentation task while having less than 20 million parameters and is capable of detecting different classes, such as roads, cars, sidewalks, vegetation, or buildings.

In Table 2, we show the first group of classes that yielded higher accuracy than 76%, ordered in descending order by the IoU metric for every class.

Table 2. Classes with higher accuracy than 76% in descending order.

Class	IoU (%)	Accuracy (%)
Road	97.93	98.83
Sky	94.54	98.15
Car	94.46	98.01
Vegetation	92.43	96.84
Building	92.16	96.43
Sidewalk	83.39	91.63
Person	80.90	91.60
Bus	80.04	85.36
Truck	79.17	84.65
Traffic Sign	78.83	85.16
Bicycle	76.71	87.72

The table above demonstrates that classes such as roads, cars, sky, vegetation, and buildings yield accuracy above 90%. Others, like buses, people, sidewalks, trucks, traffic signs, and bicycles, yield lower accuracy. Nonetheless, these are satisfactory results given the lightweight model approach, the quantity of data used, and the difficulty in distinguishing classes like traffic signs in different distance conditions. In the following, Table 3 lists the classes with an mIoU below 76%, indicating that they are more challenging to accurately segment.

Table 3. Classes with lower accuracy than 76%, which are more difficult to detect.

Class	IoU (%)	Accuracy (%)
Traffic Light	70.19	81.64
Motorcycle	65.86	76.66
Train	64.97	69.36
Terrain	63.02	69.89
Pole	62.83	72.81
Fence	59.84	70.10
Rider	59.45	71.76
Wall	55.14	62.86

Table 3 indicates that classes such as motorcycles, trains, traffic lights, poles, terrain, fences, walls, and riders yield lower precision. These challenges may be due to the inherent complexity and variability of these objects in different road scenes. In the future, by using more data with greater diversity, it will be possible to obtain better accuracy results for these classes. Here, it was challenging to detect these classes accurately, because distinction is needed between concrete walls, buildings, or fences, and many state-of-the-art models lack accuracy for these classes. Overall, the computer vision model, with 17.2 million parameters, detects different objects and environmental elements, such as cars, roads, and sidewalks, with good accuracy, while its lower performance in certain classes highlights opportunities for future improvements.

3.3. Global Metrics

To evaluate the final performance of our computer vision model with the Cityscapes validation set, we collected different accuracy metrics. The following accuracy metrics were obtained after the final evaluation phase:

- Mean IoU (mIoU): 76.41%;
- Mean Accuracy (mAcc): 83.66%;
- Overall Accuracy (aAcc): 95.87%.

The above metrics demonstrate that the computer vision model, having 17.2 million parameters, is capable of reaching a global mean intersection over union of 76.41%. This metric reflects the average overlap between the predicted segments and the ground truth values across all classes, and it is widely used in evaluating the performance of computer vision models. The data provided above also show that the model achieved a mean accuracy (mAcc) value of 83.66%. This metric represents the average classification accuracy per pixel for each class. Lastly, the overall accuracy (aAcc) of 95.87% is reached. This metric shows the ratio of correctly classified pixels to the total number of pixels in the Cityscapes validation set.

3.4. Qualitative Analysis

In Figure 4, we show an example image of a road scene, where the original image is presented on the left and the segmentation output produced by the system is shown on the right. From this side-by-side comparison, it is clearly shown that the model can perform semantic segmentation on road scene images, detecting different objects in the digital image with good accuracy. From this comparison, the following observations emerge.

- Accurate detection of close objects: The model is capable of detecting near-field classes with good accuracy, including roads, cars, pathways, and buses, when these objects are at an approximately 50–100 m distance.
- Challenges with distant objects: It is more challenging to detect distant objects like traffic signs, indicating potential areas for further improvement.
- Overall performance: Under ideal weather conditions, the segmentation quality is good, and the model clearly defines boundaries for different objects, although minor inaccuracies can appear in more complex or distant regions.

The segmentation results provided above demonstrate that the computer vision model developed can effectively detect and segment close objects, ensuring the clear and detailed recognition of the main elements on the road. It should be mentioned that lower accuracy metrics were reached for more distant objects such as traffic signs, poles, or traffic lights. This is a possible area for improvement in the future, where the main focus could be on improving the ability of the computer vision model to detect fine details in more distant regions. This could be achieved by modifying the model architecture or by using additional

datasets. In this way, it would be possible to further enhance the performance of the system in different real-world scenes on the road.

Figure 4. Road scene example with the original image on the left and the segmented one on the right.

3.5. Model Robustness Analysis

To evaluate the robustness of our segmentation model, we also tested it on images of a different dataset (Mapillary Vistas) that were not used during training. The examples provided below include images captured in an urban city area, which highlight the model's performance in a different environment, as well as another image taken on a highway to demonstrate the model's ability to accurately segment unseen data across different scenarios.

In Figure 5, we show the original image on the left and the corresponding segmented result on the right. This visual presentation demonstrates the ability of the model to detect major objects and the environment with good overall accuracy. It also indicates that there is room for improvement in the segmentation of small objects, such as traffic signs or traffic lights, and that the accuracy at larger distances could be improved. For example, in the highway image, where the distance is more than 100 meters, the model lacks the accuracy to differentiate cars from the road, but, with closer objects and environmental elements, the accuracy is good.

Figure 5. Segmentation results with unseen images from another dataset.

As demonstrated in Table 4, the highest performance metrics are obtained for the main environment and object classes, such as roads, cars, buildings, and vegetation, which indicate good segmentation performance for these dominant regions. Meanwhile, classes

such as people or traffic signs result in lower performance, which can be attributed to their smaller sizes and greater variability in their appearance.

Table 4. Class performance metrics.

Class	IoU (%)	Accuracy (%)
Road	96.72	98.29
Sky	93.05	96.60
Car	90.58	96.42
Vegetation	89.44	96.43
Building	89.18	94.64
Sidewalk	75.57	84.58
Person	67.59	83.94
Traffic Sign	62.63	69.76

3.6. Computational Performance Analysis

The computer vision model, with 17.2 million parameters, was evaluated on an NVIDIA GTX 1060 GPU, chosen specifically to represent the low-end sector of Cuda-capable devices (released in 2016). The test was performed using a single image with a batch size of 1, with different input resolutions. The computational performance metrics, such as the computational cost (GFLOPs), inference time (ms), and throughput (FPS) at different resolutions, are provided in Table 5, illustrating the model's resource requirements under the different conditions. As the resolution increases, both the computational cost (GFLOPs) and inference time increase, while the throughput (FPS) decreases. Specifically, at a 256 × 256 resolution, the model requires 9.1 GFLOPs and achieves 42.4 FPS with an inference time of 24 ms. With the highest resolution of 2048 × 1024, the computational cost reaches 419.4 GFLOPs, and the inference time increases to 769 ms. The data provided in the table demonstrate the trade-offs between the resolution, computation, and performance, which must be carefully considered when a different resolution is needed for resource-constrained applications.

Table 5. Computational performance at different input resolutions.

Input Resolution	GFLOPs	Inference Time (ms)	Throughput (FPS)
256 × 256	9.1	24	42.4
512 × 512	37.9	37	26.9
1024 × 1024	176.4	120	8.3
2048 × 1024	419.4	769	1.3

In Table 6, we present the computational performance of our segmentation model on the NVIDIA Jetson Orin Nano 8 GB embedded platform. The performance metrics are reported for three different input resolutions, capturing the inference speed (FPS) and GPU memory usage.

Table 6. Model performance on Jetson Orin Nano 8 GB.

Input Size	Inference Speed (FPS)	GPU Memory Usage (MiB)
256 × 256	52.89	54
320 × 320	32.23	72
512 × 512	15.56	112

The results provided in Table 6 demonstrate that, at the lowest resolution, the computer vision model can achieve an inference speed of 52.89 FPS while consuming 54 MiB of GPU memory. When using a higher resolution, the inference speed decreases to 32.23 FPS at an input size of 320 × 320. At a 512 × 512 resolution, the GPU memory consumption increases to 112 MiB, and 15.56 FPS is still acceptable for many autonomous applications. However, for applications that require higher input resolutions or faster inference speeds at larger scales, a more powerful embedded device with greater computational resources may be more appropriate than the Jetson Nano embedded platform.

4. Discussion

The experimental results obtained with the computer vision road scene segmentation system demonstrate good accuracy and effectiveness when the model uses a Transformer-based MiT backbone as an encoder for feature extraction and attention-based mechanisms in the decoder module. During the training log analysis, the results revealed that, during the whole training cycle, the mode's loss parameter steadily decreased; shortly before reaching the final iterations, it demonstrated the best accuracy performance at 156,000 iterations. Furthermore, the training process confirmed that setting the training schedule at 160,000 iterations was cost-efficient and effective, because an additional experiment extending the training process was not effective. In addition, the per-class performance metrics show that the main elements of the road scene, such as the road, cars, and buildings, are segmented with good accuracy, indicating the model's strong ability to detect and distinguish different objects and environmental elements on the road.

Creating a model that maintains a lightweight design with a low parameter count while achieving competitive accuracy is a challenging task. It is always necessary to have a balance between accuracy and lightweight model design, because, while a more complex model can achieve better accuracy metrics, it also has a significantly larger number of parameters and is consequently not as lightweight. This balance is critical, as increasing model complexity often improves the performance, but it can compromise the efficiency and suitability of the model for deployment in memory-limited environments.

According to the global evaluation metrics of the developed model, the results validate the good performance of the system and its capacity to detect and differentiate objects in digital road scene images, achieving a mean IoU of 76.41% with 17.2 million parameters in the network. Despite its compact size, the computer vision system is capable of maintaining competitive segmentation performance, making it suitable for use in memory-limited resource environments. The visual examples provided in Figure 6 show that the model is successful in segmenting near-field objects within 50–100 m, and classes like roads, cars, and pathways are detected with good accuracy. However, it is worth mentioning that more distant objects, such as poles, traffic lights, and traffic signs, as well as fences, yield lower precision. Thus, these classes could represent an avenue for future system enhancement by incorporating additional training data or modifying the internal system architecture to capture small-scale objects or environmental elements in distant regions. In Figure 6, we provide qualitative results, with the original RGB image on the left, the ground truth masks in the middle, and the segmentation results on the right.

Regarding other open-source computer vision approaches, models such as DSNet [35], HRNetV2 + OCR [36], DeepLabv3 + [36], CSFNet-1 [37], and EEEA-Net-C2 [38] have demonstrated good accuracy results on the Cityscapes dataset while maintaining a relatively low parameter count. The demonstrated approach has 17.2 million parameters and achieves a 76.41% mIoU. This approach also demonstrates that, by using a Transformer-based encoder for feature extraction and an attention-based mechanism to fuse multiscale features, it is possible to achieve competitive accuracy with other models while maintaining

a lightweight design and good overall performance. The developed approach also extends the knowledge provided by existing studies, showing that Transformer-based architectures can effectively segment complex urban scenes with models that have a low parameter count and are capable of achieving good accuracy in road scene semantic segmentation tasks. Table 7 shows details of several semantic segmentation models evaluated on the Cityscapes validation set, including their input resolutions, parameter counts, GFLOPs, and mIoU values, as reported in the literature.

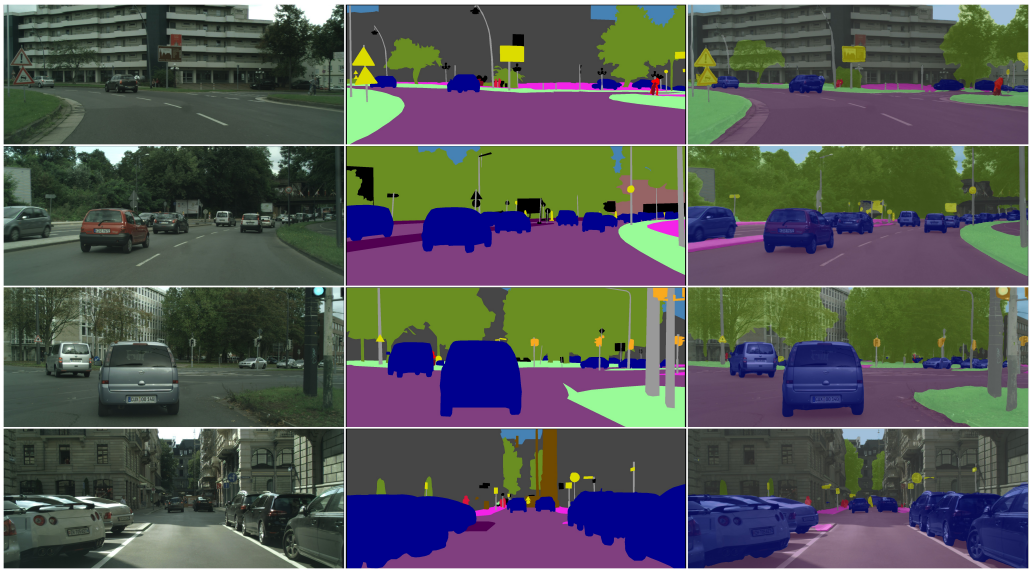

Figure 6. Qualitative segmentation results. The left column shows the original RGB image, the middle column shows the ground truth masks, and the right column shows the model's segmentation results.

Table 7. Comparison of semantic segmentation models.

Model	Input Size	Params (M)	GFLOPs	mIoU (%)
EEEA-Net-C2	512 × 512	7.34	28.7	76.8
CSFNet-1	1024 × 512	12.6	86.9	74.8
Ours	512 × 512	17.2	37.9	76.4
DSNet	2048 × 1024	37.5	226.6	82.0
DeepLabv3+	2048 × 1024	43.5	1444.6	79.6
HRNetV 2 + OCR	2048 × 1024	70.3	1206.3	81.6

The computer vision models in Table 7 are sorted by the number of parameters in the network. From the data provided in the table, we can see that our approach offers a balance between segmentation accuracy, model size, and computational efficiency. The proposed model has 17.2 million parameters and requires 37.9 GFLOPs at an input resolution of 512 × 512. Models like EEEA-Net-C2 and CSFNet-1 represent lightweight architectures, achieving mIoU accuracy metrics of 76.8% and 74.8%, respectively. In comparison, CSFNet-1 has a slightly lower parameter count but a larger computational cost. DSNet gives good accuracy results and computational efficiency, but it has more than twice as many parameters in the network. Moreover, models like DeepLabV3+ and HRNetV 2 + OCR offer better accuracy but at the cost of having much larger architectures. This comparison highlights that the proposed approach offers a compelling trade-off by delivering good segmentation performance in a lightweight architecture.

In Table 8, we compare the intersection over union (IoU) per class of our lightweight segmentation model against those of two much larger architectures: HRNet-W48 (65.9 M)

and DeepLabV3-R101 (84.7 M). Despite having roughly one-quarter to one-fifth of the number of parameters, our approach achieves only marginally lower IoU scores across many semantic categories. The side-by-side comparison makes it clear that, although larger architectures can reach slightly higher accuracy scores, our Transformer-based design with an attention-guided decoding module offers a compelling trade-off, maintaining competitive performance while drastically reducing the model complexity.

Table 8. Per-class segmentation accuracy (IoU) compared across different network architectures.

Class	DeepLabV3-R101 (84.7 M)	HRNet-W48 (65.9 M)	Ours (17.2 M)
Road	98.3	98.4	97.9
Car	95.4	95.6	94.5
Sky	94.2	95.2	94.5
Building	92.5	93.4	92.2
Vegetation	92.0	93.0	92.4
Sidewalk	85.7	86.4	83.4
Person	81.3	83.4	80.9
Traffic Sign	78.7	81.7	78.8
Bicycle	77.8	79.0	76.7
Traffic Light	69.8	72.4	70.2
Motorcycle	69.2	67.1	65.9
Terrain	63.3	66.7	63.0
Rider	63.9	62.3	59.5
Pole	58.9	69.3	62.8
Fence	62.7	66.4	59.8
Wall	53.5	59.7	55.1

In addition to the overall quantitative comparison, Table 8 shows that our Transformer-based encoder with an attention-guided decoding module matches the larger HRNet-W48 and DeepLabV3-R101 architectures almost exactly in large and homogeneous classes such as roads, cars, buildings, and vegetation, with IoU differences of less than 1.2 percentage points. The results show that the global context modeling of the MiT backbone effectively captures broad, texture-rich regions. However, all three architectures, including the larger baselines, have the greatest difficulty in accurately segmenting slender or distant objects, such as poles, fences, and walls.

In summary, the demonstrated approach confirms that a computer vision system based on the Transformer neural network architecture with an attention-guided decoding module can effectively capture contextual information for the road scene segmentation task. While the model performs well with near-field objects, future work could direct more attention to areas with more distant, small-scale objects, because many models lack accuracy in this regard, and it remains a challenging task.

5. Conclusions

In this paper, we introduce an approach that uses a Transformer-based neural network architecture to segment road scene images, using the MiT backbone as a feature extractor and an attention-guided decoder to effectively fuse multiscale features. Experimental results obtained with the Cityscapes dataset revealed that, despite the small parameter count, it can achieve a competitive mean IoU of 76.41% in different semantic segmentation tasks in urban scenes. The conducted training process confirmed that the system is able to converge reliably in 160,000 iterations, and an additional training process with the same configuration and dataset did not produce better accuracy results.

The model evaluation phase and visual data highlighted that the main objects in close proximity can be detected with good accuracy, while the classes with lower accuracy offer an avenue for future improvements. In general, the primary design goal of this work was to develop a lightweight computer vision road scene segmentation model. Our Transformer-based segmentation model has only 17.2 million parameters, which is remarkably low compared to many computer vision semantic segmentation models. The demonstrated model also has a small memory footprint, making it well suited for deployment on memory-limited devices.

Author Contributions: Conceptualization, R.M.; Formal analysis, B.L. and R.M.; Funding acquisition, R.M.; Investigation, B.L. and R.M.; Methodology, B.L.; Project administration, R.M.; Resources, B.L.; Software, B.L.; Supervision, R.M.; Validation, B.L.; Visualization, B.L.; Writing—original draft, B.L.; Writing—review and editing, R.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Data Availability Statement: All data are freely available via the indicated datasets.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional Neural Network
FLOPS	Floating Point Operations per Second
FPS	Frames per Second
GPS	Global Positioning System
IoU	Intersection over Union
mAcc	Mean Accuracy
mIoU	Mean Intersection over Union
MLP	Multilayer Perceptron
MiT	Mix Transformer
NASA	National Aeronautics and Space Administration

References

1. Rafique, Z. Improving Efficiency of Computer Vision for Autonomous Vehicles. Master's Thesis, Bournemouth University, Poole, UK, 2020. [CrossRef]
2. Fernandes, S.; Duseja, D.; Muthalagu, R. Application of Image Processing Techniques for Autonomous Cars. *Proc. Eng. Technol. Innov.* **2020**, *17*, 6074. [CrossRef]
3. Sellat, Q.; Bisoy, S.; Priyadarshini, R.; Vidyarthi, A.; Kautish, S.; Barik, R.K. Intelligent Semantic Segmentation for Self-Driving Vehicles Using Deep Learning. *Comput. Intell. Neurosci.* **2022**, *2022*, 6390260. [CrossRef]
4. NASA. Mars 2020 Perseverance Rover. 2020. Available online: https://science.nasa.gov/mission/mars-2020-perseverance/ (accessed on 19 March 2025).
5. Chen, W.; Chi, W.; Ji, S.; Ye, H.; Liu, J.; Jia, Y.; Yu, J.; Cheng, J. A Survey of Autonomous Robots and Multi-Robot Navigation: Perception, Planning and Collaboration. *Biomim. Intell. Robot.* **2025**, *5*, 100203. [CrossRef]
6. Iparraguirre-Gil, O. Computer Vision and Deep Learning based Road Monitoring towards a Connected, Cooperative and Automated Mobility. Ph.D. Thesis, Universidad de Navarra, Pamplona, Spain, 2022. [CrossRef]
7. Marasinghe, R.; Yigitcanlar, T.; Mayere, S.; Washington, T.; Limb, M. Computer Vision Applications for Urban Planning: A Systematic Review of Opportunities and Constraints. *Sustain. Cities Soc.* **2024**, *100*, 105047. [CrossRef]
8. Landsiedel, C.W. Semantic Mapping for Autonomous Robots in Urban Environments. Ph.D. Thesis, Technische Universität München, Munich, Germany, 2018.
9. Boysen, N.; Fedtke, S.; Schwerdfeger, S. Last-mile Delivery Concepts: A Survey from an Operational Research Perspective. *Or Spectr.* **2020**, *43*, 1–58. [CrossRef]

10. Ni, J.; Chen, Y.; Tang, G.; Shi, J.; Cao, W.; Shi, P. Deep Learning-Based Scene Understanding for Autonomous Robots: A Survey. *Intell. Robot.* **2023**, *3*, 374–401. [CrossRef]
11. Ogunsina, M.; Efunniyi, C.P.; Osundare, O.S.; Folorunsho, S.O.; Akwawa, L.A. Reinforcement Learning in Autonomous Navigation: Overcoming Challenges in Dynamic and Unstructured Environments. *Eng. Sci. Technol. J.* **2024**, *5*, 2724–2736. [CrossRef]
12. Adiuku, N.; Avdelidis, N.P.; Tang, G.; Plastropoulos, A. Advancements in Learning-Based Navigation Systems for Robotic Applications in MRO Hangar: Review. *Sensors* **2025**, *24*, 1377. [CrossRef]
13. Katona, K.; Neamah, H.A.; Korondi, P. Obstacle Avoidance and Path Planning Methods for Autonomous Navigation of Mobile Robot. *Sensors* **2024**, *24*, 3573. [CrossRef]
14. Janai, J.; Güney, F.; Behl, A.; Geiger, A. Computer Vision for Autonomous Vehicles: Problems, Datasets and State of the Art. *Found. Trends® Comput. Graph. Vis.* **2021**, *12*, 1–308. [CrossRef]
15. Conde, M.V. An Embarrassingly Pragmatic Introduction to Vision-based Autonomous Robots: Applications, Datasets and State of the Art. *arXiv* **2021**, arXiv:2112.05534.
16. Pelikan, H.R.M.; Reeves, S.; Cantarutti, M. Encountering Autonomous Robots on Public Streets. In Proceedings of the HRI '24: Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction, Boulder, CO, USA, 11–15 March 2024.
17. Buckeridge, S.; Carreno-Medrano, P.; Cosgun, A.; Croft, E.; Chan, W.P. Autonomous Social Robot Navigation in Unknown Urban Environments Using Semantic Segmentation. *arXiv* **2021**. [CrossRef]
18. Siegwart, R.; Nourbakhsh, I.; Scaramuzza, L. *Introduction to Autonomous Mobile Robots*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2011.
19. Dilek, E.; Dener, M. Computer Vision Applications in Intelligent Transportation Systems: A Survey. *Sensors* **2023**, *23*, 2938. [CrossRef]
20. Che, C.; Zheng, H.; Huang, Z.; Jiang, W.; Liu, B. Intelligent Robotic Control System Based on Computer Vision Technology. *Appl. Comput. Eng.* **2024**, *64*, 142–147. [CrossRef]
21. Upadhyay, D.; Upadhay, D.K.; Singh, R.; Mishra, D.; Khatri, V. Robotic Vision: Advancements in Computer Vision for Autonomous Systems. *Tuijin Jishu/J. Propuls. Technol.* **2023**, *4*. [CrossRef]
22. Sultana, F.; Sufian, A.; Dutta, P. Evolution of Image Segmentation using Deep Convolutional Neural Network: A Survey. *Knowl.-Based Syst.* **2020**, *201–202*, 106062. [CrossRef]
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All You Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
24. Bai, Y.; Mei, J.; Yuille, A.; Xie, C. Are Transformers More Robust Than CNNs? *arXiv* **2021**, arXiv:2111.05464.
25. Zhai, X.; Kolesnikov, A.; Houlsby, N.; Beyer, L. Scaling Vision Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
26. Hurtado, J.V.; Valada, A. Semantic Scene Segmentation for Robotics. *arXiv* **2024**, arXiv:2401.07589.
27. Wang, Y.; Ahsan, U.; Li, H.; Hagen, M. A Comprehensive Review of Modern Object Segmentation Approaches. *Found. Trends® Comput. Graph. Vis.* **2023**, *13*, 111–283. [CrossRef]
28. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *arXiv* **2014**, arXiv:1411.4038.
29. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
30. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
31. Zheng, Y.; Tan, X.; Zheng, Z.; Zhou, Y.; Li, B.; Yi, L. Rethinking Semantic Segmentation with Transformers. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021.
32. Xie, E.; Wang, W.; Yu, Z.; An, T.; Gao, Y.; Lu, M.; Xu, X.; Ren, T.; Zhang, C.; Xiao, T.; et al. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *arXiv* **2021**, arXiv:2105.15203.
33. Yuan, K.; Guo, S.; Liu, Z.; Zhou, A.; Yu, F.; Wu, W. Incorporating Convolution Designs into Visual Transformers. *arXiv* **2021**, arXiv:2103.11816.
34. Li, J.; Zhang, Y.; Yun, P.; Zhou, G.; Chen, Q.; Fan, R. RoadFormer: Duplex Transformer for RGB-Normal Semantic Road Scene Parsing. *IEEE Trans. Intell. Veh.* **2024**, *9*, 5163–5172. [CrossRef]
35. Guo, Z.; Bian, L.; Huang, X.; Wei, H.; Li, J.; Ni, H. DSNet: A Novel Way to Use Atrous Convolutions in Semantic Segmentation. *arXiv* **2024**, arXiv:2406.03702. [CrossRef]
36. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep High-Resolution Representation Learning for Visual Recognition. *arXiv* **2020**, arXiv:1908.07919. [CrossRef]

37. Qashqai, D.; Mousavian, E.; Shokouhi, S.B.; Mirzakuchaki, S. CSFNet: A Cosine Similarity Fusion Network for Real-Time RGB-X Semantic Segmentation of Driving Scenes. *arXiv* **2024**, arXiv:2407.01328.
38. Termritthikun, C.; Jamtsho, Y.; Ieamsaard, J.; Muneesawang, P.; Lee, I. EEEA-Net: An Early Exit Evolutionary Neural Architecture Search. *Eng. Appl. Artif. Intell.* **2021**, *104*, 104397. [CrossRef]