Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.36700

Deep Learning-Based Visual Semantic Perception Technology for Mobile Vehicles

Jiawei Sun

School of Artificial Intelligence, University of Science and Technology Beijing, Beijing, China
tuxiaoyi1@outlook.com

Abstract. With the large-scale implementation of scenarios such as logistics warehousing and park inspections, the demand for autonomous environmental perception by mobile robots continues to grow. Low-cost pure vision semantic perception has become a core technology for ensuring autonomous safe navigation of these robots. Traditional manual feature algorithms lack stability in complex road conditions and varying lighting conditions, while deep learning relies on autonomous hierarchical feature learning to break through the original performance limit. This article focuses on real-time semantic perception of embedded mobile cars, systematically sorting out three mainstream deep learning solutions: CNN, lightweight network, and visual Transformer, and comparing the adaptation differences of each architecture in segmentation accuracy, inference speed, and complex working conditions. Summarize the existing technological bottlenecks from three aspects: imbalanced real-time accuracy, weak robustness in complex environments, and difficulty in recognizing distant and occluded targets. Corresponding optimization ideas are proposed, including dynamic lightweight inference, cross-domain adaptive enhancement, convolutional attention hybrid architecture, and end-to-end collaborative deployment. The article comprehensively outlines the development trajectory of visual semantic perception technology for mobile vehicles, providing a systematic theoretical reference for the design and embedded deployment optimization of visual algorithms for mobile robots in warehousing and inspection tasks.

Keywords: mobile car, visual semantic perception, Semantic Segmentation lightweight network

1. Introduction

For mobile cars, the core premise of realizing safe and autonomous movement is to rely on visual sensors to complete environmental semantic perception. Traditional manual feature algorithms rely on manually designed feature operators, which have poor generalization ability under complex conditions such as dramatic light changes, car bumps, and indoor and outdoor scene switching, and are difficult to meet the needs of refined pixel-level scene analysis. Therefore, convolutional neural networks (CNNs) are introduced to carry out pixel-level semantic segmentation, so as to solve the problem of basic scene classification and region division, and form the first basic technical route.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.36700

However, the complete CNN model has a huge number of parameters and calculations, which can not adapt to the computing power and power consumption of the car's embedded board. Therefore, researchers made lightweight transformations at the network structure and convolution operator levels, derived a lightweight real-time semantic model, and formed the second technical route of end-to-end real-time reasoning. At the same time, CNN is limited by the local receptive field and has poor modeling effect on occluded targets and small obstacles at a long distance. In order to make up for the short board of global context reasoning, the visual transformer realizes the whole map-dependent modeling with the self-attention mechanism, forming the third technical route focusing on the perception of complex and difficult scenes. Relying on automatic hierarchical feature learning, deep learning has broken through the performance bottleneck of traditional methods and has formed three mainstream technology systems: CNN, lightweight real-time networks, and visual transformers.

It should be pointed out that most of the existing reviews focus on the general image segmentation task or the horizontal comparison of a single architecture, and lack a systematic review of the embedded computing constraints, real-time requirements, and complex unstructured road scenes faced by the specific carrier of a mobile car. In this paper, the research boundary converges to the embedded end-side scene of the mobile car, and a comparative analysis is carried out around the accuracy-real-time trade-off of the three types of technology systems under the condition of limited computing power, in order to provide a more targeted reference for the selection and deployment of the end-side semantic segmentation model.

2. Technical basis of mobile car visual semantic perception

2.1. Definition and core tasks of visual semantic perception

Visual semantic perception refers to that the mobile car collects scene images through the visual sensors and assigns semantic category labels to each pixel or local area, so as to obtain a high-level semantic description of the environment. The core task of visual semantic perception of mobile cars consists of several small tasks. Semantic segmentation is responsible for pixel-by-pixel classification of the whole image, dividing the scene into semantic regions; object detection is responsible for locating and identifying object instances in the scene; and scene semantic understanding further infers the spatial relationship between semantic elements. Together, they form the semantic perception basis of mobile cars. The subsequent core of this paper focuses on semantic segmentation, because it can output pixel-level global region division results, which is the bottom core perceptual input of mobile car path planning and global obstacle avoidance, and the related lightweight and end-to-end optimization research system is complete, which is convenient for the system to sort out the technical context of various perceptual architectures.

2.2. Deep learning perception network foundation

Deep learning relies on multi-layer nonlinear transformation stacking to realize hierarchical feature learning, which is the mainstream technology of visual semantic perception. The convolutional neural network can extract visual features efficiently by using the local receptive field and weight sharing mechanism, and complete the abstraction from texture details to high-level semantics step by step. On this basis, the full convolution network discards the fully connected layer and creates an end-to-end pixel-by-pixel semantic segmentation paradigm. Relying on the self-attention mechanism, visual transformer models the full-length distance dependence of images, gets rid of the

DOI: 10.54254/2755-2721/2026.36700

limitation of the convolution receptive field, and opens up a new direction for the global semantic understanding of scenes.

3. Research on the mainstream technology of visual semantic perception based on deep learning

3.1. Visual semantic perception method based on CNN

The deep involvement of convolutional neural networks in the field of visual semantic perception began with the paradigm shift from image classification architecture to intensive prediction tasks. The fully convolutional network (FCN), first proposed by Long et al. In 2015, established an end-to-end deep learning semantic segmentation framework [1]. FCN samples the highly compressed feature map at the end of the encoder through transpose convolution to restore spatial resolution, and designs a three-level jump structure of fcn-32s, FCN-16s, and fcn-8s. Among them, fcn-8s has the best effect, achieving 62.2% mIoU on the Pascal VOC 2012 data set, which is about 20% higher than the previous method [1]. Since then, FCN has become the reference or improved baseline model for almost all segmentation methods, and the improvement around it mainly focuses on two dimensions.

One is the recovery method of spatial resolution: SegNet uses the pooled index stored by the encoder to guide the decoder to complete the nonlinear anti-pooled upsampling, and U-Net uses skip connections to completely transfer the spatial details of each level of the encoder to the decoder. Both of them correct the edge ambiguity caused by FCN transpose convolution upsampling from the perspective of computational cost and edge fidelity. Limengyi and Zhu Dingju pointed out in the review that the mainstream segmentation framework after FCN is generally developed along several routes, such as encoder-decoder, feature fusion, context module, and image pyramid, and continues to improve around the problems exposed by FCN, such as low feature resolution, insufficient spatial consistency, and difficulty in multi-scale object modeling [2].

The second is the expansion of receptive fields and multi-scale context aggregation: the DeepLab series introduces hole convolution to realize the exponential expansion of receptive fields without increasing the parameters, and its subsequent versions use the hole space pyramid pooling module to extract multi-scale features in parallel. PSPNet uses pyramid pooling to aggregate high-level features at multi granularity. These two kinds of multi-scale modeling ideas effectively alleviate the inherent limitations of convolution architecture on the low efficiency of long-distance semantic dependency modeling. Liangxinyu and others combed the performance evolution of typical networks such as SegNet and DeepLab on commonly used data sets, and these reviews jointly outlined the complete technical pedigree of convolutional semantic segmentation since FCN [3].

It is worth noting that FCN and its improved paradigm do not stay at the level of general image segmentation, but gradually extend to the mobile car scene. The multi-scale FCN method for mobile car navigation published by Thai Viet Dang and Ngoc Tam Bui on electronics is a typical example [4]. It uses vgg-16 as the encoder to build a multi-scale full convolution network, focuses the semantic segmentation task on the binary division of the passable area and the non passable area, and carries out perspective correction on the segmentation results to build the scene elevation, so as to identify the movable area and complete obstacle avoidance navigation, and achieves an overall accuracy of 93.5% on the measured data [4]. This work shows that the end-to-end pixel-level segmentation paradigm established by FCN can adapt to the navigation perception requirements of mobile vehicles with low structural complexity, which provides a direct engineering foreshadowing for the rise of lightweight real-time segmentation networks.

DOI: 10.54254/2755-2721/2026.36700

On the whole, the convolutional semantic segmentation method based on FCN has built the first-generation technology foundation of mobile car visual semantic perception through the improvement of encoding and decoding structure, the introduction of multi-scale context aggregation, and customized adaptation for specific application scenarios. However, due to the physical limitation of the local receptive field in convolution operations, the acquisition of long-distance pixel correlation can only rely on layer-by-layer stacking and gradual expansion, and the global context modeling is always not direct enough. This structural weakness also constitutes the deep motivation for the introduction of subsequent transformer architecture.

3.2. Lightweight real-time semantic perception model

The strict requirements for reasoning speed and resource occupation in the end-to-end deployment of mobile cars make lightweight design an inevitable premise for the practical application of visual semantic perception. The related research generally follows the two routes of "operator-level lightweight" and "architecture-level lightweight". The former compresses the single-layer computational overhead from the mathematical decomposition of the convolution operation, while the latter reconstructs the balance between accuracy and speed from the overall network structure.

The core of operator-level lightweight is to reduce the redundant computation in standard convolution. The MobileNet series proposed by Google laid the foundation for depthwise separable convolution [5]. The operator splits the standard convolution into two stages: channel-by-channel depth convolution and 1×1 point-by-point convolution. It decouples the joint mapping of space and channel dimensions, so that the single-layer computation is reduced to one-eighth to one-ninth of the standard convolution and the accuracy loss is limited. Its latest version, MobileNetV4, further introduces the unified backward residual bottleneck and the multi-query attention mechanism for mobile accelerators, and achieves 87% Top-1 accuracy with a 3.8 MS delay on the Pixel 8 Edge TPU, marking that operator-level lightweight is moving from simply compressing the amount of computation to the general design of software and hardware collaboration [6]. The ShuffleNet series proposed by Kuangshi starts from packet convolution, which groups the input channels into independent convolutions to compress the amount of calculation, and uses a channel shuffling operation to break through the information barrier between groups with almost zero additional overhead [7]. ShuffleNetV2 further sums up several engineering criteria from the memory access characteristics of embedded hardware, so that the measured reasoning speed no longer depends solely on the theoretical floating-point computation [8].

On the basis of operator lightweight, architecture-level lightweight focuses on the structural reconstruction of partitioned networks, and the representative idea is "double branch parallel". Bisenet is the first to decouple spatial details and semantic understanding into two independent forward paths: the spatial path maintains high resolution to capture the edge of the object, the context path relies on the lightweight backbone to quickly down sample and aggregate the global semantics, and the two are integrated and output by the feature fusion module, which for the first time resolves the contradiction between spatial accuracy and semantic depth in real-time segmentation at the framework level [9]. BiSeNet achieved 68.4% mIoU and 105 FPS on Cityscapes, becoming a classic baseline for embedded real-time segmentation [9]. Dfanet takes multi-layer aggregation as another orientation and achieves a multi-scale effect by reusing the features extracted from the previous tower and combining them with the lightweight aggregation module. It achieves 71.3% mIoU on Cityscapes with 7.8m parameters, which can reduce the amount of computation by about 7 times compared with the previous method under similar accuracy [10]. Both of them prove that lightweight semantic perception in the mobile car scene is not a simple

DOI: 10.54254/2755-2721/2026.36700

replacement of the backbone network, but a systematic trade-off between operator decomposition, structure reconstruction, and multi-scale modeling.

In general, the lightweight real-time semantic perception model has formed a mature design system that takes efficiency as the key and takes into account accuracy, which lays an engineering foundation for the landing of deep learning perception algorithms on the low-computing-power platform of mobile cars. However, the mining of lightweight operators in this kind of method is still limited to the local computational characteristics of convolution structure, and its global context modeling ability is subject to the progressive expansion of the receptive field layer by layer, which just leaves room for the introduction of the transformer's self-attention mechanism in the field of visual perception.

3.3. Visual perception method based on transformer

The Vision Transformer (VIT) proposed by Dosovitskiy et al. Completely migrated the transformer to the image recognition task for the first time [11]. The core idea is to divide the image into image blocks with fixed size, convert each image block into a serialized token vector through linear projection, and then directly model the global dependency by multi-layer self attention, so that the network can obtain the complete receptive field of the whole image from the shallow layer, and completely get rid of the indirect path that the convolution network relies on layer by layer stacking to gradually expand the receptive field [11].

However, the native ViT always maintains a fixed feature map resolution and single-scale token representation in the forward process. The perceptual ability of different size targets is uneven, and the self-attention computing overhead increases in a square relationship with the number of input pixels, which makes it difficult to meet real-time requirements under high-resolution input. To solve these problems, Liu et al. Proposed the swing transformer, which introduced the window self attention mechanism, limited the self attention of the whole graph to a fixed size local window, and realized the alternating flow of information between windows through the cross window shift operation, to reduce the computational complexity of self attention from quadratic to linear, while retaining the output ability of hierarchical multi-scale features, and can be directly used as the backbone encoder of semantic segmentation network [12].

On this basis, the SegFormer proposed by Xie et al. Has built a simple segmentation architecture of a pure transformer. The encoder uses a layered transformer module to gradually reduce the feature resolution and expand the channel dimension to form a pyramid feature. The decoder is only a multi-layer perceptron, completely abandoning the convolution operation [13]. This regular structural design greatly reduces the loss of accuracy of model quantization. Its lightweight version, SegFormer-B0, has only obtained 76.2% of mIoU on the Cityscapes Street View dataset with 3.7m parameters, providing a feasible reference for the landing of transformers in the embedded end-to-end scene [13].

Overall, the core value of the transformer's visual semantic perception of mobile cars lies in the long-distance semantic reasoning ability given by global self-attention. When the target is blocked by a large area and only the local visible area is exposed, the network can use the scene's global context clues to make cross-region inference, significantly reducing the miscarriage and missed detection of convolution methods in such scenes, but the non-optimized self-attention computing overhead still poses a challenge to run independently and in real time on the end side.

DOI: 10.54254/2755-2721/2026.36700

4. Existing technical difficulties

4.1. Contradiction between accuracy and real-time performance

The primary contradiction of visual semantic perception of mobile cars is the difficulty of reconciling accuracy and real-time performance. Transformer obtains the semantic reasoning ability far beyond convolution method by virtue of global self-attention, but its square computational complexity makes it difficult for the end side to run in real time. Even if pruning, quantization, and knowledge distillation compression methods are used, the quantization error and distillation loss will also propagate to the segmentation boundary. Once the compression force exceeds the boundary, the accuracy will decline significantly. The accumulation of multi-layer approximation errors is difficult to avoid in the fine area. Therefore, in the case of a small vehicle with limited computational rigidity, the minimum acceptable accuracy threshold must be defined for specific tasks as the constraint boundary for compression and acceleration.

4.2. Weak perceptual robustness in complex dynamic driving environment

Strong light overexposure and heavy shadows change the local pixel distribution, and rain, snow, and fog introduce noise and weaken the contrast. The image jitter and motion blur caused by vehicle body mechanical vibration destroy the temporal and spatial consistency of frame-by-frame feature extraction. The lack of lane line and curb identification on unstructured roads is more likely to lead to miscalculation of drivable areas. In addition, most models are trained on fixed data sets. Once they experience indoor and outdoor scene switching or encounter uncovered lighting and weather conditions, the domain adaptation ability will be significantly reduced, and it is difficult to continuously output stable and reliable perception results in long-term operation.

4.3. Poor perception of long-range targets and large-area occluded scenes

Distance and occlusion are other typical difficulty that restricts the perceived reliability of mobile cars. The pixels of long-range targets are sparse, and the spatial features are severely diluted after multiple downsampling. The local receptive field of a convolutional network is not enough to capture the sparse signal. Although transformers can alleviate the problem with global attention, it is still difficult to accurately identify the target when the semantic clues are extremely scarce. In occluded scenes, convolutional networks rely on local neighborhood information reasoning, which is prone to segmentation fracture and category misjudgment. Although it can infer across regions by using the global context, when the target is completely obscured and the visible clues are extremely scarce, the network also lacks sufficient basis to restore the complete target semantics. The two kinds of problems together constitute the current direction to be solved.

5. Research prospects

In view of the above technical difficulties, the knowledge distillation strategy can be introduced, and the lightweight network training can be supervised by using the semantic prior of the high-precision large model. Under the premise of ensuring the ultra-low reasoning delay at the embedded end, the lightweight model can be controlled to segment the mIoU loss, so as to achieve the dynamic balance between accuracy and real-time performance. For the problem of poor robustness in complex environments, the pre-image enhancement algorithm of illumination normalization and dynamic blur repair can be combined to eliminate the interference of imaging distortion on feature extraction.

DOI: 10.54254/2755-2721/2026.36700

At the same time, the domain adaptive learning strategy is introduced to reduce the domain differences between indoor and outdoor scenes, day and night, cloudy and sunny scenes, and improve the generalization ability of the model through cross-scene feature alignment. In the case of long-range targets and large-area occluded scenes, the lightweight window attention and convolution feature fusion mechanism can be used to ensure real-time performance in conventional scenes by relying on efficient convolution. When occluded and long-range small targets are detected, the global attention branch can be turned on adaptively, and the long-distance dependency modeling ability of the model can be improved without significantly increasing the reasoning overhead.

6. Conclusion

In this paper, the development and application characteristics of CNN, lightweight networks, and visual transformers are systematically sorted out, focusing on the core task of semantic segmentation in the visual semantic perception scene of the embedded end of the mobile car. The research shows that the convolutional network architecture based on FCN has low deployment cost and strong stability, which is the core scheme of early car visual perception. Through operator optimization and structure reconstruction, various lightweight networks can effectively reduce the computational cost of the model and adapt to the real-time reasoning requirements of the embedded terminal. However, there are problems such as insufficient feature extraction and accuracy degradation under complex working conditions. Relying on the global self-attention mechanism, visual transformer has significant advantages in long-distance and occluded target perception tasks, effectively making up for the technical weakness of convolutional networks. However, the original model has high computational overhead and is difficult to meet the requirements of real-time landing at the end of the car. This paper further summarizes the four core problems in the current field, including the imbalance between accuracy and real-time performance, weak robustness in complex environments, low accuracy of special target recognition, and limited project landing, and combines multi-dimensional optimization and improvement ideas. This paper completely clarifies the technical iteration logic of visual semantic perception of mobile cars, which can provide reliable theoretical and technical reference for the subsequent lightweight, highly robust perceptual algorithm development and embedded engineering deployment.

References

[1] Shelhamer, E., Long, J., & Darrell, T. (2017). Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(4), 640–651. https://doi.org/10.1109/TPAMI.2016.2572683

[2] Li, M., & Zhu, D. (2021). A review of image semantic segmentation methods based on fully convolutional networks. *Computer Systems & Applications, 30*(9), 41–52. https://doi.org/10.15888/j.cnki.csa.008078 [in Chinese].

[3] Liang, X., Luo, C., Quan, J., et al. (2020). Research progress of image semantic segmentation technology based on deep learning. *Computer Engineering and Applications, 56*(2), 18–28. [in Chinese].

[4] Dang, T. V., & Bui, N. T. (2023). Multi-scale fully convolutional network-based semantic segmentation for mobile robot navigation. *Electronics, 12*(3), 533. https://doi.org/10.330/electronics12030533

[5] Howard, A. G., Zhu, M., Chen, B., et al. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications (arXiv:1704.04861). *arXiv*.

[6] Qin, D. F., Leichner, C., Delakis, M., et al. (2024). MobileNetV4: Universal models for the mobile ecosystem. In *Computer Vision – ECCV 2024*. Springer. https://doi.org/10.1007/978-3-031-73661-2_5

[7] Zhang, X., Zhou, X., Lin, M., et al. (2018). ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6848–6856). IEEE. https://doi.org/10.1109/CVPR.2018.00716

DOI: 10.54254/2755-2721/2026.36700

[8] Ma, N., Zhang, X., Zheng, H. T., et al. (2018). ShuffleNet V2: Practical guidelines for efficient CNN architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 122–138). Springer. https://doi.org/10.1007/978-3-030-01264-9_8

[9] Yu, C., Wang, J., Peng, C., et al. (2018). BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 334–349). Springer. https://doi.org/10.1007/978-3-030-01261-8_20

[10] Li, H., Xiong, P., Fan, H., et al. (2019). DFANet: Deep feature aggregation for real-time semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9514–9523). IEEE. https://doi.org/10.1109/CVPR.2019.00975

[11] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

[12] Liu, Z., Lin, Y., Cao, Y., et al. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 9992–10002). https://doi.org/10.1109/ICCV48922.2021.00986

[13] Xie, E., Wang, W., Yu, Z., et al. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*.