

A Review of Street Scene Semantic Segmentation: Current Status, Challenges, and Future Prospects

Linlin Liu*

School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, Henan 45400, P R China

* Corresponding author: Linlin Liu (Email: 15993744254@163.com)

Abstract: Street scene semantic segmentation is a core task in environmental perception for autonomous driving, which has made significant progress in recent years driven by deep learning. This paper systematically reviews the current research status in this field, analyzes technical challenges, and explores future development directions. First, it examines the technological evolution of traditional machine learning and deep learning methods, focusing on model innovations based on Convolutional Neural Networks (CNNs) and Transformer architectures. Then, it discusses cutting-edge solutions such as multimodal fusion, self-supervised learning, and 3D segmentation to address key challenges, including complex scene understanding, real-time constraints, and data efficiency. Finally, in the context of smart transportation and urban digital management, it proposes pathways to overcome technical bottlenecks, providing theoretical references for future research and practical applications.

Keywords: Deep learning; Semantic segmentation; Autonomous driving

1. Introduce

In the context of accelerating digital transformation, street scene semantic segmentation has emerged as a key technology for visual perception and intelligent decision-making, driving the development of smart cities. By enabling pixel-level scene analysis, this technology provides precise environmental perception for autonomous driving systems while constructing dynamic digital twins to support intelligent urban governance. Deep semantic segmentation networks can accurately interpret critical elements such as road topology, traffic signs, and dynamic obstacles, building high-confidence three-dimensional environmental representations[1]. Through multi-level feature fusion mechanisms, these systems can simultaneously identify both macroscopic road network structures and microscopic obstacle details, providing reliable input for path planning, obstacle avoidance, and risk prediction. Additionally, semantic analysis-based urban element digitization can be widely applied to traffic flow analysis and infrastructure health monitoring[2]. By leveraging spatiotemporal semantic correlation analysis, these systems can enhance urban governance by detecting road anomalies and optimizing public resource allocation. To address challenges such as sudden changes in lighting and adverse weather conditions in street environments, deep segmentation networks integrate deformable convolutions and attention mechanisms to enhance feature robustness. Adaptive normalization techniques effectively mitigate environmental noise, ensuring stable output even under extreme conditions such as fog, rain, and snow. Furthermore, street scenes contain diverse objects of varying scales, including buildings, pedestrians, vehicles, and traffic signs[3]. To tackle the challenge of scale variation, atrous spatial pyramid pooling and feature pyramid networks (FPN) facilitate cross-scale feature interactions, significantly improving the detection of multi-scale objects. The decoder design incorporates skip connections and detail restoration modules to compensate for spatial information loss during downsampling, thereby enhancing segmentation accuracy. To support edge computing and real-time inference, dual-branch

networks separate spatial detail processing from semantic information extraction, reducing computational complexity while maintaining accuracy. Moreover, knowledge distillation transfers the scene understanding capabilities of high-precision teacher models to lightweight student networks, enabling efficient deployment on resource-constrained devices. Despite the rapid advancements in deep learning, particularly in convolutional neural networks (CNN)[4], the complexity and variability of street scenes, along with the challenges posed by high-resolution data processing, continue to present obstacles. Key challenges include high computational costs, limited robustness, and difficulties in handling multi-scale objects. High-resolution street scene images demand significant computational resources, affecting real-time performance and making efficient deployment on low-power devices challenging. Furthermore, variations in weather, lighting conditions, and occlusions may lead to misclassification or missed detections, compromising the reliability of perception systems. Additionally, effectively integrating multi-scale features remains a critical technical challenge due to the significant size differences among objects, such as distant pedestrians and nearby buildings.

Consequently, improving the accuracy and efficiency of semantic segmentation remains a core research objective. Looking ahead, with continuous advancements in deep learning algorithms[5], computational resource optimization, and lightweight network design, street scene semantic segmentation is expected to play an increasingly vital role in autonomous driving, smart cities, and intelligent transportation, driving the widespread application of intelligent and automated technologies.

2. Research Status

2.1. Semantic Segmentation Based on Machine Learning

Before the rise of deep learning, traditional machine learning-based image segmentation methods provided preliminary solutions for street scene understanding by

combining handcrafted feature design with statistical models. The core idea of these methods is to extract low-dimensional features from images and use classifiers or clustering algorithms to achieve pixel- or region-level semantic segmentation. The technical approaches can be classified into the following three categories:

2.1.1. Region-based Segmentation Methods

Threshold Segmentation: Uses grayscale histogram analysis to set global or local thresholds, categorizing pixels into classes such as roads and buildings[6]. It is effective for simple street scenes with uniform lighting but is sensitive to complexities like shadows and reflections. **Region Growing and Split-Merge:** Iteratively aggregates adjacent regions based on pixel similarity (e.g., color, texture) to extract continuous road contours but struggles with region fragmentation caused by occlusion. **Clustering Algorithms:** Unsupervised methods such as K-means and MeanShift cluster pixels in color space. These approaches were used in early traffic sign detection to separate objects from the background but fail to distinguish semantically similar objects (e.g., vehicles and billboards).

2.1.2. Feature Engineering-Based Segmentation Methods

Handcrafted Feature Extraction: Uses Histogram of Oriented Gradients (HOG) to describe edge structures, Local Binary Patterns (LBP) for texture representation, and Scale-Invariant Feature Transform (SIFT) to capture local geometric properties, forming feature vectors at the pixel or superpixel level. **Classifier Design:** Converts segmentation tasks into per-pixel classification problems using models such as Support Vector Machines (SVM) and Random Forests for semantic annotation[7]. For instance, the HOG+SVM combination has been used for sidewalk boundary detection, though its feature representation is limited by manually designed constraints. **Conditional Random Fields (CRF):** Incorporates spatial context constraints after classification and optimizes an energy function to eliminate isolated noise points, improving segmentation continuity for roads and curbstones.

2.1.3. Graph Model-Based Segmentation Methods

Markov Random Fields (MRF): Models spatial dependencies between pixels and uses iterative belief propagation algorithms to achieve globally consistent segmentation, useful for extracting large homogeneous regions in street scenes (e.g., sky, vegetation)[8]. **Superpixel Segmentation:** The SLIC algorithm generates perceptually meaningful regions, reducing computational complexity by applying segmentation at the superpixel level instead of individual pixels. This approach has been used for the rapid extraction of lane markings in street images. **Active Contour Models:** Driven by an energy function, these models evolve curves to approximate object boundaries. In vehicle contour detection, edge gradient information is used to dynamically refine the contour, but the approach is sensitive to initialization and computationally expensive.

Although traditional machine learning methods are effective in specific scenarios, they face three fundamental limitations:

Limited Feature Representation Ability: Handcrafted low-dimensional features struggle to capture the essence of multi-scale objects and complex lighting variations in street scenes. **Semantic Gap Problem:** Low-level visual features (e.g., color, texture) lack strong associations with high-level semantics. **Computational Inefficiency:** Post-processing steps such as

CRF and MRF introduce exponentially increasing algorithm complexity, making real-time processing infeasible.

These limitations have driven a shift toward deep learning-based approaches. Convolutional Neural Networks (CNNs) enable end-to-end feature learning, automatically extracting hierarchical features while modeling both local details and global context, significantly improving segmentation accuracy and adaptability in complex street scenes. However, spatial constraints from traditional machine learning, such as CRF's smoothing priors[9], have been integrated into modern segmentation networks, leading to hybrid architectures that combine CNN backbones with graph model optimization, continuously influencing the evolution of semantic segmentation technology.

2.2. Semantic Segmentation Based on Deep Learning

Deep learning-driven semantic segmentation technology achieves fine-grained analysis of complex scenes through end-to-end feature learning and context modeling. This technology overcomes the limitations of traditional feature engineering and has demonstrated revolutionary advantages in street scene understanding and medical image analysis. The core architectures and key advancements are outlined below:

2.2.1. Evolution of Basic Network Architectures

Fully Convolutional Network (FCN): The first model to extend CNNs to pixel-level predictions, utilizing deconvolution layers to restore spatial resolution, laying the foundation for modern segmentation networks[10]. **Encoder-Decoder Structure:** U-Net introduces skip connections[11], fusing shallow detail features with deep semantic features, significantly improving segmentation accuracy for small objects. **Atrous Convolution Networks[12]:** The DeepLab series employs Atrous Spatial Pyramid Pooling (ASPP) to capture multi-scale contextual information, addressing challenges related to object scale variations[13]. **Transformer-Based Architectures:** Models such as SETR and SegFormer leverage self-attention mechanisms to establish global dependencies, enhancing long-range context modeling capabilities.

2.2.2. Key Technological Breakthroughs in Street Scene

(1) Multi-Scale Feature Fusion

Feature Pyramid Network (FPN): Constructs top-down and lateral connections to achieve cross-level feature interactions. **Dynamic Convolutions:** Networks such as CondConv and DyNet adaptively adjust convolution kernel parameters to enhance model adaptability in complex environments. **Attention-Guided Mechanisms:** The spatial/channel dual-attention module (CBAM) strengthens responses in key regions, improving segmentation accuracy[14].

(2) Real-Time Optimization Design

Dual-Branch Architecture: BiSeNet separates spatial detail processing from semantic information processing, improving speed by 3×[15]. **Knowledge Distillation:** Transfers knowledge from a teacher network to a lightweight student network (e.g., STDCNet), reducing computation while maintaining accuracy. **Hardware-Aware Compression:** Utilizes channel pruning and mixed-precision quantization to ensure efficient model deployment on embedded devices.

(3) Adaptation to Dynamic Scenes

Deformable Convolutions (DCNv2): Adaptively adjusts sampling locations to enhance feature extraction for occluded objects[16]. **Domain Adaptation:** Adversarial training is used

to align synthetic data with real-world scene feature distributions, improving model generalization. Online Adaptive Normalization: Dynamically adjusts normalization layer parameters to handle lighting variations, ensuring robust segmentation under diverse weather conditions.

2.2.3. Emerging Research Directions

3D Segmentation Networks: Models like PointNet++ and Cylinder3D integrate point cloud and image data, advancing 3D scene understanding. **Video Semantic Segmentation:** Optical flow-based feature propagation (e.g., Vid-ODE) improves temporal consistency in video segmentation. **Self-Supervised Pretraining:** Frameworks such as MAE and DINO learn universal scene representations from unlabeled data, enhancing performance with limited annotations. **Spiking Neural Networks (SNN):** Event-camera-based dynamic vision processing enables real-time responses at the microsecond level. **Neural-Symbolic Systems:** These models incorporate rule-based constraints into neural network training (e.g., Neural-Symbolic Recursive Reasoner), improving semantic consistency and interpretability.

Currently, deep learning-based semantic segmentation is evolving toward scene adaptability, computational efficiency optimization, and cognitive enhancement. Hybrid architectures combining Transformer and CNNs[17] (e.g., HRFormer), Neural Radiance Fields (NeRF) for 3D reconstruction, and large-scale models like Segment Anything Model (SAM) are continually expanding the application boundaries of semantic segmentation. These advancements are driving more robust environmental perception solutions for applications such as autonomous driving, digital twins, and smart cities.

3. Relevant Datasets and Evaluation Metrics

3.1. Datasets

3.1.1. Cityscapes Dataset

Cityscapes is a large-scale dataset for semantic understanding and object detection, specifically designed for urban scene analysis. It was jointly developed by the University of Stuttgart and the Max Planck Institute for Informatics in Germany. Released in 2016, this dataset aims to provide a large-scale, high-quality resource for advancing research in urban scene understanding. The Cityscapes dataset consists of high-resolution images collected from various cities, annotated at the pixel level to classify and identify different objects in urban environments, such as cars, pedestrians, roads, and buildings. Additionally, it includes images captured under different weather conditions (e.g., sunny, rainy, snowy) and at various times of the day (daytime and nighttime), enabling models to be trained and tested in diverse environmental conditions.

Key features of the Cityscapes dataset:

(1) Large scale and high resolution – The dataset includes a vast collection of high-resolution images, each with a resolution of 1024×2048 pixels, making it suitable for training and evaluating complex deep learning models.

(2) Pixel-level annotations – Each image is labeled at the pixel level, providing semantic information for objects such as cars, pedestrians, roads, and buildings, along with bounding box annotations for object detection.

(3) Diverse urban environments – The dataset contains images from 30 different cities, covering a wide range of urban scenarios, including city centers, suburbs, roads, and

parks.

(4) Varied environmental conditions – It includes images captured under different weather conditions (sunny, rainy, snowy) and at different times of the day (daytime and nighttime), enhancing model adaptability.

(5) Versatile applications – The Cityscapes dataset is widely used in semantic segmentation, instance segmentation, and object detection, making it a critical resource for developing urban scene understanding algorithms.

Overall, the Cityscapes dataset is an essential resource for urban scene analysis and computer vision research, with broad applications in autonomous driving, intelligent transportation, and urban planning.

3.1.2. CamVid Dataset

CamVid (Cambridge-driving Labeled Video Database) is a dataset released by the University of Cambridge for urban road scene analysis. It consists of 701 precisely annotated images designed for semantic segmentation. If used for object detection, bounding box (bbox) annotations are required.

The dataset is divided into training, validation, and test sets, providing a well-structured benchmark for evaluating segmentation models.

3.1.3. Pascal VOC2012 Dataset

The Pascal VOC2012 dataset has been widely recognized and used since its release. It was originally designed for the PASCAL Visual Object Classes Challenge (VOC Challenge) in 2012 and has since become a fundamental dataset for machine learning and deep learning research. The primary objective of this dataset is to create challenges for visual object recognition in real-world scenes. As a labeled image dataset, it represents a standard supervised learning problem.

The Pascal VOC2012 dataset consists of 11,530 images with extensive annotations covering 20 object categories. These categories reflect real-world complexity, making it a valuable resource for algorithm research and optimization.

The dataset is widely applied in various computer vision tasks, including: **Object detection** – Identifies and localizes objects within images using bounding boxes, serving as a key benchmark for detection algorithms. **Object classification** – Categorizes objects in images based on their classes, supporting classification algorithm development. **Semantic segmentation** – Assigns a class label to each pixel, enabling fine-grained segmentation. **Instance segmentation** – Differentiates between individual instances of the same object category, extending semantic segmentation to object-level segmentation.

Additionally, Pascal VOC2012 serves as a performance benchmark for evaluating detection and segmentation models. Due to its historical significance and broad adoption, it is supported by many open-source tools and frameworks (e.g., TensorFlow, PyTorch), making dataset access and utilization more convenient.

Table 3-1: Comparison of Three Datasets

Dataset	Characteristics	Application Scenarios
Cityscapes	High resolution, multi-weather conditions, pixel-level annotations	Autonomous driving, urban scene analysis
CamVid	Video sequence annotations, suitable for temporal segmentation	Dynamic scene understanding
Pascal VOC2012	Multi-category, supports object detection and instance segmentation	General segmentation benchmark

3.2. Evaluation Metrics

Semantic segmentation is a crucial task in computer vision, aiming to classify every pixel in an image into a specific category. The following metrics are commonly used to evaluate segmentation model performance:

Accuracy

Evaluation aspect: Overall classification performance.

Accuracy measures how well a model correctly classifies pixels across the entire image, while accuracy provides an overall performance measure, it may be misleading when classes are imbalanced. calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where:

TP (True Positive): Correctly classified positive pixels

TN (True Negative): Correctly classified negative pixels

FP (False Positive): Incorrectly classified positive pixels

FN (False Negative): Incorrectly classified negative pixels

Mean Intersection over Union (mIoU)

Evaluation aspect: Segmentation quality across categories.

Intersection over Union (IoU) measures the overlap between predicted and ground-truth segmentation masks. The mean IoU (mIoU) averages IoU values across all classes, making it a widely used metric in semantic segmentation:

$$IoU = \frac{TP}{TP + FP + FN}$$

Mean Precision (mP)

Evaluation aspect: Accuracy of positive class predictions.

Precision measures how many predicted positive pixels are correct. Mean Precision (mP) is the average precision across all classes:

$$Precision = \frac{TP}{TP + FP}$$

Mean Recall (mR)

Evaluation aspect: Coverage of positive class samples.

Recall indicates how many actual positive pixels were correctly classified. Mean Recall (mR) is the average recall across all classes:

$$Recall = \frac{TP}{TP + FN}$$

F1 Score

Evaluation aspect: Balance between precision and recall.

F1 Score is the harmonic mean of precision and recall, useful when both are equally important:

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Dice Coefficient

Evaluation aspect: Similarity between predicted and ground-truth segmentation.

The Dice coefficient is similar to the F1 Score and measures the overlap between predicted and actual masks. A higher Dice score indicates better segmentation performance:

$$Dice\ Coefficient = \frac{2 \times TP}{2 \times TP + FP + FN}$$

4. Strategies for Overcoming Challenges

4.1. Small Object Detection and Occlusion Handling

In the field of computer vision, small object detection and occlusion handling are among the most challenging problems, directly impacting the reliability of applications such as autonomous driving and video surveillance. The difficulty in small object detection arises from limited pixel coverage, leading to weak feature representation. Traditional Convolutional Neural Networks (CNNs) often lose fine-grained details due to downsampling operations. For example, in aerial images, vehicles may occupy only 20×20 pixels, making their texture and shape features almost indistinguishable from background noise. Occlusion issues stem from the discontinuity of visible regions, such as individuals in a dense crowd revealing only partial limbs, making it difficult for traditional detectors to infer the complete object from local fragments.

These two problems are often intertwined—small objects are more susceptible to occlusion, and occlusion further exacerbates the difficulty of extracting small object features.

For street scene elements like traffic signs and signal lights, high-resolution feature maps combined with contextual information can be used. Specifically, preserving high-resolution details at shallow network layers while capturing contextual information at deeper layers allows for more accurate small object localization. For occlusion handling, multi-scale prediction and instance segmentation approaches can be employed, first detecting occluded object instances and then using contextual inference to classify occluded pixels.

Technological Advancements and Solutions:

4.1.1. Multi-Scale Feature Enhancement:

Feature Pyramid Network (FPN): Uses lateral connections to integrate shallow high-resolution features with deep semantic features, improving small object detection performance by 5.3% AP on the COCO dataset. Atrous Spatial Pyramid Pooling (ASPP): Captures multi-scale contextual features using different dilation rates, enhancing occluded object boundary prediction in Cityscapes dataset. Super-Resolution Reconstruction: Generative Adversarial Networks (GANs) perform 4× image upscaling, leading to an 18% improvement in small vehicle detection recall on the KITTI dataset.

4.1.2. Attention-Guided Mechanisms:

Spatial-Channel Dual Attention: The SENet module dynamically calibrates channel weights, while CBAM further enhances spatial attention for key regions, reducing false detection rates of small objects by 22% in the VisDrone UAV dataset. Temporal Attention: VideoBERT leverages frame-to-frame correlation modeling and motion cues to infer the locations of occluded objects, reducing ID switch occurrences by 37% in MOTChallenge tracking tasks.

4.1.3. Occlusion Modeling and Completion:

Deformable Convolution Network (DCN): Adaptively adjusts convolution kernel sampling positions to capture irregular object contours, achieving 74.6% mAP on the Occluded-Vehicles dataset. Partial Occlusion Inference: CenterNet++ introduces an occlusion-aware branch, predicting visible region probability maps, improving occluded object detection accuracy by 9.8 percentage points

in the Pascal VOC dataset. Physical Simulation-Based Augmentation: Blender-generated synthetic data with randomized occlusions reduced miss detection rates by 13.2% in the Waymo Open Dataset.

4.1.4. Frontier Research Directions:

Neural Radiance Fields (NeRF): Enables 3D scene reconstruction for occlusion reasoning, recovering 62% of severely occluded objects in the KITTI-360 dataset. Spiking Neural Networks (SNNs): Mimics the biological vision system's dynamic information processing, achieving 83.4% IoU in DAVIS dynamic occlusion segmentation tasks. Cross-Modal Fusion: Integrating LiDAR point cloud data with RGB images improves detection robustness in heavily occluded environments, boosting stability by 41% in the nuScenes dataset.

4.2. Real-Time Processing Requirements

In autonomous driving scenarios, semantic segmentation must provide real-time feedback. On one hand, this requires optimizing model architectures by adopting lightweight convolutional models such as MobileNet, reducing parameter size. On the other hand, knowledge distillation transfers knowledge from complex high-capacity models to compact student models, maintaining accuracy while improving inference speed, ensuring real-time processing of street scenes.

In applications like autonomous driving, industrial inspection, and mobile robotics, real-time segmentation is a critical constraint for system deployment. The fundamental challenge lies in achieving a balance between computational efficiency and segmentation accuracy under limited hardware resources, ensuring the model's output strictly adheres to real-time constraints.

The main difficulties in real-time semantic segmentation include:

Computational-Accuracy Tradeoff: Large deep models require significant computational resources, making them unsuitable for edge computing devices. Robustness in Complex Environments: Lighting variations and extreme weather conditions cause feature distribution shifts, reducing segmentation reliability. High-Resolution Memory Overhead: Downsampling layers in standard networks lead to the loss of fine-grained features, while high-resolution inputs significantly increase GPU memory consumption.

Optimized Real-Time Strategies:

To tackle these challenges, real-time segmentation solutions focus on algorithmic innovations:

(1) Lightweight Architecture Design: Reducing computational overhead through efficient backbone architectures. Dual-branch networks separate spatial detail processing from semantic extraction, allowing adaptive computation based on scene complexity.

(2) Attention-Guided Distillation: Teacher-student models guide lightweight networks to focus on critical features, improving segmentation efficiency. Multi-dimensional attention modules enhance feature selection and filtering, reducing computational redundancy.

(3) Hardware-Aware Compression: Low-precision quantization and structured model pruning minimize computational costs while preserving model performance. These techniques, when combined, form a systematic optimization pipeline that ensures efficient semantic segmentation in real-world, high-speed applications.

5. Future Prospects

5.1. Lifelong Learning and Continuous Adaptation

Street scenes evolve over time, requiring segmentation models to possess lifelong learning capabilities to continuously learn from new data and update their knowledge. Meta-learning-based approaches enable models to rapidly adapt to new urban areas and changes in traffic regulations, reducing the cost of retraining while improving segmentation performance in dynamic street environments.

5.2. Quantum Computing for Model Evolution

With advancements in quantum computing, its immense computational power is expected to accelerate deep model training, particularly for processing large-scale street scene data. This breakthrough has the potential to overcome traditional computational bottlenecks, enabling the exploration of more complex model architectures for ultra-precise and ultra-fast semantic segmentation, ushering in a new era of street scene understanding.

Conclusion: Street scene semantic segmentation has undergone a long journey of evolution, transitioning from handcrafted features to deep learning-driven intelligent segmentation. Despite numerous challenges, multi-domain integration and technological innovations continue to drive breakthroughs. In the future, as emerging technologies become integrated, semantic segmentation will play a pivotal role in various smart city applications, reshaping the perception and application paradigms of street scene analysis.

References

- [1] Feng D, Haase-Schütz C, Rosenbaum L, et al. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(3): 1341-1360.
- [2] Tremblay M, Arjona-Medina J, Unterthiner T, et al. Speeding up semantic segmentation for autonomous driving[J]. 2016.
- [3] Garcia-Garcia A, Orts-Escolano S, Oprea S, et al. A review on deep learning techniques applied to semantic segmentation[J]. arXiv preprint arXiv:1704.06857, 2017.
- [4] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25.
- [5] Jiang F, Grigorev A, Rho S, et al. Medical image semantic segmentation based on deep learning[J]. Neural Computing and Applications, 2018, 29: 1257-1265.
- [6] Pohle R, Toennies K D. Segmentation of medical images using adaptive region growing[C]//Medical Imaging 2001: Image Processing. SPIE, 2001, 4322: 1337-1346.
- [7] Halder A, Dey D, Sadhu A K. Lung nodule detection from feature engineering to deep learning in thoracic CT images: a comprehensive review[J]. Journal of digital imaging, 2020, 33(3): 655-677.
- [8] Seghers D, Hermans J, Loeckx D, et al. Model-based segmentation using graph representations[C]//Medical Image Computing and Computer-Assisted Intervention-MICCAI 2008: 11th International Conference, New York, NY, USA, September 6-10, 2008, Proceedings, Part I 11. Springer Berlin Heidelberg, 2008: 393-400.

- [9] Chen S, Gamechi Z S, Dubost F, et al. An end-to-end approach to segmentation in medical images with CNN and posterior-CRF[J]. *Medical image analysis*, 2022, 76: 102311.
- [10] Dai J, Li Y, He K, et al. R-fcn: Object detection via region-based fully convolutional networks[J]. *Advances in neural information processing systems*, 2016, 29.
- [11] Sahli H, Ben Slama A, Labidi S. U-Net: A valuable encoder-decoder architecture for liver tumors segmentation in CT images[J]. *Journal of X-ray science and technology*, 2022, 30(1): 45-56.
- [12] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[J]. *arXiv preprint arXiv:1706.05587*, 2017.
- [13] Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. *IEEE transactions on pattern analysis and machine intelligence*, 2017, 40(4): 834-848.
- [14] Liu T, Luo R, Xu L, et al. Spatial channel attention for deep convolutional neural networks[J]. *Mathematics*, 2022, 10(10): 1750.
- [15] Feng H, Liu W, Xu H, et al. A lightweight dual-branch semantic segmentation network for enhanced obstacle detection in ship navigation[J]. *Engineering Applications of Artificial Intelligence*, 2024, 136: 108982.
- [16] Wang W, Dai J, Chen Z, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions[C]//*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023: 14408-14419.
- [17] Yuan Y, Fu R, Huang L, et al. Hrformer: High-resolution vision transformer for dense predict[J]. *Advances in neural information processing systems*, 2021, 34: 7281-7293.