ISSN: 2688-8645

Analysis of Pedestrian Semantic Segmentation Technology in Autonomous Driving Scenarios under Occlusion Conditions

Yingxin He

School of Professional Studies, New York University, New York, NY 10003, USA

yh5909@nyu.edu

Abstract

Semantic segmentation has become a cornerstone of visual scene understanding, particularly in safety-critical domains such as autonomous driving, robotics, and urban surveillance. Recent advances in convolutional and Transformer-based deep learning models have yielded strong performance on standard benchmarks under ideal conditions. However, traditional semantic segmentation methods struggle in real-world scenes characterized by occlusions, overlaps, and partial visibility, which often result in prediction failures and poor generalization. In response to these limitations, occlusion-aware segmentation has emerged as a new paradigm, incorporating visibility modeling, occlusion reasoning, structural restoration, and temporal completion strategies. This paper presents a comprehensive survey of both traditional and occlusion-aware semantic segmentation approaches, with a structured analysis of their evolution, strengths, and limitations. The review highlights the growing importance of integrating structural and contextual reasoning to improve robustness in occluded environments and identifies future directions for developing scalable, accurate, and occlusion-resilient segmentation systems.

Keywords

Semantic Segmentation; Occlusion-Aware Modeling; Instance-Level Reasoning; Boundary Refinement; Temporal Feature Alignment.

1. Introduction: Semantic Segmentation and the Challenge of Occlusion

Semantic segmentation, which assigns a categorical label to each pixel in an image, has become a foundational task in modern computer vision. Enabled by the advancement of deep learning, semantic segmentation serves as the cornerstone of various high-level scene understanding applications. In autonomous driving, it allows precise parsing of critical elements such as roads, vehicles, pedestrians, and traffic signs, thereby enhancing the reliability and safety of navigation systems. In robotics, it enables intelligent agents to understand and interact with their environment for efficient path planning and dynamic obstacle avoidance. In urban surveillance and public safety, semantic segmentation assists in people flow analysis, behavior monitoring, and anomaly detection. In the medical field, it is applied to organ and lesion localization, diagnostic assistance, and surgical planning.

Fueled by convolutional neural networks (CNNs), models such as DeepLab, Mask R-CNN, PSPNet, OCRNet, and Panoptic FPN have achieved impressive performance on benchmark datasets including PASCAL VOC, Cityscapes, ADE20K, and COCO-Stuff. These models have significantly pushed the boundary of semantic segmentation from research to real-world deployment [1-3].

However, challenges persist in complex environments where occlusion and object overlap are prevalent. In such scenes, traditional segmentation approaches often fail to distinguish heavily occluded targets, leading to incomplete recognition, unstable predictions, and poor

ISSN: 2688-8645

generalization [4,5]. For instance, pedestrian occlusion in crowded urban environments, fuzzy object boundaries, and missing part structures can severely degrade the performance of vision systems. This limitation is especially critical in safety-sensitive domains such as autonomous driving.

To address these challenges, the concept of occlusion-aware semantic segmentation has gained traction. New models integrate visibility reasoning, instance-level occlusion ordering, part-based feature enhancement, and explicit region recovery mechanisms to enhance robustness under occlusion. This paper presents a comprehensive overview of the evolution of semantic segmentation techniques, comparing traditional approaches and emerging occlusion-aware models, with a particular focus on their applicability to real-world occlusion scenarios such as pedestrian detection in autonomous driving.

2. Traditional Semantic Segmentation: Structured Progress without Occlusion Awareness

2.1. Stage I: Fully Convolutional Network (FCN) and Encoder-Based Architectures

The shift to deep learning-based semantic segmentation began with the introduction of Fully Convolutional Networks (FCNs) [1]. This seminal work restructured classification CNNs to support arbitrary-sized, end-to-end pixel-wise prediction. By incorporating skip connections to merge low-level spatial details with high-level semantics, FCNs significantly improved spatial accuracy. Later, Zhang et al. extended FCNs to domain-adaptive settings, introducing hierarchical adaptation to mitigate domain shift issues [6]. This stage established a trainable architecture template for semantic extraction and structural modeling, though limited by fixed receptive fields and insufficient multi-scale context.

2.2. Stage II: Receptive Field Expansion and Atrous Convolution

DeepLabv3 addressed receptive field limitations with Atrous Spatial Pyramid Pooling (ASPP), introduced by Chen et al., which captures multi-scale context using convolutions with varying dilation rates while preserving high-resolution details [2]. Follow-up works expanded ASPP to weakly supervised tasks and specialized domains like remote sensing (e.g., OBIA, SSIM-enhanced ASPP), culminating in structural variants such as CHASPP [7-9]. This stage emphasized hierarchical and dense sampling but relied on static context stacking without dynamic scale interaction.

2.3. Stage III: Multi-Scale Structure Modeling and Fusion

To overcome static context fusion, PSPNet employed Pyramid Pooling Modules (PPM) to aggregate multi-scale features [10]. Lin et al. proposed MSCI, introducing LSTM-based inter-scale semantic flow [11]. These works marked a transition to semantic-structural interaction. However, limitations in fine structure recovery remained.

2.4. Stage IV: Structural Restoration and Boundary Refinement

Encoder-decoder architectures like DeepLabv3+ introduced lightweight decoders for multi-stage upsampling [3]. UNet++ redesigned skip connections using nested decoders, while SegFix added model-agnostic boundary refinement using direction prediction and pixel displacement [12,13]. These works emphasized structure restoration and prepared the ground for fine-grained segmentation.

2.5. Stage V: Contextual Reasoning and Attention Mechanisms

Recent models adopt attention to enhance semantic consistency and long-range dependency modeling. OCRNet incorporated pixel-to-class center representations [14]. Transformer-based

ISSN: 2688-8645

models like DC-Swin and ISSA introduced efficient attention fusion for global reasoning with a reduced computational load [15,16]. These models emphasize semantic-structural alignment and global dependency modeling.

2.6. Limitations in Occlusion Scenarios

Despite their success, traditional models assume full visibility, leading to common issues under occlusion: occluded objects misclassified as background, blurred or missing boundaries, and NMS mistakenly discarding true positives. These challenges prompt the need for occlusion-aware modeling.

3. Occlusion-Aware Semantic Segmentation: Structural Reasoning for Robust Recognition

3.1. Visibility Modeling and Part-Aware Feature Representation

These methods enhance the representation of visible regions via pose guidance or part-based pooling. PORoI pooling splits RoIs into sub-parts and aggregates features based on visibility confidence [17]. Pose-guided attention mechanisms dynamically weight feature channels to prioritize visible regions [4]. These approaches improve performance on datasets like CityPersons and Caltech, especially under moderate occlusion. However, they often require auxiliary annotations and are limited in generalizability.

3.2. Instance Ordering and Fusion Optimization

To address fusion failures in overlapping scenes, OCFusion introduced an occlusion prediction head to infer which instance should appear in front [18]. BCNet further employed bilayer GCNs to model inter-instance occlusion as a relational graph [19]. These methods achieve higher PQ on occlusion-rich datasets but incur quadratic computational costs.

3.3. Contour-Aware Boundary Recovery

Inspired by early works such as Leibe et al. and Gao et al., these methods refine boundaries using global shape priors and contour templates [5,20]. Techniques like Chamfer matching and MDL optimization are employed for mask correction in crowded scenes. While effective for edge enhancement, these approaches are not easily integrated into end-to-end frameworks.

3.4. Temporally-Guided Occlusion Completion

Video-based methods like TFC align features from unoccluded reference frames via deformable convolutions. This strategy recovers occluded regions across time and demonstrates notable gains on OVIS. However, it is constrained to video input and is sensitive to alignment noise.

3.5. Summary

Occlusion-aware methods address the blind spots of traditional models by incorporating structural reasoning, temporal consistency, and explicit visibility cues. Despite challenges in annotation and computational load, they represent a critical advancement toward robust real-world segmentation under occlusion.

4. Conclusion

Semantic segmentation has made significant progress through the evolution of network architectures, multi-scale fusion strategies, and attention mechanisms. However, the assumption of full object visibility limits the effectiveness of traditional models in occlusion-heavy environments. Occlusion-aware segmentation methods bridge this gap by introducing visibility-aware modeling, structural reasoning, and cross-frame completion. This survey synthesizes the progression from conventional models to occlusion-specific designs, revealing

that robustness in complex real-world scenes demands a holistic integration of spatial, temporal, and semantic cues. As autonomous systems become more ubiquitous, addressing occlusion explicitly will be essential to ensuring their safety and reliability.

5. Future Work

Although occlusion-aware segmentation has shown promising results, several challenges remain. First, the reliance on detailed annotations such as visibility masks, part labels, and occlusion order increases the cost of dataset creation. Future research could focus on reducing this dependency through weakly supervised or self-supervised approaches. Second, many occlusion-aware methods suffer from scalability limitations due to high computational complexity, particularly those involving pairwise instance modeling or video frame alignment. Developing lightweight architectures with efficient reasoning capabilities remains an open problem. Finally, unifying spatial reasoning, depth estimation, and temporal fusion into a single end-to-end framework could offer a path toward comprehensive occlusion understanding. Exploring such multi-modal, multi-scale architectures—possibly leveraging 3D information or large vision-language models—may unlock the next wave of innovation in robust semantic segmentation.

Refereneces

[1] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

[2] Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587..

[3] Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV) (pp. 801-818).

[4] Zhang, S., Yang, J., & Schiele, B. (2018). Occluded pedestrian detection through guided attention in CNNs. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 6995-7003).

[5] Leibe, B., Seemann, E., & Schiele, B. (2005, June). Pedestrian detection in crowded scenes. In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) (Vol. 1, pp. 878-885). IEEE.

[6] Zhang, Y., Qiu, Z., Yao, T., Liu, D., & Mei, T. (2018). Fully convolutional adaptation networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6810-6818).

[7] Du, S., Du, S., Liu, B., & Zhang, X. (2021). Incorporating DeepLabv3+ and object-based image analysis for semantic segmentation of very high resolution remote sensing images. International Journal of Digital Earth, 14(3), 357-378.

[8] He, H., Yang, D., Wang, S., Wang, S., & Li, Y. (2019). Road extraction by using atrous spatial pyramid pooling integrated encoder-decoder network and structural similarity loss. Remote Sensing, 11(9), 1015.

[9] Lian, X., Pang, Y., Han, J., & Pan, J. (2021). Cascaded hierarchical atrous spatial pyramid pooling module for semantic segmentation. Pattern Recognition, 110, 107622.

[10] Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2881-2890).

[11] Lin, D., Ji, Y., Lischinski, D., Cohen-Or, D., & Huang, H. (2018). Multi-scale context intertwining for semantic segmentation. In Proceedings of the European conference on computer vision (ECCV) (pp. 603-619).

ISSN: 2688-8645

[12] Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2019). Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE transactions on medical imaging, 39(6), 1856-1867.

[13] Yuan, Y., Xie, J., Chen, X., & Wang, J. (2020, August). Segfix: Model-agnostic boundary refinement for segmentation. In European conference on computer vision (pp. 489-506). Cham: Springer International Publishing.

[14] Yuan, Y., Chen, X., & Wang, J. (2020, August). Object-contextual representations for semantic segmentation. In European conference on computer vision (pp. 173-190). Cham: Springer International Publishing.1.

[15] Wang, L., Li, R., Duan, C., Zhang, C., Meng, X., & Fang, S. (2022). A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images. IEEE Geoscience and Remote Sensing Letters, 19, 1-5.

[16] Huang, L., Yuan, Y., Guo, J., Zhang, C., Chen, X., & Wang, J. (2019). Interlaced sparse self-attention for semantic segmentation. arXiv preprint arXiv:1907.12273.

[17] Zhang, S., Wen, L., Bian, X., Lei, Z., & Li, S. Z. (2018). Occlusion-aware R-CNN: Detecting pedestrians in a crowd. In Proceedings of the European conference on computer vision (ECCV) (pp. 637-653).

[18] Lazarow, J., Lee, K., Shi, K., & Tu, Z. (2020). Learning instance occlusion for panoptic segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10720-10729).

[19] Ke, L., Tai, Y. W., & Tang, C. K. (2021). Deep occlusion-aware instance segmentation with overlapping bilayers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4019-4028).

[20] Gao, T., Packer, B., & Koller, D. (2011, June). A segmentation-aware object detection model with occlusion handling. In CVPR 2011 (pp. 1361-1368). IEEE.