Proceedings of the 4th International Conference on Mechatronics and Smart Systems
DOI: 10.54254/2755-2721/2026.35609

A Survey of Deep Learning-Driven Multi-Modal Perception Fusion for Autonomous Driving

Xichen Huang

RCF Experimental School, Beijing, China
huangxichen@rdfzcygj.cn

Abstract. This paper presents a comprehensive survey on deep learning based multi-modal perception fusion frameworks for autonomous driving systems. The traditional single-modal solutions have intrinsic limitations in robustness against harsh weather and dynamic environments, and the integration of complementary sensors, such as cameras, LiDAR, and millimeter-wave radar, provides a critical pathway to high-level automation. We provide a systematic review of the basic fusion paradigms, classifying them into early, late and deep (feature-level) architectures, and evaluating their trade-offs in terms of computational overhead, architectural modularity and joint feature learning. Special focus is given to state-of-the-art architectures, comparing the high representational power of the Tensor Fusion Networks (TFN) to the computationally-efficient Multi-modal Circulant Fusion (MCF) and context-aware Transformer frameworks. This study places real-world deployment challenges in the context of algorithmic paradigms, investigating the friction between “black-box” deep learning models and stringent functions. Besides algorithmic paradigms, this work puts real-world deployment challenges into context and discusses the friction between ``black-box" deep learning models and rigorous functional safety standards (e.g., ISO 26262), and energy constraints of new energy vehicles. Finally, we explore the paradigm shift to "vehicle-road synergy" (V2X) infrastructure as a critical mechanism for providing the safety redundancy and edge-computing capabilities needed for fully reliable, next-generation autonomous driving.

Keywords: Multi-Modal Signal Processing, Autonomous Driving, Signal Fusion

1. Introduction

Autonomous driving technology is the intersection of the automotive industry and artificial intelligence. Environmental perception is a key prerequisite for high levels of automation. Conventional single-modal perception solutions (e.g., based on only cameras or radar) have severe bottlenecks in robustness and reliability under the challenges of complex weather, lighting variations and occlusions.

In particular, camera vision systems can gather a lot of information but are vulnerable to extreme lighting and weather. LiDAR provides an accurate 3D geometry structure, but the performance is poor in rain and fog. Radar provides all-weather capability but it degrades texture and shape details. The limitations of these sensors necessitate integration and analysis of data from these sensors.

DOI: 10.54254/2755-2721/2026.35609

As deep learning becomes more developed, appropriate multi-modal perception solution strategies combining information from heterogeneous sensors (e.g. cameras, LiDAR, millimeter-wave radar) become one of the promising ways to improve system performance. This is not a trivial aggregation of data — they are complicated algorithms that align, weight and process hierarchical layers. Switching gears, deep learning has quickly become the go-to method for these tasks, surpassing earlier data analysis algorithms, due to its power in representation learning and end-to-end optimization. However, at present there is no consolidated summary which examines fusion approaches and their use in different driving scenarios (e.g. city navigation and highway cruise). This gap drives us to consolidate this knowledge and outline a clear research direction for both researchers and engineers.

This study proposes a systematic review and comparison of deep learning-based multi-modal fusion techniques with traditional perception methods. Its purpose is to provide some theoretical references and technical support for the design optimization of autonomous driving system on new energy vehicles, especially for the functional safety specifications (e.g. This paper will firstly do some review about the development history of autonomous driving perception technology. It will then turn to the investigation of several deep learning-based multi-modal fusion architectures, what solutions are available along with a benefit–cost analysis where tradeoffs and challenges will be elaborated. Finally, it will cover future-oriented technology trends with associated industry application cases.

The logical progression of this structure from historical context, through relevant technical analysis, and into a forward-looking synthesis is designed to build an integrated body of knowledge. It focuses to be not only a theory reference for the researchers, but also as an executable guide for engineers confronted with this challenge in real-world system developing.

2. Literature review

2.1. Traditional perception methods and their limitations

Autonomous driving systems in their infancy used hand manufactured feature extraction techniques to see sensor data. For example, methods like Histogram of Oriented Gradients (HOG) and Scale-Invariant Feature Transform (SIFT) were heavily used to identify pedestrians and cars [1]. HOG: HOG describes the distribution of local intensity gradients and edge directions within a dense grid of cells which works great in pedestrian detection in controlled environments. Instead, SIFT identifies and describes local features in images invariant to scale, rotation and affine distortion allowing for robust feature matching across different viewpoints. These approaches contain local gradients and invariant features which are the basis of object recognition. Nonetheless, they are fundamentally reliant on manual feature engineering which is a major bottleneck in making them generalizable to novel or unseen scenarios.

Though these conventional techniques are interpretable and computationally efficient, they lack scalability necessary for complex and dynamic environments, as in reality driving. They struggle to obtain semantic high-level representations and fail at robustness under occlusion, viewpoint change and variations in the environment such as lighting and weather. As a specific example, detectors based on the HOG features suffer from severe degradation in their performance during nighttime conditions or heavy rain when gradient information may be unreliable. Likewise, SIFT-based matching faces challenges with textureless surfaces and repetition patterns frequently found in highway or tunnel situations. As a result, those key constraints provide the direct impetus for

DOI: 10.54254/2755-2721/2026.35609

switching to data-driven learning paradigms that can Auto-discover useful features from the original sensor streams with no domain-specific engineering assumptions necessary.

2.2. The rise of deep learning in perception tasks

Deep learning using convolutional neural networks (CNNs): a paradigm shift in computer vision from handcrafted features to end-to-end hierarchical representations. Detection techniques Faster R-CNN and YOLO (you only look once) have developed more accurate detection results at greater speed with lower deficits [2, 3]. Faster R-CNN shares full-image convolutional features with the detection network via Region Proposal Networks (RPN), achieving almost cost-free region proposals and state-of-the-art accuracy on standard object detection benchmarks. By proposing to remove the margin inherent in sliding-window approaches, YOLO demonstrates that object detection can be simplified as a single regression problem from image pixels to bounding box coordinates and class probabilities directly. YOLO makes predictions on bounding boxes and class probabilities from full images in a single evaluation, which results in very fast processing. Deep learning models learn multi-scale spatial features and semantic abstractions from the data that are generalizable across various driving conditions than hand-crafted features. The also have the ability to learn end-to-end from raw pixels to semantic labels, avoiding handcraft intermediate representations. A recent PointNet method does a deep learning directly on the raw point sets and is invariant to input permutation [4], thus the method is suitable for 3D point clouds from LiDAR sensor. It allows training the model all at once for 3D classification and segmentation. The fully convolutional networks (FCNs) is the most common approach for the semantic segmentation of images. DeepLab reframes image segmentation as a classification problem wherein each pixel is classified as either foreground or background. For each pixel it considers local context at multiple scales without subsampling. This model is a fixed-size input, and it uses atrous convolution to get multi-scale context, which yields very impressive improvements over FCN-8s baseline.

2.3. Evolution of multi-modal fusion strategies

This consists of fusing multiple sensor modalities on a single deep learning-based perception framework, to increase robustness and perceptual coverage. Fusion strategies are commonly subdivided into early, late and deep (i.e.: feature-level) fusion [5]. These categories offer various balances between the amount of information available during training, computational complexity, and modularity.

Early fusion Concatenates/project multiple sensor signals from the input domain into a single representation. This allows one to use as much joint information as possible in learning the network. But it is sensitive to temporal and spatial misalignments, e.g., camera and LiDAR sensor mismatch [6], and difficult to debug. We note that cancellation of such alignment errors in the input domain can persist through the entire network. In contrast, all modalities are intertwined in the same network with a certain degree of 'decoupling' between them. In this context, late fusion benefits are also taking advantage of the complementary aspect of each modality, while they may be more robust to sensor failures: the failure or poor performance of one specific sensor does not completely penalize the perception system. Then it paid for joint feature learning and processed simultaneously. As long as there is a layer for late fusion, a branch can be designed independently, ignored, or even replaces entire sub networks of a target modality. In such a case, representation of image could be combined with the LiDAR point-cloud representation through a late fusion layer to produce a final outcome [7]. The advanced fusion strategies used here all operate at an intermediate level to

DOI: 10.54254/2755-2721/2026.35609

leverage the strengths of both early and late fusion. The core of these capabilities is attention and learned weights that lead to interaction between modalities at the representation level. TFN models explicit intra- and inter-modality interactions with tensor outer products involving the overparameterization of all cross-modal correlations [8], but does so at high computational cost that scales cubically in feature dimensionality. MCF mitigates the scalability issue by approximating those tensor interactions with circulant matrices that have fewer parameters but are equally expressive [9]. Attention based models and more recently Transformers have become the state of the art for deep fusion by dynamically weighting how important modalities are with respect to each other and which regions individually carry most context information [10]. These methods for deep fusion are consistent with trends in representation learning, demonstrating success as part of an end-to-end autonomous driving pipeline [11]. Notably, the author of this paper further verifies that deep fusion at feature level consistently outperforms early and late fusion strategies on standard autonomous driving benchmarks with occlusion and weather included.

3. Deep learning-based multi-modal fusion methods

3.1. Core sensor technologies and fusion fundamentals

Cameras are rich in semantic and texture information, which is crucial for classification tasks. LiDAR provides precise 3D spatial structure, which is key to localization and geometry perception. Radar provides velocity extraction and reliability in rough weather conditions [7]. Figure 1 provides visual evidence of the complementary capability profiles among these three forms, highlighting their respective strengths across six key performance dimensions.

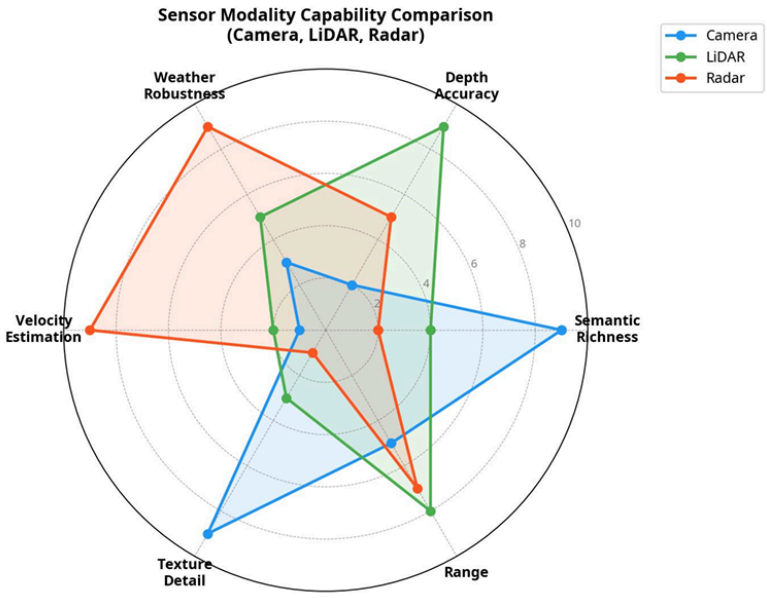

Figure 1. Sensor modality capability comparison. Radar chart illustrating the relative strengths of Camera, LiDAR, and Radar across six key dimensions

The necessity of fusion functionality comes from the embedded redundancy and complementarity. Redundant sensing strengthens reliability with overlapping coverage to ensure loss of a sensor or performance deterioration has limited effects on the overall system. By integrating information from multiple sources that cannot be captured by individual sensors, complementary

DOI: 10.54254/2755-2721/2026.35609

sensing expands the representation space and consequently enables robust perception in uncertainty. As shown in Figure 2, the three main fusion strategies — Early, Late and Deep Fusion — are characterized by essentially diverse locations and manners of combining modalities in the processing pipeline.

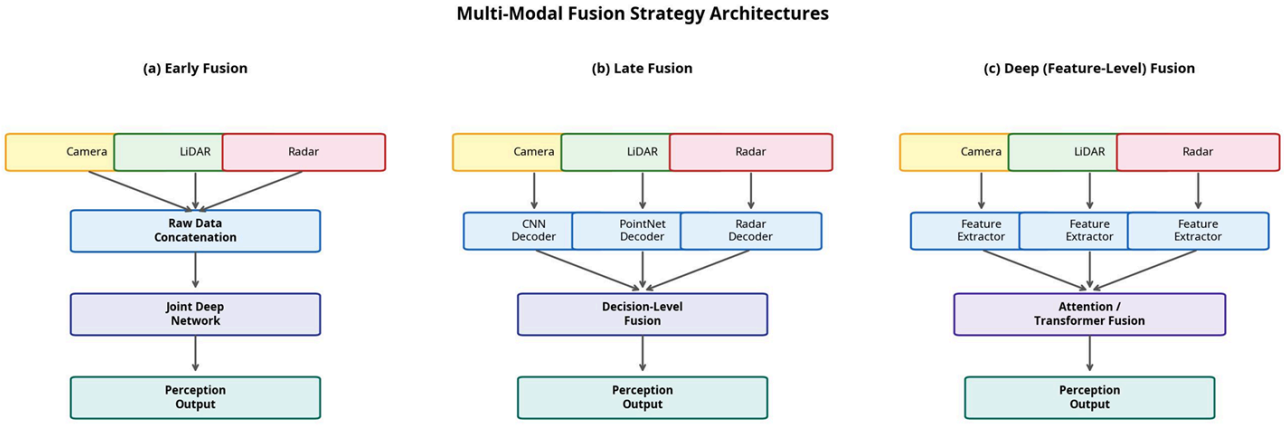

Figure 2. Multi-modal fusion strategy architectures. Comparison of (a) early fusion, (b) late fusion, and (c) deep feature-level fusion pipelines

3.2. Key deep learning fusion architectures

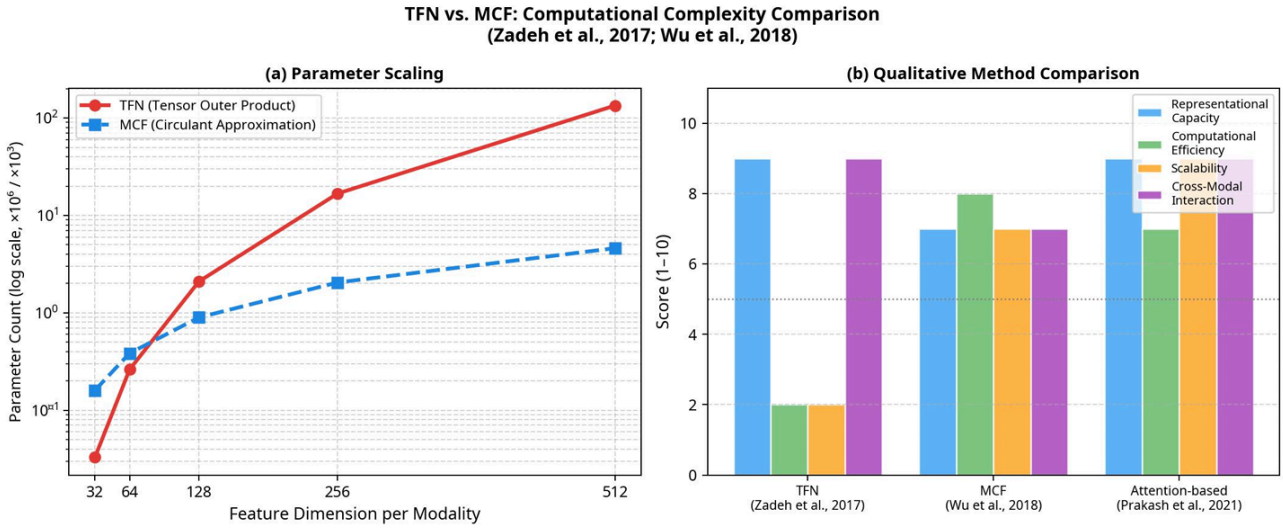

Figure 3. TFN vs. MCF computational complexity comparison. & Qualitative comparison of TFN, MCF, and Attention-based methods across four key criteria

Tensor Fusion Networks (TFN) explicitly model intra- and inter-modal interactions that capture full cross-modal correlations [8]. But, this comes with the price of exponential computational complexity since the number of parameters scales in a cubic manner w.r.t. to each modality feature dimension. However, this basic scalability limitation limits the applicability of TFN in computationally limited automotive embedded systems. To overcome this limitation, we approximate tensor interactions through circulant matrices and reduce parameters significantly while retaining capacity using Multi-modal Circulant Fusion (MCF) [9]. As illustrated in Figure 3(a), scaling of the parameter count of MCF is near-linear compared to the cubic scaling of TFN

DOI: 10.54254/2755-2721/2026.35609

rendering MCF much more effective for real world deployment. The qualitative comparison that we present in Figure 3(b) supports our findings as well since TFN scores high on representational capacity and cross-modal interaction but low on computational efficiency and scalability, while MCF is generally better across all metrics applied.

Attention-based models and Transformers represent the state-of-the-art in multi-modal fusion. By leveraging self-attention mechanisms, they dynamically assign importance weights across modalities and spatial regions, enabling context-aware fusion [10]. As illustrated in Figure 4, the architecture proposed by Prakash et al. [10] tokenizes features from each sensor modality: visual tokens from a CNN backbone, geometric tokens from a PointNet encoder [4], and velocity tokens from a radar feature extractor — and then applies cross-modal attention to produce a unified representation for downstream perception tasks. These architectures align with broader trends in representation learning and scalability, offering end-to-end trainable pipelines that can be optimized jointly across all modalities.

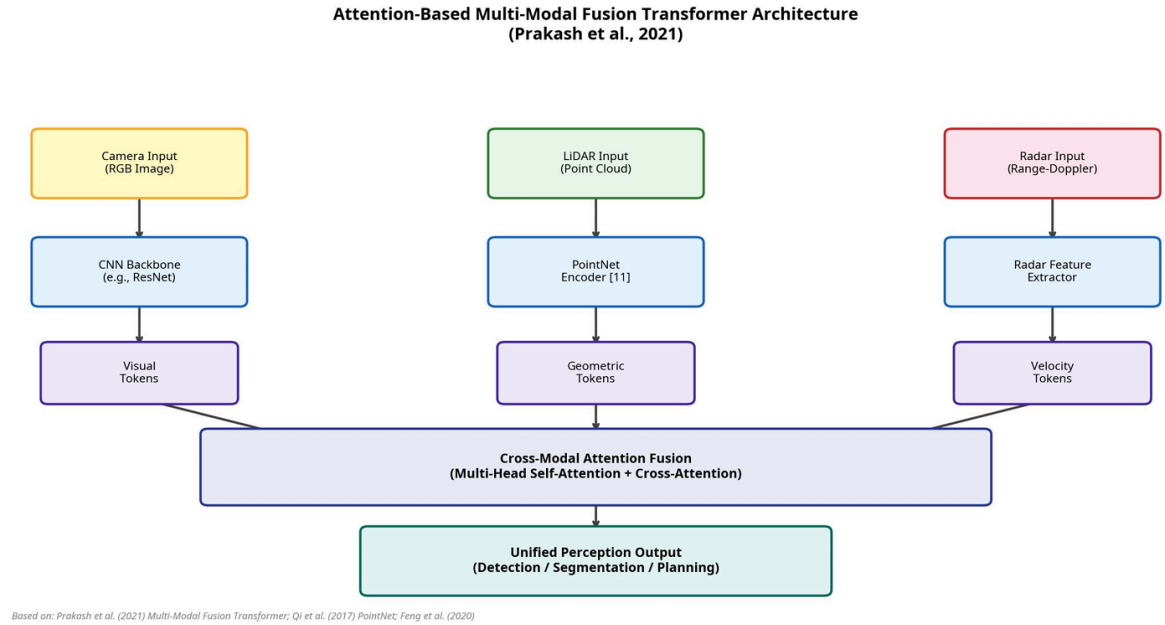

Figure 4. Attention-Based Multi-Modal Fusion Transformer Architecture. The pipeline tokenizes features from Camera (CNN), LiDAR (PointNet [4]), and Radar, then applies Cross-Modal Attention Fusion to produce a unified perception output. Based on Prakash et al. (2021) [10] and Qi et al. (2017) [4]

3.3. Task-specific fusion models

In the context of 3D object detection, PointNet [4] and its derivatives demonstrated great efficacy in processing point cloud data (a natural form for 3D data), as well as fusion with image features to increase semantic understanding. PointNet is a major contribution that employs a permutation-invariant backbone network to process unordered input in point sets and has served as a fundamental piece of many multi-modal detection pipelines. DeepLab uses atrous convolutions which are useful for semantic segmentation to capture multi-scale context [12]. Extending to multi-modal inputs, these models are shown to integrate geometric and semantic cues for improved pixel-level accuracy. Next, dense features from a camera-centered DeepLab are fused with LiDAR-derived depth maps to

DOI: 10.54254/2755-2721/2026.35609

provide both detailed understanding of the scene as well as high geometric accuracy at object boundaries, especially in occluded regions.

4. Conclusion

This paper reviews the research works on deep learning-based multi-modal perception fusion in autonomous driving. From classic methods like HOG, SIFT to Deep Neural Networks It investigates the capabilities of individual sensors (camera, LiDAR, radar), emphasizing the importance of multi-sensor data fusion for autonomous driving. The evaluation also covers various strategies for fusion. It demonstrated that the deep feature-level fusion scheme (e.g., Multi-modal Circulant Fusion, MCF and Transformer) could lead to the best performance and compression ratio trade-off with a higher robustness. Apart from algorithms, the implementation of these systems in the real world requires us to think beyond environments, norms and national strategies. Advanced autonomous drivingMhe applications of this technology face iron-clad safety regulations. "Black box" deep learning models are difficult to prove the very high functional safety requirements set forth in standards, such as ISO 26262. Governments have been clamping down on stricter data privacy and public safety regulations too. Such policies have proved extremely difficult for fully autonomous vehicles relying entirely on AI to use roads.

The Chinese government finds a compromise solution to cope with the rapid growth of AI and strict rules for its usage. In your paradigm of only using "single-vehicle intelligence" similar as Western companies, the government engages in the "vehicle-road synergy" (V2X) strategy. The government makes special pilot areas and allows businesses to test and operate self-driving vehicles but under the roof of the government. However, problems still exist. The world is full of complex cities, and we still struggle with so called "long-tail" edge cases for autonomous systems. Additionally, some complex models such as the Transformer demand significant computing resources, contradicting the energy saving objective of new energy vehicles. The renewal of infrastructure can be an attractive solution to the limits of vehicles and safety risks. This means upgrading typical roads, whether they are smart road network integration with sensors, edge computing and 5G technology. This type of infrastructure could give the safety redundancy required for fully autonomous driving by allowing the road itself to gather more information and to see farther than the sensors on a car.

References

[1] Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (Vol. 1, pp. 886–893).

[2] Ren, S., He, K., Girshick, R., et al. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems, 28*.

[3] Redmon, J., Divvala, S., Girshick, R., et al. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 779–788).

[4] Qi, C. R., Su, H., Mo, K., et al. (2017). PointNet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 652–660).

[5] Qian, H., Li, Y., Zhang, H., et al. (2025). A review of multi-sensor fusion in autonomous driving. *Sensors, 25*(19), 6033.

[6] Cui, Y., Wang, Y., Li, Y., et al. (2021). Deep learning for image and point cloud fusion in autonomous driving: A review. *IEEE Transactions on Intelligent Transportation Systems*.

[7] Feng, D., Haase-Schütz, C., Rosenbaum, L., et al. (2020). A review of deep learning-based methods for multi-modal fusion in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*.

[8] Zadeh, A., Chen, M., Poria, S., et al. (2017). Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1103–1114).

DOI: 10.54254/2755-2721/2026.35609

[9] Wu, A., Wang, W., Zong, B., et al. (2018). Multi-modal circulant fusion for video-to-language and backward. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.
[10] Prakash, A., Chitta, K., & Geiger, A. (2021). Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7077–7087).
[11] Cui, H., Radosavljevic, V., Chou, F.-C., et al. (2024). A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 1206–1226).
[12] Chen, L.-C., Papandreou, G., Schroff, F., et al. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.