Thesis Abstract

Zainab Ouardirhi

Occlusion remains a difficult problem for object detection, given that it impacts the performance of the model in real life scenarios like autonomous driving, surveillance, and robotics. Object detection models of the traditional 2D variety are particularly poor at recognizing objects when they are either partially or fully occluded because they lack depth perception and depend purely on visual features. In addition, it is true that 3D detection models incorporate spatial depth information which helps to address the problem, but their dependence on advanced domain specific computation makes them problematic as well. This thesis systematically evaluates the extent to which occlusion affects detection accuracy, comparing 2D, 3D, and multimodal fusion approaches to determine the most effective strategies for improving robustness in occlusion-heavy environments.

To overcome these challenges, FuDensityNet, a novel multimodal object detection framework is proposed in this work to address occlusion by combining 2D image data and voxelized 3D point clouds. The proposed model integrates multiple novel advancements aimed at improving detection robustness under occlusion.

A critical part of the model is the extraction of the voxel grid with the density-aware method, which aims to improve the representation of the 3D point cloud data. This approach applies voxelization in a completely different way than the traditional and outdated uniform grid subdivision. It modifies the voxel shapes according to pin density, allowing for high resolution in dense regions and reduction in low resolution regions. This method enhances feature representation in occluded scenarios and ensures better utilization of spatial depth information.

Moreover, a Voronoi diagram Neighbor Density calculation improves the assessment of occlusion considering the spatial distribution of the neighboring point clouds. Known methods for occlusion estimation make use of the quite basic depth discontinuities which are easily overcome with obstacles. The Voronoi approach implements adaptive spatial partitions, thus it becomes possible to detect occlusion severity more accurately, especially in highly urbanized areas with high object overlap.

Building on these foundations, a Multi-Scale Occlusion Rate Determination framework is introduced, integrating density-aware voxelization and Voronoi-based neighbor analysis. This approach quantifies occlusion severity across multiple scales, enabling the network to dynamically adjust feature fusion strategies based on occlusion intensity. The adaptive selection of fusion strategies ensures more reliable detection under varying occlusion levels.

The network architecture of FuDensityNet is also designed to enhance occlusion handling. A multimodal fusion strategy effectively integrates appearance-based 2D features with depth-aware 3D information, mitigating the impact of occlusion on detection accuracy. Unlike conventional fusion techniques that treat all modalities equally, FuDensityNet prioritizes features based on occlusion severity, ensuring that depth-based features receive greater attention in heavily occluded scenarios.

Beyond the architectural advancements, this thesis conducts extensive benchmarking evaluations to analyze the impact of occlusion in object detection. A systematic study investigates how occlusion affects model performance across different datasets. Results reveal that 2D detectors such as Faster R-CNN and RetinaNet suffer a 10–15% AP drop in moderate occlusions, with performance degrading up to 20% in severe occlusion cases. In contrast, 3D-based approaches such as VoxelNet demonstrate greater resilience, but still experience a 12–15% performance drop under extreme occlusions, confirming the need for enhanced fusion strategies.

The detection performance of FuDensityNet is assessed within widely used object detection datasets of KITTI and NuScenes, enabling direct comparison with other 2D and 3D detection models. The results prove that FuDensityNet surpasses single modality models and existing architectures meant for occlusion handling by 9.4% in AP over the best 3D only model and significantly outperform.

Further experiments extend this evaluation to occlusion-aware datasets, specifically OccludedPascal3D, which contains artificially generated occlusion masks to assess the robustness of detection models in extreme occlusion scenarios. Comparative analysis reveals that models trained on standard datasets struggle to adapt to occlusion-heavy environments, whereas FuDensityNet, by dynamically adjusting feature fusion based on occlusion severity, maintains a significantly higher detection rate, reducing false negatives by 12.7% compared to state-of-the-art methods.

A final comparative study systematically evaluates the trade-offs between 2D-only, 3D-only, and multimodal detection approaches. Findings suggest that while 3D detection enhances robustness, it remains computationally expensive and dependent on specialized sensors. By contrast, multimodal fusion achieves a strong balance between accuracy and efficiency, demonstrating superior occlusion handling without excessive computational overhead.

While this work significantly advances occlusion handling in object detection, future improvements focus on optimizing occlusion handling strategies further. One promising direction is the integration of learned depth estimation techniques to generate 3D point clouds from monocular images, reducing dependence on LiDAR sensors and improving scalability in real-world applications. This ongoing research aims to provide a cost-effective alternative for depth-aware detection, ensuring that occlusion handling remains feasible even in resource-constrained environments. Additionally, future work will explore end-to-end learning strategies to refine feature selection dynamically, further enhancing FuDensityNet's adaptability to varying occlusion conditions.

By introducing novel occlusion assessment techniques, multimodal fusion strategies, and extensive benchmark evaluations, this thesis establishes a robust foundation for next-generation object detection systems. The findings insist that the traditional 2D detection models need to be phased out, confirming that the inclusion of 3D spatial information either from LiDAR sensors or learned depth estimation makes a considerable difference in occlusion discrimination. FuDensityNet is an answer to such problems an efficient and adequate system to these challenges, offering a flexible detection framework for real-world occlusion-aware applications.