FuDensityNet: Fusion-Based Density-Enhanced Network for Occlusion Handling

Zainab Ouardirhi[1,2][a], Otmane Amel[1][b], Mostapha Zbakh[2][c] and Sidi Ahmed Mahmoudi[1][d]

[1] *Computer and Management Engineering Department, FPMS, University of Mons, Belgium*

[2] *Communication Networks Department, ENSIAS, Mohammed V University in Rabat, Morocco*

Keywords: Object Detection, Occlusion Handling, Voxelization, Density-Aware, Multimodal Fusion.

Abstract: Our research introduces an innovative approach for detecting occlusion levels and identifying objects with varying degrees of occlusion. We integrate 2D and 3D data through advanced network architectures, utilizing voxelized density-based occlusion assessment for improved visibility of occluded objects. By combining 2D image and 3D point cloud data through carefully designed network components, our method achieves superior detection accuracy in complex scenarios with occlusions. Experimental evaluation demonstrates adaptability across concatenation techniques, resulting in notable Average Precision (AP) improvements. Despite initial testing on a limited dataset, our method shows competitive performance, suggesting potential for further refinement and scalability. This research significantly contributes to advancements in effective occlusion handling for object detection methodologies. The abstract and conclusion highlight the substantial increase in AP achieved through our model.

1 INTRODUCTION

Accurately recognizing objects in challenging conditions is a fundamental concern in computer vision and deep learning, impacting applications like autonomous driving and surveillance systems (Pandya et al., 2023). The presence of occlusions, where objects are partially or wholly obscured, poses a significant challenge by concealing vital visual cues (Gunasekaran and Jaiman, 2023). This paper addresses the critical need for robust object recognition in the face of occlusions.

Researchers have historically tackled occlusion challenges using techniques like sliding windows and template matching, but deep learning has ushered in innovative strategies to comprehend and mitigate occlusions (Ye et al., 2023b). Despite advancements, there remains a research gap in effectively handling occlusions (Ouardirhi et al., 2022), specifically in adapting to varying degrees of occlusion. Our contribution involves developing a novel technique that adjusts the detection mechanism based on occlusion severity and rate (Nguyen et al., 2023).

The paper explores novel network designs for object detection and introduces methods for occlusion handling, including image preprocessing, voxelization, and feature fusion. A key contribution is the Voxel Density-Aware approach (VDA) for estimating occlusion rates, enhancing occlusion-aware object recognition. The proactive assessment of occlusion presence before engaging the detection network optimizes detection performance across diverse scenarios (Steyaert et al., 2023).

The proposed model's architecture and functionalities are detailed, accompanied by thorough tests and assessments on benchmark datasets. Results showcase improvements in detection accuracy, especially in challenging scenarios dominated by occlusions. Subsequent sections explore related research, trace the evolution of occlusion handling techniques, and offer a detailed exploration of our proposed strategy, including architectural insights, fusion methods, and experimental settings. The conclusion emphasizes the significance of adaptive occlusion handling, providing insights into potential future directions.

[a] https://orcid.org/0000-0001-8302-7273
[b] https://orcid.org/0009-0005-5470-362X
[c] https://orcid.org/0000-0002-1408-3850
[d] https://orcid.org/0000-0002-1530-9524

Ouardirhi, Z., Amel, O., Zbakh, M. and Mahmoudi, S.
FuDensityNet: Fusion-Based Density-Enhanced Network for Occlusion Handling.
DOI: 10.5220/0012425400003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 3: VISAPP*, pages 632-639
ISBN: 978-989-758-679-8; ISSN: 2184-4321

2 RELATED WORK

This section reviews current research on occlusion handling in object detection, exploring three main themes: 2D Object Detection in Occluded Scenes, Point Cloud-based 3D Object Detection, and Multi-Modal Fusion for Occlusion Handling.

2D Object Detection in Occluded Scenes. Researchers investigate the resilience of Deep Convolutional Neural Networks (DCNNs) in handling partial occlusions in 2D object recognition. Two-stage methods, inspired by the R-CNN series, employ an initial region proposal step followed by object refinement (Bharati and Pramanik, 2020; Zhang et al., 2021). Single-stage networks, such as YOLO series, SSD, and OverFeat, streamline the process by directly regressing bounding boxes without an explicit proposal step. However, their performance with occlusions is limited.

Occlusion-handling techniques have emerged to enhance robustness. Cutmix (Yun et al., 2019) uses a regularization method to obstruct areas in training images, improving resilience against occlusions. CompNet (Kortylewski et al., 2020) blends DCNN features with a dictionary-based approach, addressing occlusion challenges. DeepID-Net (Ouyang et al., 2015) introduces deformable pooling layers, enhancing model averaging efficacy for robust feature representation. SG-NMS in the Serial R-FCN network refines object detection through a heuristic-based approach, combining bounding boxes and suppressing overlaps (Yang et al., 2020).

Point Cloud-Based 3D Object Detection. In point cloud prediction, two prominent research tracks emphasize efficiency. One approach involves projecting point clouds onto 3D voxels, as seen in VoxNet (Maturana and Scherer, 2015), PV-RCNN++ (Shi et al., 2020) and SECOND (Yan et al., 2018). These models use strategies like cubic window attention on voxels or accelerated sparse convolutions to enhance computational efficiency. LiDARMultiNet (Ye et al., 2023a) represents a significant advancement by integrating various forms of supervision, unifying semantic segmentation, panoptic segmentation, and object recognition.

In contrast, PointNet (Zhao et al., 2017) focuses specifically on 3D point clouds, extracting permutation-invariant characteristics and contributing to the robustness of point-based 3D networks. To advance this research, PointNet++ (Qi et al., 2017) introduces a hierarchical neural network that progressively captures local characteristics at multiple contextual scales.

Multi-Modal Fusion for Occlusion Handling. The evolution of 3D sensors and their applications in environmental understanding has driven extensive research on 3D object detection. The fusion of LiDAR and camera data, explored in works such as (Li et al., 2022b) and (Wang et al., 2021), has particularly gained attention. Fusion techniques typically fall into three categories: input-level (early fusion), feature-level, or decision-level (late fusion) methods (Li et al., 2023).

Notably, a subset of studies focuses on multi-modal feature fusion during proposal generation and Region of Interest (RoI) refinement. Pioneering works like MV3D (Chen et al., 2017) and AVOD (Ku et al., 2018) employ multi-view aggregation for multi-modal detection, emphasizing the importance of integrating information from different perspectives. Other studies, as explored in (Chen et al., 2023), adopt the Transformer decoder as the RoI head, facilitating effective multi-modal feature fusion. These investigations collectively underscore the diverse methods aiming to seamlessly integrate LiDAR-camera data for enhanced object detection, particularly in addressing challenges posed by occlusions.

3 FuDensityNet

This section unveils our novel approach to addressing occlusion challenges within deep learning frameworks throughout the object detection process. Our Voxel Density Aware (VDA) method, leverages the point density derived during voxelization. This density-driven insight guides the selection of an optimal model for precise object detection. Simultaneously, we introduce our multimodal fusion network, strategically designed to synergize the strengths of both 2D and 3D features. Illustrated in Figure 1, this integration serves as a pivotal component of our proposed approach.

3.1 Occlusion Rate Assessment Using Density Analysis

In this subsection, we present our innovative approach, Occlusion Rate Assessment via Density Analysis. This technique employs density-based analysis to quantify occlusion levels, enhancing object detection accuracy.

Density-Aware Voxel Grid Extraction. Our method utilizes point cloud density data to gauge occlusion levels in a 3D scene. By measuring point density in specific regions, we discern the extent of occlusion. Higher point density values signify greater occlusion,

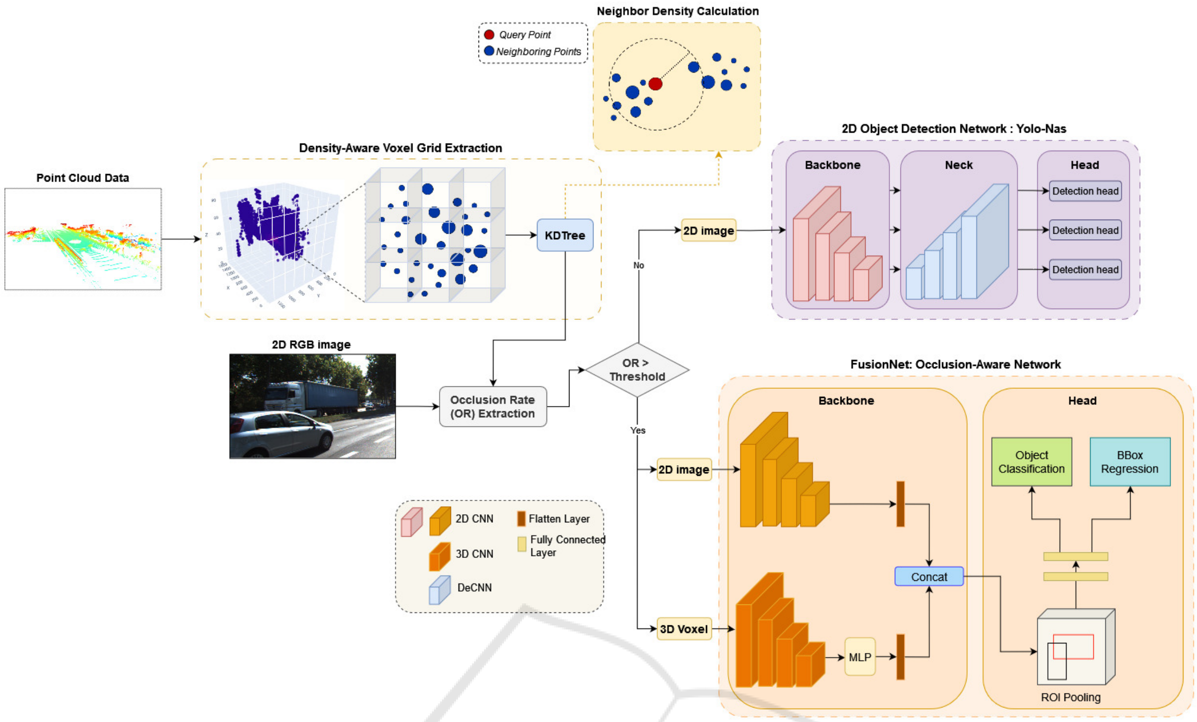


Figure 1: FuDensityNet Workflow: Point cloud voxelization extracts occluded points based on density. High-density areas trigger neighbor density computation via KDTree. The computed occlusion rate (OR) is compared to a threshold; surpassing it deploys FusionNet with voxelized point cloud and 2D image. Below the threshold, YOLO-NAS is used. FuDensityNet optimally combines FusionNet and YOLO-NAS for accurate detection across scenarios.".

while lower values indicate smaller occlusions, as depicted in Figure 2.

We represent 3D points as $P_i = (x_i, y_i, z_i)$ with i ranging from 1 to N, where N is the total point count. Utilizing a 3D grid with discrete cells or voxels, each defined by center coordinates (x_j, y_j, z_j) for j voxel indices, we assign each point P_i to the closest voxel. The voxel index (j_x, j_y, j_z) is determined by finding the closest voxel's coordinates (x_j, y_j, z_j) to P_i. This is done using the formula:

$$j_x = \left\lfloor \frac{x_i - x_{\min}}{\text{voxel\_size}} \right\rfloor, j_y = \left\lfloor \frac{y_i - y_{\min}}{\text{voxel\_size}} \right\rfloor$$

$$and \quad j_z = \left\lfloor \frac{z_i - z_{\min}}{\text{voxel\_size}} \right\rfloor \tag{1}$$

where the voxel grid's minimum coordinates are denoted as $(x_{\min}, y_{\min}, z_{\min})$, and `voxel_size` represents the dimension of each voxel. Density computation for each voxel (x_j, y_j, z_j) involves counting points within its boundaries. The density D_j for voxel is calculated as:

$$D_j = \sum_{i=1}^{N} \chi_j(P_i) \tag{2}$$

where $\chi_j(P_i)$ is an indicator function that returns 1 if point P_i falls within voxel (x_j, y_j, z_j), and 0 otherwise.

Neighbor Density Calculation. The spatial distribution revealed by initial density analysis prompts an examination of surrounding areas through neighbor density computation. This exploration distinguishes between dense patches with gaps, indicating potential occlusions, and continuous, concentrated areas suggestive of other occlusion scenarios (Figure 1). Leveraging a KDTree (Bentley, 1975) structure for efficiency, our technique involves constructing a KDTree from the entire point cloud, facilitating swift nearest neighbor searches. For voxels exceeding the density threshold (D_{voxel}), a KDTree query identifies points within a radius (r) around the voxel, determining $D_{neighbors}$. The neighbor density (ND_{voxel}), calculated as the ratio of $D_{neighbors}$ to the sphere's volume with radius r, is expressed in Equation 3:

$$ND_{voxel} = \frac{D_{neighbors}}{\frac{4}{3}\pi r^3} \tag{3}$$

Comparing ND_{voxel} with D_{voxel} yields insights into the spatial distribution around the voxel. Substantially lower ND_{voxel} signals dispersed high-density regions, implying potential occlusions with gaps. Conversely, close ND_{voxel} and D_{voxel} values

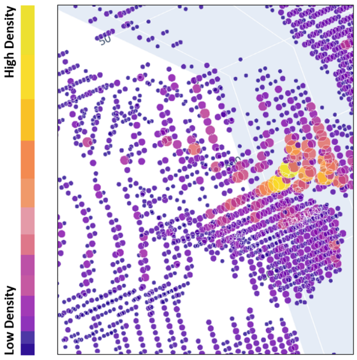


Figure 2: Visualizing Occlusion Intensity in 3D Scenes with Density-Aware Voxel Grid : The voxelized point cloud data, showcasing varying point densities in size and color to represent different occlusion intensities.

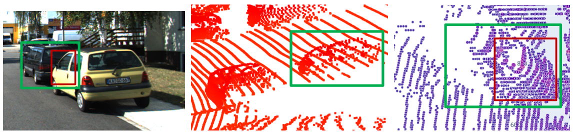
suggest a contiguous, dense region indicative of occluded objects in proximity. This assessment aids in distinguishing occlusion scenarios, facilitating informed decisions on occlusion handling (Figure 3).

Occlusion Rate Determination and Model Selection. The density-based metric guides the decision to employ our occlusion handling network or established detection models. Extracting the occlusion rate (OR) involves comparing point density in specific regions, using the metric, to a predefined threshold based on benchmarks like KITTI (Geiger et al., 2012). If OR exceeds the threshold, our occlusion handling network is used; otherwise, a state-of-the-art detection model ensures optimal object detection.

3.2 Network Architecture for Occlusion Handling

Our network seamlessly integrates 2D image and 3D point cloud data, forming a holistic solution for robust object detection, even in challenging occluded scenarios.

Backbone Networks. Our choice of backbone networks, detailed in Section 4.2.1, is based on extensive experimentation. For the 2D image backbone, we strategically selected a fine-tuned ResNet-50 (He et al., 2016) for its exceptional performance in handling occlusions during feature extraction. Although YOLO outperformed it in the overall model comparison, we chose ResNet-50 for feasibility reasons. Simultaneously, our decision for the point cloud backbone, VoxNet (Maturana and Scherer, 2015), was driven by its ability to seamlessly handle occlusion complexities by transforming voxelized point cloud data into hierarchical features. VoxNet exhibited robust performance in capturing spatial information, crucial for accurate object identification in partially occluded scenarios.



(a) 2D RGB image. (b) Point Cloud Data. (c) Voxelized Data.

Figure 3: Visualizing Data Pre- and Post-Voxel Density Aware Approach: (a) Input image with partial occlusion, (b) corresponding point cloud data reveals gaps between occluded objects, (c) post-voxelization with density indicates increased point density along the line of sight. Heightened density in the red box signifies potential occlusion, with larger points in a distinct color denoting the occluded part.

Multimodal Fusion Method. To effectively address occlusion through multimodal fusion, we employ the Low-rank Multimodal Tensor Fusion (LMF) method. LMF enhances traditional tensor fusion (Zadeh et al., 2017) by minimizing computational costs without compromising performance through the use of low-factor weights. Our experiments (Section 4.2.3) highlight LMF as the optimal choice, showcasing superior occlusion handling efficacy compared to other fusion methods. Its ability to preserve critical information while minimizing computational overhead is a key component in our approach to robust object detection in occluded scenarios.

Detection Head. After integrating the backbone networks, the key step involves applying the Faster R-CNN (F-RCNN) detection head. This begins with a RoI pooling layer, mathematically represented as:

$$\text{ROI Pooling}(x,R) = \frac{1}{|R|}\sum_{i\in R} x_i, \tag{4}$$

ensuring effective feature alignment for spatial information extraction. The pooled features then pass through two fully connected layers (FC_1 and FC_2), expressed as:

$$FC_1(x) = \sigma(W_1 x + b_1), \tag{5}$$

$$FC_2(x) = \sigma(W_2 FC_1(x) + b_2), \tag{6}$$

facilitating the learning of intricate data relationships. These learned features contribute to predicting class probabilities (P_{class}) and bounding box regressions (P_{2DBox}), given by:

$$P_{\text{class}} = \text{softmax}(FC_2(x)), \tag{7}$$

$$P_{\text{2DBox}} = FC_{\text{2DBox}}(x), \tag{8}$$

By seamlessly integrating backbone networks with the F-RCNN head, our approach provides a robust solution tailored for real-world scenarios with occlusion challenges.

4 MAIN RESULTS

In this section, we present key results. We outline our experimental setup, compare 2D networks for the optimal backbone, evaluate 3D backbones for FusionNet selection, and conduct comprehensive comparisons between FusionNet, occlusion handling techniques, and our approach, FuDensityNet. FuDensityNet strategically combines FusionNet and a specialized object detection approach based on occlusion levels for enhanced accuracy in diverse scenarios.

4.1 Experimental Setup

Resources. We utilized Google's Tensor Processing Unit (TPU) v2, a specialized hardware accelerator with 35GB RAM, for high-performance machine learning and deep learning tasks. **Datasets.** Our analysis involves FuDensityNet evaluation at three levels. First, on KITTI2D (Geiger et al., 2012) and CityPersons (Zhang et al., 2017) for benchmarking the 2D backbone. Second, on KITTI3D (Geiger et al., 2012) and OccludedPascal3D (Xiang et al., 2014) to compare 3D network performances for assessing the 3D backbone. We conduct a comparison analysis of FuDensityNet versus alternative occlusion handling methods and our occlusion-aware network, clarifying the beneficial impact of our VDA strategy on detection accuracy.

4.2 Results and Analysis

4.2.1 Comparative Analysis on the 2D Backbones

In this section, we conduct a comparative analysis of 2D backbone networks, optimizing for inference speed and accuracy in object detection across various scenarios based on the degree of occlusion. Our models were initially trained on the KITTI2D dataset, containing 7481 training images. The results in Table 1 present AP calculated on testing data from two datasets: 7481 testing images from KITTI2D and 5000 testing images from the CityPersons dataset.

1. **YOLO-NAS for Low to No Occlusion.** In low to no occlusion scenarios, our analysis (Table 1) designates YOLO-NAS as the optimal model for efficient and accurate object detection. Recognized for its lightweight architecture, YOLO-NAS provides rapid and precise detection, making it a top-performing choice in situations with minimal occlusion. This strategic selection optimizes both speed and accuracy without the need for complex feature extraction.
2. **Custom Backbone for Moderate to High Occlusion.** Our analysis (Table1) consistently highlights ResNet50 as a superior 2D backbone for object detection. Renowned for its advanced feature extraction, ResNet50 significantly contributes to addressing occlusion challenges. Augmenting this with our chosen 3D backbone creates a customized model, enhancing resilience against occluded objects.

4.2.2 Comparative Analysis on the 3D Backbones

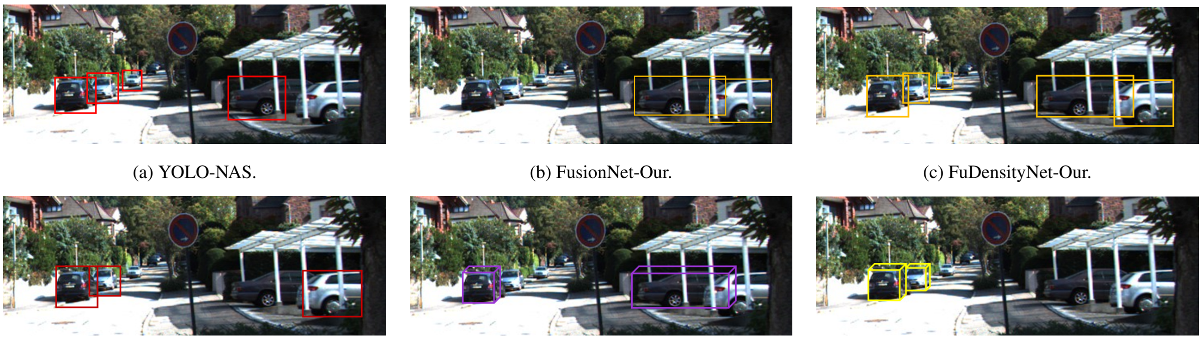
Our analysis (Table 2) demonstrates VoxNet's superior performance across KITTI3D and OccludedPascal3D datasets. Despite training on the KITTI3D dataset with 7481 point clouds, the results in Table 2 showcase AP on testing data from KITTI3D (7481 point clouds) and OccludedPascal3D (2073 point clouds). VoxNet, with a limited dataset of 100 point clouds, achieves a notable 40% AP, underscoring its potential for robust 3D object detection. This motivates our choice to leverage VoxNet's strengths for enhanced object detection in occluded scenarios.

4.2.3 Comparative Analysis on Fusion Techniques

In our multimodal fusion exploration for occlusion handling, we investigated four late fusion methods:

Table 1: Object Detection AP Results for KITTI 2D and CityPersons Datasets.

Model	AP(%) (KITTI 2D)			(CityPersons)
	Car	Pedestrian	Cyclist	Person
F-RCNN (Sharma et al., 2023)	71.2	67.4	66.7	85.7
ResNet50-F-RCNN (He et al., 2016)	**76.8**	**69.4**	**67.8**	**87.3**
MobileNetv2-F-RCNN (Sandler et al., 2018)	57.2	53.8	48.5	79.3
vgg16-F-RCNN (Simonyan and Zisserman, 2014)	59.2	58.4	47.6	80.9
SSD (Liu et al., 2016)	66.7	64.4	58.1	84.1
RetinaNet (Lin et al., 2017)	65.6	63.3	58.4	82.5
YOLOv5s (Sozzi et al., 2022)	89.9	87.7	83.8	88.9
YOLOv6s (Li et al., 2022a)	92.2	88.1	85.7	92.1
YOLOv7 (Wang et al., 2023)	90.2	86.5	84.1	90.5
YOLOv8s (Huang et al., 2023)	93.7	91.3	87.2	93.7
YOLO-NAS (Sharma et al., 2023)	**95.4**	**92.8**	**91.1**	**95.0**



(a) YOLO-NAS.

(b) FusionNet-Our.

(c) FuDensityNet-Our.

(d) CompNet.

(e) MV3D.

(f) YOLO3D.

Figure 4: Visual Examples for Comparing Occlusion Handling in Object Detection Models Using KITTI Dataset.

Table 2: Object Detection AP Results on OccludedPascal3D and KITTI3D Datasets for Different Models.

Model	**AP (%) (OccludedPascal3D)**									**AP(%) (KITTI3D)**		
	aeroplane	***bicycle***	***boat***	***bottle***	***bus***	***car***	***motorbike***	***train***	***tvmonitor***	***Car***	***Pedestrian***	***Cyclist***
SECFPN (Radosavovic et al., 2020)	28.8	27.3	27.7	27.0	28.3	28.8	28.5	27.0	26.0	36.8	34.2	31.5
PointNet++ (Qi et al., 2017)	27.6	26.1	26.5	25.9	27.1	27.6	27.5	26.0	25.0	34.7	32.1	29.4
SSN (Zhu et al., 2020)	26.4	24.9	25.3	24.8	26.0	26.4	26.6	25.0	24.0	32.6	30.0	27.3
ResNeXt-152-3D (He et al., 2016)	23.9	22.5	23.0	22.6	23.6	23.9	24.6	23.0	22.0	22.4	20.2	18.1
VoxNet (Maturana and Scherer, 2015)	**30.0**	**28.5**	**29.7**	**28.1**	**29.5**	**30.0**	**29.6**	**28.0**	**27.0**	**40.0**	**38.5**	**35.7**

Table 3: Comparison analysis of fusion methods.

Fusion Method	**AP(%) (KITTI 3D)**		
	Car	**Pedestrian**	**Cyclist**
Concatenation	40.5	39.3	29.6
Arithmetic Fusion (Addition)	38.5	36.4	27.4
Arithmetic Fusion (multconcat)	42.3	38.8	29.4
Sub-space Concat	41.3	38.2	29.1
Low Rank Tensor Fusion	**43.5**	**39.9**	**31.4**

Concatenation (Ramachandram and Taylor, 2017), Arithmetic Fusion (Addition (RODRIGUES et al.,), Multconcat (Amel and Stassin, 2023)), Sub-space Concat (Ramachandram and Taylor, 2017), and Low Rank Tensor Fusion (LMF) (Zadeh et al., 2017). LMF, detailed in Section 3.2, minimizes computational cost while exhibiting notable scalability. Table 3 illustrates LMF's (Zadeh et al., 2017) superiority in generating multimodal representations that effectively capture cross-modality interactions to address occlusion challenges. Results, though with a limited dataset, show promise, prompting the need for further comprehensive assessment.

4.2.4 Comparative Analysis of Global Network

The evaluation of FuDensityNet against YOLO3D and MV3D reveals its superior performance across various occlusion levels (Table 4). YOLO-NAS excels in low occlusion scenarios, achieving high AP values (52.3% in Easy, 51.2% in Moderate, and 49.1% in Hard for Car detection).

FuDensityNet strategically combines YOLO-NAS's strengths in low occlusion with FusionNet's capabilities in moderate to high occlusion, yielding competitive AP values. For Car detection, FuDensityNet outperforms YOLO-NAS in the Easy category (52.3%) and maintains strong performance in Moderate (39.9%) and Hard (36.6%) occlusion scenarios. Similar trends are observed for Pedestrian and Cyclist detection.

FuDensityNet's proficiency in leveraging multimodal data contributes to its competitive performance, particularly in conjunction with YOLO-NAS (Figure 4). These results validate our approach and position FuDensityNet as a solution for occlusion-aware object detection in real-world scenarios, warranting further investigations for comprehensive understanding and broader applications.

5 CONCLUSION

In this study, we introduced an innovative occlusion handling approach that integrates 2D images and 3D point cloud data, utilizing advanced preprocessing and novel network architectures to enhance accuracy in detecting obscured objects. A key aspect involves preprocessing input data, incorporating density-based occlusion assessment through voxelization, providing insights into scene complexity. The pivotal in-

Table 4: Object Detection AP Results on KITTI Dataset for Occlusion Analysis.

Network	Data	**Car**			**Pedestrian**			**Cyclist**		
		Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
YOLO-NAS (Sharma et al., 2023)	2D	**52.3**	**51.2**	**49.1**	**50.2**	**48.3**	**45.8**	**49.3**	**47.4**	**43.5**
YOLO3D (Ali et al., 2018)	LIDAR	32.3	29.5	24.2	31.7	27.2	23.5	28.5	21.6	17.0
CompNet (Kortylewski et al., 2020)	2D	40.2	32.6	31.2	36.5	32.3	29.8	28.3	24.1	21.2
MV3D (Chen et al., 2017)	2D+LIDAR	41.3	40.2	38.1	35.2	31.3	29.8	37.3	31.4	30.5
FusionNet-Our	2D+LIDAR	40.5	39.9	36.6	39.3	33.8	30.1	29.6	29.5	27.8
FuDensityNet-Our	2D+LIDAR	**52.3**	39.9	36.6	**50.2**	33.8	30.1	**49.3**	29.5	27.8

novation lies in fusing 2D and 3D data using well-designed network architectures, achieving superior detection accuracy, even in challenging occluded scenarios. Acknowledging limitations is crucial, as initial testing used a limited dataset, necessitating further experimentation with more extensive datasets for generalizability. Our occlusion handling approach demonstrates a significant advancement in object detection, evidenced by quantifiable improvements in AP across different occlusion levels, yielding substantial increases compared to existing methods. In conclusion, our study establishes a robust occlusion handling approach, a noteworthy advancement in object detection technology, anticipating broader applicability and potential contributions to advancing object detection technology. Future work will address identified limitations, including dataset expansion and continued refinement of proposed methodologies.

REFERENCES

Ali, W., Abdelkarim, S., Zidan, M., Zahran, M., and El Sallab, A. (2018). Yolo3d: End-to-end real-time 3d oriented object bounding box detection from lidar point cloud. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0.

Amel, O. and Stassin, S. (2023). Multimodal approach for harmonized system code prediction. In *Proceedings of the 31st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 181–186.

Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517.

Bharati, P. and Pramanik, A. (2020). Deep learning techniques—r-cnn to mask r-cnn: a survey. *Computational Intelligence in Pattern Recognition: Proceedings of CIPR 2019*, pages 657–668.

Chen, X., Ma, H., Wan, J., Li, B., and Xia, T. (2017). Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1907–1915.

Chen, X., Zhang, T., Wang, Y., Wang, Y., and Zhao, H. (2023). Futr3d: A unified sensor fusion framework for 3d detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 172–181.

Geiger, A., Lenz, P., and Urtasun, R. (2012). Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE.

Gunasekaran, K. P. and Jaiman, N. (2023). Now you see me: Robust approach to partial occlusions. *arXiv preprint arXiv:2304.11779*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Huang, Z., Li, L., Krizek, G. C., and Sun, L. (2023). Research on traffic sign detection based on improved yolov8. *Journal of Computer and Communications*, 11(7):226–232.

Kortylewski, A., Liu, Q., Wang, H., Zhang, Z., and Yuille, A. (2020). Combining compositional models and deep networks for robust object classification under occlusion. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1333–1341.

Ku, J., Mozifian, M., Lee, J., Harakeh, A., and Waslander, S. L. (2018). Joint 3d proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8. IEEE.

Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., Ke, Z., Li, Q., Cheng, M., Nie, W., et al. (2022a). Yolov6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*.

Li, Y., Qi, C. R., Zhou, Y., Liu, C., and Anguelov, D. (2023). Modar: Using motion forecasting for 3d object detection in point cloud sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9329–9339.

Li, Y., Yu, A. W., Meng, T., Caine, B., Ngiam, J., Peng, D., Shen, J., Lu, Y., Zhou, D., Le, Q. V., et al. (2022b). Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17182–17191.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer.

Maturana, D. and Scherer, S. (2015). Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ international conference on* *intelligent robots and systems (IROS)*, pages 922–928. IEEE.

Nguyen, A. D., Pham, H. H., Trung, H. T., Nguyen, Q. V. H., Truong, T. N., and Nguyen, P. L. (2023). High accurate and explainable multi-pill detection framework with graph neural network-assisted multimodal data fusion. *Plos one*, 18(9):e0291865.

Ouardirhi, Z., Mahmoudi, S. A., Zbakh, M., El Ghmary, M., Benjelloun, M., Abdelali, H. A., and Derrouz, H. (2022). An efficient real-time moroccan automatic license plate recognition system based on the yolo object detector. In *International Conference On Big Data and Internet of Things*, pages 290–302. Springer.

Ouyang, W., Wang, X., Zeng, X., Qiu, S., Luo, P., Tian, Y., Li, H., Yang, S., Wang, Z., Loy, C.-C., et al. (2015). Deepid-net: Deformable deep convolutional neural networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2403–2412.

Pandya, S., Srivastava, G., Jhaveri, R., Babu, M. R., Bhattacharya, S., Maddikunta, P. K. R., Mastorakis, S., Piran, M. J., and Gadekallu, T. R. (2023). Federated learning for smart cities: A comprehensive survey. *Sustainable Energy Technologies and Assessments*, 55:102987.

Qi, C. R., Yi, L., Su, H., and Guibas, L. J. (2017). Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30.

Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K., and Dollár, P. (2020). Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436.

Ramachandram, D. and Taylor, G. W. (2017). Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6):96–108.

RODRIGUES, L. S., Sakiyama, K., Takashi Matsubara, E., Marcato Junior, J., and Gonçalves, W. N. Multimodal fusion based on arithmetic operations and attention mechanisms. *Available at SSRN 4292754*.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520.

Sharma, P., Gupta, S., Vyas, S., and Shabaz, M. (2023). Retracted: Object detection and recognition using deep learning-based techniques. *IET Communications*, 17(13):1589–1599.

Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., and Li, H. (2020). Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10529–10538.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sozzi, M., Cantalamessa, S., Cogato, A., Kayad, A., and Marinello, F. (2022). Automatic bunch detection in white grape varieties using yolov3, yolov4, and yolov5 deep learning algorithms. *Agronomy*, 12(2):319.

Steyaert, S., Pizurica, M., Nagaraj, D., Khandelwal, P., Hernandez-Boussard, T., Gentles, A. J., and Gevaert, O. (2023). Multimodal data fusion for cancer biomarker discovery with deep learning. *Nature Machine Intelligence*, 5(4):351–362.

Wang, C., Ma, C., Zhu, M., and Yang, X. (2021). Pointaugmenting: Cross-modal augmentation for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11794–11803.

Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. (2023). Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7464–7475.

Xiang, Y., Mottaghi, R., and Savarese, S. (2014). Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE winter conference on applications of computer vision*, pages 75–82. IEEE.

Yan, Y., Mao, Y., and Li, B. (2018). Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337.

Yang, C., Ablavsky, V., Wang, K., Feng, Q., and Betke, M. (2020). Learning to separate: Detecting heavily-occluded objects in urban scenes. In *European Conference on Computer Vision*, pages 530–546. Springer.

Ye, D., Zhou, Z., Chen, W., Xie, Y., Wang, Y., Wang, P., and Foroosh, H. (2023a). Lidarmultinet: Towards a unified multi-task network for lidar perception. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3231–3240.

Ye, H., Zhao, J., Pan, Y., Cherr, W., He, L., and Zhang, H. (2023b). Robot person following under partial occlusion. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7591–7597. IEEE.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032.

Zadeh, A., Chen, M., Poria, S., Cambria, E., and Morency, L.-P. (2017). Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Zhang, J., Wang, J., Xu, D., and Li, Y. (2021). Hcnet: a point cloud object detection network based on height and channel attention. *Remote Sensing*, 13(24):5071.

Zhang, S., Benenson, R., and Schiele, B. (2017). Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3221.

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890.

Zhu, X., Ma, Y., Wang, T., Xu, Y., Shi, J., and Lin, D. (2020). Ssn: Shape signature networks for multi-class object detection from point clouds. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 581–597. Springer.