A novel approach for recognizing occluded objects using Feature Pyramid network based on occlusion rate analysis

1st Zainab Ouardirhi
FPMS, University of Mons
Computer and Management Engineering Department
ENSIAS, Mohammed V University in Rabat
Communication Networks Department
Rabat, Morocco
Zainab.OUARDIRHI@umons.ac.be

2nd Sidi Ahmed Mahmoudi
FPMS, University of Mons
Computer and Management Engineering Department
Mons, Belgium
Sidi.MAHMOUDI@umons.ac.be

3rd Mostapha Zbakh
ENSIAS, Mohammed V University in Rabat
Communication Networks Department
Rabat, Morocco
mostapha.zbakh@ensias.um5.ac.ma

***Abstract*—In response to the rising adoption of smart video surveillance systems (SVS), this paper addresses a common challenge: occlusion in object detection. Existing object recognition methods often overlook the relative occlusion of neighboring objects, leading to real-world SVS systems encountering significant issues. Recent occlusion-handling approaches have limitations, such as data type inflexibility and difficulty distinguishing objects. In this study, we propose an end-to-end solution. Our approach utilizes the Feature Pyramid Network (FPN) for small and overlapping object detection, replaces Grey Level Co-Occurrence Matrices (GLCM) with point cloud density analysis for accurate occlusion rate determination, and integrates depth information with RGB data to improve occluded object separation. This holistic method aims to enhance object detection performance, particularly in challenging occlusion scenarios.**

***Index Terms*—Occlusion rate, Object recognition, CNN, Voxel Density, Feature pyramid, Feature fusion**

Fig. 1. High partial occlusion levels in complex object detection scenario in fields and crowded environment.

I. Introduction

A wide range of practical applications, including robotics, smart video surveillance, and automated driving, heavily rely on object recognition and action/event detection [1], [2]. Additionally, it is a need for tasks like person search, pose estimation, and multi-target tracking [3], [4] , [5].

These techniques, in particular, are essential building blocks for smart video surveillance systems, which can be developed by using image processing algorithms that perform various transformations (pre-processing, convolutional morphological operations, etc.) to extract important features and objects and by using an automatic (Machine Learning) or deep learning (DL) process to train models that can replicate human knowledge using a set of data.

Recent developments in the fields of high-speed computation and Big Data have made it feasible to train models using massive datasets, which has improved the performance of object detection systems. However, one of these learning strategies' greatest current problems is the occlusion of objects. As shown in figure I, the occlusion impacts negatively on the clarity of targeted objects and the inefficiency of the SVS which can lead to detection confusion, low localization performance, incomplete obstacle detection and an increase of accident risks in the fields.

Particularly, partial occlusion represents a formidable hurdle for computer vision systems. In real-world environments, objects frequently find themselves surrounded and concealed by other elements, generating a data distribution characterized by intricate overlaps in shape, appearance, and position. This inherent complexity makes it arduous to encapsulate within fixed training data.

Deep learning algorithms, while exhibiting remarkable ca-

pabilities in various computer vision tasks, still encounter challenges in recognizing partially occluded objects that humans can effortlessly discern. This limitation persists even when deep networks are trained with high levels of partial occlusion. It underscores a fundamental inadequacy in contemporary computer vision systems, one that demands innovative solutions.

In this paper, we present an innovative approach to address the enduring issue of occlusion in computer vision. Our methodology leverages a Fusion Pyramid Network (FPN) architecture that seamlessly integrates information from two distinct sources. The first source processes 2D images, while the second taps into a 3D network utilizing point cloud data to extract depth information from objects. This fusion of modalities allows us to tackle occlusion challenges more effectively, enhancing object recognition even under challenging conditions.

The remainder of this paper is organized into four sections: the following section reviews related works, shedding light on existing approaches and their limitations. Subsequently, we present our experimental methodology, including the dataset and evaluation metrics, before conducting a thorough comparison with state-of-the-art techniques on the KITTI [6] dataset. The penultimate section introduces our novel recommendation framework and its constituent modules, detailing the architecture and rationale behind our FPN-based solution. Finally, the concluding section encapsulates our findings, underscores the significance of our approach in advancing the field of computer vision, particularly in addressing occlusion challenges, and discusses potential future research directions.

II. Related works

One of the key challenges in object detection continues to be deformation and occlusion. Most earlier research focused on the benefits and drawbacks of object detection algorithms based either on extracting the background objects of the occluded objects or exploiting the depth of targeted objects [7]. When less than 10% of the object is obscured by the occlusion, traditional learning algorithms often work quite well [8].

Deformation and occlusion remain two of the main obstacles to object detection. The majority of early studies concentrated on the advantages and disadvantages of object recognition algorithms that either extracted the background objects of the occluded objects or took advantage of the depth of the targeted ones [7]. Traditional learning methods frequently perform well when less than 20% of the object is hidden by the occlusion [8]. Unfortunately, as the occlusion percentage increases, the detection failure rate also increases. In fact, as the degree of opaqueness approaches 50%, object recognition in fact gets fairly challenging [9]. In contrast, generative models can clearly discriminate between the depiction of background context and the targeted objects, which significantly helps with the occlusion problem.

Timur et al. [10] proposed a model based on image decoding into a collection of people detection in a crowded environment utilizing POM [11], a multi-camera generative detection technique, as the base model to manage occlusions but built to synthesis depth maps instead of binary images which they refer to as DPOM. POM explicitly manages complicated occlusion interactions between detected individuals and employs a sophisticated technique based on a generative model to estimate the probability of occupancy. DPOM modifies the original input images in order to extract the depth of the objects, as a result it loses a lot of information about these objects, making it impossible to distinguish between them. This constraint has a negative impact on the recognition phase, leading to unrecognizably formed objects.

Sun et al. used in [12] CompNet [13] a Bayesian generative model with neural network features to replace the fully-connected classifier in a CNN. This model applies the probability distribution to describe the image's features, such as the object classes and amodal segmentation to accurately classify images of partly occluded objects. Although this method improved the model's resistance to occlusion, it required significant form priors and is thus only suitable for rigid objects such as vehicles.

Chen et al. [7] proposed an end-to-end multi-view 3D CNN-based multi-view 3D object detection network (MV3D), that uses Laser Imaging Detection and Ranging (LIDAR) data to tackle the occlusion problem by getting precise depth information about targeted objects. It combines both 2D and 3D information using a regional fusion network which helps the network to perform better during hard occlusion. Although MV3D performed well with occlusion, it is heavy and it leads to longer training prosses, which is not great to be used in real time scenarios.

Ali et al. [14] proposed a solution, based on the extension of the YOLOv4 [15] generic object detector called YOLO3D, that follows the "LIDAR-only" paradigm where it only uses the projected LIDAR point cloud as a special bird's eye view grid to retain 3D information. YOLO3D extends the one-shot regression meta-architecture, which has been successful in the 2D perspective image space, to produce oriented 3D object bounding boxes. However, the accuracy of the YOLO3D model's outcome of detection is still incomparable to that of MV3D.

Yang et al. [16] proposed the Semantics-Geometry Non-Maximum-Suppression (SG-NMS) algorithm, a heuristic-based approach that combines bounding-boxes in accordance with detection scores obtained from a CNN-based network called Serial R-FCN [17], suppressing overlapping bounding-boxes with lower detection scores.

Takahashi et al. [18] proposed the expandable YOLO (E-YOLO) technique, which is an improved version of YOLOv3 based on edge detection and frame differences. In order to achieve high quality 3D object recognition, the model adheres to the "Camera-only" paradigm, which works well during occlusion, by using prior knowledge about the size of 2D objects and attempting to predict the 3D bounding box using a stereo camera. As result, the approach has a high-speed detection characteristic that has a promising, wide-ranging commercial

viewpoint. The results, however, also shown that this model, when compared to DCNNs, is much less discriminating in detecting non-occluded objects.

Jenkins et al. [19] presented a novel, regression-based framework for counting densely spaced objects in 3D, called CountNet3D, showing that regression-based 3D counting methods outperform state-of-the-art 3D object detectors.

Other recent occlusion handling approaches of objects in closed environments were also proposed. Reynolds et al. [20] published a novel salient object recognition dataset based on pictures taken in an actual use case where photographers with vision impairments submitted images in order to get help understanding the visual content.

Omeed et al. [21] developed an agricultural imaging system using a high resolution stereo camera pair and active illumination source that is resistant to daytime lighting variability, and estimated the amount of error caused by sizing apples in images using 2D fruit shape versus 1D fruit diameter in a controlled experiment to determine the effectiveness of 2D sizing.

Summary Occlusion is a substantial barrier to object detection. The effectiveness of the different occlusion handling methods varies depending on the strategy and the kind of data, but it is undeniable that when occlusion is present, the detection performance is still far from perfect. There are a number of ways to deal with occlusion, as we discussed in the previous subsection (II), including extracting the depth using LIDAR data or stereo cameras, generative models that can easily distinguish between the depiction of background context and the targeted objects, and deep learning techniques that can either rely on data augmentation strategies that can enrich the diversity of datasets, to enhance the robustness and generalization of the frameworks.

Even though numerous solutions to the problem of occlusion have been put forth, there are still many obstacles that must be overcome, creating a significant gap between detectors and humans. These obstacles include detectors' rigidity with regard to any type of data (dimension, environment, object sizes, etc.), their inability to distinguish between various targeted objects, and—most importantly—their inability to analyze occlusion rates on images. We will provide our framework for handling occlusion in the section that follows. This framework will assist in addressing latency issues and enhancing model performance when faced with occlusion.

III. Proposed approach

Our framework (FuDensityNet) is designed to enhance the efficiency of image analysis in various scenarios, particularly when occlusions are present. In essence, we introduce a versatile Framework (depicted in figure III) capable of processing input images, determining their data type (2D or 3D), performing essential data preprocessing, and assessing the occlusion rate. This occlusion rate evaluation serves as a pivotal decision point for selecting the most suitable model within our system.

If the calculated occlusion rate falls below a predefined Threshold value, the Framework leverages the state-of-the-art object detection model available today for optimal results. However, when the occlusion rate surpasses this threshold, the Framework seamlessly switches to our proposed occlusion handling approach, which will be elaborated upon in the subsequent subsection. This strategic adaptation ensures that our system consistently delivers superior outcomes, even in challenging scenarios characterized by occluded objects.

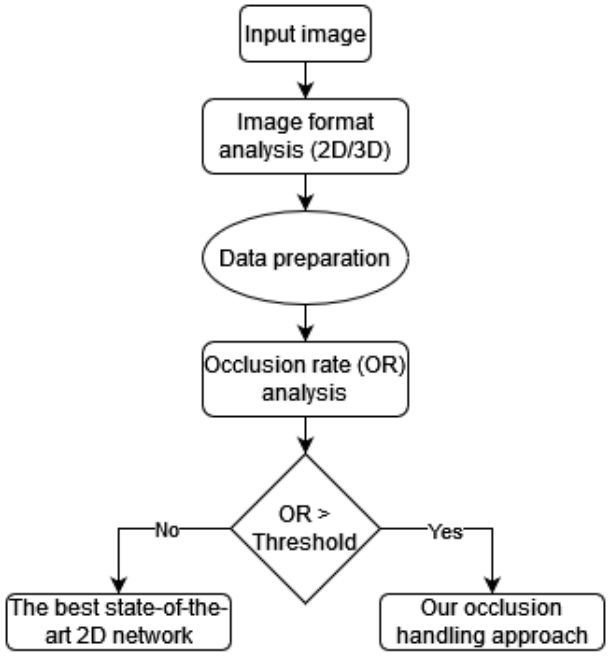

Fig. 2. FuDensityNet Diagram: Our proposed occlusion handling framework.

A. Data representation

Model performance may be significantly influenced by factors other than training models, such as data quality, quantity, and dimension, such as 2D or 3D images. This is particularly relevant during occlusion. We recognized, in particular, that before building the model, consideration must be given to the dimension of the neural network's input data. In fact, we found that using 3D data really aids in the occlusion problem solving. Since objects that are closer to the camera have a lower depth than those that are farther away, this may be used to separate blobs when two or more objects overlap and form a clustered region.

This will ultimately aid in object separation during occlusion. On the other hand, the network is unable to distinguish between various objects using just their depth; 2D information is also crucial since it enables the extraction of feature maps that describe the intended objects (cite: ning2021survey). This gave us the idea of combining the two pieces of data; instead of having a single input, we will have two: the RGB image and the aerial LIDAR point cloud data, commonly referred to as the "LIDAR-camera fusion" paradigm.

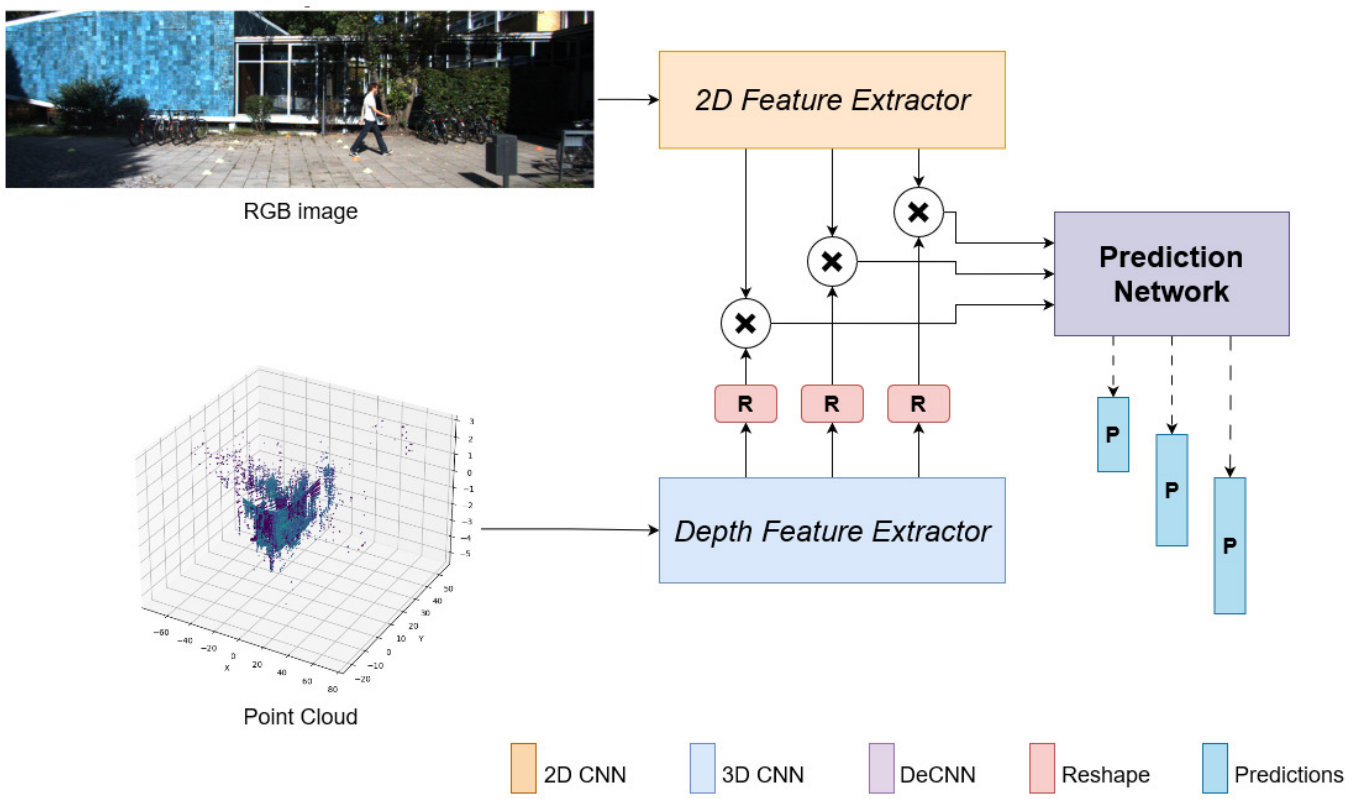

Fig. 3. Architecture of our proposed FPN based occlusion handling DCNN approach.

B. Occlusion rate analysis

To enhance the performance of our system, it is crucial to analyze the occlusion rate before determining the object detection model to be used. In this additional image processing stage, we extract the characteristics of the targeted objects and mathematically distinguish between them using their pixels.

Previously, we employed GLCM [22] techniques to extract texture features from the images, which characterized the texture of an image by calculating how often pairs of pixels with specific values and in a specified spatial relationship occurred in an image. This method allowed us to create a GLCM and then extract statistical measures from this matrix. While this technique proved valuable in handling certain occlusion scenarios, our extensive testing and experimentation revealed its limitations.

Through our rigorous analysis, we determined that GLCM did not provide the depth of insight into occlusion patterns required for robust object detection in a broader range of real-world scenarios. Its effectiveness was constrained by its focus on texture features, which sometimes inadequately captured the nuances of occlusion, particularly in 3D scenes.

As a result, we have transitioned to a more data-driven approach that leverages point cloud density data to evaluate occlusion levels within specific 3D scenes. By measuring the density of points in discrete regions, this method allows us to quantitatively assess occlusion rates. Higher point density values within regions indicate more substantial occlusions, while lower point density values highlight smaller occlusions. This approach offers a comprehensive means of quantifying occlusion, enhancing our adaptability in selecting the appropriate object detection model.

Following the extraction of the occlusion rate through density-based voxelization of point cloud data, we compare it to a predefined criterion established to govern object detection model selection. Through extensive experimentation using the KITTI dataset [6], which offers a comprehensive collection of both 2D and 3D images, we have determined that the threshold for occlusion rates that do not significantly impede detection performance lies at 20%. State-of-the-art object recognition approaches, which do not explicitly address the occlusion problem, exhibit superior performance when the occlusion rate falls below this threshold.

By employing point cloud density data and adopting a threshold-based approach, we have forged a robust methodology for occlusion rate analysis that complements our overall system. This approach ensures that our system dynamically adapts its object detection strategy, optimizing performance in response to varying occlusion levels within the environment.

C. Proposed occlusion handling approach

Taking inspiration from the Feature Pyramid Networks (FPNs) [23] framework initially designed for 2D object de-

tection, we have devised a novel deep convolutional neural network (DCNN) feature extractor tailored for the unique challenges posed by LIDAR point clouds and RGB images. Our innovation lies in the ability to generate high-resolution feature maps from these disparate data sources, a capability that empowers us to precisely localize even small object classes within the scene.

As depicted in Figure 3, our network adopts the 'LIDAR-camera fusion' paradigm, seamlessly integrating information from both modalities. The network takes a 2D image as input, routing it through a dedicated 2D neural network backbone. This backbone efficiently extracts 2D feature maps that describe the salient characteristics of the targeted objects within the image.

Simultaneously, the LIDAR data from the same image is channeled into a 3D neural network backbone, which specializes in capturing the 3D spatial information, specifically the depth of objects within the scene. These extracted 3D feature maps provide invaluable insights into the objects' physical relationships and relative positions within the environment.

Notably, these two backbones represent the foundational modules of our network. They capture a comprehensive range of information, ranging from fine-grained visual details to depth-related spatial cues. The integration of these diverse features is achieved through a fusion process that transmits the rich semantic information contained within the high-level feature maps to the lower levels. This recursive connection and fusion are executed in a specific manner, allowing us to harmoniously combine the high-level and low-level feature maps.

By fusing the insights from both the 2D and 3D domains, our network possesses the capacity to robustly analyze complex scenes and accurately localize objects of interest. This fusion of multimodal information equips our system with the comprehensive context necessary to address challenging scenarios, including occlusions and small object classes, further enhancing the effectiveness of our object detection capabilities.

IV. Comparative Analysis

A. Experimental Setup

In our experiments, we leveraged the computational prowess of the Tensor Processing Unit (TPU) v2, which provides an ideal environment for high-performance machine learning and deep learning applications. Equipped with a substantial 35GB of RAM capacity, the TPU v2 played a pivotal role in ensuring efficient model training and evaluation. Additionally, we conducted our experiments using the KITTI dataset, a well-established benchmark in the computer vision community.

Our experimental design entailed a rigorous comparative analysis involving FuDensityNet, alternative occlusion handling methods, and our occlusion-aware network. This enabled us to systematically evaluate the performance of our proposed approach across various occlusion scenarios.

B. Evaluation of 2D Object Detection Models

In our quest to optimize object detection performance in the face of varying occlusion levels, we embarked on a comprehensive comparative analysis of diverse object detection models, as detailed in Table I. This analysis culminated in the formulation of a strategic approach tailored to address different degrees of occlusion encountered in real-world scenarios.

Model	Class AP(%) (KITTI 2D)		
	Car	Pedestrian	Cyclist
F-RCNN	71.2	67.4	66.7
ResNet50-F-RCNN	*76.8*	*69.4*	*67.8*
MobileNetV2-F-RCNN	57.2	53.8	48.5
vgg16-F-RCNN	59.2	58.4	47.6
SSD	66.7	64.4	58.1
RetinaNet	65.6	63.3	58.4
YOLOv5s	89.9	87.7	83.8
YOLOv6s	92.2	88.1	85.7
YOLOv7	90.2	86.5	84.1
YOLOv8s	93.7	91.3	87.2
YOLO-NAS	**95.4**	**92.8**	**91.1**

TABLE I
Object Detection AP Results for KITTI 2D Dataset

When confronted with scenarios characterized by low to no occlusion, our analysis pinpointed YOLO-NAS [24] as the standout performer. YOLO-NAS, distinguished by its lightweight architecture, demonstrated remarkable efficiency and accuracy in object detection. It excelled in scenarios where occlusion was minimal, making it the ideal choice for rapid and precise object detection without the need for intricate feature extraction. This strategic selection harmonizes inference speed and accuracy in scenarios where occlusion presents minimal challenges.

In scenarios with moderate to high occlusion, a custom backbone in Faster R-CNN [17], integrating ResNet50 [25] and our custom 3D backbone, proves effective. This integration harnesses the image feature extraction capabilities of ResNet50 and spatial understanding from 3D data, enhancing the model's resilience against occluded objects. This combined approach enables Faster R-CNN to accurately locate objects despite occlusion, further bolstered by its robust region proposal mechanism.

C. Performance Assessment of FuDensityNet

Our performance evaluation of FuDensityNet (Table II), juxtaposed against YOLO3D, CompNet, and MV3D, revealed its notable superiority over YOLO3D, while it demonstrated performance levels closely aligned with CompNet. Notably, FuDensityNet's performance ranked lower than MV3D and CompNet, with MV3D consistently exhibiting superior performance across all occlusion levels.

This comparative analysis provides valuable insights into the effectiveness of each approach in handling diverse occlusion scenarios. While MV3D consistently outperformed other

Network	Car			Pedestrian			Cyclist		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
CompNet	85.3	78.5	74.9	82.6	75.2	71.8	80.4	68.7	64.5
YOLO3D	64.0	57.8	49.2	61.2	52.7	45.9	56.4	43.0	35.0
MV3D	88.0	**76.1**	**73.5**	85.2	**72.4**	**69.8**	81.2	**65.5**	**62.1**
YOLO-NAS	**95.4**	72.2	70.1	**92.8**	69.3	70.8	**91.1**	64.4	61.5
FuDensityNet-Our	**95.4**	69.8	64.5	**92.8**	68.0	62.3	**91.1**	61.6	58.2

TABLE II
OBJECT DETECTION AP RESULTS ON KITTI DATASET FOR OCCLUSION ANALYSIS

models, it's essential to interpret these results considering contextual factors that may influence the outcomes. FuDensityNet, with its competitive performance, positions itself as a promising solution for occlusion-aware object detection, bridging the gap between efficiency and effectiveness in complex real-world environments.

CONCLUSION

In our pursuit of advancing object detection, we delved into occlusion handling strategies. This research explored cutting-edge methodologies and identified a critical need for addressing occlusion challenges comprehensively. Our response was the development of an end-to-end framework that employs novel techniques, including point cloud density analysis, in place of GLCM, for precise occlusion rate determination. We integrated the Feature Pyramid Network (FPN) strategy to enhance object detection, especially in scenarios with small or overlapping objects. The fusion of depth information and RGB imagery further improved occluded object separation.

While our model, FuDensityNet, didn't claim the top spot, it showcased competitive performance, surpassing one state-of-the-art model. Our ongoing mission is to fine-tune this model, achieving an ideal balance between accuracy and speed for effective occlusion-aware object detection. In the future, we plan to integrate this model into person search tasks, recognizing the paramount importance of occlusion handling in real-world computer vision applications.

In sum, our research represents a significant stride toward robust occlusion handling in object detection, and we are committed to pushing the boundaries of performance in this dynamic field.

REFERENCES

[1] D. T. Nguyen, W. Li, and P. O. Ogunbona, "Human detection from images and videos: A survey," *Pattern Recognition*, vol. 51, pp. 148–175, 2016.

[2] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?," in *Proceedings of the iEEE conference on computer vision and pattern recognition*, pp. 1259–1267, 2016.

[3] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "Motchallenge 2015: Towards a benchmark for multi-target tracking," *arXiv preprint arXiv:1504.01942*, 2015.

[4] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pp. 483–499, Springer, 2016.

[5] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 466–481, 2018.

[6] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361, IEEE, 2012.

[7] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1907–1915, 2017.

[8] F. Pérez-Hernández, S. Tabik, A. Lamas, R. Olmos, H. Fujita, and F. Herrera, "Object detection binary classifiers methodology based on deep learning to identify small objects handled similarly: Application in video surveillance," *Knowledge-Based Systems*, vol. 194, p. 105590, 2020.

[9] A. Kortylewski, J. He, Q. Liu, and A. L. Yuille, "Compositional convolutional neural networks: A deep architecture with innate robustness to partial occlusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8940–8949, 2020.

[10] T. Bagautdinov, F. Fleuret, and P. Fua, "Probability occupancy maps for occluded depth images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2829–2837, 2015.

[11] J. Berclaz, A. Shahrokni, F. Fleuret, J. Ferryman, and P. Fua, "Evaluation of probabilistic occupancy map people detection for surveillance systems," in *Proceedings of the IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, no. CONF, 2009.

[12] Y. Sun, A. Kortylewski, and A. Yuille, "Amodal segmentation through out-of-task and out-of-distribution generalization with a bayesian model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1215–1224, 2022.

[13] A. Kortylewski, Q. Liu, A. Wang, Y. Sun, and A. Yuille, "Compositional convolutional neural networks: A robust and interpretable model for object recognition under occlusion," *International Journal of Computer Vision*, vol. 129, no. 3, pp. 736–760, 2021.

[14] W. Ali, S. Abdelkarim, M. Zidan, M. Zahran, and A. El Sallab, "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*, pp. 0–0, 2018.

[15] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

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

[17] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[18] M. Takahashi, Y. Ji, K. Umeda, and A. Moro, "Expandable yolo: 3d object detection from rgb-d images," in *2020 21st International Conference on Research and Education in Mechatronics (REM)*, pp. 1–5, IEEE, 2020.

[19] P. Jenkins, K. Armstrong, S. Nelson, S. Gotad, J. S. Jenkins, W. Wilkey, and T. Watts, "Countnet3d: A 3d computer vision approach to infer counts of occluded objects," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3008–3017, 2023.

[20] J. Reynolds, C. K. Nagesh, and D. Gurari, "Salient object detection for images taken by people with vision impairments," *arXiv preprint arXiv:2301.05323*, 2023.

[21] O. Mirbod, D. Choi, P. H. Heinemann, R. P. Marini, and L. He, "On-tree apple fruit size estimation using stereo vision with deep learning-based occlusion handling," *Biosystems Engineering*, vol. 226, pp. 27–42, 2023.

[22] B. Sebastian V, A. Unnikrishnan, and K. Balakrishnan, "Gray level co-occurrence matrices: generalisation and some new features," *arXiv preprint arXiv:1205.4831*, 2012.

[23] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

[24] J. Terven and D. Cordova-Esparza, "A comprehensive review of yolo: From yolov1 and beyond. arxiv 2023," *arXiv preprint arXiv:2304.00501*.

[25] B. Koonce and B. Koonce, "Resnet 50," *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*, pp. 63–72, 2021.