Volume: 15 Issue: 06 Pages: 6239– 6245 (2024) ISSN: 0975-0290

Advanced Obstacle Detection and Distance Estimation for Forklift Operations through Integrated Deep Learning Networks

Xinwei Liu [1], Yimin Song [1], Guoyun Ye [2], Min Fu [2], and Wen Liu [3]
[1] Tianjin University
[2] Ningbo Ruyi Joint Stock Co. Ltd
[3] Zhejiang Wanli University

----------ABSTRACT----------

The safety and efficiency of forklift operations in industrial settings are critically dependent on the accurate detection and precise distance measurement of obstacles. This study introduces an innovative deep learning framework that synergizes advanced computer vision methods for obstacle detection with a novel approach to distance estimation using monocular imaging. By harnessing the capabilities of these techniques, the proposed system significantly enhances the safety protocols during forklift navigation. Our comprehensive experimental evaluation demonstrates notable advancements in the accuracy of obstacle identification and the reliability of distance calculations across a range of obstacle sizes and environmental conditions. The outcomes position this research as a pivotal step towards the automation and optimization of forklift operations.

Keywords - **Obstacle Detection, Distance Estimation, Forklift Operation.**

Date of Submission: May 24, 2024 Date of Acceptance: June 16, 2024

I. INTRODUCTION

Obstacle detection and distance estimation are essential components of industrial forklift operations, critical for maintaining a safe working environment. In the context of material handling, forklifts are required to navigate through often crowded and busy spaces, where the presence of static and dynamic obstacles poses a constant risk of collisions. The ability to accurately identify and assess the distance of these obstacles is paramount to prevent accidents, ensure the well-being of workers, and protect equipment from damage.

The precision of obstacle detection systems significantly enhances the operator's awareness of their surroundings. This heightened awareness is particularly beneficial in settings where visual obstructions or poor lighting conditions might otherwise hinder the operator's ability to see potential hazards. With the aid of such systems, operators can make informed decisions about their movements, leading to safer navigation and reducing the likelihood of incidents. Moreover, the efficiency of forklift operations directly correlates with the effectiveness of obstacle detection and distance estimation. By minimizing the time spent on cautious maneuvering and the potential for accidents, forklifts can operate at optimal productivity levels. This not only improves the overall workflow but also contributes to a more confident and focused operation, as the operators can rely on technology to assist with safety measures.

Compliance with safety regulations is another key aspect where obstacle detection and distance estimation play a crucial role. Industries are subject to strict safety standards, and the implementation of advanced safety features such as these systems helps companies adhere to these standards. By doing so, they avoid the financial and legal repercussions associated with workplace accidents and demonstrate a commitment to worker safety. The adaptability of obstacle detection systems to dynamic industrial conditions is also noteworthy. As warehouse layouts change and activity levels fluctuate, a robust system can adapt to these changes, ensuring that safety measures remain effective despite the evolving environment. This adaptability is particularly important as industries strive to maintain a competitive edge by regularly updating their operational processes.

In the emerging era of automation, obstacle detection and distance estimation are becoming integral to the development of autonomous forklifts. These technologies provide the spatial intelligence necessary for autonomous vehicles to operate safely without human intervention, marking a significant step forward in the automation of industrial processes. While skilled operators are vital to the safe operation of forklifts, the integration of advanced obstacle detection systems serves to minimize the risk of human error. These systems offer an additional layer of safety that is not subject to fatigue or distractions, ensuring a consistent level of vigilance and further reducing the potential for accidents.

Therefore, the integration of obstacle detection and distance estimation technologies into forklift operations is a critical strategy for enhancing safety, efficiency, and regulatory compliance. As industrial environments continue to evolve, the importance of these systems in maintaining a safe and productive workplace will only continue to grow.

This paper is organized in the following: Section 2 reviews related work, while Section 3 details our proposed method. Section 4 discusses the training strategy, and Section 5 presents experimental results. Section 6 concludes the paper.

Volume: 15 Issue: 06 Pages: 6239– 6245 (2024) ISSN: 0975-0290

II. BACKGROUND AND RELATED WORK

Traditional methods for obstacle detection and distance estimation have been extensively researched and developed to enhance the safety and efficiency of autonomous vehicles and industrial machinery, including forklifts. These methods can be categorized into sensor-based and computer vision-based techniques, each with unique characteristics and applications.

Sensor-based methods have long been employed for their ability to provide reliable and robust detection in various environmental conditions. Ultrasonic sensors, for example, are known for their simplicity and cost-effectiveness. They operate by emitting sound waves and measuring the time delay of the echo to determine the distance to an obstacle. Despite their widespread use, ultrasonic sensors have limited resolution and can be sensitive to noise, which may affect their performance in noisy industrial environments [1]. Infrared sensors offer another means of detection, often used for their ease of use and resistance to environmental conditions. They function by emitting infrared light and measuring the reflection to identify obstacles. However, infrared sensors may struggle with accuracy in environments with varying temperatures or diverse surface materials, as these factors can influence the reflection of infrared light [2]. Laser-based systems, such as LIDAR, have become a standard in many high-precision applications due to their ability to generate detailed, three-dimensional maps of the environment. LIDAR systems are highly accurate and capable of detecting a wide range of obstacles, but they tend to be more expensive and require significant computational power for data processing and interpretation [3].

Computer vision-based methods have gained popularity with the advancement of digital image processing and machine learning techniques. Monocular vision systems, which use a single camera, rely on algorithms to estimate depth from a sequence of images. These systems can be less expensive and more flexible than sensor-based approaches, but they may suffer from inaccuracies in low-textured environments or under poor lighting conditions [4]. Stereo vision systems, which utilize two cameras to capture and compare images, offer improved depth estimation by calculating the disparity between the two images. This method can provide more accurate and reliable results than monocular systems, but they require careful calibration and synchronization between the cameras [5].

Overall, traditional methods for obstacle detection and distance estimation have made significant strides, offering a range of solutions for different applications. Sensor-based methods provide robust and reliable detection, while computer vision techniques offer high-resolution and flexible solutions. However, each method has its limitations, and ongoing research aims to address these challenges and further improve the performance of these systems.

2.1. Deep Learning based Methods

The evolution of deep learning in object detection has significantly impacted the field of computer vision, particularly in the context of monocular vision for distance estimation. Deep learning-based approaches have transitioned from handcrafted feature extraction methods to powerful convolutional neural networks (CNNs) that can automatically learn complex representations from data. Initially, deep learning models like AlexNet and VGGNet revolutionized object detection by demonstrating the effectiveness of CNNs in feature extraction and classification tasks [6], [7]. These models, however, were primarily designed for image classification and had to be adapted for object detection in more complex scenes.

The introduction of YOLO (You Only Look Once) marked a turning point in real-time object detection, as it provided a single-pass approach to detecting objects in images with bounding boxes and class probabilities [8]. YOLO's success led to a series of improvements, including YOLOv2, YOLOv3, and YOLOv4, each enhancing the previous version in terms of speed, accuracy, and the ability to handle more intricate object interactions [9], [10], [11], [12]. The role of monocular vision in distance estimation has also seen significant advancements with the integration of deep learning. Early attempts at monocular depth estimation relied on traditional machine learning techniques and were limited by their ability to generalize to unseen scenes [13]. With the advent of deep learning, models could be trained on large datasets to learn depth cues from monocular images, leading to more accurate depth predictions [14], [15].

The combination of object detection and distance estimation in monocular vision systems has been further advanced by the development of architectures that can jointly learn both tasks. These models leverage the spatial and semantic information from the monocular images to estimate not only the location of objects but also their depth and distance from the camera [16]. Despite these advancements, challenges remain in the area of monocular depth estimation, such as dealing with ambiguous depth cues in large open scenes or under varying lighting conditions. Current research continues to explore ways to improve the robustness and accuracy of these systems, including the use of attention mechanisms, contextual information, and multi-task learning strategies [17].

The integration of deep learning into object detection and distance estimation using monocular vision has opened up new possibilities for applications in autonomous vehicles, robotics, and augmented reality. The ongoing development of these techniques promises to further enhance the capabilities of monocular vision systems, bringing them closer to the performance of stereo and other more complex sensing systems.

2.2. Small Objects Detection in Complex Scenes

The accurate detection of small objects in complex scenes has been a challenging task in the field of computer vision. However, with the advent of deep learning, significant progress has been made in developing neural network architectures that can effectively address this issue. These advancements are crucial for applications such as surveillance, autonomous driving, and industrial monitoring,

Volume: 15 Issue: 06 Pages: 6239– 6245 (2024) ISSN: 0975-0290

where detecting and identifying small objects amidst cluttered backgrounds is essential.

One of the key advancements in neural network architectures that have improved small object detection is the development of feature pyramid networks (FPNs). FPNs address the issue of object scale variation by creating a hierarchy of feature maps at different resolutions. This allows the network to maintain and utilize high-resolution information from earlier layers while benefiting from the semantic context provided by deeper layers. By doing so, FPNs enhance the detection of small objects by capturing fine details without losing spatial information [18].

The use of attention mechanisms has also been a notable development in recent neural network architectures. Attention modules, such as the non-local neural network, enable the model to dynamically focus on relevant parts of the input image, which is particularly useful for small object detection. By allocating more computational resources to the regions of interest, these attention-based models can better capture the details of small objects and improve detection accuracy [19].

Furthermore, the concept of transfer learning and the use of pre-trained models have played a significant role in enhancing small object detection. Models pre-trained on large-scale datasets, such as ImageNet, can be fine-tuned for specific tasks involving small objects. This approach leverages the rich feature representations learned from vast and diverse datasets, enabling the detection models to generalize better to small objects in complex scenes [20].

Advancements in neural network architectures, such as the development of FPNs, the incorporation of attention mechanisms, and the use of transfer learning, have collectively contributed to the improved detection of small objects in complex scenes. These innovations have led to more accurate and efficient detection systems, which are vital for a wide range of applications where the accurate identification of small objects is critical.

III. Proposed Method

In this work, we present a novel deep learning framework that unifies the tasks of obstacle detection and distance estimation, with a particular emphasis on small objects. The proposed method is designed to overcome the limitations of existing approaches, which often treat these tasks separately, leading to suboptimal performance in complex, real-world scenarios. The structure of the proposed method is illustrated in Fig. 1.

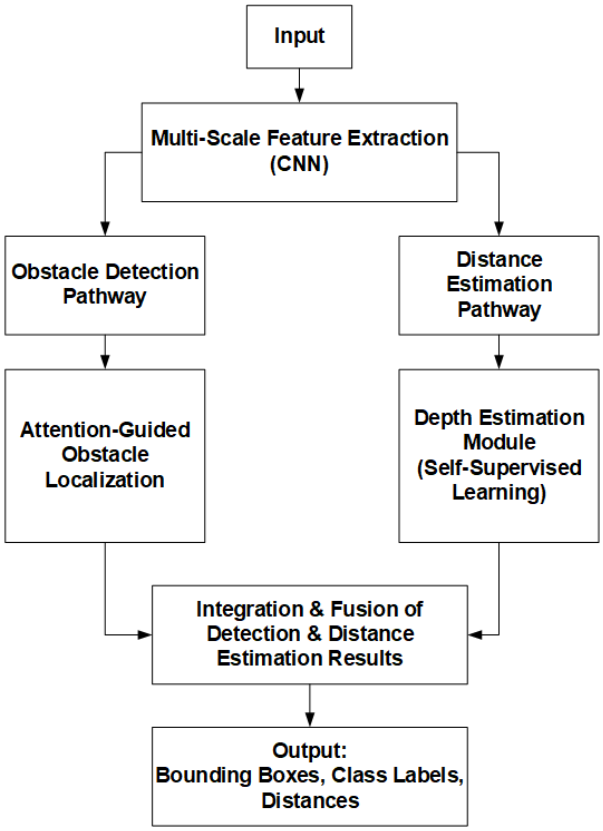

Fig. 1. The structure of the proposed method.

3.1 Backbone Network

The backbone network is composed of a series of convolutional layers that serve to extract features at various scales. The initial layers employ larger kernel sizes (e.g., 7x7) with corresponding strides to reduce the spatial dimensions of the input while capturing low-level features. Subsequent layers utilize smaller kernel sizes (e.g., 3x3) with overlapping strides to refine these features and capture more intricate patterns. The network incorporates skip connections, inspired by ResNet architecture, which allow for the fusion of low-level features with high-level abstractions. This design choice is critical for preserving spatial information, which is essential for the accurate detection of small objects. The backbone is shown in Fig. 2.

3.2 Object Detection Pathway

The object detection pathway consists of a sequence of specialized layers designed to identify and localize obstacles within the scene. This pathway includes a region proposal network (RPN) that scans the feature maps produced by the backbone to identify regions of interest (RoIs). These RoIs are then processed by a series of convolutional layers equipped with non-linear activation functions to predict bounding boxes and class probabilities. To enhance the focus on small objects, the pathway incorporates a feature pyramid structure that aggregates features from different layers. This pyramid structure ensures that the network does not lose sight of small object details when processing larger context. The object detection pathway is shown in Fig. 3.

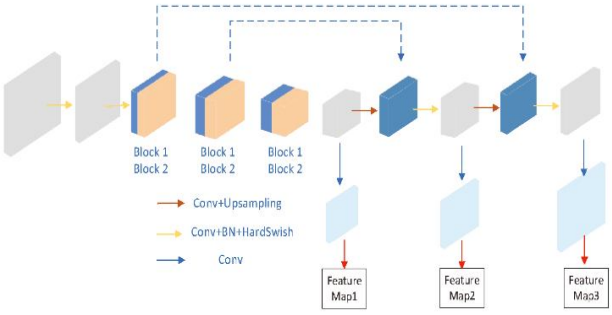

Fig. 2. The structure of the backbone.

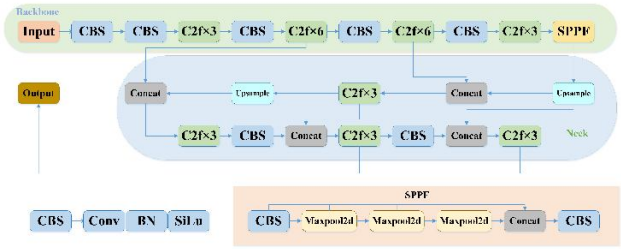

Fig. 3. The structure of the object detection pathway.

3.3 Distance Estimation Pathway

The distance estimation pathway leverages the temporal consistency of video frames to infer depth information. This pathway is equipped with a self-supervised learning module that estimates the relative motion between the camera and the objects in the scene. The module uses optical flow estimation techniques to analyze the movement of feature points between consecutive frames, which is then used to predict the depth of the objects. A novel aspect of this pathway is the dynamic scale factor adjustment, which adapts the depth estimation based on the size of the detected objects. This mechanism prevents small objects from being underestimated in distance and ensures that the network's depth estimation is accurate across a range of object sizes. The distance estimation is shown in Fig. 4.

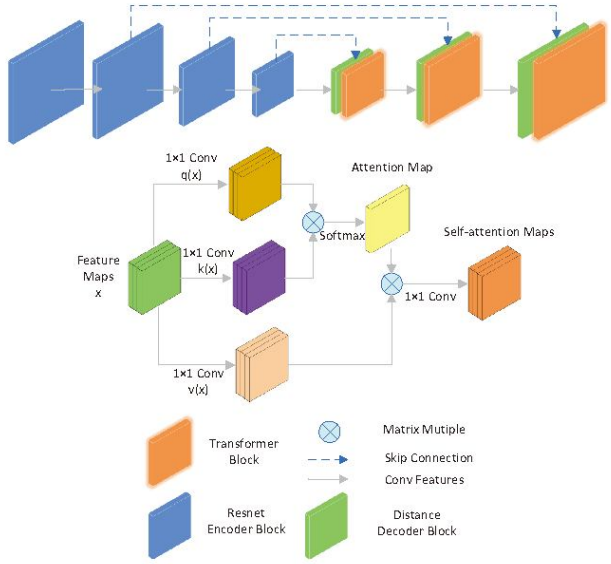

Fig. 4. The structure of the distance estimation pathway.

3.4 Integration of Pathways

The integration of the object detection and distance estimation pathways is facilitated through a series of fusion layers. These layers combine the output of the object detection pathway, which includes the bounding boxes and class labels, with the depth estimates from the distance estimation pathway. The fusion process is designed to maximize the synergy between the two tasks, leading to improved performance in both detection and distance estimation. The integrated network is shown in Fig. 5.

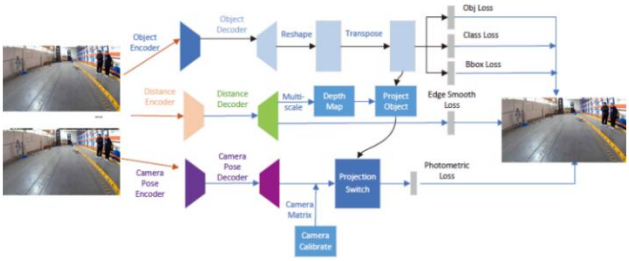

Fig. 5. The structure of the integrated network.

3.5 Training Strategy

The network is trained using a combined loss function that optimizes for both object detection accuracy and distance estimation precision. The loss function includes components for bounding box regression, class probability, and depth estimation. The training process involves a large dataset of annotated images and video sequences, with data augmentation techniques employed to increase the robustness of the network to variations in object size, lighting, and scene complexity.

IV. Experimental Results

4.1 Evaluation Metrics

The objective of our study is to accurately estimate the bounding boxes of objects and their distances with a high degree of precision, aligning as closely as possible with the actual facts. To this end, we employ "precision" and "recall" as the primary indicators to assess the precision of object detection. For depth prediction, we utilize four distinct metrics for evaluation: the absolute relative difference (Abs Rel), the squared relative difference (Squa Rel), the root of the mean squared errors (RMSE), and the root of the mean squared errors calculated from the logarithms of the predicted and ground truth distances (RMSElog). Here, d_i^{gt} represents the ground truth distance, while d_i signifies the predicted distance. The computation of these errors is based on the following five equations:

$$Threshold: \%\ of\ d_i\ s.t. \max\left(\frac{d_i}{d_i^{gt}}, \frac{d_i^{gt}}{d_i}\right) = \sigma < Thresh.$$

$$Abs\ Rel: \frac{1}{N}\sum_{d\in N}\frac{|d - d^{gt}|}{d^{gt}}$$

$$Squa\ Rel: \frac{1}{N}\sum_{d\in N}\frac{||d - d^{gt}||^2}{d^{gt}}$$

$$RMSE(linear): \sqrt{\frac{1}{N}\sum_{d\in N}||d_i - d_i^{gt}||^2}$$

$$RMSE(log): \sqrt{\frac{1}{N}\sum_{d\in N}||logd_i - logd_i^{gt}||^2}$$

4.2 Evaluation Resultss

We created a private dataset contains videos from forklift operation environment for the performance evaluation.

Volume: 15 Issue: 06 Pages: 6239– 6245 (2024) ISSN: 0975-0290

Table 2. Quantitative performance comparison of our network with other self-supervised monocular methods for distance estimation.

Approach	Lower is better			Higher is better		
	Abs Rel	Squa Rel	RMSE Log	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$
Zhou	0.183	1.295	0.270	0.734	0.902	0.959
GeoNet	0.149	1.060	0.226	0.796	0.935	0.975
Struct2depth	0.141	1.026	0.215	0.816	0.945	0.979
PackNet-SfM	0.120	0.892	0.193	0.864	0.954	0.980
Monodepth2	0.115	0.903	0.193	0.877	0.959	0.981
FisheyeNet	0.117	0.867	0.190	0.869	0.960	0.982
SymDistNet	0.109	0.718	0.180	0.896	0.973	0.986
Shu	0.104	0.729	0.179	0.893	0.965	0.984
Proposed	**0.008**	**0.694**	**0.171**	**0.901**	**0.978**	**0.993**

For object detection network, we used 3000 training images and 3000 test images extracted from videos in our dataset, comprising a total of 100,000 labeled objects, then grouped the categories in the target detection dataset into just three categories: people, vehicle, and obstacle, and used these three categories for both training and validation. We compare the proposed method with famous and most commonly used object detection network YOLOv5, and the proposed method has better precision and recall rates than YOLOv5. The final results for obstacle detection are shown in Table 1.

Table 1. Detection results compared to YOLOv5 network.

Class	Method	Precision(%)	Recall(%)
People	Proposed	96.7	95.6
	YOLOv5	95.4	93.5
Vehicle	Proposed	95.2	92.8
	YOLOv5	94.4	90.2
Obstacle	Proposed	92.9	88.8
	YOLOv5	90.3	81.5

For the distance estimate network, before training, to split raw data, we use the same method as obstacle detection. The training data contains 8000 images, validation data contains 4000 images, and test data contains 500 images. Then, we filter static frames using the default camera matrix for all images. The focal length is averaged. Further, we add channel spatial and self-attention modules to the depth decoder. Both of them can enhance the effect. The self-attention module performs better than the CBAM module because our data are sequential, and similar objects can be better distinguished and further focused. The actual distance is then acquired. We test our method and other classical methods, and the bold numbers in Table 2 shows that our proposed models are able to predict distances with lower absolute errors. Our proposed method has lower Abs Rel, Squa Rel, RMSE compared to the other selected classical methods. And the proposed method also has higher σ values then the other approaches in Table 2.

Several examples of object detection and distance estimation in videos frames are illustrated in Fig. 5. From Fig. 5 we can notice that, the proposed method can very well detect both big and small obstacles on the working path of the forklift, while their distances to the vehicle are accurately estimated.

Fig. 5. Examples of object detection and distance estimation results in video frames.

Volume: 15 Issue: 06 Pages: 6239– 6245 (2024) ISSN: 0975-0290

V. CONCLUSION

This study introduces an innovative deep learning framework for obstacle detection and distance estimation in forklift operations. The system integrates advanced computer vision with monocular imaging, significantly improving safety and efficiency during navigation. Experiments demonstrate high accuracy in obstacle identification and reliable distance calculations across various conditions. Our approach, particularly effective for small object detection in complex scenes, positions this research as a key advancement towards the automation of forklift operations.

The proposed method outperforms existing techniques, showing lower errors in distance estimation and higher precision and recall in object detection compared to YOLOv5. This research highlights the transformative potential of deep learning in industrial safety, suggesting that as environments evolve, the integration of such systems will be increasingly vital for maintaining safe and productive workplaces.

REFERENCES

[1] Y. V. Chavan, P. Y. Chavan, A. Nyayanit, and V. S. Waydande, "Obstacle detection and avoidance for automated vehicle: a review," *J Opt*, vol. 50, no. 1, pp. 46–54, Mar. 2021, doi: 10.1007/s12596-020-00676-6.

[2] T. S. Arulananth and M. Baskar, "IR Sensor based obstacle detection and avoiding robot," *Palarch's Journal of Archaeology of Egypt/Egyptology*, vol. 17, no. 9, pp. 3328–3336, 2020.

[3] A. N. Catapang and M. Ramos, "Obstacle detection using a 2D LIDAR system for an Autonomous Vehicle," in *2016 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, IEEE, 2016, pp. 441–445. Accessed: May 23, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7893614/

[4] T.-J. Lee, D.-H. Yi, and D.-I. "Dan" Cho, "A monocular vision sensor-based obstacle detection algorithm for autonomous robots," *Sensors*, vol. 16, no. 3, p. 311, 2016.

[5] N. Bernini, M. Bertozzi, L. Castangia, M. Patander, and M. Sabbatelli, "Real-time obstacle detection using stereo vision for autonomous ground vehicles: A survey," in *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, IEEE, 2014, pp. 873–878. Accessed: May 23, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6957799/

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012, Accessed: May 23, 2024. [Online]. Available: https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[7] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition." arXiv, Apr. 10, 2015. Accessed: May 23, 2024. [Online]. Available: http://arxiv.org/abs/1409.1556

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788. Accessed: May 23, 2024. [Online]. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Redmon_You_Only_Look_CVPR_2016_paper.html

[9] J. Zhang, M. Huang, X. Jin, and X. Li, "A real-time chinese traffic sign detection algorithm based on modified YOLOv2," *Algorithms*, vol. 10, no. 4, p. 127, 2017.

[10] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection." arXiv, Apr. 22, 2020. Accessed: May 23, 2024. [Online]. Available: http://arxiv.org/abs/2004.10934

[11] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "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*, 2023, pp. 7464–7475. Accessed: May 23, 2024. [Online]. Available: http://openaccess.thecvf.com/content/CVPR2023/html/Wang_YOLOv7_Trainable_Bag-of-Freebies_Sets_New_State-of-the-Art_for_Real-Time_Object_Detectors_CVPR_2023_paper.html

[12] A. Jain, R. Singh, and P. Singh, "Enhancing Outlier Detection and Dimensionality Reduction in Machine Learning for Extreme Value Analysis," *Int. J. Advanced Networking and Applications*, vol. 15, no. 06, pp. 6204–6210, 2024.

[13] A. Saxena, M. Sun, and A. Y. Ng, "Make3d: Learning 3d scene structure from a single still image," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 5, pp. 824–840, 2008.

[14] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2650–2658. Accessed: May 23, 2024. [Online]. Available: http://openaccess.thecvf.com/content_iccv_2015/html/Eigen_Predicting_Depth_Surface_ICCV_2015_paper.html

[15] H. Li *et al.*, "Convolution Serialization Recommendation with Time Characteristics and User Preferences," *Int. J. Advanced Networking and Applications*, vol. 15, no. 06, pp. 6156–6162, 2024.

[16] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131. Accessed: May 23, 2024. [Online]. Available: http://openaccess.thecvf.com/content_ECCV_2018/html/Ningning_Light-weight_CNN_Architecture_ECCV_2018_paper.html

[17] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 10, pp. 2024–2039, 2015.

Volume: 15 Issue: 06 Pages: 6239– 6245 (2024) ISSN: 0975-0290

[18] 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*, 2017, pp. 2117–2125. Accessed: May 23, 2024. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2017/html/Lin_Feature_Pyramid_Networks_CVPR_2017_paper.html

[19] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803. Accessed: May 23, 2024. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Wang_Non-Local_Neural_Networks_CVPR_2018_paper.html

[20] X. Xu, H. Zhang, Y. Ma, K. Liu, H. Bao, and X. Qian, "Transdet: Toward effective transfer learning for small-object detection," *Remote Sensing*, vol. 15, no. 14, p. 3525, 2023.