iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IX **Month of publication:** September 2023

DOI: https://doi.org/10.22214/ijraset.2023.55885

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IX Sep 2023- Available at www.ijraset.com

Real-Time Pedestrian Detection with YOLOv7 and Intel MiDaS: A Qualitative Study

Dashmesh Yashasvi Singh[1], Sumesh Sood[2], Arvind Kalia[3]
Department of Computer Science, Himachal Pradesh University, Shimla

Abstract: ***In the realm of Advanced Driving Assistance Systems (ADAS), the accurate assessment of pedestrian proximity is of paramount importance. This paper introduces a qualitative methodology that integrates YOLOv7-pose for object detection and pose estimation with the MiDaS (Monocular Depth Estimation in Real-Time with Deep Learning) model for monocular depth estimation. The main objective is to qualitatively assess pedestrian proximity to the camera within the ADAS framework. This procedure involves classifying pedestrians as "near" or "far" based on an inverse depth threshold that has been predetermined. In addition, the paper performs a qualitative comparative analysis of the results produced by the MiDaS Small, Hybrid, and Large variants to learn more about the performance of depth estimation in these contexts, particularly in relation to the presence of pedestrians. The evaluation emphasises this approach's qualitative potential for achieving situationally appropriate and context-aware pedestrian proximity assessment. The safety and adaptability of ADAS systems can be improved with the help of such insights, which have numerous applications in robotics, surveillance, and autonomous vehicles.***
Keywords: YOLOv7; MiDaS; Pedestrian Proximity Detection; Pedestrian Prediction

I. INTRODUCTION

The potential for autonomous vehicles and Advanced Driving Assistance Systems (ADAS) to provide significant advantages in safety, reducing traffic congestion, and improving fuel efficiency has sparked a recent increase in interest in these technologies. The phrase "safety envelope" has become popular in this context which refers to the clear space around a machine. This is especially important for autonomous systems, where the coexistence of humans is a major concern to ensure safe operation [1].

In the past, human operators were primarily in charge of upholding this safety perimeter. There is, however, a fundamental shift in this responsibility as technology develops, particularly with the incorporation of cutting-edge models like YOLO for object detection and various models for depth or proximity estimation [2].

YOLO models have been used in several fields, including agriculture, to identify and categorize crops, diseases and pests, as well as to perform tasks like face detection and recognition in the fields of biometrics and security [3]. However, the real-time object detection capabilities of YOLO models enable them to implement quick identification and monitoring of a variety of entities like automobiles and pedestrians [4], [5], as well as bicycles and various impediments [3]. These real-time object detection capabilities were highly improved by YOLOv7 [6] in 2022 and therefore it was established as the focus of this research.

Similarly, understanding the importance of depth estimation, it becomes evident that this capability plays a pivotal role in driving scenarios to maintain the safety envelope[1]. The complicated task of determining depth with a single lens and only RGB input channels instead of an additional depth channel ‘D’, is called Monocular Depth Estimation. One such promising monocular depth estimation deep learning-based model is MiDaS [2]. Ranfl et al. during the development of MiDaS created novel loss functions that could handle a range of data sources and scales [2]. When they looked into the best methods for combining datasets during training, they found that a multi-objective optimization strategy produced the best results [7], which resulted in the advancement of the state-of-the-art generic monocular depth estimation.

One can determine the separation between a camera and people or objects by combining these two powerful methods. This ability opens a wide range of possibilities in fields like surveillance, robotics, and most importantly autonomous vehicles. Therefore, in this study, we will explore the potential of combining YOLOv7-pose detections with Midas monocular depth estimation to estimate the proximity of objects and pedestrians.

A. Objectives

In pursuit of a cohesive vision, the following key objectives are stated for this integrated framework:

1) *Framework Development:* To create an integrated framework using Intel MIDAS and YOLOv7 for object detection and depth estimation.
2) *Qualitative Assessment:* To qualitatively evaluate the system's performance showcasing the outcome achieved.

B. Outline

The paper unfolds in six distinct sections. In the introduction, we set the stage for our study. The methodology section outlines our approach to combining Intel MIDAS and YOLOv7. We delve into the intricacies of model integration in the subsequent section. Then, we shift focus to the qualitative evaluation of our integrated system, dissecting its performance in real-world scenarios. In the practical applications section, we illustrate how this integration can revolutionize fields like autonomous vehicles, robotics, and surveillance. Finally, we summarize our results and suggest new lines of inquiry for further study.

II. LITERATURE REVIEW

In the literature review section, two streams of research are navigated through: the first encompasses studies related to YOLOv7, focusing on its evolution and contributions to object detection. The second stream delves into the world of MIDAS, discussing its significance in monocular depth estimation. These insights form the foundation for our integrated framework.

A. Related to YOLOv7

Joseph Redmon et al. first presented the YOLO model in 2015 in their paper "You Only Look Once: Unified, Real-Time Object Detection" [8]. Their goal was to streamline the detection of objects into a single stage and reduce the time-consuming inference processes involved with multistage methods. The model partitioned input images into grids, assessing the likelihood of object presence within each grid, and then amalgamated adjacent high-value probability grids to form distinct objects [9]. Non-Max Suppression (NMS) was employed to filter out low-value predictions, and model training involved comparing ground truth with the centre of each detected object thereby facilitating accuracy assessment and weight adjustments [8].

Subsequently, in their 2016 paper "YOLO 9000: Better, Faster, Stronger" [10], Joseph Redmon and Ali Farhadi presented YOLOv2, capable of detecting over 9000 different item categories. YOLOv2 introduced the concept of anchor boxes, predetermined areas indicating ideal object positions within an image. These anchor boxes were determined through dimension clustering of training data, ensuring accurate representation during model training and ultimately enhancing accuracy [10].

The evolution continued with YOLOv3, detailed in Joseph Redmon and Ali Farhadi's 2018 paper titled "YOLOv3: An Incremental Improvement" [11]. While slightly larger, YOLOv3 maintained commendable speed and accuracy. Notably, it incorporated 75 convolutional layers without fully connected or pooling layers, reducing model size and weight. This version leveraged residual models from the ResNet architecture, implementing feature pyramid networks (FPN) for multi-level feature learning while preserving rapid inference times, striking a balance between accuracy and efficiency [9].

In their 2020 publication, "YOLOv4: Optimal Speed and Accuracy of Object Detection"[10], Bochkovskiy, et al. introduced YOLOV4. YOLOv4 introduced the ideas of the "bag of freebies" (techniques that enhance model performance without raising the cost of inference) and the "bag of specials" (techniques that enhance accuracy while boosting the cost of computing). The “bag of freebies” consists of data augmentation techniques, bounding box regression loss, regularization etc whereas “bag of specials” consists of Spatial attention modules (SAM), Non-max suppression (NMS) etc[9].

Glen Jocher, the creator and CEO of Ultralytics, launched YOLOv5 a few months after YOLOv4 in 2020[11]. It was created in Pytorch rather than Darknet and makes use of many of the advancements mentioned in the YOLOv4 section. The Ultralytics algorithm AutoAnchor is used in YOLOv5. This pre-training programme examines anchor boxes and modifies them if necessary to make them better fit the training and dataset parameters, such as image dimensions. It first uses the k-means function on the dataset labels to establish the foundation for a Genetic Evolution (GE) method. The GE algorithm then uses the fitness functions CIoU loss [12] and Best Possible Recall to evolve these anchors across 1000 generations.

Meituan Vision AI Department released YOLOv6 [13] in 2022. The network design consists of a PAN topology neck, an effective decoupled head with a hybrid-channel strategy, and an efficient backbone with blocks made of RepVGG or CSPStackRep. The research also introduces improved quantization methods that lead to quicker and more precise detectors, such as post-training quantization and channel-wise distillation. Overall, YOLOv6 performs better on accuracy and speed metrics than earlier state-of-the-art models.

YOLOv7 [6] was put forward by Bochkovskiy, et al. And in the 5 to 160 frame per second range, YOLOv7 outperformed all other object detectors in terms of speed and accuracy. It was trained without using pre-trained backbones, just like YOLOv4, using only the MS COCO dataset. The accuracy was improved by YOLOv7's recommended adjustments to the architecture and a number of "bag-of-freebies," which simply increased training time and had no impact on inference speed.

B. Related to MiDaS

In 2020, some tools were introduced in [2] that enhanced monocular depth estimation by combining diverse datasets (mainly 3-D Movies). These tools included a versatile loss function and a systematic approach to dataset amalgamation called *zero-shot-cross-dataset* transfer [2]. Comprehensive evaluation showcased superior performance, establishing a new state-of-the-art in monocular depth estimation. These advancements promised practical deployment of models tailored to various applications.

In 2021, the authors of [14] introduced a groundbreaking architectural concept known as dense vision transformers. These innovations replaced traditional convolutional networks with vision transformers as the architectural backbone for dense prediction tasks. Tokens from different stages of the vision transformer were artfully combined to generate image-like representations at various resolutions. A convolutional decoder was then employed to progressively merge these representations into full-resolution predictions. This approach yielded superior predictions, with substantial performance improvements, particularly when ample training data was accessible. Notably, in monocular depth estimation, the architecture outperformed state-of-the-art fully-convolutional networks by up to 28%. In semantic segmentation, the dense vision transformers set a new state-of-the-art benchmark with a 49.02% mean Intersection over Union (mIoU) on the ADE20K dataset. Additionally, the architecture displayed adaptability by achieving state-of-the-art results when fine-tuned on smaller datasets like NYUv2, KITTI, and Pascal Context. The dense prediction transformer (DPT) emerged as a powerful neural network architecture, especially effective when trained on extensive datasets.

C. Insights

With its superior performance in terms of speed and accuracy, YOLOv7 has since emerged as a leader in the field of object detection. Notably, it outperformed all real-time object detectors operating at 30 FPS or higher on GPU V100, achieving a remarkable accuracy of 56.8% Average Precision (AP). A variant, YOLOv7-E6, showed off its abilities by outpacing the transformer-based detector SWIN-L Cascade-Mask R-CNN by an amazing 509% in speed and achieving a 2% boost in accuracy [6].

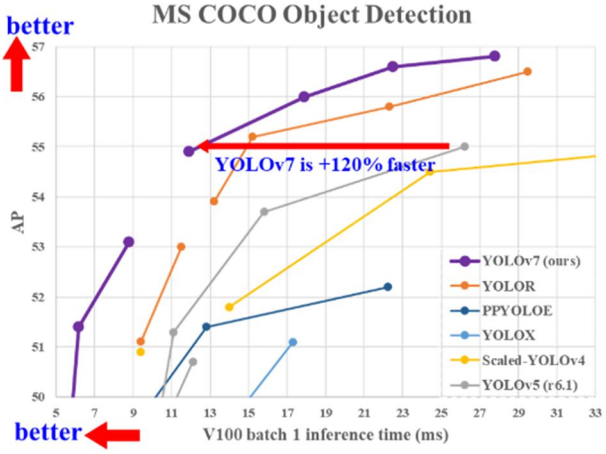


Fig. 1 Comparison of YOLOv7 with other realtime models [6]

The YOLOv7-E6 detector also showed a 551% speed improvement over the convolutional-based detector ConvNeXt-XL Cascade-Mask R-CNN, along with a 0.7% accuracy improvement. The comparative analysis in Fig. 1 shows that YOLOv7 outperformed a wide range of object detectors in terms of both speed and accuracy, including YOLOR, YOLOX, Scaled-YOLOv4, YOLOv5, DETR, Deformable DETR, DINO-5scale-R50, ViT-Adapter-B, and others. It's interesting to note that YOLOv7, despite having fewer parameters and computations, was still able to achieve higher average precision (AP) than YOLOv4 and YOLOR-CSP.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IX Sep 2023- Available at www.ijraset.com

Furthermore, it significantly reduced parameters and computations while maintaining an equivalent AP to YOLOv4-tiny-31. The proposed trainable bag-of-freebies method improved this already outstanding performance by increasing the YOLOv7 object detection accuracy [6]

III. METHODOLOGY

In the methodology section, the approach to integrating Intel MIDAS and YOLOv7 for comprehensive object detection and depth estimation is presented starting with a bird's eye view of the methodology followed.

A. High-Level Approach Followed

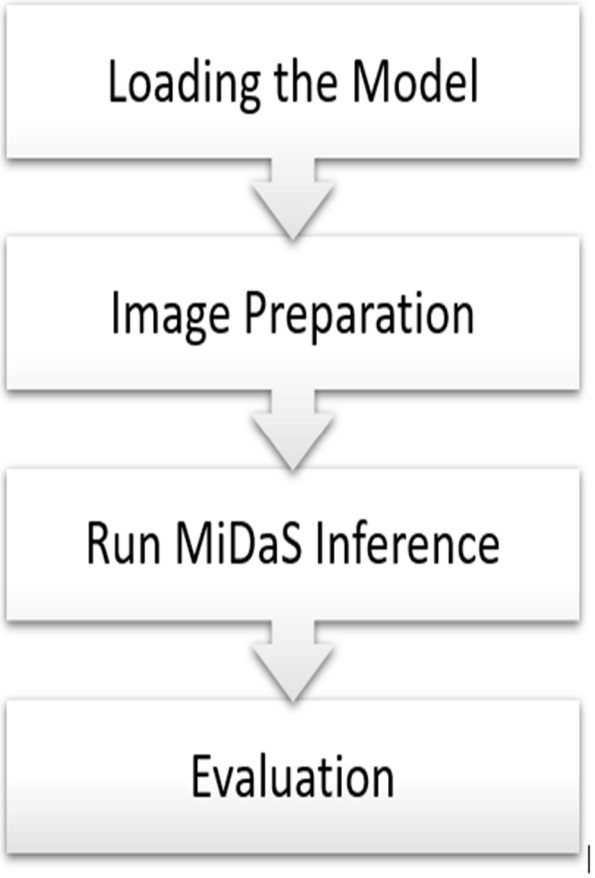


Fig. 2 The overall approach followed in the study

Step 1: Model Setup
Load the MiDaS model (small/hybrid/large) and YOLOv7 models.

Step 2: Image Preparation and Transformation
Load the input image, convert it to RGB, apply YOLOv7-w6-pose model to calculate the keypoints, and resize it to 500x600 pixels

Step 3: Run MiDaS Inference
Use the MiDaS model for depth prediction on the transformed image, upsample the result to the original image size, and store it in the 'output' variable.

Step 4: Evaluation

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IX Sep 2023- Available at www.ijraset.com

B. Tools and Libraries utilised during implementation:

TABLE 1
RESOURCES USED FOR IMPLEMENTATION

S No.	Tools	Libraries
1	Google Colab	Cuda
2	Jupyter Lab	Matplotlib
3	Github	pyTorch
4	Google Drive	pyTorch Lightning
5		Numpy
6		OpenCV
7		timm

IV. IMPLEMENTATION DETAILS

A. Object Detection and Pose Estimation with YOLOv7-pose:

1) In this phase, the YOLOv7-pose model is employed for both object detection and pose estimation.
2) The model's path, which is yolov7-w6-pose.pt, is specified, and then the YOLOv7-pose model is loaded using the function loading_yolov7_model().
3) A confirmation message that the given model is ready for usage is printed if the model is loaded successfully.

B. Processing the Input Image:

1) This step focuses on preparing the input image for further analysis.
2) The YOLOv7-pose model is used to process the input image after calling the run_inference() function. A list of keypoint coordinates is produced as a result of this operation, containing details on the detected objects and their poses.
3) The draw_keypoints() function is used to display the detected keypoints on the processed image. It requires as inputs the processed image, the keypoint coordinates, and the threshold settings.

C. Running MiDaS Inference on Detected Boxes:

1) This stage involves assessing the depth of the detected objects, particularly pedestrians, using the MiDaS models.
2) Based on a predetermined threshold inverse depth value, pedestrians are classified as either "near" or "far" in this essential decision. Inverse depth values less than 0.5 are categorised as "near" pedestrians, whereas values greater than 0.5 are categorised as "far" pedestrians.
3) Each pedestrian's classification is determined based on their specific inverse depth value.
4) An organised CSV file is kept with the findings of this depth assessment, including detection positions and inverse depth values, for later analysis and research.

D. Evaluation:

1) Beyond the core functionality, the system's performance is assessed under various scenarios, including video and webcam streams.
2) The possibility of improving depth sensing's effectiveness and resilience is an important topic that is covered in the debate. This can be accomplished by extrapolating depth measurements from both a stereo camera configuration and a monocular camera.
3) Qualitatively compare the outputs of MiDaS Small, Hybrid, and Large variants, focusing on their performance in estimating depth, particularly in relation to pedestrian proximity to the camera.

V. RESULTS AND DISCUSSION

A. Running Inference On The Model

First, the MiDaS model's small variant is loaded using PyTorch hub and deployed on the GPU for evaluation. Then, MiDaS transforms are loaded, and an input image taken from the JAAD dataset [15] is read using cv2.imread().

After conversion to RGB and resizing, the image is processed through the MiDaS pipeline on the GPU. Finally, depth predictions are generated, Fig. 3, which are upsampled, and saved in a .csv file [16].

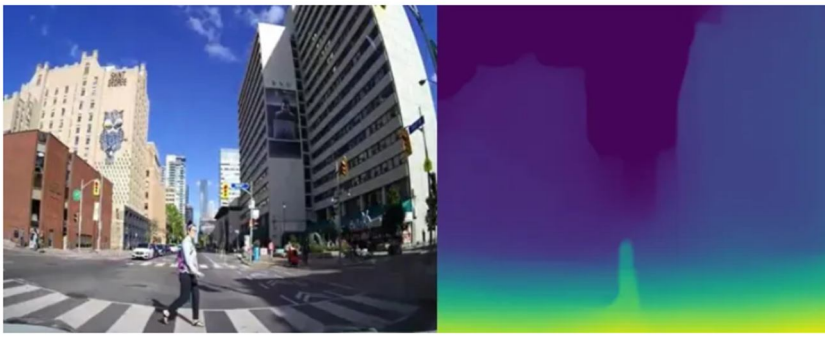


Fig. 3 The depth map of the image produced after MiDaS pipeline.

Then YOLOv7-pose model is then employed for detecting the object and estimating its pose on the same input image. By specifying the model file path, loading it with "loading_yolov7_model()" function, and confirming successful loading, inference is proceeded with by "run_inference()", which processes the image, yielding keypoint coordinates [16]. Then, "draw_keypoints()" from YOLO's utils code visually annotates keypoints on the image, taking coordinates and thresholds as input as shown in the Fig. 4.

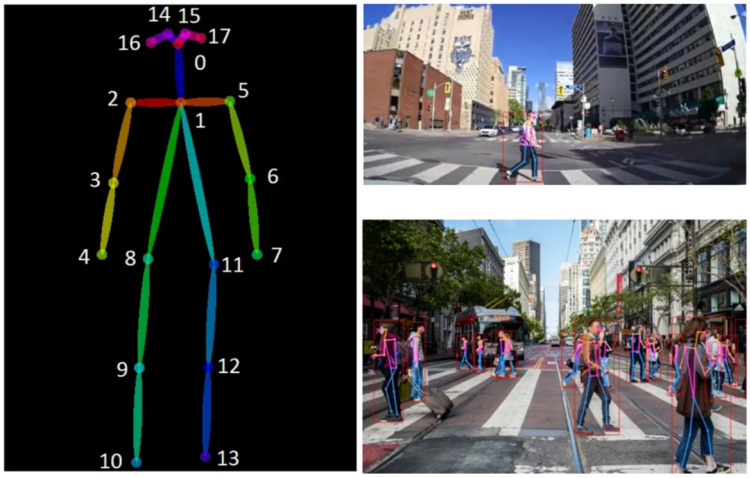


Fig. 4 Annotated Keypoints on the Test Image using YOLOv7-pose

Finally, MiDaS inference is run on the annotated images containing bounding boxes. To finally perform pedestrian proximity classification, two classes "near" and "far" are created. The classification is determined by an application-specific inverse depth threshold, e.g., 0.5. Individuals which have a value of less than 0.5 for inverse depth are classified as "near," while those exceeding this value are designated as "far." These classifications are visually represented in the obtained results as shown in the Fig. 5.

Fig. 5 Classification of the detected pedestrians as “near” or “far”

When assessing the aforementioned pipeline for video and webcam streaming, it is noted that the system performs admirably at lower frame rates but may exhibit diminished accuracy and performance at higher frame rates.

B. Qualitative Comparison of the available MiDaS Models:

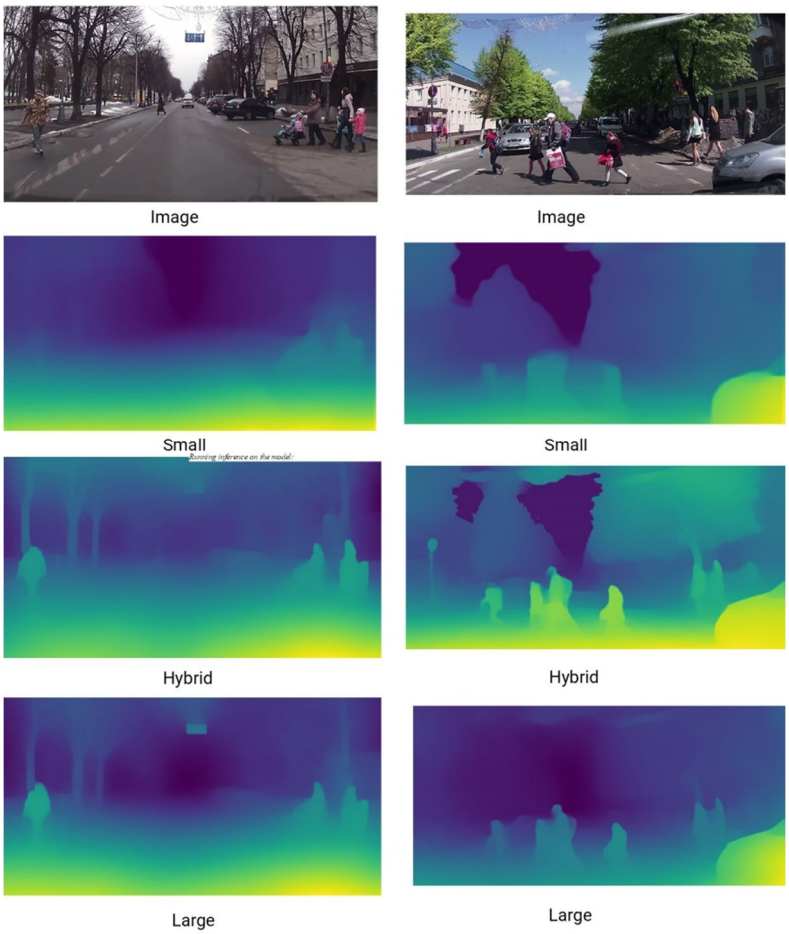


Fig. 6 Comparison of the available MiDaS Models.

The distinction in the level of detail resolution among the three models, MiDaS Small, Hybrid, and Large, is readily apparent. However, this advantage is accompanied by trade-offs, specifically, slower inference speeds and larger model sizes (Small: 82MB, Hybrid: 470MB, Large: 1.2GB).

VI. CONCLUSION AND FUTURE PROSPECTS

A. Conclusion

In summary, this study provides a comprehensive examination of the fusion between YOLOv7-pose and Intel MiDaS, offering real-time capabilities for object detection, pose estimation, and depth analysis. This integration of state-of-the-art models holds immense potential across numerous domains, especially in the realm of guaranteeing safety within autonomous vehicles and Advanced Driving Assistance Systems (ADAS). By harnessing the strengths of YOLOv7-pose for object identification and pose estimation, coupled with MiDaS for monocular depth estimation, the study investigates the prospect of accurately gauging the proximity of objects and pedestrians.

The assessment underscores the remarkable swiftness and precision achieved by YOLOv7-pose, while also delineating the differing proficiencies of MiDaS in its Small, Hybrid, and Large versions when it comes to depth estimation.

These findings underscore the practicality of this amalgamated framework, demonstrating its applicability in fields like surveillance, robotics, and notably, autonomous driving systems.

It's noteworthy that the system's performance may exhibit variations across diverse scenarios and frame rates, indicating promising avenues for further research and enhancement, particularly when dealing with challenging conditions.

B. Future Prospects

This research unlocks several prospects for forthcoming exploration:

1) Ongoing endeavours can be channelled into refining the integrated framework to elevate its real-time functionality and precision. This entails the fine-tuning of model parameters and the exploration of hardware acceleration alternatives.

2) Subsequent investigations can concentrate on techniques for augmenting data to bolster the system's robustness under a wide array of environmental circumstances, encompassing low-light situations, unfavourable weather conditions, and diverse terrains.

3) The integration of semantic segmentation techniques could empower the system not only to detect objects but also to comprehend their attributes. This is particularly valuable in intricate urban settings.

4) The execution of field trials and real-world deployments is imperative to validate the system's performance in practical settings, further refining its precision and dependability.

In conclusion, the fusion of YOLOv7-pose and Intel MiDaS marks a significant stride in the domain of computer vision, with far-reaching implications spanning diverse industries. Subsequent research will inevitably build upon this foundation, propelling innovations that enhance safety, efficiency, and reliability within autonomous systems.

REFERENCES

[1] A. Tupper and R. Green, “Pedestrian Proximity Detection using RGB-D Data,” in 2019 International Conference on Image and Vision Computing New Zealand (IVCNZ), 2019, pp. 1–6. doi: 10.1109/IVCNZ48456.2019.8961013.

[2] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, “Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 3, pp. 1623–1637, Mar. 2022, doi: 10.1109/TPAMI.2020.3019967.

[3] J. Terven and D. Cordova-Esparza, “A Comprehensive Review of YOLO: From YOLOv1 and Beyond.” arXiv, Aug. 07, 2023. doi: 10.48550/arXiv.2304.00501.

[4] W. Lan, J. Dang, Y. Wang, and S. Wang, “Pedestrian Detection Based on YOLO Network Model,” in 2018 IEEE International Conference on Mechatronics and Automation (ICMA), IEEE, Aug. 2018, pp. 1547–1551. doi: 10.1109/ICMA.2018.8484698.

[5] W.-Y. Hsu and W.-Y. Lin, “Adaptive Fusion of Multi-Scale YOLO for Pedestrian Detection,” IEEE Access, vol. 9, pp. 110063–110073, 2021, doi: 10.1109/ACCESS.2021.3102600.

[6] 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.” arXiv, Jul. 06, 2022. doi: 10.48550/arXiv.2207.02696.

[7] O. Sener and V. Koltun, “Multi-Task Learning as Multi-Objective Optimization,” in Neural Information Processing Systems, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:52957972

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91.

[9] G. Maindola, “A Brief History of YOLO Object Detection Models From YOLOv1 to YOLOv5,” MLK - Machine Learning Knowledge, Aug. 26, 2021. https://machinelearningknowledge.ai/a-brief-history-of-yolo-object-detection-models/ (accessed Sep. 25, 2023).

[10] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection.” arXiv, Apr. 22, 2020. doi: 10.48550/arXiv.2004.10934.

[11] X. Wang et al., “A Lightweight Traffic Lights Detection and Recognition Method for Mobile Platform,” Drones, vol. 7, no. 5, p. 293, Apr. 2023, doi: 10.3390/drones7050293.

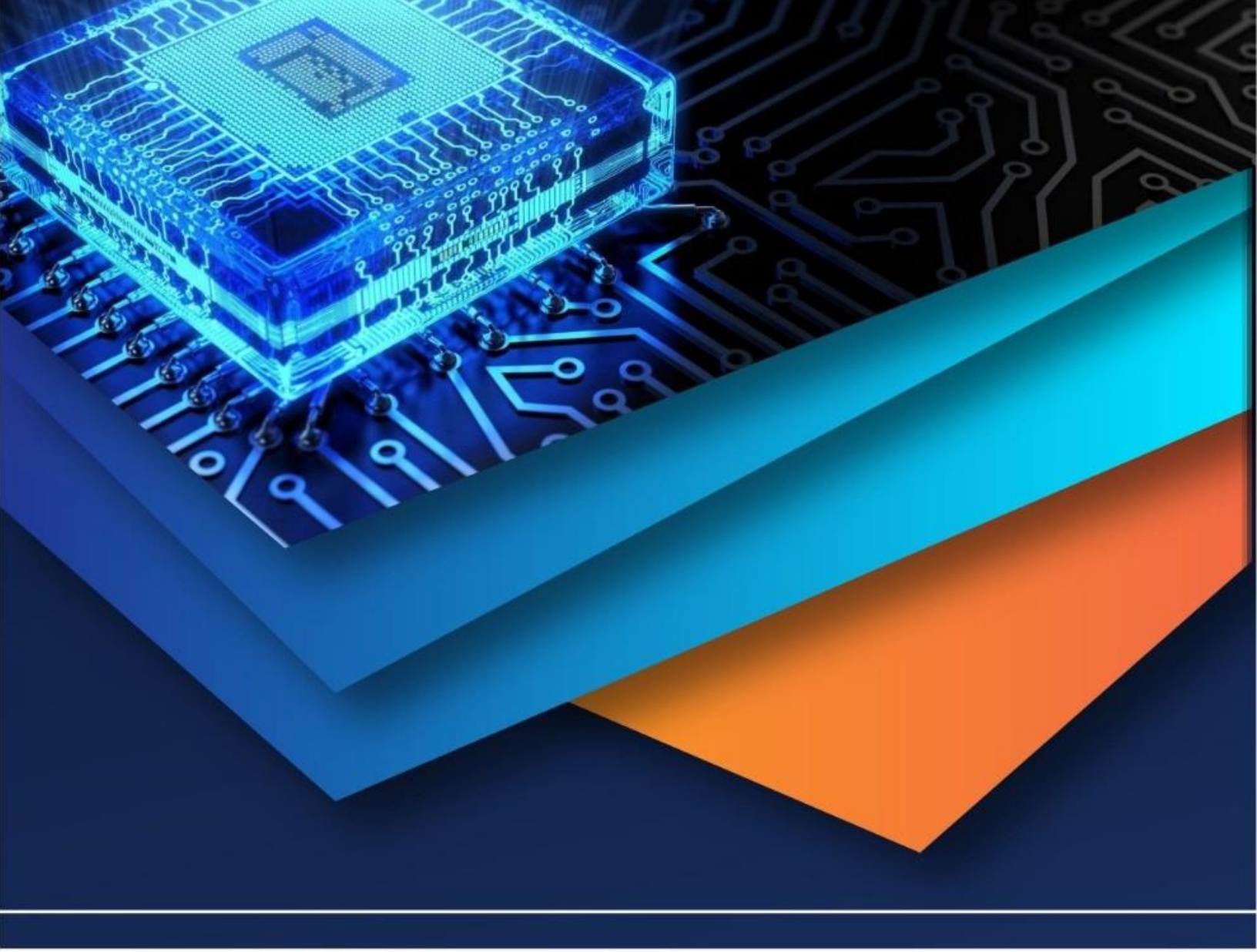
[12] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, “Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression,” Proc. AAAI Conf. Artif. Intell., vol. 34, no. 07, pp. 12993–13000, Apr. 2020, doi: 10.1609/aaai.v34i07.6999.

[13] C. Li et al., “YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications.” arXiv, Sep. 07, 2022. doi: 10.48550/arXiv.2209.02976.

[14] R. Ranftl, A. Bochkovskiy, and V. Koltun, “Vision Transformers for Dense Prediction.” arXiv, Mar. 24, 2021. doi: 10.48550/arXiv.2103.13413.

[15] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, “Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior,” in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 206–213.

[16] “Estimating pedestrian proximity using MiDaS and YOLOv7 | by Jaimin-k | Medium.” https://medium.com/@jaimin-k/estimating-pedestrian-proximity-using-midas-and-yolov7-14b5a08b8740 (accessed Sep. 25, 2023).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY