High Precision Single Shot Object Detection in Automotive Scenarios

Soumya A[1], C Krishna Mohan[1] and Linga Reddy Cenkeramaddi[2]
[1]*Department of Computer Science and Engineering, Indian Institute of Technology, Hyderabad, India*
[2]*Department of Information and Communication Technology, University of Agder, Grimstad, 4879, Norway*

Keywords: Deep Learning, Convolutional Neural Network, Object Detection, Multi-Class Classification, Computer Vision.

Abstract: Object detection in low-light scenarios is a challenging task with numerous real-world applications, ranging from surveillance and autonomous vehicles to augmented reality. However, due to reduced visibility and limited information in the image data, carrying out object detection in low-lighting settings brings distinct challenges. This paper introduces a novel object detection model designed to excel in low-light imaging conditions, prioritizing inference speed and accuracy. The model leverages advanced deep-learning techniques and is optimized for efficient inference on resource-constrained devices. The inclusion of cross-stage partial (CSP) connections is key to its effectiveness, which maintains low computational complexity, resulting in minimal training time. This model adapts seamlessly to low-light conditions through specialized feature extraction modules, making it a valuable resource in challenging visual environments.

1 INTRODUCTION

Object detection using deep learning is a fundamental task in the realm of computer vision that involves identifying and localizing objects of interest within an image or video. Object detection holds a crucial role in computer vision systems, finding applications across various fields such as video surveillance (Gajjar et al., 2017), medical imaging (Adel et al., 2010), (Li et al., 2019b), autonomous driving (Li et al., 2019a), and robot navigation (Truong et al., 2015), (Karaoguz and Jensfelt, 2019). The advent of deep learning, particularly convolutional neural networks (CNNs), has led to significant advancements in the accuracy and efficiency of object detection. This literature review explores the key contributions and trends in object detection using deep learning techniques.

Two-stage detectors and one-stage detectors represent distinct methods for object detection. Two-stage detectors, such as Faster R-CNN, employ a two-step process for object detection. In the first stage, they generate a set of region proposals using a region proposal network (RPN). Region proposals are refined and classified in the second stage to obtain the final detections. This two-stage architecture provides more accurate object localization and is well-suited for complex scenes and small objects, but it has a very high inference time and is computationally expensive. In comparison, single-stage object detectors perform region proposal and object detection in a single pass through the network. The (You Only Look Once) YOLO (Redmon et al., 2016) introduced a single-stage end-to-end object detection approach. It makes predictions for bounding boxes and class probabilities in a single pass by analyzing the entire image once. YOLO achieved real-time inference speed and demonstrated competitive accuracy. Subsequent versions, such as YOLO v3 (Redmon and Farhadi, 2018), YOLO v4 (Bochkovskiy et al., 2020), and YOLO v6 (Li et al., 2022), further improved accuracy and extended the model's capabilities. One-stage detectors are faster than two-stage detectors but relatively less accurate.

The proposed model is better than the other advanced single-shot detectors, leveraging state-of-the-art methods to enhance precision while reducing computational complexity. Its architecture, comprising a backbone, neck, and head, is designed for efficiency and effectiveness. The backbone, with its lower computational demands and cross-stage partial (CSP) connections, ensures smoother gradient flow. The neck excels at integrating features across diverse scales, facilitating semantic and spatial information sharing. Meanwhile, the head streamlines the prediction of classifications and bounding box coordinates.

A, S., Krishna Mohan, C. and Cenkeramaddi, L.
High Precision Single Shot Object Detection in Automotive Scenarios.
DOI: 10.5220/0012383100003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 2: VISAPP*, pages 604-611
ISBN: 978-989-758-679-8; ISSN: 2184-4321

A key advantage lies in adopting a state-of-the-art loss function from the literature, which accounts for bounding box overlap and size similarity, resulting in faster convergence and superior accuracy.

Our research paper introduces a novel architecture for object detection, meticulously designed to optimize efficiency and effectiveness, with the following key contributions:

- A carefully crafted backbone network, drawing inspiration from Inception ResNetV2 and incorporating CSP connections for superior performance in image-related tasks.
- A multi-block approach features three distinct block types (A, B, and C), each tailored to extract features at different resolutions, enabling effective object detection across various scales and complexities.
- Including cross-stage partial (CSP) connections in all block types ensures smooth gradient flow during training, improving convergence and model performance.
- Multi-scale object detection capability allows our architecture to adapt to objects of varying sizes and spatial distributions dynamically.

2 RELATED WORKS

This section summarizes the recent advancements in object detection. Several works were proposed for object detection, and the effectiveness of convolutional neural networks (CNN) classifiers has been shown in (Coman et al., 2018) to outperform traditional machine learning techniques focused on feature extraction. In the context of object detection, the faster region-based convolutional neural network (Faster-RCNN) with InceptionV2 architecture is utilized in (Galvez et al., 2018a) to identify five individuals and one quadrotor within the given image. In (Galvez et al., 2018b), the authors have presented a low-shot transfer detector using a deep architecture and a controlled transfer learning framework to address the challenges of limited training data in object detection. An object detection approach was introduced in (Xu et al., 2018) with a region selection network for selecting regions from which to consider features and a gatting network to transform the feature maps. A novel object detection model in (Zeng et al., 2013) is designed to train multi-stage classifiers. In (Liu et al., 2016), a single-stage detector (SSD) has been designed, incorporating convolutional outputs for bounding boxes connected to several feature maps within the network. Enhancing small object detection through contextual information fusion within the faster R-CNN framework is presented in (Fang and Shi, 2018). In (Beery et al., 2020) Context, R-CNN presented with the attention to access a camera-specific memory bank and improve object detection by incorporating contextual information from previous frames.

3 PROPOSED METHOD

The proposed architecture, comprising a backbone, neck, and head, is carefully designed to optimize the efficiency and effectiveness of object detection.

3.1 Backbone Network

The proposed architecture's backbone draws inspiration from the highly effective Inception ResNetV2 model while incorporating cross-stage partial (CSP) connections, renowned for its exceptional performance in image-related tasks. The architectural design in Figure 1 consists of a meticulously designed backbone that is crucial to the entire model. At its core, the backbone comprises several important elements, each with a specific purpose. It all begins with the stem, which serves as the initial feature extractor. Strategically, it reduces the spatial dimensions of the input image by a factor of 8. This dimension reduction proves instrumental in capturing essential features while efficiently processing the input data. Moving forward, Block A takes center stage, featuring an inception module with shortcut connections. What sets Block A apart is the incorporation of CSP connections, which involve the deliberate splitting of feature maps.

This architecture includes ten Block A units, excelling at extracting high-resolution features from the input, which is crucial for subsequent stages. Following Block A, the reduction Block A comes into play, effectively reducing the spatial resolution by a stride of 16. This strategic reduction enhances the receptive field, enabling more comprehensive feature analysis in subsequent stages.

Block B shares the idea of Block A but focuses on mid-resolution feature maps. A total of 20 Block B units contribute to extracting vital mid-level features. Subsequently, reduction Block B follows suit, reducing spatial resolution with a larger stride of 32. This strategic choice enables the model to detect objects of varying sizes and scales efficiently. Block C emerges as a critical component, specializing in refining low-resolution features, ultimately optimizing the model for detecting objects with fine details and spatial complexity.

All three block types (A, B, and C) feature CSP connections, ensuring smooth gradient flow during training and facilitating improved convergence. The architecture captures outputs at three distinct scales after blocks A, B, and C to enable multi-scale object detection. This enables the model to adapt dynamically to diverse object sizes and spatial distributions.

3.2 Neck and Decoupled Head

The neck component in Figure 2 processes feature maps from different scales, effectively combining spatial richness and semantic enrichment. The fusion of these features, through up-sampling and down-sampling, ensures the sharing of crucial semantic information and spatial resolution across all three scales, enhancing the model's comprehensiveness and robustness in object detection. Following the neck component, the decoupled head, featuring dedicated convolutional layers, takes center stage. It is designed explicitly to predict classification scores and bounding box coordinates for each of the three scales. This architecture employs three separate decoupled heads, one for each scale, ensuring precise object detection and accurate localization. This holistic design showcases a well-coordinated flow that optimizes the entire model's ability to detect and identify objects effectively across multiple scales and complexities.

3.3 Training

The proposed object detection model was employed with the optimization algorithm, stochastic gradient descent (SGD), with a learning rate of 0.01. The model was trained with 11 epochs, each processing batches of size 32. To aid in convergence and optimization, the SGD optimizer was configured with a momentum of 0.9 and a weight decay of 0.0005, which helps control the magnitude of weight updates during training.

Mixed precision training was adopted to enhance the training process further and accelerate computations. This method optimizes memory usage and speeds up training by reducing computational costs while maintaining adequate numerical precision.

Using mixed precision training allowed for faster convergence while maintaining the model's performance quality. The chosen hyperparameters, including the learning rate and batch size, were selected to balance the trade-off between model convergence and optimizing computational efficiency. The model follows an anchor point-based approach and incorporates task-assignment learning. Anchor points are predefined reference points used during the object detection process to facilitate the efficient localization of objects. Task assignment learning optimizes the assignment of bounding boxes to anchor points, further enhancing the model's overall performance.

4 DATASET AND IMAGE COLLECTION

The image collection dataset available in IEEE Dataport (Gao et al., 2022) is considered for the automotive object detection scenario. This dataset comprises camera images corresponding to six classes with varied dimensions. The dataset contains 19,740 images and labels. We randomly selected 15,777 images for training and 1800 images for validation and testing. The camera image of size $1440 \times 1080 \times 3$ pixels is resized to $416 \times 416 \times 3$. There may be one or more objects in one image, so the location of each object is pre-annotated. All the objects in the dataset are categorized into six distinct groups: person, car, cyclist, bus, truck, and motorbike. Although the dataset's author mentions six classes, there seem to be only four classes (pedestrian, bicycle, car, and truck). Therefore, the dataset is highly imbalanced.

Annotating Dataset: Since the dataset consists of a few inconsistent labels, the necessity for annotation arises. Instead of manual annotation, we employed a pre-trained Faster RCNN for object detection, and subsequently, we recorded the associated bounding box coordinates, storing them in a CSV file. This method allowed us to annotate the entire dataset seamlessly.

5 EVALUATION OF THE STATE-OF-THE-ART CNNs

In this section, we perform a comprehensive assessment of various deep-learning benchmark models using the Automotive dataset (Gao et al., 2022). We evaluate the YOLOv5n, YOLOv6n, YOLOv8n, and RT-DETR models, all of which are designed for single-stage object detection. The models are trained to predict bounding boxes and class probabilities directly from the entire image, enabling real-time detection. It's worth noting that our evaluation identified specific dataset-related challenges. Notably, the dataset contains four unique classes, but the mAP calculation was conducted for six classes, which can potentially lead to inaccuracies. To provide a more

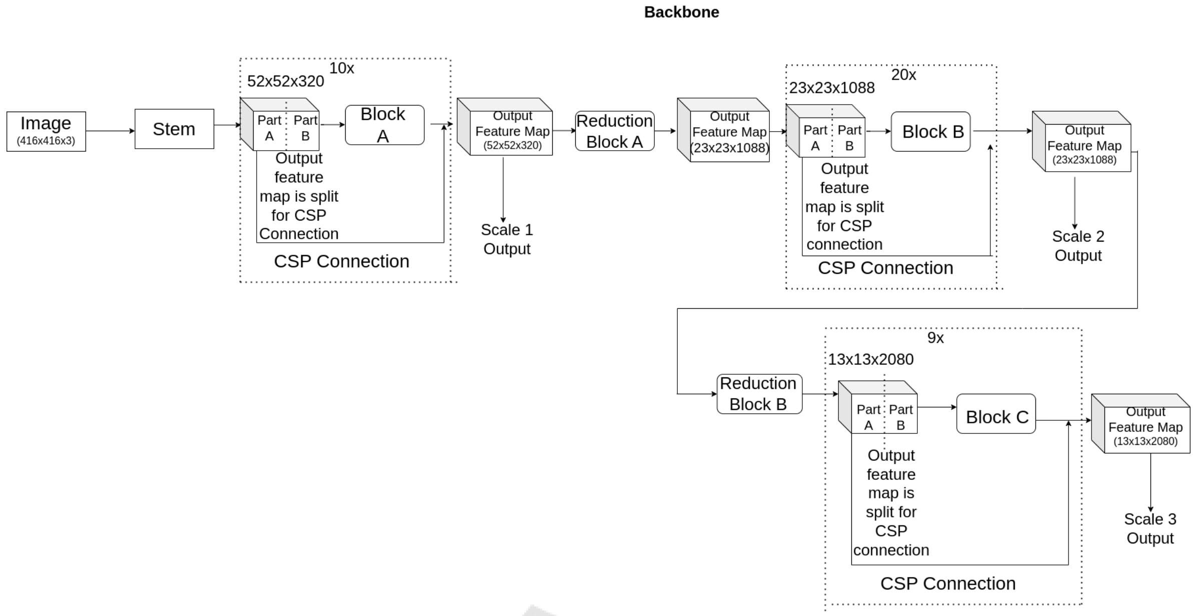

Figure 1: Proposed Architecture Backbone.

comprehensive understanding of our model's effectiveness, especially for individual classes, we present the evaluation results in Table 1, including precision, recall, mAP50, and mAP50-95 metrics.

5.1 Evaluation Metrics

Proposed object detection model performance is typically assessed by measuring its accuracy using metrics such as Average Precision (AP) or Mean Average Precision (mAP). These metrics involve calculating the average of AP scores across all object classes.

We employed the mean average precision (mAP) as the evaluation metric to assess the performance of the proposed object detection model. Mean average precision (mAP) extends AP for multi-class or multi-label scenarios commonly found in object detection. AP is computed for each class or label, and then mAP is calculated by averaging these AP values. In object detection, for instance, we calculate AP for each object class to measure how well the model identifies objects of that class. mAP then offers an overall performance score, considering the precision-recall performance across all object classes. Like AP, higher mAP signifies better detection accuracy across different classes. A high mAP means a model has a low false negative and a low false positive rate. We provide a detailed breakdown of our model's performance shown in Table 2, which is an essential reference point critical for a deeper understanding of its effectiveness across various object classes and detection scenarios.

To train the model, we utilize two distinct loss functions:

Classification Loss- Varifocal loss (Zhang et al., 2021): The varifocal loss shown in Eq. 1 is utilized as the classification loss function. This loss function effectively tackles the issue of class imbalance between positive and negative samples, thereby enhancing the classification performance.

$$VFL(p,q) = \begin{cases} -q(qlog(p) + (1-q)log(1-p) & \text{if } q > 0 \\ -q\alpha^{\gamma} plog(1-p) & \text{if } q = 0 \end{cases} \quad (1)$$

where p is the predicted IoU-aware classification score, and q is the target score. For a foreground point belonging to its respective ground-truth class, the target score q is determined as the intersection over union (IoU) between the generated bounding box and its associated ground truth. If the point does not belong to its ground-truth class, the target score is set to 0.

Bounding box loss - Complete IoU loss (CIOU) (Zheng et al., 2020): The complete IoU (CIOU) loss is employed for the bounding box regression task. CIOU considers both box overlaps and size similarity, leading to more accurate bounding box predictions essential for precise object localization.

5.2 Evaluation Results

Our model's performance evaluation was conducted on the Automotive dataset, comprising 15,777 images for training, 1,800 for validation, and 1,800 testing

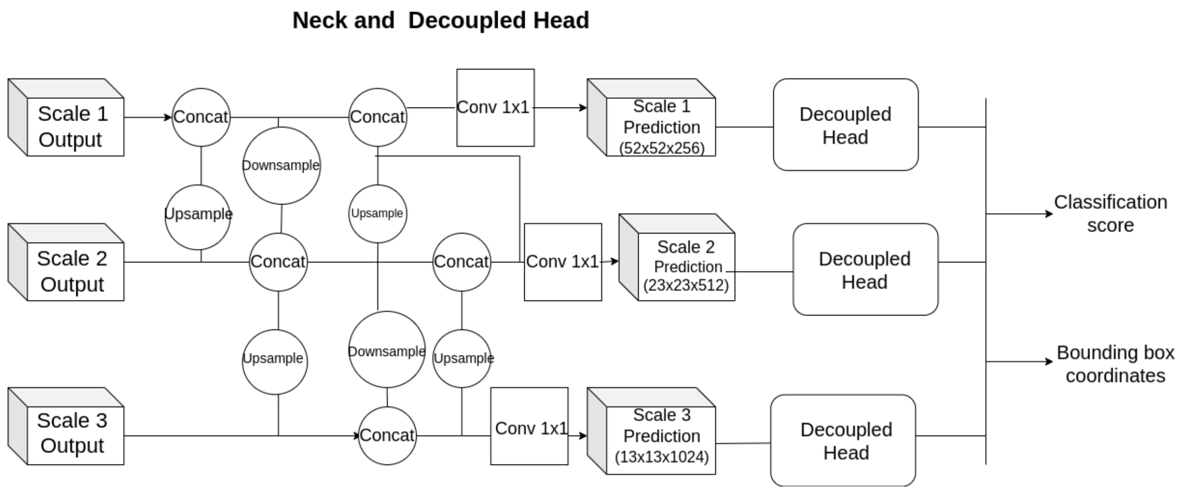

Figure 2: Proposed Architecture Neck and Decoupled head.

Table 1: Performance for all the state-of-the-art models.

Model	Precision	Recall	mAP50	mAP50-95
YOLO v8n (yolov8, 2023)	0.811	0.831	0.83	0.68
YOLO v6n (Li et al., 2022)	0.792	0.732	0.773	0.575
YOLO v5n (yolov5, 2023)	0.811	0.831	0.83	0.68
RT-DETR (Lv et al., 2023)	0.863	0.897	0.915	0.781
Proposed SSD Model	0.859	0.408	0.406	0.257

Table 2: Class-wise performance evaluation table for the proposed single shot detection model.

Class	Images	Precision	Recall	F1-Score	mAP-50	mAP50-95
Person	2378	0.94	0.45	0.6	0.45	0.30
Car	2378	0.85	0.4	0.544	0.4	0.28
Bicycle	2378	0.827	0.38	0.53	0.39	0.26
Truck	2378	0.85	0.44	0.56	0.4	0.25

images, providing a diverse representation of real-world scenarios. Object detection metrics were employed to gauge our model's efficacy, including the mean average precision (mAP), precision, recall, and F1 score. It is worth noting that the dataset presents certain challenges, primarily stemming from an imbalanced class distribution and limited samples available for classes such as motorbike and bus, which can introduce bias into the model's performance evaluation. As a result, the model might exhibit relatively strong performance for classes with larger sample sizes while encountering challenges in accurately detecting and classifying instances of motorbikes and buses. Upon testing, it was observed that the dataset contains only four unique classes, but the mAP calculation was conducted for six classes. This discrepancy in class count could lead to inaccuracies and misleading results during evaluation. Consequently, the mAP, though a widely used metric, might not accurately depict the full extent of our model's accuracy. To address this limitation, we offer a comprehensive breakdown of performance metrics for each class. By highlighting precision, recall, and F1 scores for every category, we shed light on our model's specific strengths and weaknesses. The resultant confusion matrix is shown in Figure 3, and the precision-recall (PR) curve is shown in Figure 4. Sample inferences obtained from distant objects are shown in Figure 5, from crowded areas are shown in Figure 6, and from shadows and low light conditions are shown in Figure 7. The model presents visual insights through sample inferences, showcasing our model's robust performance in complex real-world scenarios. Despite the challenges posed by the dataset, our model's adaptability and resilience, particularly in low-light conditions, make it a promising solution for a wide range of practical object detection tasks. It is important to mention that the proposed method is trained on unbalanced datasets, which is also an influential factor in the real-time performance of object detection tasks.

A confusion matrix is depicted with a total of seven classes, and we identified that the dataset is im-

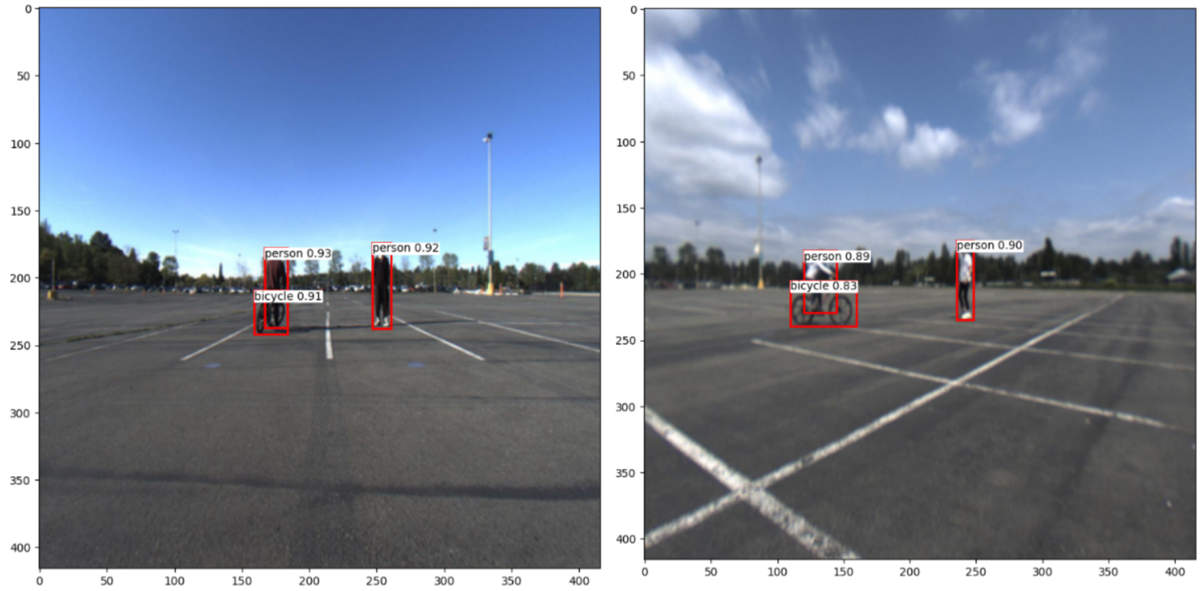

Figure 5: Sample Inferences: Predictions of distant objects.

balanced and does not have any samples of motorbikes, and the bus class is not included in our annotations generated with faster RCNN as it contains very few samples in the dataset. This lack of samples is causing the variation in the resulting matrix presented in Figure 3 and the results presented in Table 4. Also, the test set we used for the model does not include the truck, resulting in the PR curve in Figure 4 being plotted with only 3 classes. The proposed model was tailored to improve inference speed on resource-constrained devices by incorporating CSP connections. This strategic integration significantly enhances the model's ability to perform rapid inferences.

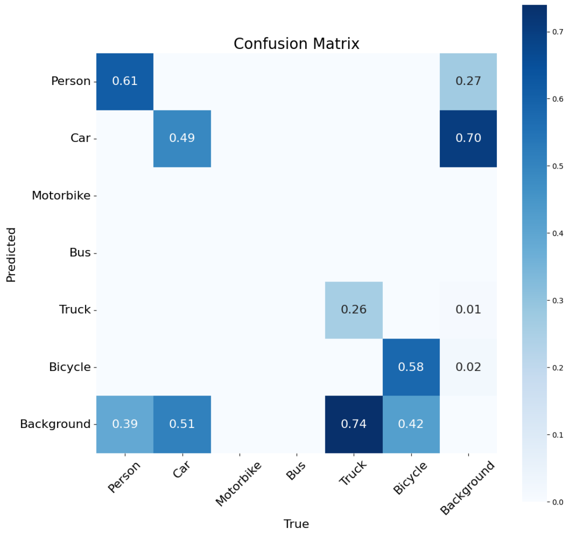

Figure 3: Confusion matrix of the proposed model with 7x7 class accuracies.

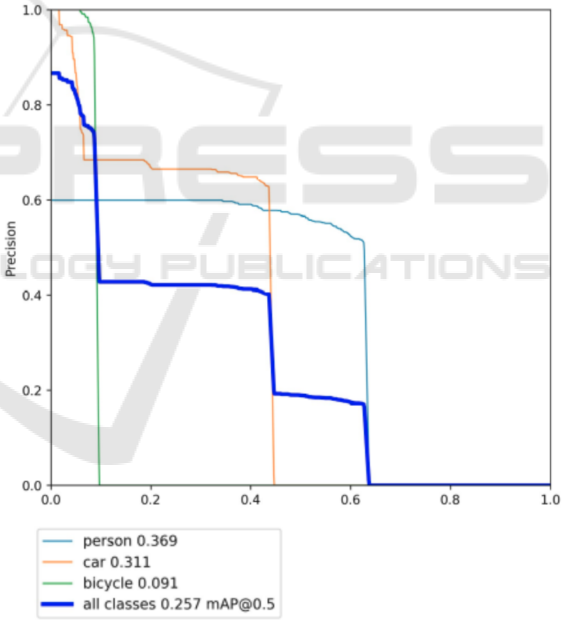

Figure 4: Precision-Recall Curve.

6 CONCLUSION

Leveraging convolutional neural networks (CNNs), the proposed high-precision single-shot object detection model excels at optimizing precision and computational efficiency.

The proposed model demonstrates adaptability by detecting objects across various scales and sizes, paving the way for practical implementation. We

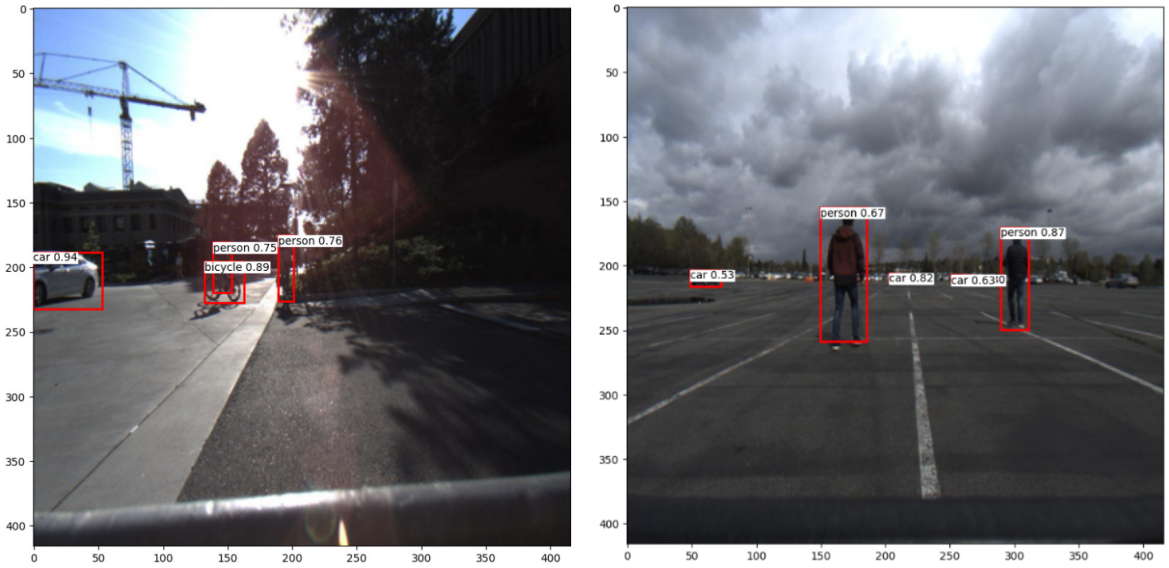

Figure 6: Sample Inferences: Predictions at crowded areas and groups of objects.

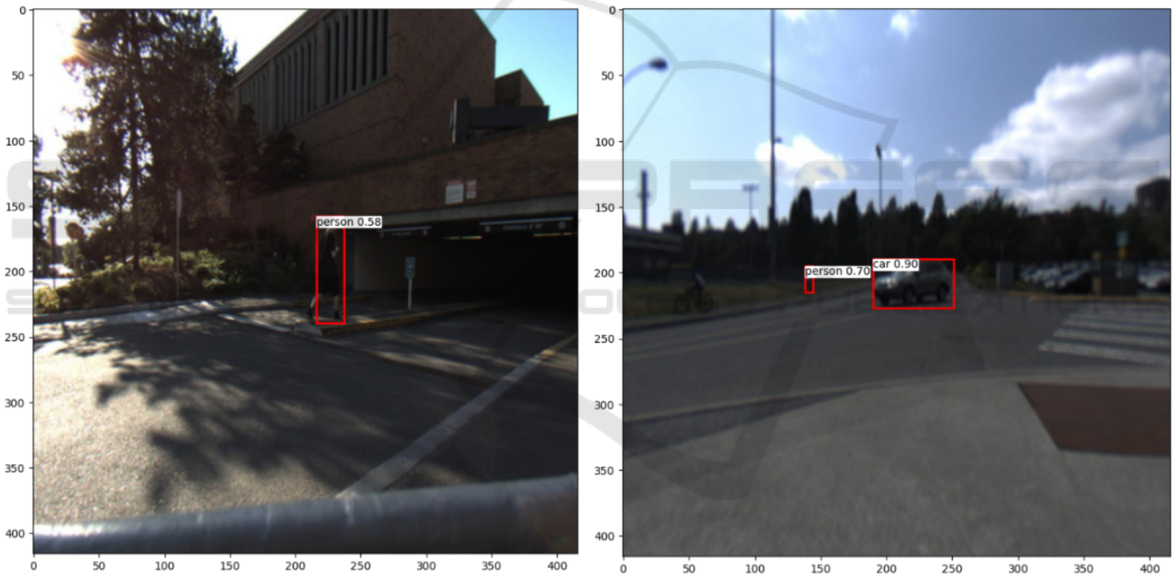

Figure 7: Sample Inferences: Predictions shadows and low light conditions.

evaluated prominent benchmark models, including YOLOv5n, YOLOv6n, YOLOv8n, and RT-DETR models. Notably, the proposed approach effectively addresses challenges inherent to the dataset, such as class discrepancies, imbalanced data distribution, and the impact of low lighting conditions, ensuring robust object detection even in less-than-ideal visibility scenarios.

The core strength of our object detection model lies in its sophisticated architecture. The seamless coordination among the backbone, neck, and decoupled head components enables the detection of objects in diverse and complex scenarios. The proposed model was optimized for efficient resource-constrained device inference, ensuring shorter training times by incorporating CSP (cross-stage partial) connections. Integrating advanced loss functions like varifocal loss and complete IoU loss for classification and bounding box regression further enhances the model's accuracy and robustness.

ACKNOWLEDGMENT

This work was supported by the Indo-Norwegian Collaboration in Autonomous Cyber-Physical Systems (INCAPS) project: 287918 of the International Partnerships for Excellent Education, Research and Innovation (INTPART) program and the Low-Altitude UAV Communication and Tracking (LUCAT) project: 280835 of the IKTPLUSS program from the Research Council of Norway.

REFERENCES

Adel, M., Moussaoui, A., Rasigni, M., Bourennane, S., and Hamami, L. (2010). Statistical-based tracking technique for linear structures detection: Application to vessel segmentation in medical images. *IEEE Signal Processing Letters*, 17(6):555–558.

Beery, S., Wu, G., Rathod, V., Votel, R., and Huang, J. (2020). Context r-cnn: Long term temporal context for per-camera object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Coman, C. et al. (2018). A deep learning sar target classification experiment on mstar dataset. In *2018 19th International Radar Symposium (IRS)*. IEEE.

Fang, P. and Shi, Y. (2018). Small object detection using context information fusion in faster r-cnn. In *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, pages 1537–1540. IEEE.

Gajjar, V., Gurnani, A., and Khandhediya, Y. (2017). Human detection and tracking for video surveillance: A cognitive science approach. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*.

Galvez, R. L., Bandala, A. A., Dadios, E. P., Vicerra, R. R. P., and Maningo, J. M. Z. (2018a). Object detection using convolutional neural networks. In *TENCON 2018 - 2018 IEEE Region 10 Conference*.

Galvez, R. L., Bandala, A. A., Dadios, E. P., Vicerra, R. R. P., and Maningo, J. M. Z. (2018b). Object detection using convolutional neural networks. In *TENCON 2018-2018 IEEE Region 10 Conference*, pages 2023–2027. IEEE.

Gao, X., Luo, Y., Xing, G., Roy, S., and Liu, H. (2022). Raw adc data of 77ghz mmwave radar for automotive object detection.

Karaoguz, H. and Jensfelt, P. (2019). Object detection approach for robot grasp detection. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE.

Li, B., Ouyang, W., Sheng, L., Zeng, X., and Wang, X. (2019a). Gs3d: An efficient 3d object detection framework for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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

Li, Z., Dong, M., Wen, S., Hu, X., Zhou, P., and Zeng, Z. (2019b). Clu-cnns: Object detection for medical images. *Neurocomputing*, 350.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer.

Lv, W., Xu, S., Zhao, Y., Wang, G., Wei, J., Cui, C., Du, Y., Dang, Q., and Liu, Y. (2023). Detrs beat yolos on real-time object detection. *arXiv preprint arXiv:2304.08069*.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Truong, X.-T., Yoong, V. N., and Ngo, T.-D. (2015). Rgb-d and laser data fusion-based human detection and tracking for socially aware robot navigation framework. In *2015 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 608–613. IEEE.

Xu, H., Lv, X., Wang, X., Ren, Z., Bodla, N., and Chellappa, R. (2018). Deep regionlets for object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 798–814.

yolov5, u. (2023). ultralytics comprehensive guide. https://docs.ultralytics.com/yolov5/. Accessed: 2023-9-1.

yolov8, u. (2023). comprehensive guide ultralytics. https://docs.ultralytics.com/. Accessed: 2023-9-1.

Zeng, X., Ouyang, W., and Wang, X. (2013). Multi-stage contextual deep learning for pedestrian detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Zhang, H., Wang, Y., Dayoub, F., and Sunderhauf, N. (2021). Varifocalnet: An iou-aware dense object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8514–8523.

Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., and Ren, D. (2020). Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34.