Ms. Rachna Shetty, et. al. International Journal of Engineering Research and Applications
www.ijera.com
ISSN: 2248-9622, Vol. 14, Issue 3, March, 2024, pp: 76-84

RESEARCH ARTICLE OPEN ACCESS

Enhancing Communication and Accessibility: Recognition of Indian Sign Language (ISL) Signs Using Faster R-CNN for (D&M) Inclusivity in India.

Ms. Rachna Shetty*, Prof. Sofia Francis.
Department of Computer Engineering, Mukesh Patel School of Technology, Management & Engineering, SVKM's NMIMS Deemed-to-be University, Mumbai, Maharashtra 400056, India.

Abstract
Purpose: The primary objective of this research paper is to examine the application of Faster R-CNN, a Region-based Convolutional Neural Network, for the identification of Indian Sign Language (ISL) signs. The overarching aim is to evaluate the effectiveness of Faster R-CNN in the domain of ISL sign recognition, taking into account the linguistic diversity within India and the need for inclusive communication. The study aims to contribute to the development of a model capable of discerning the intricacies associated with various dialects of ISL.
Methods: The research employs a comprehensive comparative analysis, focusing on the performance of Faster R-CNN. The investigation encompasses the incorporation of the YOLO V8 model for bounding box detection. This comparison is conducted in relation to other Convolutional Neural Network (CNN) models that serve the same purpose. The study addresses both stationary and dynamic image contexts in order to ensure robust performance in real-world situations encountered by the Deaf and Dumb (D&M) community. The methods involve assessing the ability of the model to effectively identify ISL signs, taking into consideration the diverse linguistic landscape and the subtleties of different dialects.
Results: The study demonstrates the performance of Faster R-CNN in conjunction with YOLO V8 model integration for bounding box detection in comparison to other CNN models used for similar purposes. The results highlight the effectiveness of the proposed model in recognizing ISL signs, emphasising its significance in both stationary and dynamic image contexts. The findings underscore the potential contributions of the model to education, employment, and social interactions for the Indian D&M community.
Conclusion: The importance of inclusive communication in a linguistically diverse environment such as India, combined with the wide range of dialects within ISL, underscores the necessity of designing a model capable of discerning subtleties. The study not only addresses the need for robust performance in real-world situations, but also highlights the broader impact on communication barriers faced by the Indian D&M community. The research advocates for the continued application of the developed model in promoting inclusivity and dismantling communication barriers, aligning with the imperative for enhanced communication in the year 2023.
Keywords - Faster R-CNN, Region-based Convolutional Neural Network, Indian Sign Language (ISL), YOLO V8 model, Bounding box detection, CNN models, Linguistic diversity. Inclusive communication.

Date of Submission: 09-03-2024 Date of acceptance: 23-03-2024

I. Introduction

Sign Language Recognition, an ever-evolving field in the realm of human-computer interaction, is on the cusp of ground-breaking advancements that have the potential to revolutionise communication by providing automated and accessible solutions. Recent advancements in the domains of image processing and artificial intelligence, driven by a range of diverse methodologies, have placed a particular emphasis on tackling the intricacies of Indian Sign Language (ISL), which is characterised by the use of dual-handed gestures. While the regional variations within ISL present their own set of unique challenges, ongoing endeavours such as DeepSign are dedicated to the development of an ISL Recognition Application, with the aim of bridging communication gaps and promoting inclusivity within the multifaceted linguistic landscape of India.

According to Shaoqing Ren, Et. Al., the architectural innovation, known as "Faster Region-Convolutional Neural Network," holds a notable place within the R-CNN family, driven by a distinctive and ambitious objective. In contrast to its predecessors, Faster R-CNN seeks to deliver a comprehensive solution that not only detects objects within an image but also accurately determines the spatial location of these objects. This dual pursuit of velocity and precision relies on an intricate amalgamation of potent concepts within deep learning, leveraging the capabilities of convolutional neural networks (CNNs) and pioneering region proposal networks (RPNs). The harmonious integration of these diverse components within Faster R-CNN engenders a model that not only streamlines the process of object detection but also enhances its overall efficiency, thereby opening up novel possibilities across a broad spectrum of applications in object recognition and localization. By synergistically harnessing the strengths of these fundamental constituents, Faster R-CNN not only signifies a significant advancement in the field of object detection but also establishes itself as a benchmark for subsequent advancements at the intersection of deep learning and computer vision.

In this context, the utilisation of Faster R-CNN for the purpose of sign language recognition proves to be advantageous. As a Region-based Convolutional Neural Network, Faster R-CNN excels in the efficient detection of objects and the precise localization of said objects, which is of paramount importance when dealing with intricate gestures such as those found within ISL. The integration of convolutional neural networks facilitates the extraction of meaningful features, capturing the subtle nuances present in signs, while region proposal networks streamline the identification of relevant regions, thus enhancing accuracy and expediting real-time recognition. This comprehensive solution for complex sign recognition cannot be seamlessly replicated by alternative models, making it an invaluable asset.

II. Related Work

A. Tyagi and S. Bansal, outline a hybrid technique for fast feature extraction in the recognition of Indian Sign Language (ISL). This technique combines the Fast Accelerated Segment Test (FAST), Scale-Invariant Feature Transform (SIFT), and Convolutional Neural Networks (CNN). The proposed model, named FiST_CNN, has been validated for 24 ISL alphabets and 10 digits, each with 200 images for every gesture. The CNN model is composed of seven convolution layers and max-pooling layers, followed by two dense, fully connected layers. The efficacy of the proposed model has been assessed using a confusion matrix. The results demonstrate that this approach achieves high accuracy in the recognition of ISL gestures, while requiring less computation time compared to alternative methods. The article also discusses relevant research, provides a concise overview of FAST, SIFT, and CNN, and elucidates the functioning of the hybrid approach. Finally, the article concludes with a discussion on the effectiveness of the proposed approach and potential future applications. (FAST, SIFT, and CNN)

O. Koller, Et. Al., introduced the Deep Sign methodology, an innovative technique for continuous sign language recognition that combines the merits of Convolutional Neural Networks (CNNs) and Hidden Markov Models (HMMs). Initially, the authors present the integration of a CNN into an HMM, treating the CNN outputs as true Bayesian posteriors and training the system as a hybrid CNN-HMM. Furthermore, the authors provide several other contributions, including a theoretical elucidation of the hybrid approach, an examination of the impact of CNN and HMM structure on the hybrid approach, and experiments that explore the utilization of out-of-domain data and ensembles of hybrid CNN-HMMs. The findings demonstrate that the Deep Sign methodology achieves a significant relative enhancement of more than 15% compared to the current state-of-the-art on three challenging standard benchmark continuous sign language recognition datasets. The authors conclude by discussing the implications of this study for enhancing communication and accessibility for individuals who rely on sign language. (CNN & HMMs)

A. Al-Shaheen, Et. Al., delved into the utilization of the YOLOv4 technique for the purpose of discerning American Sign Language (ASL) gestures and signs. The authors elaborate on the significance of this technology in facilitating communication between individuals who are deaf and mute, and those who possess the ability to hear and speak. They provide comprehensive information regarding the dataset employed for the training and testing of the model, as well as the methodology and outcomes of their experiments. The authors draw the conclusion that the YOLOv4 technique exhibits promise in accurately detecting and recognizing ASL gestures and signs, and propose potential applications for this technology that extend beyond communication with individuals who are deaf and mute.

A. A. Barbhuiya, Et. Al., investigate the utilization of deep learning-based convolutional neural networks to effectively represent and model stationary signs within the context of sign language recognition. The authors introduce an innovative technique for extracting distinctive features and categorizing American Sign Language (ASL) signs by utilising modified versions of the AlexNet and VGG16 models. The ASL Fingerspelling Dataset, comprising 870 images of 29 distinct signs, is employed for both training and evaluating the models. The experimental outcomes demonstrate that the proposed approach achieves a notable level of accuracy in recognizing signs, surpassing other advanced methods currently in use. The paper concludes by discussing the limitations inherent in the proposed approach and offering suggestions for future research endeavours.

Model Architecture

Faster RCNN:

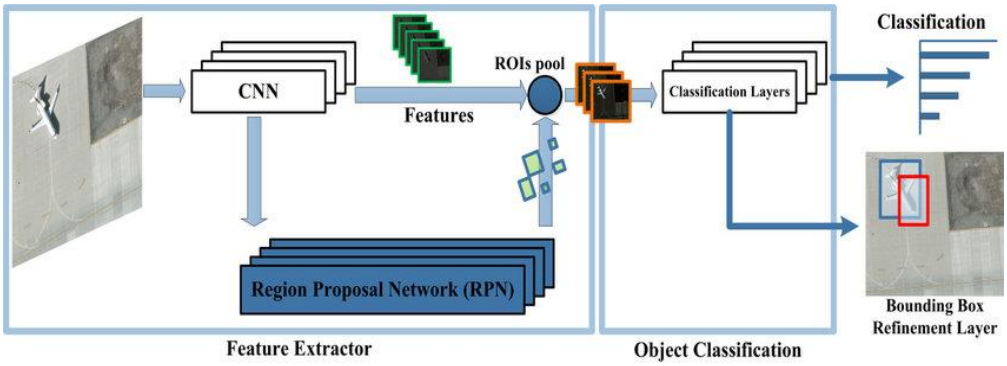

Fig. 1: Model Architecture. Source: Adapted from Researchgate.net.

The Faster R-CNN methodology consists of a multitude of components that collaborate harmoniously to successfully execute the complex task of object detection and recognition. The initiation of this process commences with the Feature Extraction component, where a Convolutional Neural Network (CNN) meticulously analyses the input image to extract vital visual features, such as edges, textures, and patterns. By generating feature maps that emphasise various aspects of the image, the CNN is able to discern and comprehend intricate patterns and shapes. The Region Proposal Network (RPN) diligently identifies potential object locations within the image by leveraging the extracted visual features. This network proposes multiple bounding boxes as plausible candidates for objects. The Region of Interest Pooling (ROIs Pool) component expertly refined and standardised these proposals, ensuring that all subsequent stages possess uniform spatial dimensions. Consequently, the unique features pertaining to the proposed regions are meticulously extracted, thus providing a rich and detailed understanding of the objects residing within said regions. The Classification Layer effectively determines the presence of objects within the proposed regions by skilfully classifying extracted features into distinct categories. Simultaneously, the Bounding Box Refinement Layer adeptly adjusts the initial bounding boxes to achieve enhanced precision in localization. This comprehensive process, encompassing feature extraction, region proposal, refinement, classification, and bounding box adjustment, is of utmost importance for tasks such as image recognition and object detection in the realm of computer vision.

YOLOv8:

Despite not having a formal document, the YOLOv8 framework introduces significant advancements that enhance the effectiveness of object detection. One notable innovation is the transition to anchor-free detection, which directly predicts the centres of objects instead of the offsets from anchor boxes. This simplifies predictions and streamlines post-processing steps such as Non-Maximum Suppression, ultimately leading to improved accuracy and faster processing by reducing the number of box predictions. In terms of structure, YOLOv8 modifies the convolutional operations by replacing the initial 6x6 convolution in the stem with a 3x3 convolution and refining the building blocks, including the introduction of C2f. The Bottleneck structure, which is similar to the 2015 ResNet block but with a 3x3 kernel for the first convolution, simplifies the architecture by enabling direct feature concatenation, reducing the number of parameters and the overall size of the tensor. The training routine incorporates mosaic

augmentation, which combines four images to expose the model to diverse scenarios, thus enhancing its ability to generalise across complex scenes. The research emphasises a nuanced approach to augmentation and advises against indiscriminate use of mosaic augmentation throughout training to avoid potential performance degradation. This meticulous consideration of architectural enhancements and training strategies establishes YOLOv8 as a sophisticated and versatile solution for object detection tasks.

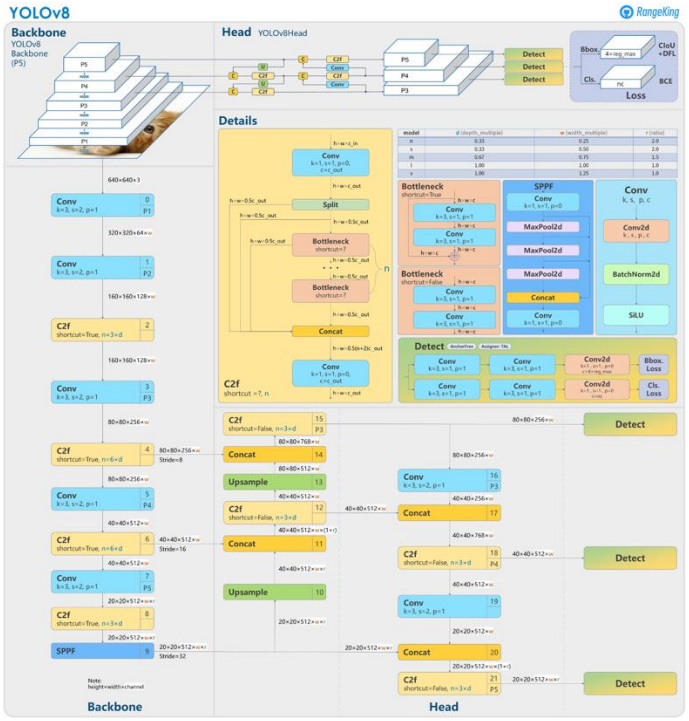

Fig. 2: YOLOv8 Architecture. Source: Adapted from GitHub user RangeKing

Dataset

The Indian Sign Language Image Dataset is a comprehensive collection designed for sign language recognition, specifically focusing on the Indian Sign Language (ISL). With 10,067 images spanning 35 classes, including the English alphabet and numeric digits, the dataset is divided into training (72%), validation (15%), and testing (13%) subsets. Pre-processing involves auto-orientation and resizing to a uniform 640 x 640 pixel resolution, without augmentation. Accompanying XML files detail image information such as class, path, bounding box coordinates, and size. Annotations, crucial for bounding box regression in object detection and localization tasks, predict and enhance bounding box coordinates, capturing sign language gestures. This dataset is a valuable resource for researchers and developers interested in sign language recognition, offering extensive coverage of alphabet, digits, and supplementary characters for diverse applications. Meticulous division and consistent pre-processing ensure the dataset's reliability, making it suitable for developing and evaluating machine learning models. Researchers and practitioners can utilise this dataset to construct and assess models accurately interpreting sign language gestures.

Implementation

In our investigation into the advancement of object detection capabilities, we implemented the Faster RCNN Resnet50 FPN V2 model in PyTorch. This model, which demonstrates superior precision in comparison to its predecessor, underwent significant improvements in its ResNet50 backbone and Faster RCNN object detection modules. Crucial optimizations, such as the adjustment of learning rates, extension of training durations, and the integration of MixUp and CutMix enhancements, contributed to the enhancement of its performance. Additionally, modifications, which included FPN with batch normalisation, two convolutional layers in the Region Proposal Network (RPN), and four convolutional layers with Batch Normalisation followed by a linear layer, further bolstered the model. Throughout the training process, utilising a batch size of 8 over the course of 6 epochs, Stochastic Gradient Descent (SGD) was employed as the optimizer with a learning rate of 0.001 and

momentum of 0.9, effectively striking a balance between model convergence and computational efficiency. Upon completion of training, evaluation revealed a mean Average Precision (mAP) at IoU 0.5 to 0.95 of 0.79, effectively highlighting the model's effectiveness in accurately localising and classifying objects.

In the pursuit of robust object detection, our study successfully employed the YOLOv8 model, implementing it by directly cloning the Ultralytics repository. Training was conducted with a batch size of 16, ensuring computational efficiency and model convergence over 50 epochs. The AdamW optimizer was utilised to facilitate efficient weight updates. Evaluation following the training phase yielded compelling outcomes, with a mean Average Precision (mAP) for validation at IoU 0.5 to 0.95 measuring 0.913, effectively showcasing the model's accuracy in accurately localising and categorising objects across a range of IoU thresholds. Remarkably, the mAP at IoU 0.5 reached an impressive value of 0.989, effectively highlighting the model's exceptional precision, particularly in scenarios with less stringent IoU requirements. These validation metrics serve to underscore the proficiency of the YOLOv8 model, establishing it as a potent solution for high-performance object detection tasks across a variety of contexts.

III. Results

Faster RCNN:

The execution of our Faster RCNN model produced admirable outcomes, effectively identifying and producing precise bounding boxes for all categories found in the test dataset. In the provided images, two exemplary instances demonstrate the model's expertise, where the characters 'D' and 'N' were accurately recognized with high confidence scores of 0.97. The bounding boxes, intricately outlining the spatial extent of the detected objects, are accompanied by associated confidence scores, offering a numerical assessment of the model's certainty in its predictions. This successful identification across diverse categories, as exemplified by the featured results, highlights the robustness and dependability of our Faster RCNN model in real-world scenarios involving object recognition.

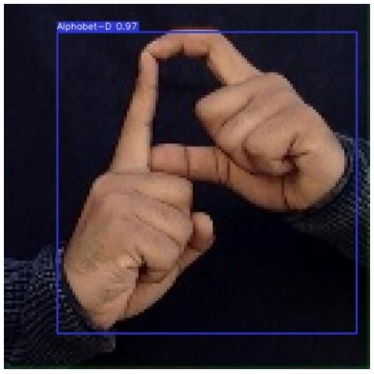

Fig. 3 and 4: Result Accuracy

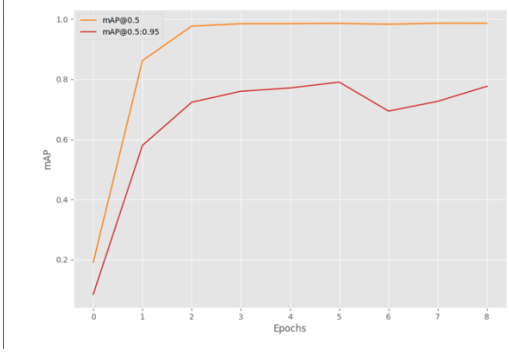

Fig. 5: mAP scores for 6 epochs

YOLOv8:

Upon subjecting our model to rigorous evaluation utilising the test dataset, we have obtained noteworthy results, which are visually represented by images that accurately display generated bounding boxes and corresponding confidence scores for the identified categories. The presented findings demonstrate the model's proficiency in identifying and localising specific objects, as demonstrated by the successful identification of the characters 'Y' and 'V' in the showcased images. These detections were accompanied by high confidence scores of 0.96 and 0.91, respectively, confirming the precision and reliability of the model in determining the presence of distinct categories in various visual contexts. The incorporation of confidence scores offers a quantitative measure of the model's certainty in its predictions, emphasising the practical applicability of our approach in real-world situations.

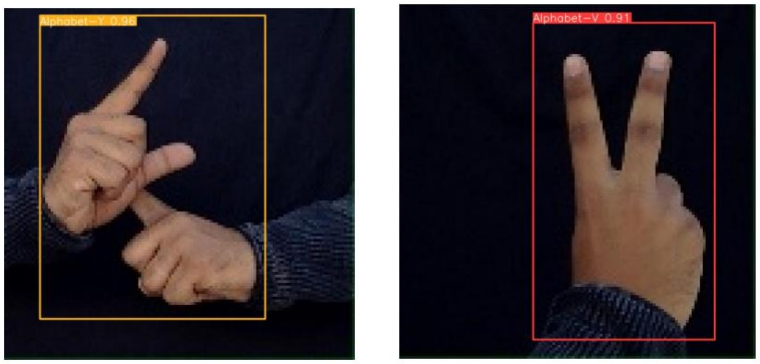

Fig. 6 and 7: Result Accuracy

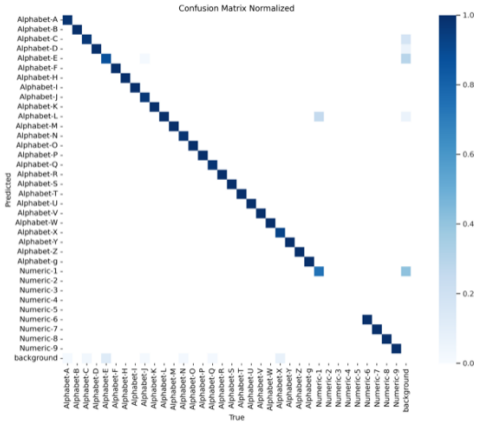

Fig. 8: Normalised Confusion Matrix for all classes

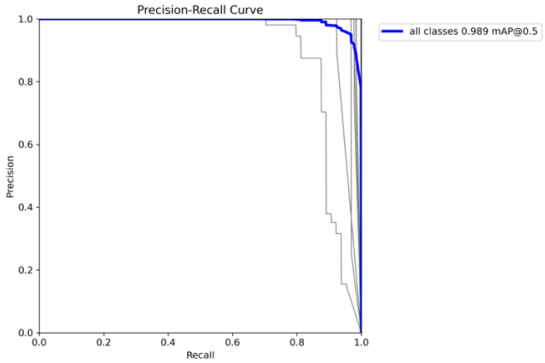

Fig. 9: Precision-Recall Curve after testing

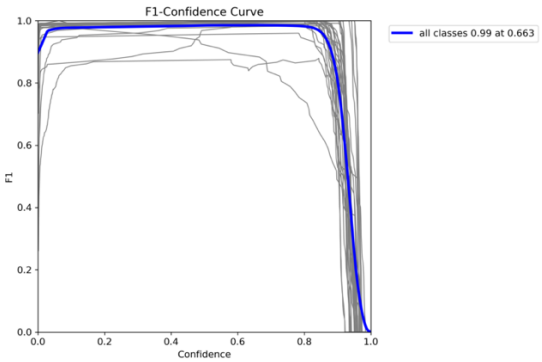

Fig. 10: F1-Confidence Curve after testing

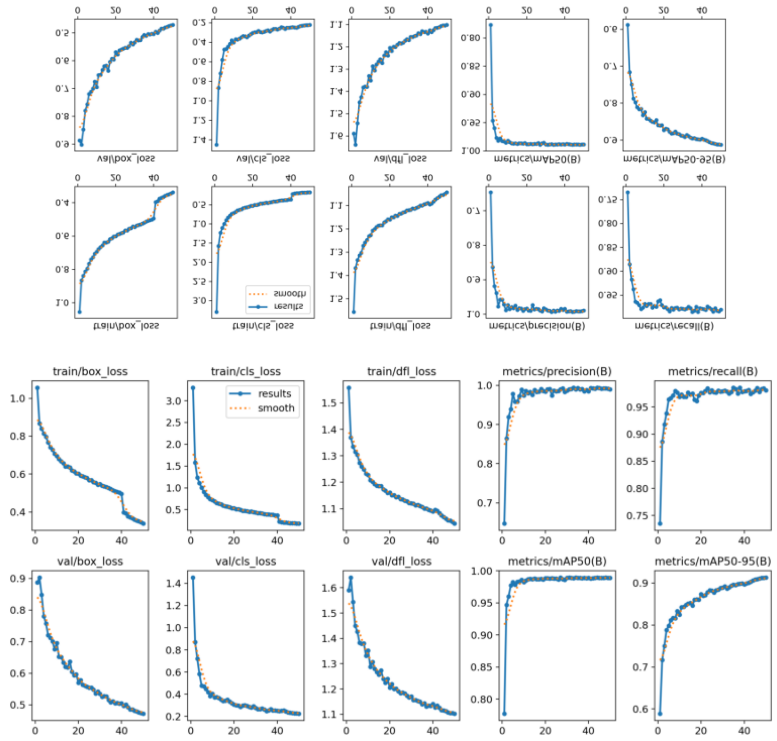

Fig. 11: Precision, Recall, mAP@0.50, mAP@0.5:0.95 graphs for all epochs

Limitations

While both the Faster RCNN Resnet50 FPN V2 and YOLOv8 models have demonstrated remarkable achievements in object detection tasks, it is essential to acknowledge inherent limitations in their implementations. The Faster RCNN model, with its sophisticated architecture, faces potential challenges in computational demands, especially with a batch size of 8 during training, hindering widespread adoption for researchers with limited computational resources. The reliance on SGD optimization, while effective, introduces sensitivity to hyperparameter choices, requiring careful tuning. On the other hand, the YOLOv8 model, despite its impressive accuracy, may encounter interpretability challenges due to its anchor-free detection approach, making it harder to understand specific predictions and impacting model explainability. The mosaic augmentation strategy, effective in diversifying the training dataset, could introduce complexities later in training, as suggested by the recommendation to disable it for the final ten epochs, necessitating a careful balance between augmentation advantages and downsides. In both cases, further exploration into model generalisation to unseen data and robustness in real-world conditions is essential, as addressing these limitations would significantly contribute to enhancing the overall reliability and applicability of these object detection models in practical scenarios.

Future Scope

The successful implementation and evaluation of the Faster RCNN Resnet50 FPN V2 and YOLOv8 models open avenues for future research and development in object detection. Investigating continuous refinement and optimization of model architectures and training strategies, such as exploring advanced backbone architectures and optimization algorithms, holds promise for achieving higher accuracy and generalisation. Adapting these models to handle real-world challenges, like varying lighting conditions and complex scenes, remains a fruitful area for exploration, with potential enhancements through advanced data augmentation and contextual information integration. Improving model interpretability, particularly addressing YOLOv8's anchor-free detection, could enhance transparency in decision-making processes. Scalability considerations, crucial for resource-constrained environments, prompt exploration into model compression and acceleration without performance compromise, broadening applicability across devices. Finally, extending evaluations to larger, diverse datasets and benchmarking against other object detection models would provide a comprehensive understanding of comparative strengths and weaknesses. These accomplishments set the stage for ongoing investigations, fostering

advancements in more robust, interpretable, and scalable object detection methodologies with broader real-world applications. Continued exploration in these directions will contribute to the evolution of object detection technology, impacting various domains.

IV. Conclusion

In conclusion, our research validates the effectiveness of the Faster RCNN Resnet50 FPN V2 and YOLOv8 models in object detection. The Faster RCNN model, with enhanced precision stemming from critical improvements to the ResNet50 backbone and object detection modules, demonstrates resilient performance through key optimizations like learning rate adjustments and augmentations. Evaluation post-training affirms its accuracy, with a mean Average Precision (mAP) at IoU 0.5 reaching an impressive 0.986, even at lower IoU thresholds. Similarly, the YOLOv8 model, prioritising computational efficiency and model convergence, exhibits exceptional accuracy, with a validation mAP at IoU 0.5 reaching 0.989. High-confidence detections of characters 'Y' and 'V' further attest to the models' precision in diverse visual contexts.

The practical applicability of both models in real-world scenarios is evident from their ability to identify and generate precise bounding boxes for different categories, showcasing their robustness and dependability in object recognition tasks. These accomplishments and insights form a robust foundation for future advancements in object detection methodologies, pushing the boundaries of accuracy, efficiency, and applicability across diverse contexts.

Statements and Declarations

ACKNOWLEDGEMENT
I would like to acknowledge the invaluable assistance provided by our mentors and colleagues throughout the duration of this research. It is of utmost importance to accentuate that there are no conflicting interests among the authors with respect to the dissemination of this research publication.

COMPETING INTERESTS
Regarding the research conducted for this publication, the author declares the nonexistence of any conflicting interests.

AUTHOR'S CONTRIBUTION
Rachna Shetty, the dedicated author of this research paper, played a pivotal role in shaping the conception and design of the study. The author's active engagement went beyond mere participation, encompassing the meticulous development and revision of the work. This involved extensive literature exploration, comprehensive data analysis, and nuanced interpretation of the results. The invaluable guidance and supervision provided by Prof. Sofia Francis were accompanied by insightful advice generously given at every stage of the study. The author thoroughly evaluated and sanctioned the final manuscript, taking first authorship, and making a significant contribution to its intellectual content.

DATA AVAILABILITY STATEMENT
Every piece of information incorporated in this research paper, including case studies, theoretical frameworks, and literature sources, is readily available to the public and meticulously referenced in the designated section. This inquiry abstained from the creation or examination of any datasets, ensuring a focus on existing and accessible knowledge.

Compliance with Ethical Standards

FUNDING INFORMATION
This publication outlines research that was carried out in an independent manner, devoid of any external aid. The study was autonomously performed by the author, without any financial backing from any external funding entity or organisation.

RESEARCH INVOLVING HUMAN AND/OR ANIMAL PARTICIPANTS
The foundational underpinnings of this research endeavour were shaped by the theoretical examination, examination of case studies, and thorough evaluation of the extensive body of literature. Direct involvement of human or animal subjects was not a component of this study. Consequently, the acquisition of ethical approval and the procurement of permission papers from both human and animal participants were rendered unnecessary.

INFORMED CONSENT
Formal informed permission was deemed unnecessary for the execution of this study, as it exclusively relied on theoretical analysis and existing literature, without the involvement of human or animal subjects. The authors refrained from collecting any original data related to people or animals; instead, they exclusively utilised publicly accessible materials.

References

[1]. "Faster R-CNN," GeeksforGeeks, 27-Feb-2020. [Online]. Available:

https://www.geeksforgeeks.org/faster-r-cnn-ml/. [Accessed: 29-Oct-2023]

[2]. Researchgate.net. [Online]. Available: https://www.researchgate.net/figure/Faster-RCNN-framework-for-object-detection-and-classification-new-network-model-to-form_fig5_346764538. [Accessed: 29-Oct-2023].

[3]. O. Koller, S. Zargaran, H. Ney, and R. Bowden, "Deep sign: Enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs," Int. J. Comput. Vis., vol. 126, no. 12, pp. 1311–1325, 2018.

[4]. A. Tyagi and S. Bansal, "Hybrid FiST_CNN approach for feature extraction for vision-based Indian sign language recognition," Int. Arab J. Inf. Technol., vol. 19, no. 3, 2022.

[5]. A. Al-Shaheen, M. Çevik, and A. Alqaraghuli, "American sign language recognition using YOLOv4 method," International Journal of Multidisciplinary Studies and Innovative Technologies, vol. 6, no. 1, p. 61, 2022.

[6]. A. A. Barbhuiya, R. K. Karsh, and R. Jain, "CNN based feature extraction and classification for sign language," Multimed. Tools Appl., vol. 80, no. 2, pp. 3051–3069, 2021.

[7]. J. Huang, W. Zhou, H. Li, and W. Li, "Sign Language Recognition using 3D convolutional neural networks," in 2015 IEEE International Conference on Multimedia and Expo (ICME), 2015.

[8]. Sabeenian, S. S. Bharathwaj, and M. M. Aadhil, "Sign language recognition using deep learning and computer vision," J. Adv. Res. Dyn. Control Syst., vol. 12, no. 05-SPECIAL, pp. 964–968, 2020.

[9]. YOLOv8 Architecture, visualisation made by GitHub user RangeKing https://github.com/RangeKing

[10]. "Indian Sign Language object detection dataset and pre-trained model by ISR1," Roboflow. [Online]. Available: https://universe.roboflow.com/isr1/indian-sign-language-3e2qh. [Accessed: 30-Oct-2023].

Tables

List of Figures