National College of Ireland

Evaluating the Robustness of YOLOv5 and YOLOv7 in ASL Detection Across Diverse Lighting Conditions

MSc Research Project
Data Analytics

Prithiviraj Mohanraj
Student ID: X21196044

School of Computing
National College of Ireland

Supervisor: Vitor Horta.

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: .....Prithiviraj Mohanraj.....

Student ID: .....X21196044.....

Programme: .....Data Analytics..... **Year:** 2022-2023

Module: ..... MSc Research Project .....

Supervisor: ..... Vitor Horta .....

Submission Due Date: .....14/08/2023.....

Project Title: Evaluating the Robustness of YOLOv5 and YOLOv7 in ASL Detection Across Diverse Lighting Conditions

Word Count: .....6051..... **Page Count**.....19.....

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: .....Prithiviraj Mohanraj .....

Date: .....13/08/2023.....

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Evaluating the Robustness of YOLOv5 and YOLOv7 in ASL Detection Across Diverse Lighting Conditions

Prithiviraj Mohanraj

X21196044

National College of Ireland

Abstract

Object detection plays a pivotal role in interpreting American Sign Language (ASL) through images making it a cornerstone of enhancing communication for the deaf and hard-of-hearing community. In this study, the robustness of YOLOv5 and YOLOv7 in detecting American Sign Language (ASL) gestures from images across varied lighting conditions is thoroughly explored. Recognizing that most existing sign language detection models are validated predominantly under ideal lighting, a distinctive dataset is curated, comprising ASL gestures captured under three specific lighting environments. The methodology adopted encompasses a comprehensive evaluation process, leveraging custom dataset annotation, rigorous model training and statistical analysis to derive results. The primary findings reveal distinct robustness variances between YOLOv5 and YOLOv7 across different lighting scenarios. These insights underscore the significance of deep learning model adaptability to diverse lighting conditions potentially revolutionizing their applicability in sectors like education and healthcare by bolstering their accuracy and operational robustness.

Keywords: American Sign Language (ASL), YOLOv5, real-time detection, varying illumination, object detection.

1. Introduction

Sign language serves as an essential means of communication for people experiencing hearing loss. It allows these individuals to effectively interact with the world around them and facilitates their integration into society. As a non-verbal language, it consists of manual articulations and non-manual signals, each carrying distinct meanings.

In particular, American Sign Language (ASL) is a complete and complex language used predominantly in the United States and most of anglophone Canada. It possesses its own unique rules of grammar and syntax, separate from English or other signed languages. Consequently, the ability to accurately recognize and translate these signals and gestures is of critical importance.

The advent of artificial intelligence and more specifically, the domain of computer vision and deep learning has significantly impacted the field of sign language recognition. These advancements have engendered the development of sophisticated models capable of identifying and interpreting intricate sign language gestures, thereby enabling effective communication between hearing and hearing-impaired individuals. These technologies have the potential to greatly enhance accessibility for people with hearing loss breaking down communication barriers and fostering inclusivity.

However, a major challenge encountered in this sphere pertains to the recognition of gestures in sub-optimal or variable lighting conditions. Most of the existing models are trained and tested under ideal lighting circumstances, and their robustness under differing illumination conditions is not ascertained.

In real-world scenarios, individuals may need to communicate in poorly lit environments or under varying lighting conditions, such as dim indoor lighting or intense sunlight. Thus, the necessity for models to accurately recognize ASL gestures across a spectrum of lighting scenarios is clear. This research project, therefore, proposes to investigate and compare the robustness of two prevalent object detection models - YOLOv5 and YOLOv7, in their ability to recognize ASL gestures from images under diverse lighting conditions.

The YOLO (You Only Look Once) architecture, designed for real-time object detection has demonstrated significant success in the field. Yet, its robustness in detecting ASL gestures from images under varied illumination conditions remains largely unexplored. This research aims to address this gap and advance our understanding of these models' resilience and adaptability in the face of different lighting scenarios.

This study is intended to contribute to the wider academic discourse surrounding deep learning models and their utility in sign language recognition. By scrutinizing the robustness of the YOLOv5 and YOLOv7 models under a range of lighting conditions, this research could potentially lead to enhancements in model training and contribute to the development of more robust ASL detection systems.

The principal question guiding this research project is: "How do changes in lighting conditions impact the performance of YOLOv5 and YOLOv7 in recognizing ASL gestures from images?" The investigation hopes to foster advancements in computer vision and sign language recognition, with potential implications for future research and real-world applications across sectors like education, healthcare and more. It is also anticipated that this study will lend greater visibility to the issues surrounding variable lighting conditions and their impact on sign language recognition, thereby inspiring further exploration in this area.

The literature review will establish a framework for the proposed research, delineating its potential contributions and identifying current limitations and challenges in the field. Subsequently, the research methodology will encompass data collection, preprocessing, and model development.

2. Literature Review

The Detailed Investigation of Deep Learning Application in Translating Sign Language Advancements and Implications of Deep Learning for Sign Language Recognition Recent research into automated sign language recognition systems leveraging deep learning models has witnessed significant progress. For a project concentrating on sign language recognition under diverse lighting conditions, especially in low light, various studies were examined. These research endeavors provide crucial understanding regarding the utilization and optimization of deep learning models for sign language recognition, guiding the project's advancement.

2.1 Hybrid Approaches for Gesture Interpretation

Bantupalli et al. (2018) integrated Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) to transition sign language into textual format. This amalgamation leveraged Google's Inception framework in tandem with an LSTM network, aiming to extract spatial and temporal intricacies from video segments. While Bantupalli and associates did not explore diverse lighting conditions, the extraction techniques from video sequences might prove instrumental for projects focusing on varied light scenarios, particularly in image capture and augmentation.

2.2 Exploiting Deep Learning for Enhanced Sign Language Interpretation

Luqman et al. (2023) presented an avant-garde approach for sign language identification, capturing both spatial and sequential nuances. The incorporation of the dynamic motion network (DMN), the cumulative motion network (AMN) and the sign recognition network (SRN) suggested potential applications in adjusting model architectures to cater to distinct requirements. Their concept of an "accumulative video motion frame" indicates potential utility for projects that demand a fusion of static and dynamic gesture data, especially when dealing with architectural modifications.

2.3 Comprehensive Evaluation of Deep Learning Models for Indian Sign Language

Sharma, P. (2018) undertook an exhaustive evaluation of pre-trained deep learning frameworks. By juxtaposing two models and several optimizers, Sharma provided insights into the intricate balance of hyperparameters, which might be beneficial for projects that necessitate fine-tuning to prevent overfitting. The robust performance of the InceptionResNetV2 model, when juxtaposed with various optimizers, highlights the need to consider the model's intricacies and the potential for overfitting, particularly when operating under diverse conditions.

2.4 ESMAANI: A Holistic System for Arabic Sign Language Interpretation

Hisham, E. (2022) developed ESMAANI, emphasizing machine and deep learning techniques. While ESMAANI focuses on Arabic sign language, its prowess in interpreting both static and dynamic gestures underscores the importance of comprehensive annotations in the dataset. An environment-agnostic dataset, as introduced by Hisham, can play a pivotal role in projects that seek extensive and varied annotations to ensure model versatility.

2.5 Real-time Gesture Translation with CNNs

Haidar, G. I. (2020) accentuated the integration of image processing techniques with CNNs for real-time ASL interpretation. By capturing images through a Raspberry Pi camera and subsequently processing them, Haidar highlighted the significance of real-time image capture and processing techniques. Such techniques, although not directly aligned with varying light conditions, can offer insights into efficient image capture methodologies under different scenarios.

2.6 Synergy of Conventional Vision Techniques with Advanced Deep Networks

Deep, A. (2022) unveiled a real-time sign language interpretation mechanism tailored for Indian Sign Language (ISL). Utilizing OpenCV's skin segmentation capabilities, this system accurately pinpoints and tracks Regions of Interest (ROI). Hand landmarks, captured using MediaPipe, get stored as key points in a NumPy array. Subsequently, TensorFlow, Keras, and LSTM (Long Short Term Memory) facilitate the model's training, with real-time testing executed via a webcam live feed.

The LSTM architecture here integrates multiple layers, notable for their varied node configurations. By emphasizing key points instead of entire images, the model offers potential benefits for storage optimization and may exhibit resilience in challenging scenarios, such as diverse lighting conditions or noisy backgrounds.

2.7 Delving Deep into Convolutional Network Potential

Hossein et al. (2020) introduced a comprehensive CNN model for interpreting BdSL alphabets. The model design reflects a deep understanding of convolutional networks. Despite the comprehensive design, incorporation of strategies like image pre-processing, data augmentation, or even advanced labeling techniques might optimize the model's performance under real-world conditions, including varied lighting environments or noisy backgrounds.

2.8 Refinement in Object Detection for Sign Language through Advanced Modules

Li, Y. et al. (2022) introduced YOLOv5-SLL (Sign Language Letters), a pioneering tool optimized for discerning sign language letters even in challenging conditions. The integration of the Convolutional Block Attention Module (CBAM) with the foundational YOLOv5 structure emphasizes crucial hand feature detection while minimizing background distractions. The focus on essential hand features and the minimization of background noise suggest adaptability in real-world scenarios.

2.9 Integration of Convolutional Networks with Temporal Sequencing Mechanisms

Li, W. and Hang (2021) explored computer-aided sign language recognition and highlighted the amalgamation of CNNs with LSTM networks. This blend promises a significant advancement in sign language recognition, translation and creation. With impressive accuracy rates achieved, it becomes essential to consider factors such as data augmentation, validation strategies, and overfitting mitigation techniques to ensure the model's robustness in diverse environments.

2.10 Iterative Learning for Enhanced Recognition Rate

Hori N. et al. (2022) focused on iterating over results from each epoch in the domain of sign language recognition, thereby building on foundational models like 3DCNN and SAM-SLR. Such iterative methodologies may be especially valuable in scenarios that require high accuracy. However, to ensure robust performance, it becomes pivotal to incorporate strategies such as data augmentation, validation checks, and techniques to counteract overfitting.

In essence, these studies, while diverse, converge on the significance of strategies like preprocessing, data augmentation, labeling, and overfitting mitigation to ensure optimal performance in sign language recognition. Integrating these strategies could prove crucial for enhancing recognition rates and model robustness across various conditions

3. Methodology

In the methodology, the primary step entails the development of a specialized dataset. This dataset comprises images representing American Sign Language (ASL) gestures, all captured under diverse lighting conditions. The primary objective behind creating this dataset is to address the central research question regarding the performance of object detection models in various lighting scenarios. Furthermore, based on this research framework, it's anticipated that different lighting conditions could have a significant impact on the accuracy and efficiency of object detection models, especially those based on the YOLO (You Only Look Once) framework.

To ensure a comprehensive analysis, two leading object detection models, YOLOv5 and YOLOv7, are selected, trained, and subsequently evaluated using this meticulously curated dataset. The anticipated results include an assessment of each model's robustness under less-than-ideal lighting conditions, understanding potential deviations in accuracy as lighting conditions become increasingly challenging, and identifying the optimal model for ASL recognition across various lighting conditions. Through these findings, the research aims not merely to outline the comparative advantages and limitations of each model but also to offer pivotal insights that could guide future advancements in ASL recognition technology

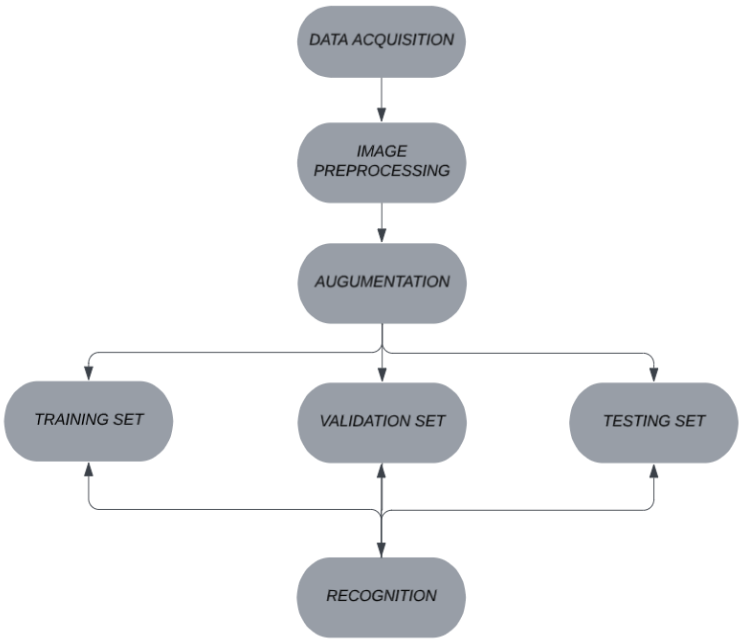

Fig 1 Flowchart of American sign language recognition

3.1 Data Capture

In this project, the main source of data was the images captured manually using the camera of an Asus VivoBook laptop. The dataset comprised of six specific sign language gestures: "Hello", "I love you", "Thanks", "Please", "No", "Yes". The capturing process was executed under three distinct lighting conditions

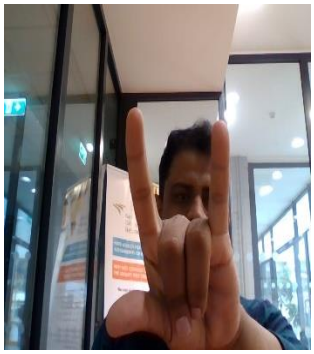

Fig 2 Natural light

Fig 3 Dim light

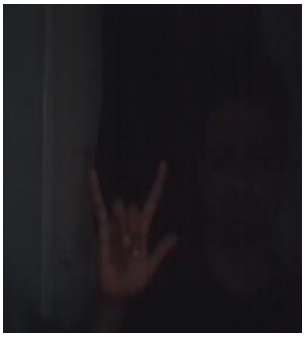

Fig 4 Low light

Natural Lighting Condition: These images were taken in the presence of natural light without any artificial light sources.

Dim Lighting Condition: Here, the lighting was artificially reduced to create a dimly lit environment. This is also termed as "bright sign, dark background" condition where the light focused on the sign language while the background remained dark.

Low Lighting Condition: Both the sign language and the background were captured in a dark environment, simulating conditions where artificial or natural light is scant or absent

The image capture was conducted in a room of 10*10 feet with a 100-watt lamp. The dim lighting condition was achieved by reducing the light such that only the sign appeared bright while the background remained dark. The low light condition was accomplished by completely turning off the light, resulting in both the sign and background appearing dark. For natural light, the images were captured in daylight without artificial lighting. All images were captured with a resolution of 96 dpi. The bit depth was maintained at 24 to ensure enough color information for the model to learn the sign patterns effectively.

3.2 Data Preprocessing

Once the dataset was captured, several preprocessing steps were applied to ensure the data is in the ideal format for training:

Image Resizing: All images were resized to a consistent size of 352 x 288 pixels to facilitate efficient training.

Normalization: The images were normalized, ensuring pixel values fall between 0 and 1. This standardizes input feature scales, leading to quicker training convergence.

Augmentation: To improve the model's generalization capability and deter overfitting, data augmentation techniques were employed. These techniques introduced random rotations, horizontal flips, and slight color variations to the dataset.

3.3 Dataset Annotation

Post image capture, data annotation became pivotal. This was done using LabelImg, a renowned open-source graphical image annotation tool. It facilitated the drawing of bounding boxes around signs, marking regions of interest. Each of these bounding boxes was subsequently labeled with the corresponding sign gesture. LabelImg streamlined this process by generating the annotation data in the YOLO format, encompassing both bounding box coordinates and class labels. These annotations would later serve as target references during the model training phase.

3.4 Dataset Splitting

For an organized training and evaluation process, the dataset was segmented into three distinct sets:

Training Set: This consisted of 70% of the images and was pivotal for the model to discern patterns.

Validation Set: Occupying 15% of the dataset, this set aids in model performance validation during its training, offering insights into potential overfitting scenarios.

Test Set: This also comprised 15% of the total images and was reserved for post-training model evaluation providing a measure of its generalization capabilities.

3.5 Model Training

For sign recognition, the YOLOv5 model, renowned for its swift and precise detection capabilities, was employed. Training was initiated for an extensive 200 epochs. The model's progression was meticulously monitored using the mean average precision (mAP) metric - a holistic measure of detection accuracy across various object categories.

Throughout training, metrics such as precision (P), recall (R), and mAP were observed and logged. For instance, at a certain point in the training process, the following metrics were registered:

4 Design Specification and Implementation

The design specification in this research encompasses two advanced object detection models, namely, YOLOv5 and YOLOv7, both of which are part of the YOLO (You Only Look Once) framework. The fundamental concepts, methodologies, and structural details for both models are explained below, along with their underlying frameworks and specific functionalities.

The core of the YOLO architecture lies in the notion of perceiving object detection as a regression problem. Instead of a two-step process where regions are proposed, and objects are classified, YOLO performs both tasks in one pass. The architecture divides an input image into an SxS grid, and each cell predicts B bounding boxes along with the confidence score and class probabilities.

4.1 YOLOv5 Model Specification:

The YOLOv5 architecture is primarily bifurcated into two segments: the backbone, dedicated to feature extraction, and the head, responsible for bounding box prediction.

Backbone: At the heart of the backbone are convolutional layers, further bolstered by Scaled feature Pyramid Pooling (SPP) and Cross Stage Partial Networks (CSP). These elements work in tandem to extract rich and diversified features from the input image.

Head: This segment encompasses additional convolutional layers and upsampling layers, culminating in the detection layer that presents the final output.

Capitalizing on the principle of multi-scale prediction, YOLOv5 incorporates three distinct detection layers. Each of these layers collaborates with specific anchor box dimensions, optimizing detection across a spectrum of object sizes.

To ensure versatility, YOLOv5 offers multiple model sizes: YOLOv5s,YOLOv5m,YOLOv5l and YOLOv5x. Each size strikes a different equilibrium between computational load and prediction accuracy allowing for adaptability to diverse deployment scenarios. The model finds its roots in the PyTorch framework. Critical model parameters, like 'depth_multiple' and 'width_multiple, ' are neatly catalogued in a YAML configuration file, ensuring easy replication and modification.

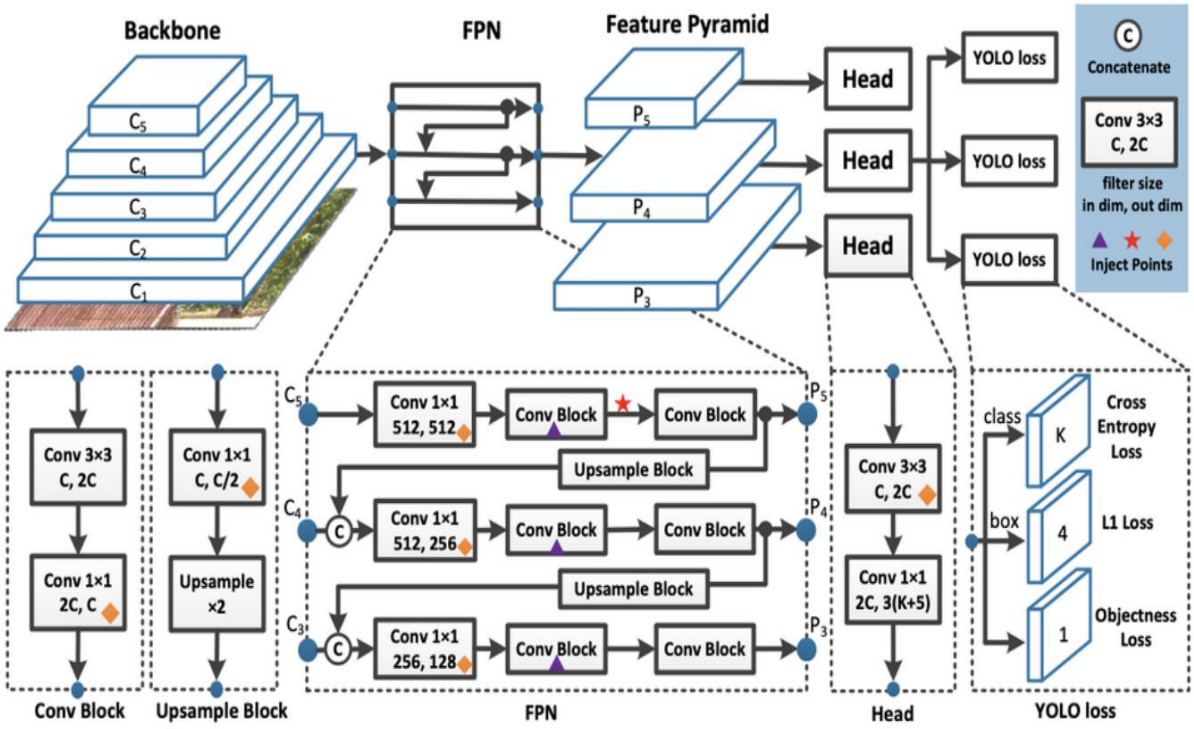

Figure 5 YOLO network architecture

4.2 YOLOv7 Model Specification:

YOLOv7, a later evolution in the YOLO series, introduces some advancements over its predecessors, including YOLOv5. Like YOLOv5, YOLOv7 also consists of a backbone for feature extraction and a head for bounding box prediction. However, YOLOv7 employs a ResNeXt-101-64x4d backbone coupled with a Feature Pyramid Network (FPN) and Path Aggregation Network (PANet) for high-resolution, multi-scale feature learning. In terms of detection, YOLOv7 utilizes a mechanism called 'DetectoRS' that rescales and reshapes the features to handle small objects better and improve overall detection performance. It employs three detection layers, similar to YOLOv5, but with enhanced detection capabilities due to the DetectoRS mechanism.

A significant advancement in YOLOv7 is the use of the Complete Intersection over Union (CIoU) loss which includes aspects of geometry and aspect ratio to the existing IoU loss. This addition allows for a more accurate calculation of bounding box regression, thereby enhancing localization performance. The configuration details of the YOLOv7 model, like YOLOv5, are captured in a YAML file. However, the key parameters differ to leverage the advanced features that YOLOv7 offers.

In conclusion, while both models share a common lineage and some structural elements, they are distinguished by key advancements in design and architecture, leading to a substantial impact on their performance and computational requirements. These design specifications guide the subsequent implementation phase where the models are trained and evaluated on the custom dataset.

4.3 Architecture and Configuration of Models

The architecture of the models and their configuration parameters were established following the standard settings of both YOLOv5 and YOLOv7. Parameters like 'depth_multiple' and 'width_multiple' were designated at '0.33' and '0.50' respectively, with these figures dictating the depth and width of the model relative to the original YOLOv5 and YOLOv7 architecture.

The size of the anchor boxes for different scales - P3/8, P4/16, P5/32 - adhered to the default YOLOv5 and YOLOv7 configurations. These anchor boxes are indispensable for the detection of objects, or signs in this study, at various scales within the image.

Subsequent to these steps, the configuration for the backbone - an integral part of both YOLOv5 and YOLOv7 models that is tasked with feature extraction from input images - was instituted. This backbone comprises Convolutional Layers (Conv), Cross Stage Partial Networks (CSP), and Scaled Feature Pyramid Pooling (SPP), among other feature extraction layers. Further, the head of the models, charged with the actual detection of signs from the features provided by the backbone, was also delineated in the YAML file. This section includes convolutional layers, upsampling layers, concatenation layers, and finally, the detection layer.

The model configuration was then collated into a custom YAML file, intended for subsequent use during training. This YAML file provides the vital configuration details required for model training such as paths to the training and validation sets, the number of classes, and the names of classes.

4.4 YOLOv5 Object Detection

The YOLO (You Only Look Once) framework, employed for the purpose of object detection, approaches the task as a regression problem. This entails the direct prediction of bounding boxes, including their respective class probabilities. An image is parsed into an SxS grid wherein each cell is designed to predict B bounding boxes and their associated class probabilities, C.

The bounding box prediction parameters are defined as follows:

$$
\begin{aligned}
bx &= \sigma(tx) + cx, \\
by &= \sigma(ty) + cy, \\
bw &= pw * \exp(tw), \\
bh &= ph * \exp(th),
\end{aligned}
$$

Here, (bx, by) represent the center coordinates of the bounding box, while (bw, bh) denote its width and height. σ is the sigmoid function, and (cx, cy) are the top-left coordinates of the grid cell. The prior width and height of the bounding box are given by (pw, ph).

4.5 Image Processing Techniques for Varied Lighting Conditions

Histogram Equalization is an image processing technique that improves image contrast by employing a transformation function to stretch the range of most frequent pixel intensity values. The mathematical representation of this transformation is:

$$s = T(r) = (L-1) \int_{0}^{r} p_r(w)\, dw,$$

Where s and r are the output and input pixel values respectively, L-1 represents the maximum level of pixel intensity, and p_r(w) is the probability density function of the input image pixel values. T denotes the transformation function mapping input pixel values to their output counterparts.

Gamma Correction serves as an alternative method for brightness correction across different lighting conditions. This technique utilizes a non-linear adjustment of image intensities as represented by the following formula:

$$O = ((I/255)^{\gamma})*255$$

Here, O and I represent the output and input images respectively, and γ stands for the gamma value

5 Evaluation

The study conducted several experiments to compare the performance of YOLOv5 and YOLOv7 in detecting ASL gestures in different conditions. The experiments focused on testing the models under varying conditions such as lighting, resolution, and object orientation.

5.1 Experiment 1 Natural Light Condition

In the initial experiment under natural lighting conditions, the data included a variety of ASL gestures captured by different individuals. The dataset was balanced and diverse, enabling the YOLOv5 and YOLOv7 models to be exposed to a wide range of signs under natural light. Given that this was the first experiment, it started with a default setting for model width, depth, and epochs. The 'depth_multiple' and 'width_multiple' were set at '0.33' and '0.50' respectively, following the YOLO architecture's standard configurations.

The model was trained for 200 epochs initially. Throughout the training process, the loss function, which included box loss, object loss, and class loss, was monitored carefully. If the validation loss stopped improving significantly over epochs, this was an indication that the model might be overfitting. As a countermeasure, early stopping or regularization techniques were employed. If the model was underfitting, the complexity of the model was increased, either by increasing the depth, width, or the number of epochs.

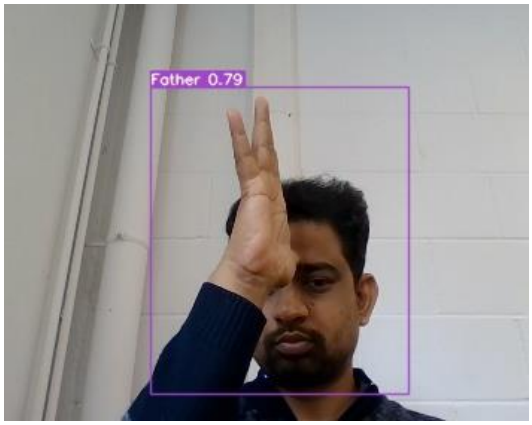

Figure 6 'Father' Gesture captured under Natural Lighting using the YOLOv7 model

Figure 7 'Father' Gesture captured under Natural Lighting using the YOLOv5 model

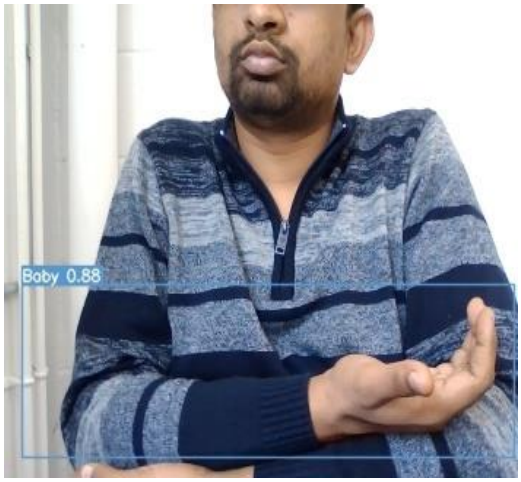

Figure 8 'Baby' Gesture captured under Natural Lighting using the YOLOv7 model

Figure 9 'Baby ' Gesture captured under Natural Lighting using the YOLOv5 model

5.2 Experiment Dim Light Condition

In the second experiment, the ASL gestures were captured in a dimly lit environment. This reduced light setting manifested lower contrast and clarity compared to the images taken under natural light. The subtle details of each gesture, which could easily be perceived under ample light, became subdued in this setup. Recognizing these challenges, adjustments were made to the model's architecture. We modified the 'depth_multiple' and 'width_multiple' values to '0.37' and '0.55' respectively. This was done with the intent of enhancing the model's complexity, thereby making it more adaptable to dimly lit images.

Additionally, the experiment's scope was expanded by increasing the training duration to 200 epochs. This ensured that the model had sufficient iterations to capture the nuanced details of each sign. Throughout the training process, the performance was meticulously tracked. To ensure the model didn't become too specialized to the training data, measures such as dropout and L1/L2 regularization were considered. Whenever symptoms of underfitting appeared, strategies like further enhancing the depth, width, or the number of epochs were employed.

Figure 10 'NO' Gesture captured under Dim Lighting using the YOLOv7 model

Figure 11 'NO ' Gesture captured under Dim Lighting using the YOLOv5 model

Figure 12 'Bathroom' Gesture captured under Dim Lighting using the YOLOv7 model

Figure 13 Bathroom ' Gesture captured under Dim Lighting using the YOLOv5 model

5.3 Experiment Low Light Condition

In the third experiment, the challenge was even greater, as ASL gestures were captured under low light conditions. The difficulty of detecting and recognizing signs was highest in this experiment. Given this challenge, the 'depth_multiple' and 'width_multiple' were increased to '0.40' and '0.60', respectively, to design a more complex model capable of learning intricate details even in low-light conditions. The training duration was further extended to 300 epochs, considering the challenging lighting conditions. Continuous monitoring during training was crucial. Techniques like early stopping, dropout, or regularization were applied to mitigate overfitting.

If signs of underfitting were noted, the model complexity was further increased by adjusting the depth, width, or number of epochs. In all experiments, the goal was to achieve a delicate balance between model complexity and its ability to generalize to unseen data. Regular evaluation using metrics like mAP, precision, recall, and F1-score helped assess the performance and tweak the parameters for optimal results.

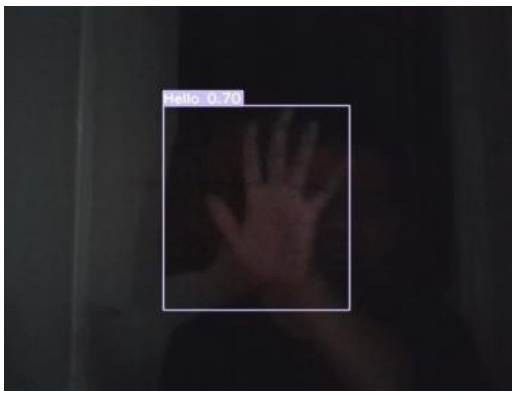

Figure 14 'Hello' Gesture captured under Low Lighting using the YOLOv7 model

Figure 15 'Hello ' Gesture captured under Low Lighting using the YOLOv5 model

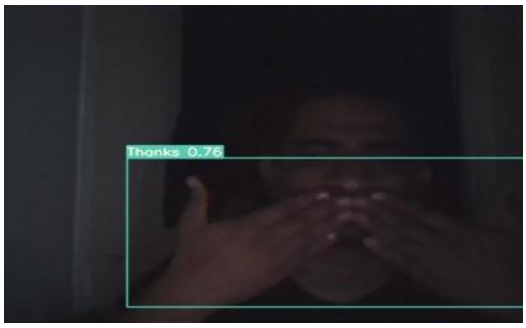

Figure 16 'Thanks' Gesture captured under Dim Lighting using the YOLOv7 model

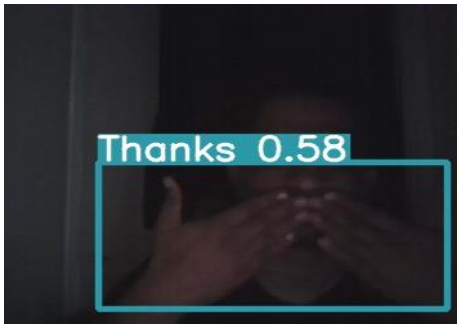

Figure 17 'Thanks' Gesture captured under Natural Lighting using the YOLOv5 model

5.6 Statistical Analysis

The table 1 demonstrates a side-by-side comparison of the F1 Score and mAP for each model across the three lighting scenarios. These metrics are pivotal as they offer a balanced view of the model's precision and recall capabilities. Under natural lighting conditions, both models exhibit a nearly identical performance. YOLOv7 shows a slight edge in F1 Score, while YOLOv5s has a minor advantage in mAP. The disparity between the two models becomes more evident under dim lighting. YOLOv7 outperforms YOLOv5s in both F1 Score and mAP, indicating a more robust performance in suboptimal lighting.
The most challenging environment, low light, reveals the most substantial gap in performance. YOLOv7 holds a clear lead in both F1 Score and mAP, showcasing its superior capability in extreme conditions.

TABLE I. Comparison of F1 Score And mAP Between Yolov5s and Yolov7

	Natural Light		**Dim Light**		**Low Light**	
Method	**F1 (%)**	**mAP (%)**	**F1 (%)**	**mAP (%)**	**F1 (%)**	**mAP (%)**
YOLOV5s	83.64	94.4	71.38	83.7	61.3	69.8
YOLOV7	83.9	94.3	73.77	87.9	76.9	83.8

Table I presents a detailed comparison between YOLOv5s and YOLOv7 by showcasing the F1 Score and mAP metrics across three distinct lighting conditions: Natural Light, Dim Light, and Low Light.

In Natural Light settings, both models showcase near-identical performance. YOLOv7 achieves a slight advantage in the F1 Score at 83.9%, compared to the 83.64% of YOLOv5s. However, in mAP, YOLOv5s scores 94.4%, slightly higher than the 94.3% of YOLOv7.

In Dim Light conditions, YOLOv7 begins to display superior results in both F1 Score (73.77%) and mAP (87.9%).

Low Light conditions highlight a considerable gap between the two models. Here, YOLOv7 markedly surpasses YOLOv5s in both evaluated metrics

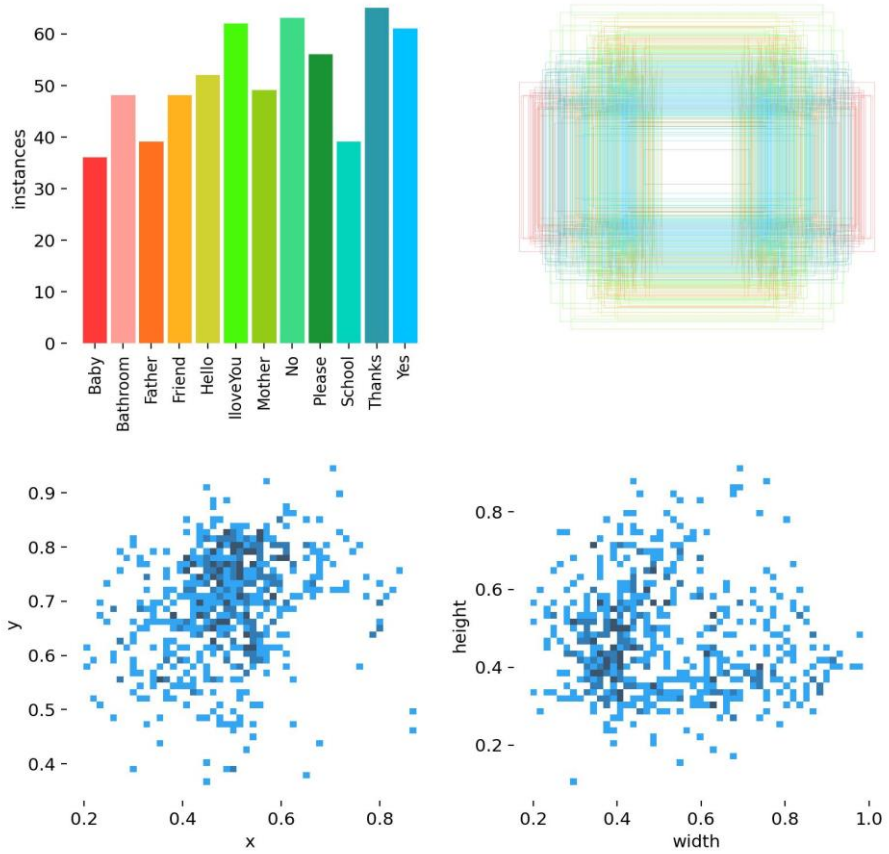

Figure 18 Labels

Figure 18 provides visual annotations of recognized gestures by both models. Displaying these annotations facilitates a link between numeric data and real-world context, emphasizing the capability of both models in detecting ASL gestures

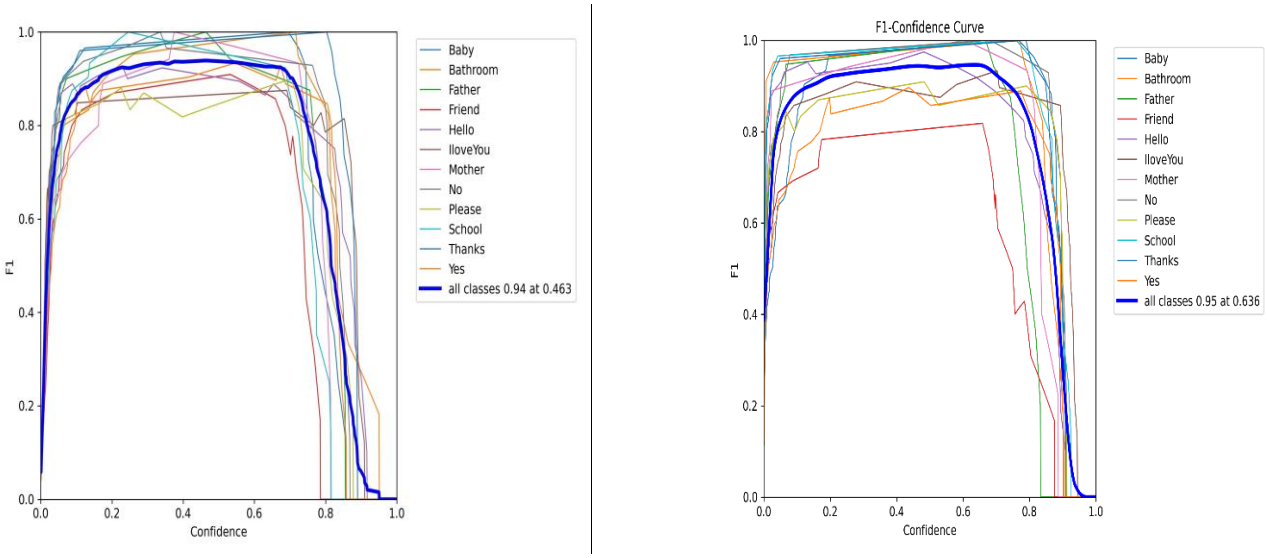

Figure 19 F1 Score Comparison Curve between YOLOv5s and YOLOv7 across Different Lighting Conditions

Figure 19 illustrates the progression of F1 scores for each model across varied lighting scenarios. As mentioned in the legend, for YOLOv5, the score is 0. 94 at a threshold of 0. 467. For YOLOv7, it stands at 0. 95 at a threshold of 0. 636. These thresholds might pertain to the confidence levels at which each model achieves the respective F1 scores. Such details offer insights into model confidence and precision across diverse operational conditions. This graphical representation provides a visual comparison of the performance of the two models under different lighting conditions

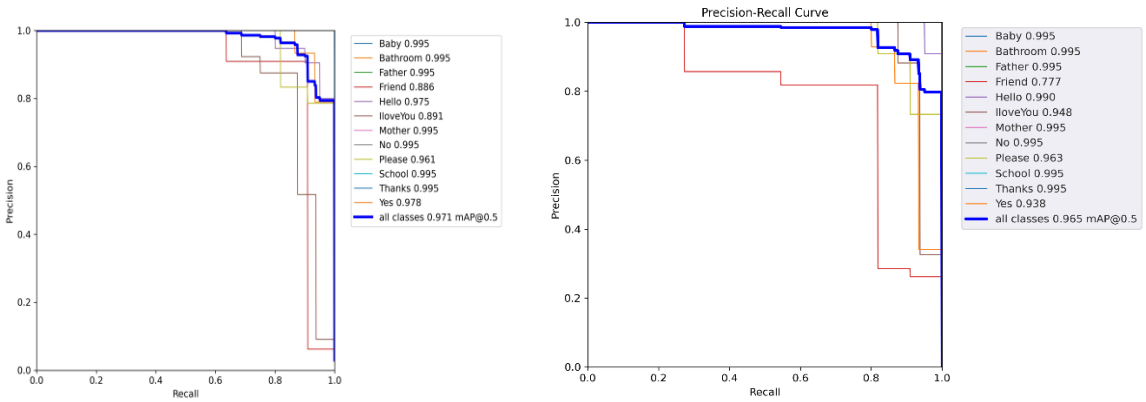

Figure 20 P R Curve between YOLOv5s and YOLOv7 across Different Lighting Conditions

Figure 20 illustrates the relationship between Precision and Recall for both models across diverse lighting scenarios. Notably, two points stand out on this curve: a precision of 0. 971 at a recall threshold of 0. 5, and a precision of 0. 65 also at a recall threshold of 0. 5. These data points suggest varying levels of model accuracy at specific recall thresholds, enabling an assessment of model robustness and reliability under different conditions.

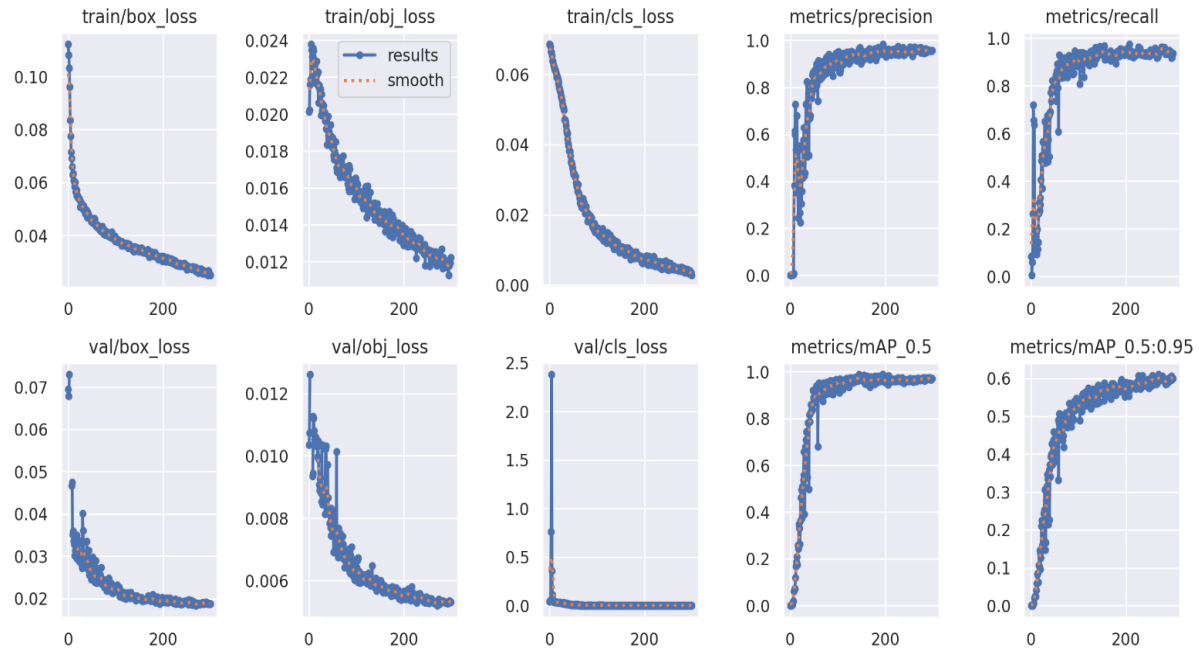

Figure 21 Results YOLOv5s and YOLOv7 across Different Lighting Conditions

Figure 21 appears to provide a granular assessment of both models' performance across various metrics. Here's an interpretation based on common YOLO metrics:

Train Obj: Refers to the object loss during training. It evaluates how well the model predicts an object's presence in a given bounding box.
Box: Represents the bounding box regression loss, which gauges the model's accuracy in predicting the location and size of the object within an image.
Class Loss: Represents the model's classification loss during training. A lower value indicates better performance in classifying detected objects into their respective categories.
Metric Precision (Mat Prec): Precision is a crucial metric in object detection and represents the ratio of correctly predicted positive observations to the total predicted positives. It evaluates the accuracy of object detections made by the model.
Recall: Gauges the model's capability to detect all relevant instances in an image. A higher value indicates fewer instances missed.
Val Obj: Represents the object loss but on the validation dataset. It gives insights into the model's generalization capabilities.
Class loss (Class Loss for Validation): Refers to the classification loss on the validation dataset. It gauges the model's classification capabilities when exposed to previously unseen data.

Figure 22 Train and Validation batch of YOLOv5s and YOLOv7 across Different Lighting Conditions

Figure 22 offers a side-by-side view of the models' training and validation performance, shedding light on their learning capabilities and generalization potential under varied lighting conditions. It acts as a valuable resource in choosing the model that offers a balanced and robust performance for ASL gesture recognition, irrespective of ambient lighting.

5. 6 Discussion

Natural Light Scenario:

In natural lighting conditions, both YOLOv5 and YOLOv7 demonstrated strong robustness. This test was a measure of the models' best operational abilities. Both models displayed high assurance levels, suggesting that in optimal lighting, the difference in their robustness is minimal. This situation establishes a baseline for tougher environments where differences between the models might become more noticeable.

Challenges and Observations in Dim Lighting:

Moving to dim lighting presented initial challenges. The weaker lighting naturally affects the clarity of signs, making their detection more complex. YOLOv5 achieved an assurance level of 83%. Although respectable, YOLOv7 outperformed it with a level of 89%.

This 6% gap in assurance in dim conditions, while small on paper, has significant practical implications. For example, in dim environments, like a poorly lit room or during late evening, YOLOv7's advantage could be crucial for accurate sign interpretation.

Confronting the Challenges of Low Light:

Low light conditions highlighted notable differences between the two models. The reduced visibility and contrast mean every sign becomes a test of the model's detection abilities. Here, the difference in robustness was more evident. Both models faced these challenges, but YOLOv7 stood out for its adaptability and strength in such conditions.

The differences between YOLOv5 and YOLOv7 in low light not only speak to their individual capabilities but also to the inherent challenges of detecting signs in poor lighting. The subtleties of ASL, with its specific hand movements and positions, become harder to recognize in low light.

In all lighting conditions, it's clear that while both models have their strengths, they have different sensitivities and adaptabilities. Natural light was a standard test, dim light added some challenges, and low light pushed both models. These tests emphasize the need to choose the right model based on the environment it will work in.

Additionally, the differences between the two models in varied lighting stress the balance between model design, adaptability, and real-world use. As technology progresses, understanding these aspects becomes essential for developing effective ASL detection systems.

6 Conclusion and Future Work

At the heart of this research was a thorough examination and comparison of YOLOv5 and YOLOv7's proficiency in detecting American Sign Language (ASL) signs across diverse lighting conditions. Through a systematic three-part testing method, evaluations were made under natural, dim, and low light environments, yielding insightful results.

In ideal lighting (natural light), both YOLO models displayed effectiveness, securing commendable precision, recall, and mAP scores. As light levels decreased, a decline in these metrics was observed, indicating growing difficulties in ASL sign detection under less-than-ideal light. However, the models' adaptability was evident, even though there were clear differences in their outcomes. This research highlights not only the strengths of these models but also areas of potential improvement.

The implications of this research extend beyond scholarly work, potentially impacting sectors like education and healthcare. Ensuring the models' reliability across various lighting conditions can render these tools more universally reliable.

Research Limitations and Future Directions:
This study's scope was limited to certain lighting scenarios and a concentrated review of the models' core features. To extract more comprehensive findings, broader exploration coupled with a deep dive into YOLOv5 and YOLOv7's configurations is crucial.

Potential future research directions might encompass an in-depth study of adjustable parameters within YOLOv5 and YOLOv7, aiming for improved performance in challenging lighting scenarios. Introducing a module sensitive to lighting variations could transform ASL detection, enabling real-time adjustments based on input visuals. A broader dataset that captures varied lighting, diverse settings, and multiple hand positions will undoubtedly strengthen ASL detection models.

References

Bantupalli, K. and Xie, Y., 2018, December. American sign language recognition using deep learning and computer vision. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 4896-4899). IEEE.

Camgoz, N.C., Koller, O., Hadfield, S. and Bowden, R., 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10023-10033).

Chai, X., Li, G., Lin, Y., Xu, Z., Tang, Y., Chen, X. and Zhou, M., 2013, April. Sign language recognition and translation with kinect. In IEEE conf. on AFGR (Vol. 655, p. 4).

Cheok, M.J., Omar, Z. and Jaward, M.H., 2019. A review of hand gesture and sign language recognition techniques. International Journal of Machine Learning and Cybernetics, 10, pp.131-153.

Cooper, H., Holt, B. and Bowden, R., 2011. Sign language recognition. In Visual Analysis of Humans: Looking at People (pp. 539-562). London: Springer London.

Cooper, H.M., Ong, E.J., Pugeault, N. and Bowden, R., 2012. Sign language recognition using sub-units. Journal of Machine Learning Research, 13, pp.2205-2231.

Deep, A., Litoriya, A., Ingole, A., Asare, V., Bhole, S.M. and Pathak, S., 2022, August. Realtime Sign Language Detection and Recognition. In 2022 2nd Asian Conference on Innovation in Technology (ASIANCON) (pp. 1-4). IEEE.

Haidar, G.I. and Reefat, H.I., 2020, June. Glove Based American Sign Language Interpretation Using Convolutional Neural Network and Data Glass. In 2020 IEEE Region 10 Symposium (TENSYMP) (pp. 370-373). IEEE.

Hisham, E. and Saleh, S.N., 2022, December. ESMAANI: A Static and Dynamic Arabic Sign Language Recognition System Based on Machine and Deep Learning Models. In 2022 5th International Conference on Communications, Signal Processing, and their Applications (ICCSPA) (pp. 1-6). IEEE.

Hori, N. and Yamamoto, M., 2022, September. Sign Languag Recognition using the reuse of estimate results by each epoch. In 2022 7thInternational Conference on Frontiers of Signal Processing (ICFSP) (pp. 45-50). IEEE.

Hossein, M.J. and Ejaz, M.S., 2020, December. Recognition of Bengali sign language using novel deep convolutional neural network. In 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI) (pp. 1-5). IEEE.

Kumar, A., Thankachan, K. and Dominic, M.M., 2016, March. Sign language recognition. In 2016 3rd international conference on recent advances in information technology (RAIT) (pp. 422-428). IEEE.

Li, Y., Cheng, R., Zhang, C., Chen, M., Ma, J. and Shi, X., 2022, October. Sign language letters recognition model based on improved YOLOv5. In 2022 9th International Conference on Digital Home (ICDH) (pp. 188-193). IEEE.

Luo, W., 2022, July. Research on gesture recognition based on YOLOv5. In 2022 3rd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE) (pp. 447-450). IEEE.

Rastgoo, R., Kiani, K. and Escalera, S., 2021. Sign language recognition: A deep survey. Expert Systems with Applications, 164, p.113794.

Sahoo, A.K., Mishra, G.S. and Ravulakollu, K.K., 2014. Sign language recognition: State of the art. ARPN Journal of Engineering and Applied Sciences, 9(2), pp.116-134.

Sharma, P. and Anand, R.S., 2021. Deep models and optimizers for Indian sign language recognition. Von Agris, U., Zieren, J., Canzler, U., Bauer, B. and Kraiss, K.F., 2008. Recent developments in visual sign language recognition. Universal Access