Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 10 No. 2 (2026) 495 - 506 e-ISSN: 2580-0760

Comparative Analysis of YOLOv8 Segmentation Variants for Indonesian Sign Language (SIBI) Recognition

Desi Fatkhi Azizah[1], Anik Nur Handayani[2]*, Aji Prasetya Wibawa[3], Osamu Fukuda[4]
[1,2,3]Department of Electrical Engineering and Informatics, Faculty of Engineering, Universitas Negeri Malang, Malang, Indonesia
[4]Saga University, Saga, Japan

*aniknur.ft@um.ac.id

Abstract

The Indonesian Sign Language System (SIBI) is the officially recognized communication medium for deaf communities in Indonesia, yet its limited public use continues to create barriers in education, healthcare, and public services. Automatic sign language recognition powered by artificial intelligence provides a promising pathway to reduce these inequities. This study presents a comprehensive comparative evaluation of YOLOv8 segmentation variants for SIBI recognition, aiming to identify models that stabilize accuracy and efficiency for real-time deployment. A mono-background dataset of SIBI alphabet gestures was annotated using instance segmentation, and five YOLOv8-seg models (n, s, m, l, x) were trained and tested across multiple data-split scenarios. Performance was assessed through precision, recall, F1-score, mAP50, mAP50–95, and inference time. Results show that YOLOv8m-seg consistently achieved the best trade-off (F1-score 0.972; mAP50 0.982), while YOLOv8n-seg delivered the fastest inference speed (5.163 ms), making it suitable for resource-constrained devices. Visualization further demonstrated the models' ability to capture hand contours and distinguish gestures accurately. Beyond technical benchmarking, this research highlights the potential of YOLOv8-based SIBI recognition as an inclusive assistive technology for bridging communication gaps in schools and clinics where interpreters are often unavailable. It also identifies governance challenges, including privacy protection, misclassification risks, and equitable access, which must be addressed for actual adoption. The findings, therefore, provide not only a contribution to computer vision research but also practical guidance for policymakers and service providers, positioning SIBI recognition systems as socially embedded technologies aligned with the goals of disability inclusion and sustainable development.

Keywords: assistive communication technology; disability inclusion; public services; SIBI; sign language recognition; YOLOv8

How to Cite: D. F. Azizah, A. N. Handayani, A. P. Wibawa, and O. Fukuda, "Comparative Analysis of YOLOv8 Segmentation Variants for Indonesian Sign Language (SIBI) Recognition", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 10, no. 2, pp. 495 - 506, Apr. 2026.
Permalink/DOI: https://doi.org/10.29207/resti.v10i2.7500

Received: January 16, 2026
Accepted: April 5, 2026
Available Online: April 29, 2026

This is an open-access article under the CC BY 4.0 License
Published by Ikatan Ahli Informatika Indonesia

1. Introduction

Sign language is the primary means of communication for deaf communities around the world [1-2], with variations in each country [3-4], such as American Sign Language (ASL) used in America [5]. British Sign Language (BSL) is used in the United Kingdom [6], etc. In Indonesia, the formally recognized sign language system is the Indonesian Sign Language System (SIBI) [7], designed to support literacy and communication skills. However, a major challenge remains: public understanding of SIBI is very limited, resulting in a communication gap between deaf people and their surroundings. This situation has led to limited access to education, social barriers, and marginalization across various aspects of life.

From a broader perspective, communication barriers faced by deaf communities represent not only linguistic challenges but also limitations in access to education, healthcare, and public services, which have been widely discussed in previous studies on disability inclusion and assistive technologies [8-9]. According to the World Health Organization's International Classification of Functioning, Disability and Health (ICF), assistive technology plays an important role in supporting functional activities and participation in daily life [10]. In contexts such as Indonesia, where the availability of professional SIBI interpreters in schools, clinics, and public service settings remains limited, technology-based solutions are often highlighted as a practical means to improve access to communication [11-12]. While these societal considerations provide important

motivation, the core technical challenge lies in developing accurate and efficient vision-based sign language recognition systems that operate reliably under real-world constraints, ensuring that advances in computer vision are translated into deployable assistive solutions [13]. Therefore, evaluating YOLOv8-seg for SIBI recognition is not only a matter of improving computer vision accuracy, but also of understanding how such systems can be responsibly governed, integrated into public services, and aligned with disability inclusion policies. This perspective situates the present study within the broader discourse of technology adoption, governance, and societal impact.

1.1 Research Background and Motivation

Advances in artificial intelligence (AI) and computer vision technology provide a great opportunity to address these issues through automatic sign language recognition systems [14-15]. To support this inclusive space, various computer vision-based approaches have been developed, ranging from Convolutional Neural Networks (CNN) to Transformer-based architectures [16]. However, challenges such as complex backgrounds, lighting variations, and similar gestures mean that the system's accuracy still needs to be improved.

Modern object detection methods, such as You Only Look Once (YOLO), have proven capable of real-time detection with high accuracy [17-18], and YOLOv8, as the latest generation, has a segmentation variant that can detect and segment objects simultaneously [19-20]. This makes it highly relevant for use in the context of SIBI gesture recognition, which requires great visual detail in the hand area. YOLOv8 provides several model variants, namely YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x, which differ in terms of architectural complexity, parameter size, and inference speed [21].

Lightweight models such as YOLOv8n and YOLOv8s are suitable for devices with limited resources [22], [23], while larger models such as YOLOv8l and YOLOv8x offer higher accuracy at a greater computational cost [24], [25]. Given these variations, it is important to conduct a comparative analysis to determine which model is most suitable for SIBI recognition. YOLOv8-seg also has the potential to improve detection accuracy by focusing attention on the hand area and reducing interference from the background.

Previous research on SIBI detection has been conducted, one of which used a Hand Keypoint-based CNN approach for classification using the ResNet-50, EfficientNet, and InceptionV3 architectures [26]. SIBI detection was performed by applying the Inflated 3D CNN (I3D) model combined with transfer learning from ImageNet and Kinetics [27]. SIBI detection using the Decision Tree and C4.5 algorithms for SIBI recognition using a daily gesture dataset found that the Decision Tree performed better in terms of accuracy, precision, recall, and F1-score metrics, and was easier to interpret [28]. This condition opens new research opportunities to evaluate all YOLOv8-seg variants for SIBI recognition. Based on the identified research gaps, this study conducts a systematic comparative evaluation of all YOLOv8-seg variants on the SIBI dataset. The analysis incorporates multiple evaluation metrics, including confusion matrix, precision, recall, F1-score, mAP50, mAP50–95, and inference time, to examine both segmentation accuracy and computational efficiency. By adopting a multi-scenario data partitioning strategy, this work aims to provide a comprehensive assessment of model robustness and deployment feasibility for real-time SIBI recognition systems.

1.2 Research Contribution

This study makes several scientific and practical contributions. First, it provides the first comprehensive benchmarking of YOLOv8-seg for SIBI recognition; unlike prior studies that focused on single architectures such as CNN, 3D-CNN, or classical machine learning methods, this research presents a systematic comparative evaluation of all YOLOv8-seg variants for SIBI alphabet recognition using instance segmentation.

Second, to enhance robustness and reduce bias from a single train–validation–test split, the study adopts a multi-scenario data partitioning approach by evaluating four different data distribution scenarios, thereby offering broader evidence of model stability and generalization capability.

Third, the study emphasizes a segmentation-oriented evaluation for fine-grained gesture recognition, demonstrating that instance segmentation with pixel-level hand masking improves the model’s focus on detailed finger structures, reduces background interference, and enhances the recognition of visually similar SIBI gestures.

Fourth, it presents an accuracy–efficiency trade-off analysis for real-time deployment by not only reporting performance metrics such as Precision, Recall, F1-score, mAP50, and mAP50–95, but also systematically analyzing inference time to determine the most suitable model for real-time assistive applications, highlighting the balanced performance of YOLOv8m-seg and the high-speed advantage of YOLOv8n-seg. Fifth, the study includes error patterns and misclassification analysis through confusion matrix evaluation, providing deeper insights into performance limitations caused by hand shape similarities and dataset constraints, thereby contributing to a better understanding of model behavior. Finally, this research bridges technical benchmarking with the broader context of assistive technology by situating SIBI recognition within real-world applications, particularly in inclusive educational and healthcare settings, thus extending its relevance beyond purely algorithmic evaluation.

2. Methods

To strengthen the methodological foundation, this study reviews several relevant previous studies as comparative material and references. The results of the review indicate that previous studies have explored this topic using diverse approaches, encompassing various objects, methods, and findings. However, there are still contextual differences and limitations that warrant further exploration.

Previous research on Chinese Sign Language (CSL) utilizes a CNN architecture combined with BiLSTM and multi-head attention, which shows high effectiveness through internal cross-validation, but has not been evaluated across devices or real-world environments [29]. Studies on Indian Sign Language (ISL) based on Automatic Speech Recognition (ASR) present adequate experiments from a multimedia perspective, but limited data access makes it difficult to thoroughly evaluate background variations, lighting, and generalization capabilities [30].

In American Sign Language (ASL), several studies use an RNN (LSTM) approach combined with Leap Motion sensor-based k-NN and achieve high real-time accuracy, although validation still relies on internal cross-validation and static alphabet gestures [31]. Another approach uses an Adaptive Graph Convolutional Network (AGCN) with webcams and MediaPipe to extract distance and angle features, offering a simple and inexpensive solution, but its performance varies and is still limited to static alphabets without testing cross-subject and cross-environment robustness [32].

Research on BISINDO using MobileNetV2 combined with gamma correction shows a clear and easily replicable methodology, but it is still limited by relatively narrow data scenarios and has not undergone strong generalization validation [33]. In ASL, a 3D-CNN architecture was also proposed to capture spatiotemporal features and was reported to outperform baseline approaches, although its evaluation was still limited to small datasets and isolated scenarios [34]. Another study proposed a signer-independent CNN method on a single-frequency dataset, which helps improve generalization to new users, but does not yet accommodate temporal variations and complex real-world conditions [35].

For British Sign Language (BSL), the Multiple Instance Learning approach with Variational Noise Contrastive Estimation utilizes weak supervision from subtitles and sign language dictionaries, but its dependence on the English ecosystem and domain gaps limits cross-language generalization and scenarios without subtitles [36]. Furthermore, a wearable system based on flexible strain sensors and IMUs combined with CNNs is capable of recognizing CSL vocabulary in real-time with high accuracy, but it is still limited by data scope and laboratory conditions, so its generalization to real-world environments and user variations is not yet fully guaranteed [37].

The detection process using YOLOv8-seg, as shown in Figure 1, consists of four stages, namely Step 1 data preprocessing, Step 2 annotation, Step 3 YOLOv8-seg, and Step 4 evaluation. Each stage plays an important role in ensuring that the detection process runs smoothly, starting from preparing the images to be analyzed, adding annotations as additional information, applying the YOLOv8-seg model, and evaluating the results to ensure the accuracy and effectiveness of the system.

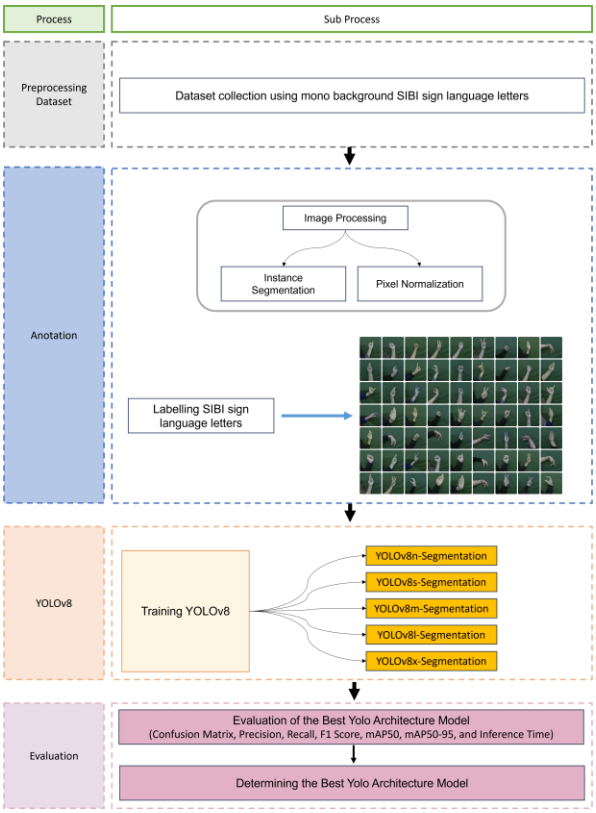


Figure 1. The proposed method

2.1 Data Preprocessing

During the preprocessing stage, one important step is the collection of SIBI sign language data used for model training, which consists of a total of 6240 images representing 24 SIBI alphabet classes (all letters except J and Z, as dynamic gestures in video form represent these two letters). Each class contains approximately 260 images, resulting in a relatively balanced class distribution. All images were captured using a mono background to maintain consistency in visual representation and to facilitate object–background separation, which is crucial for the image segmentation process. The dataset includes images of SIBI letters in various poses and hand-shape variations to enhance the model's robustness across different visual conditions.

Data distribution in this research test scenario was conducted with four variations [38]70% of the data for training, 15% for validation, and 15% for testing. The second scenario divides 70% of the data for training,

20% for validation, and 10% for testing. The third scenario allocates 75% of the data for training, 10% for validation, and 15% for testing. The fourth scenario uses 80% of the data for training, 10% for validation, and 10% for testing. This data distribution aims to evaluate the model more comprehensively, considering the balance between training, validation, and testing data to improve the model's generalization ability. This data distribution also supports a more robust approach in assessing model performance in various scenarios.

2.2 Annotation

Annotation is an important stage in the dataset preparation process, because the quality of labels directly affects the performance of detection and segmentation models [39-40]. In this study, annotations were made on SIBI sign language images to provide information on the position and shape of the hand area, which is the main object. The annotation process used an instance segmentation approach, so that each hand gesture was not only limited to a bounding box, but also equipped with a pixel mask that represented the shape of the object in more detail.

The steps taken in annotating the SIBI mono background dataset involve image processing to prepare the data to suit the model's requirements. This stage begins with the application of instance segmentation, which aims to separate the hand gesture area from the background so that the model focuses on the relevant parts. With segmentation, the system can reduce interference from unnecessary visual elements and improve detection accuracy [41-42]. Next, pixel normalization is performed, which is the process of scaling pixel values to a specific range so that the data distribution is more uniform. This normalization serves to speed up the training process, stabilize convergence, and prevent the dominance of certain pixel intensities over the training results [43-44]. The combination of instance segmentation and pixel normalization is expected to produce cleaner and more optimal data representation for the YOLOv8 segmentation model training stage.

Labeling was performed using Roboflow Annotate, a graphics-based annotation software that supports segmentation formats. Each image is labeled according to the letter categories in SIBI, for example, “A”, ‘B’, “C”, and so on, as shown in Figure 2. Labeling the SIBI dataset ensures that the dataset is well organized. This annotation process ensures that the YOLOv8-seg model can learn the unique characteristics of each gesture based on the hand area that has been separated from the background.

In addition, to maintain consistency, standards are provided in the form of annotation guidelines containing rules for determining mask boundaries, class grouping, and handling images with low quality or varying lighting conditions [45-46]. With this standard, it is hoped that subjective variations in labeling can be minimized. All annotation results are then exported into a format compatible with YOLOv8-seg, namely COCO JSON with polygon representation. This careful annotation stage is an important foundation in research as it determines how well the model is able to recognize and distinguish SIBI sign language gestures.



Figure 2. Labeling Dataset SIBI

2.3. YOLOv8 Architecture

YOLOv8 is the latest architecture from the You Only Look Once (YOLO) family, released by Ultralytics in 2023 as an improvement on the previous generation [47]. This architecture is designed to combine high inference speed with competitive detection accuracy, making it ideal for use in real-time applications such as sign language recognition. One of the main advantages of YOLOv8 is its flexibility, which supports three main computer vision tasks: object detection, instance segmentation, and image classification [48-49]. In the context of this research, the focus is on instance segmentation because it provides a more detailed visual representation of hand gestures than simple object detection with bounding boxes.

The architectural structure of YOLOv8 consists of three main components, namely the backbone, neck, and head [50-51]. The backbone extracts important features from the input image through a series of convolutions and C2f (Cross Stage Partial with Faster Aggregation) blocks, which are an extension of the CSP module in YOLOv5 [52]. The neck section uses the Path Aggregation Network (PANet) mechanism to combine features from various depth levels, thereby supporting better detection of both small and large objects [53]. Meanwhile, the head section of YOLOv8 has been updated with an anchor-free design, which allows for more adaptive detection box predictions without relying on traditional anchor box schemes [54]. The architecture of YOLOv8 can be seen in Figure 3.

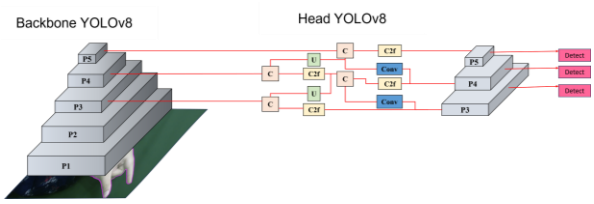


Figure 3. Architecture YOLOv8

In this study, all variants of the YOLOv8 model will be tested [55-56]. The main differences between the variants lie in the number of parameters and the architecture's complexity, as shown in Table 1. YOLOv8-seg parameters.

2.4 Evaluation

The evaluation stage was conducted to assess the performance of the YOLOv8-seg model in accurately and efficiently detecting and recognizing SIBI sign language gestures. A comprehensive evaluation was necessary so that the research results would not only show the model's accuracy level but also assess its reliability, prediction balance, and ability to be implemented in real-time applications.

Table 1. YOLOv8-seg Parameters

Model	Depth Factor	Width Factor	Number of Parameters	FLOPs (G)	Description
YOLOv8n-seg	0.33	0.25	~3.4M	~12.6	Lightweight, fast, suitable for real-time mobile segmentation.
YOLOv8s-seg	0.33	0.50	~11.8M	~42.0	More accurate, still efficient for mid-range devices.
YOLOv8m-seg	0.67	0.75	~27.3M	~110.2	Balance between segmentation accuracy and speed.
YOLOv8l-seg	1.00	1.00	~46.0M	~220.0	High accuracy requires a more powerful GPU.
YOLOv8x-seg	1.00	1.25	~71.8M	~350.0	Largest, highest accuracy, but slower inference.

One of the evaluation methods used is the Confusion Matrix, which provides information about the distribution of model predictions against actual classes [57-58]. The Confusion Matrix allows for the identification of the types of errors made by the model, such as false positives and false negatives, thereby providing a detailed overview of the model's classification capabilities. From this Confusion Matrix, two important metrics can be calculated, namely Precision and Recall. Precision measures the accuracy of positive predictions generated by the model [59], whereas Recall measures the extent to which the model successfully detects all objects in the dataset [60]. To obtain a more balanced evaluation between Precision and Recall, the F1 Score metric is used, which is the harmonic mean of the two [61]. The F1 Score is very important in this study because it can show the model's performance in dealing with the trade-off between positive and negative prediction errors.

In addition, this study also uses the Mean Average Precision (mAP) metric in two scenarios, namely mAP50 and mAP50-95. The mAP50 metric measures the average precision at an Intersection over Union (IoU) threshold of 0.5 [62], whereas mAP50-95 calculates the average precision at various IoU thresholds ranging from 0.5 to 0.95 [63]. These two metrics are used to assess the accuracy of the model in detecting gestures with varying degrees of closeness to the actual annotation. Finally, the efficiency of the model is assessed using Inference Time, which is the average time required for the model to process an image and generate a prediction [64-65]. This metric is important in the context of real-time sign language applications, where fast response is crucial for successful communication. With this combination of quantitative metrics, this study is expected to provide a comprehensive assessment of the performance of each YOLOv8-seg variant on the SIBI dataset, both in terms of accuracy and speed. The evaluation equation can be seen in Equations 1-6.

$$Confusion\ Matrix = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \quad (5)$$

$$Inference\ Time = \frac{\sum_{i=1}^{n} T_i}{n} \quad (6)$$

The evaluation metrics used in this study include Precision, Recall, and F1-score, which are derived from the confusion matrix. Precision is defined as the ratio of true positive predictions to the total predicted positives, while Recall represents the ratio of true positive predictions to all actual positive instances. The F1-score is calculated as the harmonic mean of Precision and Recall to provide a balanced evaluation. In this context, true positive (TP) refers to correctly identified hand gestures, false positive (FP) indicates incorrect positive predictions, and false negative (FN) represents instances where the model fails to detect the actual gesture.

2.5 Training Configuration and Reproducibility

All experiments were conducted on a system equipped with an Intel® Core™ Ultra 7 155H processor, 32GB RAM, and an NVIDIA GeForce RTX 4070 GPU running Windows 11. The models were trained for 50 epochs with an input resolution of 640 × 640 pixels. The default YOLOv8 optimizer (SGD/Adam as configured by Ultralytics) and learning rate scheduling were applied.

The batch size was set according to GPU memory capacity to ensure stable convergence. To enhance reproducibility, a fixed random seed was applied during training initialization. All experiments were conducted using the Ultralytics YOLOv8 framework, with consistent hyperparameter settings across model variants to ensure a fair comparison.

2.6 Application Scenario Simulation

To connect technical benchmarking to real-world needs, we designed a simulation scenario that illustrates how YOLOv8-seg, within the SIBI recognition system, could be used as an assistive communication tool. In this simulation, hand gestures captured by a standard

camera are detected and segmented by the YOLOv8 model, and the recognized sign is automatically translated into text. The text output can then be displayed on a screen or further converted into synthesized speech. This pipeline demonstrates how the proposed system could potentially support communication in inclusive educational settings or healthcare consultations, especially when professional interpreters are unavailable.

Although this study did not yet involve direct trials with deaf users or service providers, the simulation provides a conceptual bridge between technical performance and practical applications for disability inclusion. By demonstrating this potential pathway, the study situates its findings not only in the field of computer vision but also in the broader context of assistive technology for accessibility and inclusive communication.

3. Results and Discussions

3.1 Overall Performance Comparison

This study successfully implemented YOLOv8-seg for Indonesian Sign Language (SIBI) recognition, focusing on hand-area segmentation to improve gesture-detection accuracy. The training process was conducted by comparing several model variants (YOLOv8n-seg, YOLOv8s-seg, YOLOv8m-seg, YOLOv8l-seg, and YOLOv8x-seg) against predetermined scenarios, yielding the results presented in Table 2.

Table 2. Test Result for Each Scenario

Scenario	Model	Precision	Recall	F1-score	mAP50	mAP50-95	Inference Time
1	YOLOv8n-seg	0.955	0.953	0.951	0.971	0.971	6.645
	YOLOv8s-seg	0.931	0.966	0.946	0.978	0.978	12.187
	YOLOv8m-seg	0.960	0.967	0.962	0.975	0.975	28.674
	YOLOv8l-seg	0.957	0.964	0.959	0.976	0.976	48.573
	YOLOv8x-seg	0.946	0.944	0.940	0.972	0.972	75.074
2	YOLOv8n-seg	0.943	0.967	0.953	0.979	0.979	5.163
	YOLOv8s-seg	0.948	0.960	0.952	0.981	0.981	12.192
	YOLOv8m-seg	0.967	0.980	0.972	0.982	0.982	25.940
	YOLOv8l-seg	0.957	0.964	0.959	0.976	0.976	48.544
	YOLOv8x-seg	0.941	0.970	0.955	0.978	0.978	73.626
3	YOLOv8n-seg	0.943	0.923	0.925	0.975	0.975	7.107
	YOLOv8s-seg	0.947	0.918	0.926	0.965	0.965	15.451
	YOLOv8m-seg	0.956	0.928	0.935	0.969	0.969	27.887
	YOLOv8l-seg	0.945	0.928	0.931	0.958	0.958	46.706
	YOLOv8x-seg	0.933	0.953	0.939	0.969	0.969	77.107
4	YOLOv8n-seg	0.905	0.920	0.905	0.958	0.958	7.421
	YOLOv8s-seg	0.958	0.944	0.945	0.970	0.970	12.839
	YOLOv8m-seg	0.921	0.922	0.911	0.967	0.967	27.774
	YOLOv8l-seg	0.940	0.954	0.943	0.977	0.977	47.915
	YOLOv8x-seg	0.946	0.944	0.940	0.972	0.972	76.355

Based on the test results summarized in Table 2, the performance of YOLOv8-seg shows consistent variation across all scenarios and model sizes. In general, the lightweight YOLOv8n-seg model offers advantages in inference speed, making it most suitable for real-time applications with limited resources, while large-capacity models such as YOLOv8l-seg and YOLOv8x-seg tend to produce high accuracy with the consequence of longer inference times.

Among all variants, YOLOv8m-seg consistently shows the best balance between accuracy and efficiency, with stable performance across various data partitioning scenarios, making it the most optimal candidate for practical implementation. These results confirm the classic trade-off between accuracy and speed in deep learning-based systems, and show that the selection of the YOLOv8-seg model needs to be tailored to the application's requirements, whether it is oriented towards maximum accuracy or real-time response.

3.2 Average Performance and Robustness Analysis

To provide a more comprehensive evaluation of model stability beyond individual-scenario results, the average performance across the four data-partitioning configurations was computed. This aggregation enables a clearer assessment of robustness and generalization capability under varying training proportions.

Table 3 presents the average F1-score and average inference time for each YOLOv8-seg variant. The results indicate that YOLOv8l-seg achieved the highest average F1-score (0.948), closely followed by YOLOv8m-seg (0.945) and YOLOv8x-seg (0.944). The relatively small performance gap among medium and large variants suggests that increasing model capacity beyond a certain threshold yields diminishing returns in recognition accuracy under mono-background conditions.

Table 3. Average Performance of YOLOv8-seg Variants Across Four Data Partitioning Scenarios

Model	Average F1-score	Average Inference Time (ms)
YOLOv8n-seg	0.934	6.83
YOLOv8s-seg	0.942	13.17
YOLOv8m-seg	0.945	27.57
YOLOv8l-seg	0.948	47.93
YOLOv8x-seg	0.944	75.54

In contrast, YOLOv8n-seg recorded the lowest average F1-score (0.934), yet maintained a significantly lower inference time (approximately 6–7 ms). This highlights

the inherent trade-off between accuracy and computational efficiency. While lightweight models sacrifice a small margin of accuracy, they offer substantial gains in processing speed, which may be critical for real-time assistive communication systems.

The aggregated results are further illustrated in Figure 4, where the bar graph visually emphasizes the marginal differences in average F1-score across model variants and the substantial increase in inference time with model scale. The consistency of F1-scores across splits indicates that segmentation-based learning contributes to stable hand-region feature extraction, reducing sensitivity to variations in training data distribution.

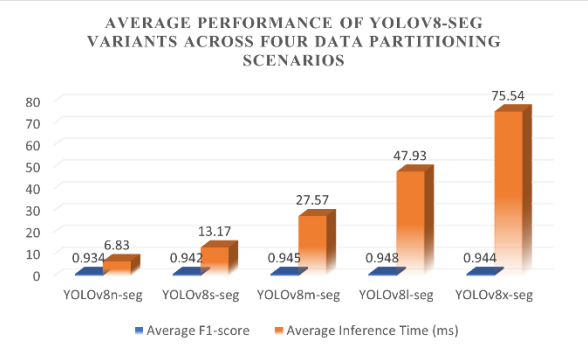


Figure 4. Diagram Average Performance

Overall, the cross-scenario aggregation confirms that medium-capacity architectures, particularly YOLOv8m-seg and YOLOv8l-seg, provide a favorable balance between robustness and efficiency. These findings strengthen the reliability of the experimental results and support informed model selection for deployment-oriented SIBI recognition systems.

3.3 Model Capacity and Performance Interpretation

The observed performance differences among YOLOv8-seg variants can be further explained by analyzing model capacity and architectural complexity. Each variant differs in depth factor, width factor, parameter count, and computational cost (FLOPs), which directly influence feature extraction capability and inference time.

Lightweight models such as YOLOv8n-seg (approximately 3.4 million parameters) prioritize computational efficiency. While this architecture enables extremely fast inference, its limited representational capacity may restrict its ability to capture subtle finger articulations and fine-grained contour variations. This explains the slightly lower average F1-score observed across scenarios.

In contrast, large-capacity models such as YOLOv8l-seg and YOLOv8x-seg contain significantly more parameters (approximately 46M and 71.8M, respectively), enabling deeper feature representation. However, under the controlled mono-background dataset used in this study, the additional capacity does not yield proportional improvements in segmentation accuracy. The relatively small performance gap among YOLOv8m-seg, YOLOv8l-seg, and YOLOv8x-seg indicates diminishing returns: increasing architectural complexity leads to substantial increases in inference time without substantial gains in recognition performance.

YOLOv8m-seg appears to represent a balanced configuration, offering sufficient feature richness to discriminate subtle hand-shape differences while maintaining manageable computational overhead. This balance explains its consistent performance across multiple data partitioning scenarios and supports its suitability for deployments that require both robustness and real-time feasibility.

These findings highlight that model selection for SIBI recognition should not be driven solely by maximum accuracy metrics, but also by computational constraints and dataset characteristics. Under controlled mono-background conditions, medium-capacity segmentation models provide an optimal trade-off between representational power and efficiency.

3.4 Segmentation vs Detection Ablation Comparison

To provide a clearer methodological progression, this study compares three experimental settings [66]: (1) detection without augmentation, (2) detection with data augmentation, and (3) instance segmentation with multi-split evaluation. This comparison illustrates the additional impact of increased data diversity and architectural improvements on the performance of SIBI recognition, as shown in Table 4.

Table 4. Comparative Performance Across Experimental Settings

Approach	Dataset Size	Evaluation Strategy	Best/Avg F1	mAP50	Inference times
Detection (Non-Aug)	1440	Single split	0.819	0.987	21.62
Detection (Aug)	3465	Single split	0.98	0.988	9.82
Segmentation	6240	Multi-split	0.945	0.982	27.57

The non-augmentation detection model achieved an F1-score of 0.819, indicating moderate generalization under limited data variability. When data augmentation was introduced, the F1-score increased substantially to 0.98. This corresponds to an approximate 19.7% relative improvement. This result confirms that synthetic data variation significantly enhances model robustness under controlled mono-background conditions.

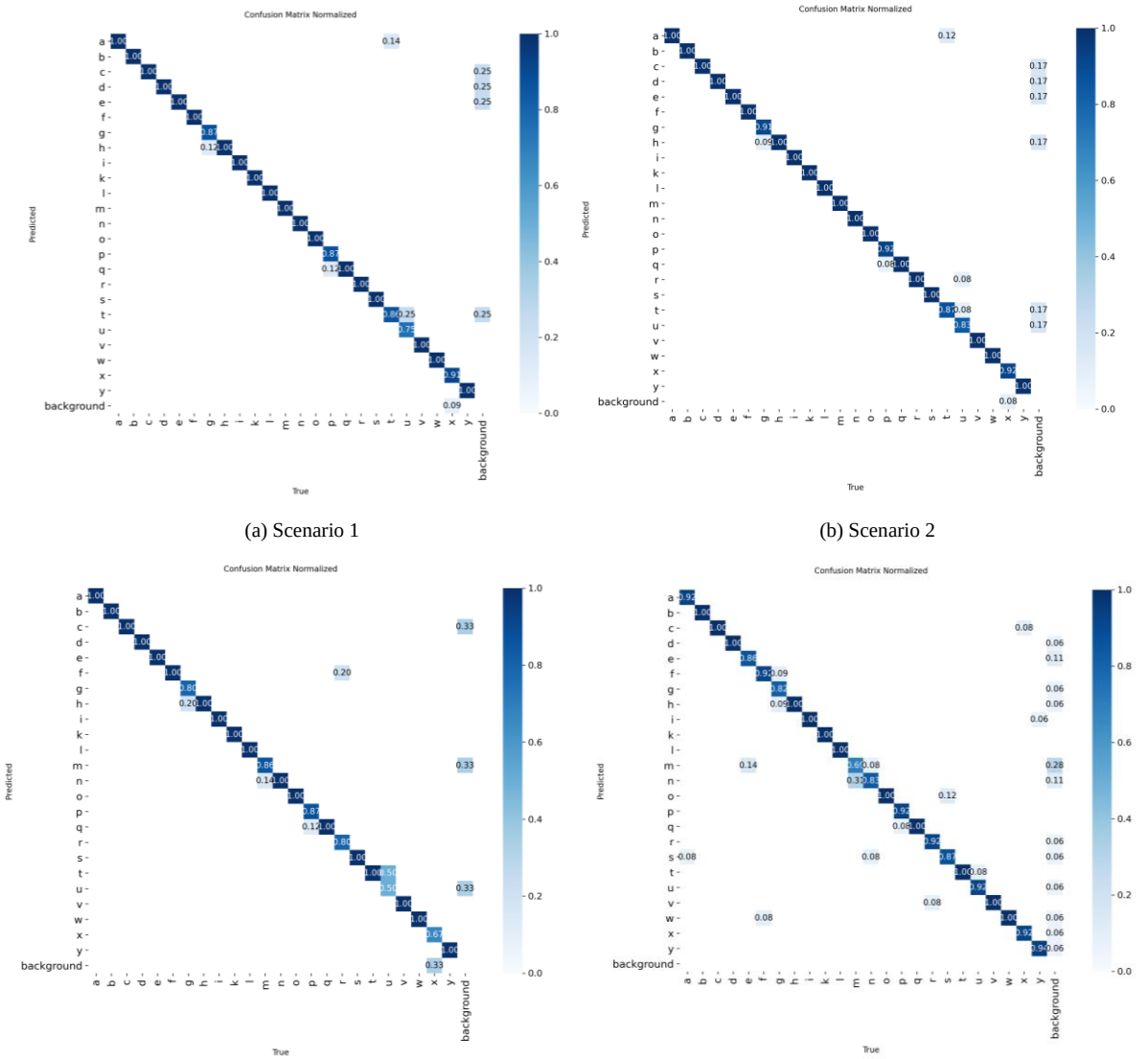
When comparing the augmented detection model with the segmentation-based framework, the segmentation model achieves an average F1 Score of 0.945 across multiple split scenarios. Compared with the peak augmented detection performance (0.98), this represents a modest 3.6% decrease. However, this comparison must be interpreted carefully. The segmentation framework was evaluated on a dataset more than four times larger (6,240 images) and across multiple train–validation–test configurations. These stricter evaluation conditions reduce dependence on a single favorable data partition and provide a more

comprehensive assessment of generalization stability. Importantly, compared with detection without augmentation, the segmentation approach achieves an overall 15.4% relative improvement in F1 Score. Beyond numerical performance, segmentation enables pixel-level mask prediction, allowing finer modeling of hand contours and improved interpretability. While detection with augmentation achieves high peak accuracy under controlled conditions, segmentation strengthens methodological rigor by using larger datasets, representing spatial masks, and evaluating cross-scenario robustness.

Overall, the comparison highlights three stages of improvement: performance gains from data augmentation, enhanced spatial representation through instance segmentation, and robustness validation via multi-split benchmarking. Thus, the contribution of this study lies not only in accuracy but also in methodological strengthening and deployment-oriented evaluation for scalable SIBI recognition systems.

3.5 Error and Misclassification Analysis

Although the overall quantitative metrics indicate high recognition performance, a deeper examination through confusion matrix analysis reveals several systematic misclassification patterns. Figure 5 presents the confusion matrix of the best-performing model (YOLOv8m-seg), which provides class-wise prediction distribution across the SIBI alphabet dataset.



(a) Scenario 1

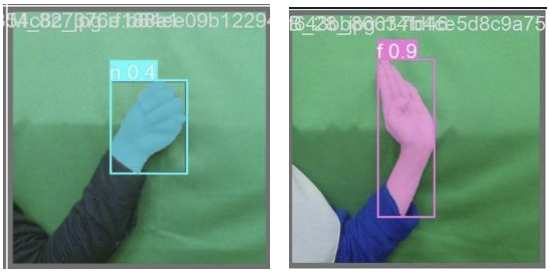
(b) Scenario 2

(c) Scenario 3 (d) Scenario 4

Figure 5. Confusion Matrix

The confusion matrix shows that most gesture classes achieve near-perfect classification, indicating strong discriminative capability under mono-background conditions. However, a small number of visually similar gestures occasionally misclassify. These errors primarily occur between gesture pairs that differ only in subtle finger articulation or minimal contour variation. For example, “M” and “N” differ mainly in slight finger

folding configurations. It can be difficult to distinguish when segmentation boundaries are imperfect. From a spatial modeling perspective, these misclassifications are attributable to fine-grained contour ambiguity rather than global detection failure. In several instances, partial finger overlap or slight hand rotation led to minor irregularities in the segmentation mask, which influenced class prediction. This observation is further illustrated in Figure 6, which shows example detection outputs that include both correct and incorrect predictions. While the segmentation masks effectively isolate the hand region from the background, small boundary distortions can alter the perceived finger structure.



(a) m is read n (b) b is read f

Figure 6. Detection Error

The error distribution suggests that the remaining inaccuracies stem from intrinsic gesture similarity rather than background interference. This indicates that the segmentation approach successfully minimizes external noise, yet distinguishing highly similar static gestures remains a challenging fine-grained classification problem.

These findings highlight potential directions for improvement. Incorporating higher-resolution input images, employing advanced augmentation that emphasizes subtle finger articulation, or integrating temporal modeling for dynamic gesture transitions may further reduce ambiguity. Nevertheless, the relatively low frequency of misclassification confirms that the YOLOv8-seg framework provides robust spatial feature learning under controlled experimental conditions.

3.6 Computational Efficiency and Runtime Analysis

In addition to segmentation accuracy, computational efficiency plays a critical role in the practical deployment of SIBI recognition systems. The YOLOv8-seg variants differ in parameter count and floating-point operations (FLOPs), which directly influence inference time and hardware requirements. The results show a non-linear increase in computational cost as model capacity grows: YOLOv8n-seg achieves an average inference time of about 6–7 ms per image (≈140–160 FPS), YOLOv8m-seg operates at around 27–28 ms (≈35–37 FPS), and YOLOv8x-seg exceeds 75 ms (≈13 FPS). Although all models remain capable of real-time processing (>10 FPS), the accuracy gains of larger models are relatively modest compared to the increase in runtime, reflecting the diminishing-return behavior discussed in Section 3.3.

From a deployment perspective, lightweight models such as YOLOv8n-seg are suitable for embedded or mobile environments with limited computational resources, while medium-capacity models such as YOLOv8m-seg provide a more balanced trade-off between segmentation accuracy and real-time performance. Larger variants (YOLOv8l-seg and YOLOv8x-seg) may be more appropriate for GPU-based systems prioritizing maximum accuracy. Overall, these results highlight that selecting a model for SIBI recognition should consider both segmentation performance and computational feasibility to support responsive assistive communication systems.

3.7 Deployment and Practical Implications

In addition to technical benchmarking, this study briefly considers the practical implications of deploying SIBI recognition systems in real-world settings. An exploratory stakeholder inquiry involving teachers in inclusive schools, university students from the deaf community, and healthcare professionals was conducted to understand perceived communication barriers, expectations toward AI-based sign language recognition tools, and potential institutional adoption challenges.

The responses indicate strong potential for such systems to reduce communication gaps in educational and healthcare environments, while also highlighting concerns related to data privacy, misclassification risks, and the need for user training prior to large-scale deployment. Although this study did not include controlled trials with deaf participants, future work will involve formal ethics approval and usability evaluation using standardized instruments (e.g., System Usability Scale or UMUX-LITE). These considerations suggest that SIBI recognition should be approached not only as a computer vision task but also as an assistive technology requiring responsible implementation and institutional support for sustainable adoption.

4. Conclusions

This study presented a comparative analysis of YOLOv8 segmentation variants for Indonesian Sign Language (SIBI) alphabet recognition under multiple data partitioning scenarios. The experimental results show that segmentation-based YOLOv8 models effectively capture fine-grained hand features required for accurate SIBI recognition. Among the evaluated variants, YOLOv8m-seg consistently provided the best balance between accuracy and inference efficiency, while YOLOv8n-seg demonstrated superior inference speed, making it suitable for real-time applications on resource-constrained devices.

Larger models such as YOLOv8l-seg and YOLOv8x-seg achieved high accuracy but were limited by longer inference times. The findings confirm the importance of instance segmentation in reducing background interference and improving recognition robustness. From a technical perspective, this work provides

practical guidance for selecting YOLOv8 segmentation models based on accuracy–efficiency trade-offs. Although the primary contribution of this study is technical, the results indicate potential applicability in assistive communication systems. Future work will focus on extending the dataset to dynamic gestures and more diverse environments to improve real-world robustness.

Acknowledgements

The authors would like to thank all parties who contributed to this research, particularly those who provided support during data preparation, experimentation, and manuscript development. Any opinions, findings, and conclusions expressed in this article are those of the authors and do not necessarily reflect the views of any affiliated institutions.

References

[1] B. Joksimoski *et al.*, "Technological Solutions for Sign Language Recognition: A Scoping Review of Research Trends, Challenges, and Opportunities," *IEEE Access*, vol. 10, pp. 40979–40998, 2022, doi: 10.1109/ACCESS.2022.3161440.

[2] B. A. Al Abdullah, G. A. Amoudi, and H. S. Alghamdi, "Advancements in Sign Language Recognition: A Comprehensive Review and Future Prospects," *IEEE Access*, vol. 12, pp. 128871–128895, 2024, doi: 10.1109/ACCESS.2024.3457692.

[3] D. Lillo-Martin and J. A. Hochgesang, "Signed languages – Unique and ordinary: A commentary on Kidd and Garcia (2022)," *First Lang.*, vol. 42, no. 6, pp. 789–793, Dec. 2022, doi: 10.1177/01427237221098858 .

[4] L. Ferrara *et al.*, "A cross-linguistic comparison of reference across five signed languages," *Linguistic Typology*, vol. 27, no. 3, pp. 591–627, Oct. 2023, doi: 10.1515/lingty-2021-0057.

[5] S. Hassan *et al.*, "ASL-Homework-RGBD Dataset: An annotated dataset of 45 fluent and non-fluent signers performing American Sign Language homeworks," Jul. 2022, [Online]. Available: http://arxiv.org/abs/2207.04021

[6] G. E. Shields, K. D. Rogers, A. Young, S. Dedotsi, and L. M. Davies, "Health State Values of Deaf British Sign Language (BSL) Users in the UK: An Application of the BSL Version of the EQ-5D-5L," *Appl. Health Econ. Health Policy*, vol. 18, no. 4, pp. 547–556, Aug. 2020, doi: 10.1007/s40258-019-00546-8.

[7] Y. K. Suyitno, I. W. Sudiarsa, I. N. B. Hartawan, and I. D. P. G. W. Putra, "Implementation of Sibi Sign Language Realtime Detection Program (Case Studi At Sekolah Luar Biasa Negeri 1 Tabanan)," *Journal of Computer Networks, Architecture and High Performance Computing*, vol. 6, no. 3, pp. 1431–1441, Jul. 2024, doi: 10.47709/cnahpc.v6i3.4405.

[8] J. S. Izquierdo-Condoy *et al.*, "Exploring healthcare barriers and satisfaction levels among deaf individuals in Ecuador: A video-based survey approach," *Disabil. Health J.*, vol. 17, no. 3, p. 101622, Jul. 2024, doi: 10.1016/j.dhjo.2024.101622.

[9] J. A. Manganello, "Shining a light on disability and health communication," *J. Commun. Healthc.*, vol. 15, no. 4, pp. 309–312, Oct. 2022, doi: 10.1080/17538068.2022.2142443.

[10] H. F. da Silva, J. F. da Silva, V. V. Cruz, N. M. A. de Figueiredo, A. C. S. Sé, and W. C. A. Machado, "Aplicabilidade da Tecnologia Assistiva para pessoas com lesão medular: Uma revisão integrativa," *Research, Society and Development*, vol. 10, no. 8, p. e0210815387, Jul. 2021, doi: 10.33448/rsd-v10i8.15387.

[11] N. G. A. P. Mayuni and I. A. O. Martini, "Bridging Health and Sustainability: Evaluating Siloam Hospital Denpasar's Role in Advancing Medical Tourism Aligned with SDG 3," *Journal of Lifestyle and SDGs Review*, vol. 5, no. 1, p. e04323, Jan. 2025, doi: 10.47172/2965-730X.SDGsReview.v5.n01.pe04323.

[12] M. P. Bhandari, "What is next for the sustainable development goals, what are the challenges concerning SDG 10—reduced inequalities?," *Sustainable Earth Reviews*, vol. 7, no. 1, p. 23, Jul. 2024, doi: 10.1186/s42055-024-00093-8.

[13] M. Yu, J. Jia, C. Xue, G. Yan, Y. Guo, and Y. Liu, "A review of sign language recognition research," *Journal of Intelligent & Fuzzy Systems*, vol. 43, no. 4, pp. 3879–3898, Aug. 2022, doi: 10.3233/JIFS-210050.

[14] B. A. Al Abdullah, G. A. Amoudi, and H. S. Alghamdi, "Advancements in Sign Language Recognition: A Comprehensive Review and Future Prospects," *IEEE Access*, vol. 12, pp. 128871–128895, 2024, doi: 10.1109/ACCESS.2024.3457692.

[15] H. ZainEldin *et al.*, "Silent no more: a comprehensive review of artificial intelligence, deep learning, and machine learning in facilitating deaf and mute communication," *Artif. Intell. Rev.*, vol. 57, no. 7, p. 188, Jun. 2024, doi: 10.1007/s10462-024-10816-0.

[16] Y. Liu, P. Nand, M. A. Hossain, M. Nguyen, and W. Q. Yan, "Sign language recognition from digital videos using feature pyramid network with detection transformer," *Multimed. Tools Appl.*, vol. 82, no. 14, pp. 21673–21685, Jun. 2023, doi: 10.1007/s11042-023-14646-0.

[17] Y. Zhou, "A YOLO-NL object detector for real-time detection," *Expert Syst. Appl.*, vol. 238, p. 122256, Mar. 2024, doi: 10.1016/j.eswa.2023.122256.

[18] M. Sohaib, M. Arif, and J.-M. Kim, "Evaluating YOLO Models for Efficient Crack Detection in Concrete Structures Using Transfer Learning," *Buildings*, vol. 14, no. 12, p. 3928, Dec. 2024, doi: 10.3390/buildings14123928.

[19] J. Ma, Y. Zhao, W. Fan, and J. Liu, "An Improved YOLOv8 Model for Lotus Seedpod Instance Segmentation in the Lotus Pond Environment," *Agronomy*, vol. 14, no. 6, p. 1325, Jun. 2024, doi: 10.3390/agronomy14061325.

[20] A. Wang *et al.*, "NVW-YOLOv8s: An improved YOLOv8s network for real-time detection and segmentation of tomato fruits at different ripeness stages," *Comput. Electron. Agric.*, vol. 219, p. 108833, Apr. 2024, doi: 10.1016/j.compag.2024.108833.

[21] M. Hussain, "YOLOv1 to v8: Unveiling Each Variant–A Comprehensive Review of YOLO," *IEEE Access*, vol. 12, pp. 42816–42833, 2024, doi: 10.1109/ACCESS.2024.3378568.

[22] H.-I. Liu *et al.*, "Lightweight Deep Learning for Resource-Constrained Environments: A Survey," *ACM Comput. Surv.*, vol. 56, no. 10, pp. 1–42, Oct. 2024, doi: 10.1145/3657282.

[23] Q. Wang, G. Feng, and Z. Li, "A Lightweight Person Detector for Surveillance Footage Based on YOLOv8n," *Sensors*, vol. 25, no. 2, p. 436, Jan. 2025, doi: 10.3390/s25020436.

[24] Y. Liu, C. Zhang, W. Wu, B. Zhang, and F. Zhou, "MiniYOLO: A lightweight object detection algorithm that realizes the trade-off between model size and detection accuracy," *International Journal of Intelligent Systems*, vol. 37, no. 12, pp. 12135–12151, Dec. 2022, doi: 10.1002/int.23079.

[25] B. Zhao, Y. Zhou, R. Song, L. Yu, X. Zhang, and J. Liu, "Modular YOLOv8 optimization for real-time UAV maritime rescue object detection," *Sci. Rep.*, vol. 14, no. 1, p. 24492, Oct. 2024, doi: 10.1038/s41598-024-75807-1.

[26] A. N. Handayani, S. Amaliya, M. I. Akbar, M. Z. Wiryawan, Y. W. Liang, and W. C. Kurniawan, "Hand Keypoint-Based CNN for SIBI Sign Language Recognition," *International Journal of Robotics and Control Systems*, vol. 5, no. 2, pp. 813–829, 2025, doi: 10.31763/ijrcs.v5i2.1745.

[27] Suharjito, N. Thiracitta, and H. Gunawan, "SIBI Sign Language Recognition Using Convolutional Neural Network Combined with Transfer Learning and non-trainable Parameters," *Procedia Comput. Sci.*, vol. 179, pp. 72–80, 2021, doi: 10.1016/j.procs.2020.12.011.

[28] A. Z. Nugraha, R. F. Salsabila, A. N. Handayani, A. P. Wibawa, E. Hitipeuw, and K. Arai, "Decision tree based algorithms for Indonesian Language Sign System (SIBI) recognition," *Applied Engineering and Technology*, vol. 3, no. 2, pp. 86–101, Aug. 2024, doi: 10.31763/aet.v3i2.1536.

[29] W. Hao, C. Hou, Z. Zhang, X. Zhai, L. Wang, and G. Lv, "A sensing data and deep learning-based sign language recognition approach," *Computers and Electrical Engineering*, vol. 118, p. 109339, Aug. 2024, doi: 10.1016/j.compeleceng.2024.109339.

[30] A. S. Dhanjal and W. Singh, "An optimized machine translation technique for multi-lingual speech to sign language notation,"

Multimed. Tools Appl., vol. 81, no. 17, pp. 24099–24117, Jul. 2022, doi: 10.1007/s11042-022-12763-w.
[31] C. K. M. Lee, K. K. H. Ng, C.-H. Chen, H. C. W. Lau, S. Y. Chung, and T. Tsoi, "American sign language recognition and training method with recurrent neural network," *Expert Syst. Appl.*, vol. 167, p. 114403, Apr. 2021, doi: 10.1016/j.eswa.2020.114403.
[32] J. Shin, A. Matsuoka, Md. A. M. Hasan, and A. Y. Srizon, "American Sign Language Alphabet Recognition by Extracting Feature from Hand Pose Estimation," *Sensors*, vol. 21, no. 17, p. 5856, Aug. 2021, doi: 10.3390/s21175856.
[33] L. N. Hayati *et al.*, "Improving Indonesian Sign Alphabet Recognition for Assistive Learning Robots Using Gamma-Corrected MobileNetV2 Lilis Nur Hayati (Improving Indonesian Sign Alphabet Recognition for Assistive Learning Robots Using Gamma-Corrected MobileNetV2)," *Buletin Ilmiah Sarjana Teknik Elektro*, vol. 7, no. 3, pp. 350–361, 2025, doi: 10.12928/biste.v7i3.13300.
[34] S. Sharma and K. Kumar, "ASL-3DCNN: American sign language recognition technique using 3-D convolutional neural networks," *Multimed. Tools Appl.*, vol. 80, no. 17, pp. 26319–26331, Jul. 2021, doi: 10.1007/s11042-021-10768-5.
[35] T. Liu, T. Tao, Y. Zhao, M. Li, and J. Zhu, "A signer-independent sign language recognition method for the single-frequency dataset," *Neurocomputing*, vol. 582, p. 127479, May 2024, doi: 10.1016/j.neucom.2024.127479.
[36] G. Varol, L. Momeni, S. Albanie, T. Afouras, and A. Zisserman, "Scaling Up Sign Spotting Through Sign Language Dictionaries," *Int. J. Comput. Vis.*, vol. 130, no. 6, pp. 1416–1439, Jun. 2022, doi: 10.1007/s11263-022-01589-6.
[37] Y. Liu *et al.*, "A wearable system for sign language recognition enabled by a convolutional neural network," *Nano Energy*, vol. 116, p. 108767, Nov. 2023, doi: 10.1016/j.nanoen.2023.108767.
[38] R. Wulanningrum, A. N. Handayani, and H. W. Herwanto, "Comparative Analysis of Yolov-8 Segmentation for Gait Performance in Individuals with Lower Limb Disabilities," *International Journal of Robotics and Control Systems*, vol. 5, no. 1, pp. 516–529, Jan. 2025, doi: 10.31763/ijrcs.v5i1.1731.
[39] C. Agnew, C. Eising, P. Denny, A. Scanlan, P. Van De Ven, and E. M. Grua, "Quantifying the Effects of Ground Truth Annotation Quality on Object Detection and Instance Segmentation Performance," *IEEE Access*, vol. 11, pp. 25174–25188, 2023, doi: 10.1109/ACCESS.2023.3256723.
[40] L. He, Y. Zhou, L. Liu, and J. Ma, "Research and Application of YOLOv11-Based Object Segmentation in Intelligent Recognition at Construction Sites," *Buildings*, vol. 14, no. 12, p. 3777, Nov. 2024, doi: 10.3390/buildings14123777.
[41] F. Tan, M. Zhai, and C. Zhai, "Foreign object detection in urban rail transit based on deep differentiation segmentation neural network," *Heliyon*, vol. 10, no. 17, p. e37072, Sep. 2024, doi: 10.1016/j.heliyon.2024.e37072.
[42] T. Li, J. Bai, and Q. Wang, "Enhancing medical text detection with vision-language pre-training and efficient segmentation," *Complex & Intelligent Systems*, vol. 10, no. 3, pp. 3995–4007, Jun. 2024, doi: 10.1007/s40747-024-01378-3.
[43] X. Yuan, X. Shen, S. Mehta, T. Li, S. Ge, and Z. Zha, "Structure injected weight normalization for training deep networks," *Multimed. Syst.*, vol. 28, no. 2, pp. 433–444, Apr. 2022, doi: 10.1007/s00530-021-00793-7.
[44] H. Lai, L. Zhang, and S. Zhang, "Improving Network Training on Resource-Constrained Devices via Habituation Normalization," *Sensors*, vol. 22, no. 24, p. 9940, Dec. 2022, doi: 10.3390/s22249940.
[45] R. Jalaboi, O. Winther, and A. Galimzianova, "Explainable Image Quality Assessments in Teledermatological Photography," *Telemedicine and e-Health*, vol. 29, no. 9, pp. 1342–1348, Sep. 2023, doi: 10.1089/tmj.2022.0405.
[46] T. Rädsch *et al.*, "Labelling instructions matter in biomedical image analysis," *Nat. Mach. Intell.*, vol. 5, no. 3, pp. 273–283, Mar. 2023, doi: 10.1038/s42256-023-00625-5.
[47] E. Bellou, I. Pisica, and K. Banitsas, "Aerial Inspection of High-Voltage Power Lines Using YOLOv8 Real-Time Object Detector," *Energies (Basel).*, vol. 17, no. 11, p. 2535, May 2024, doi: 10.3390/en17112535.
[48] S. Qi, X. Song, T. Shang, X. Hu, and K. Han, "MSFE-YOLO: An Improved YOLOv8 Network for Object Detection on Drone View," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024, doi: 10.1109/LGRS.2024.3432536.
[49] Q. Zhou, Z. Wang, Y. Zhong, F. Zhong, and L. Wang, "Efficient Optimized YOLOv8 Model with Extended Vision," *Sensors*, vol. 24, no. 20, p. 6506, Oct. 2024, doi: 10.3390/s24206506.
[50] M. Yue, L. Zhang, Y. Zhang, and H. Zhang, "An Improved YOLOv8 Detector for Multi-Scale Target Detection in Remote Sensing Images," *IEEE Access*, vol. 12, pp. 114123–114136, 2024, doi: 10.1109/ACCESS.2024.3444606.
[51] X. Zhao, W. Zhang, H. Zhang, C. Zheng, J. Ma, and Z. Zhang, "ITD-YOLOv8: An Infrared Target Detection Model Based YOLOv8 for Unmanned Aerial Vehicles," *Drones*, vol. 8, no. 4, p. 161, Apr. 2024, doi: 10.3390/drones8040161.
[52] X. Zuo, Z. Shao, J. Wang, X. Huang, and Y. Wang, "A cross-stage features fusion network for building extraction from remote sensing images," *Geo-spatial Information Science*, vol. 28, no. 2, pp. 387–401, Mar. 2025, doi: 10.1080/10095020.2024.2307922.
[53] Z. Zhang, X. Qiu, and Y. Li, "SEFPN: Scale-Equalizing Feature Pyramid Network for Object Detection," *Sensors*, vol. 21, no. 21, p. 7136, Oct. 2021, doi: 10.3390/s21217136.
[54] H. Wang, C. Liu, Y. Cai, L. Chen, and Y. Li, "YOLOv8-QSD: An Improved Small Object Detection Algorithm for Autonomous Vehicles Based on YOLOv8," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–16, 2024, doi: 10.1109/TIM.2024.3379090.
[55] N. U. A. Tahir, Z. Long, Z. Zhang, M. Asim, and M. ELAffendi, "PVswin-YOLOv8s: UAV-Based Pedestrian and Vehicle Detection for Traffic Management in Smart Cities Using Improved YOLOv8," *Drones*, vol. 8, no. 3, p. 84, Feb. 2024, doi: 10.3390/drones8030084.
[56] L. Ramos, E. Casas, E. Bendek, C. Romero, and F. Rivas-Echeverría, "Hyperparameter optimization of YOLOv8 for smoke and wildfire detection: Implications for agricultural and environmental safety," *Artificial Intelligence in Agriculture*, vol. 12, pp. 109–126, Jun. 2024, doi: 10.1016/j.aiia.2024.05.003.
[57] E. Asci *et al.*, "A Deep Learning Approach to Automatic Tooth Caries Segmentation in Panoramic Radiographs of Children in Primary Dentition, Mixed Dentition, and Permanent Dentition," *Children*, vol. 11, no. 6, p. 690, Jun. 2024, doi: 10.3390/children11060690.
[58] A. Hinterreiter *et al.*, "ConfusionFlow: A Model-Agnostic Visualization for Temporal Analysis of Classifier Confusion," *IEEE Trans. Vis. Comput. Graph.*, vol. 28, no. 2, pp. 1222–1236, Feb. 2022, doi: 10.1109/TVCG.2020.3012063.
[59] M. Owusu-Adjei, J. Ben Hayfron-Acquah, T. Frimpong, and G. Abdul-Salaam, "A systematic review of prediction accuracy as an evaluation measure for determining machine learning model performance in healthcare systems," Jun. 04, 2023. doi: 10.1101/2023.06.01.23290837.
[60] Z. Mawardi, D. Gautam, and T. G. Whiteside, "Utilization of Remote Sensing Dataset and a Deep Learning Object Detection Model to Map Siam Weed Infestations," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 17, pp. 18939–18948, 2024, doi: 10.1109/JSTARS.2024.3465554.
[61] R. Srilatha, Y. D. G, Y. Perumalla, D. S. Yaganti, and S. Yegoti, "Machine Learning Approaches in Financial Fraud Detection: Performance Evaluation and Statistical Insights," in *Digital Crossroads: Integrating Humanities, Science and Technology Edition 1*, BP International, 2025, pp. 47–61. doi: 10.9734/bpi/mono/978-93-48859-10-5/CH5.
[62] H. Ma, W. Zhao, B. Liu, and W. Chen, "Multi-Scale Target Detection in Autonomous Driving Scenarios Based on YOLOv5-AFAM," *Applied Sciences*, vol. 14, no. 11, p. 4633, May 2024, doi: 10.3390/app14114633.
[63] H. Peng and S. Yu, "A Systematic IoU-Related Method: Beyond Simplified Regression for Better Localization," *IEEE Transactions on Image Processing*, vol. 30, pp. 5032–5044, 2021, doi: 10.1109/TIP.2021.3077144.
[64] S. Gao, "Exploration and evaluation of faster R-CNN-based pedestrian detection techniques," *Applied and Computational Engineering*, vol. 32, no. 1, pp. 185–190, Jan. 2024, doi: 10.54254/2755-2721/32/20230208.
[65] C. Tian, Z. Liu, H. Chen, F. Dong, X. Liu, and C. Lin, "A Lightweight Model for Shine Muscat Grape Detection in

Complex Environments Based on the YOLOv8 Architecture," *Agronomy*, vol. 15, no. 1, p. 174, Jan. 2025, doi: 10.3390/agronomy15010174.

[66] D. F. Azizah, A. N. Handayani, and A. P. Wibawa, "YOLOv8 Performance for SIBI Sign Language Mono Background Detection," in *2025 International Conference on Computer Engineering, Network and Intelligent Multimedia (CENIM)*, IEEE, Nov. 2025, pp. 185–190. doi: 10.1109/CENIM67940.2025.11326292.