

# End-to-end system for translating *bahasa isyarat Indonesia* sign language gestures into Indonesian text

Satria Putra, Erdefi Rakun

Department of Computer Science, Faculty of Computer Science, University of Indonesia, Depok, Indonesia

## Article Info

### Article history:

Received Dec 10, 2024

Revised Jul 7, 2025

Accepted Oct 14, 2025

### Keywords:

*Bahasa isyarat Indonesia*

Object detection

Sign language translation

Threshold conditional random fields

YOLOv7 YOLOv8

## ABSTRACT

This study addresses critical challenges in developing an end-to-end *bahasa isyarat Indonesia* (BISINDO) SLT by integrating advanced deep learning techniques to overcome complex background interference, transitional gesture recognition, and limitations in dataset availability. While existing SLT systems struggle with isolated word recognition and manual preprocessing, our work introduces three key innovations: (1) implementation of YOLOv8 for optimized object detection, achieving 88% mAP and reducing WER to 11.40%, outperforming YOLOv5/v7 in handling complex backgrounds; (2) automated removal of transitional gestures using Threshold conditional random fields (TCRF), which attained 95.68% accuracy, significantly improving upon MobileNetV2's performance (WER: 6.89% vs. 93.53%); and (3) end-to-end BISINDO SLT by expansion of the BISINDO dataset to 435 word labels, enabling comprehensive sentence-level translation. Experimental results demonstrate the system's robustness, with 8.31% of WER, 84.13% of SAcc, and 87.08% of SacreBLEU after dataset expansion and redundancy elimination through grouping methods. The proposed framework operates without manual intervention, marking a substantial advancement toward real-world applicability.

This is an open access article under the [CC BY-SA](#) license.



## Corresponding Author:

Erdefi Rakun

Department of Computer Science, Faculty of Computer Science, University of Indonesia

16424 Depok, Indonesia

Email: efi@cs.ui.ac.id

## 1. INTRODUCTION

Developing inclusive communication tools for the deaf community represents an urgent necessity, as they face complex challenges in daily interactions. Hearing impairment, whether congenital or acquired [1], often forces deaf individuals to rely primarily on visual communication methods, such as sign language, for effective communication [2]. However, these methods encounter significant obstacles, including limited public understanding and a lack of practical communication aids. Current sign language translation (SLT) solutions based on CNNs [3]–[8], remain limited to recognizing individual letters/words without full sentence translation capabilities. These limitations underscore the crucial need for an end-to-end system that can translate complete sentences in real-time across various conditions without requiring human intervention. Implementing such a solution would be transformative for the deaf community, enabling improved access to education and employment, facilitating social interactions, and promoting greater societal inclusion. This advanced system would not only overcome existing solution shortcomings but also establish a more natural communication bridge between the deaf community and the hearing society.

Two main sign languages are used in Indonesia: *Sistem isyarat Bahasa Indonesia* (SIBI) and *bahasa isyarat Indonesia* (BISINDO). SIBI, the official sign language recognized by law, follows the grammatical structure of spoken Indonesian. In contrast, BISINDO has evolved organically within deaf communities,

featuring a more intuitive grammar and regional variations [9]. Despite its widespread use, BISINDO remains under-researched compared to SIBI, highlighting a critical gap in the literature. Hence, this research aims to develop an end-to-end application to translate BISINDO sign language into a sentence in real-world conditions.

SLT presents three significant challenges. First, transitional gestures, which are words in a sentence that carry no specific meaning, can be challenging to recognize due to their variability, which reduces translation accuracy. For example, in the BISINDO sentence "*Dia Lari Karena Takut*" ("He runs because he is scared"), these meaningless transitions occur at the start, between words, and at the end (Figure 1). Since they act as irrelevant inputs in continuous signing, automated removal of such gestures can enhance the system's word recognition performance.

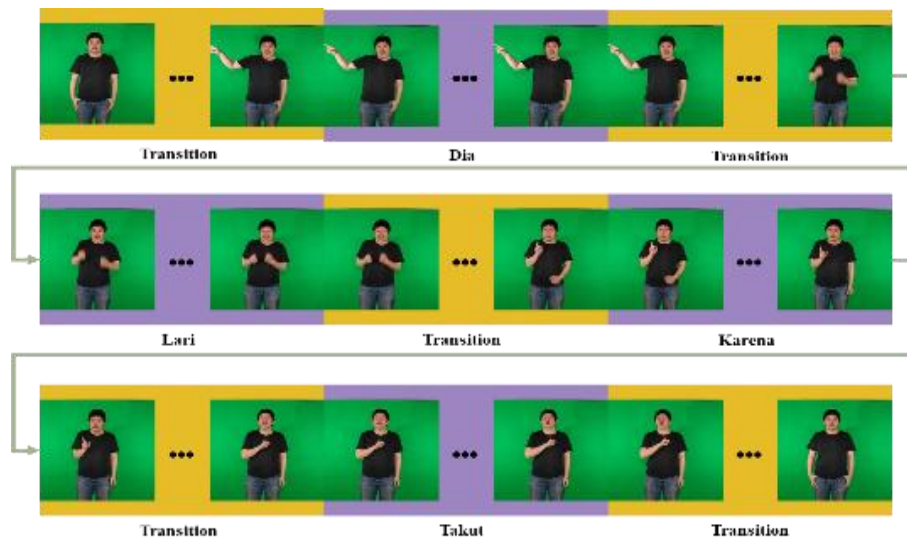


Figure 1. Transitional gestures

Addressing the issue of transitional gestures, research by [10] used the Automated Epenthesis Removal method with TCRF in SIBI, significantly reducing the word error rate (WER) to 3.367% and improving sentence accuracy (SAcc) to 80.18%. However, this method has not yet been applied to BISINDO.

The second challenge for SLT is the presence of complex backgrounds. In real-world scenarios, sign language is often performed in environments with varying backgrounds, such as shopping malls, schools, or public spaces. These complex backgrounds can significantly interfere with accurately segmenting key features, such as facial expressions and hand movements, which are critical for interpreting sign language. In contrast, a plain green background provides clear and distinct features for segmentation, and complex environments introduce noise and visual distractions that reduce the system's accuracy (Figure 2). This noise can lead to misinterpretations or failures in recognizing signs, ultimately hindering the system's performance.



Figure 2. Complex background

Research by [11] addressed the challenge of complex backgrounds in BISINDO using the Faster R-CNN [12] and YOLOv5 [13] object detection methods. Each method achieved a mean average precision (mAP) score of 71.3% and 74.1%, though YOLOv5 outperformed Faster R-CNN in WER at 16.42%, SacreBLEU at 67.77%, and SAcc at 49.29%. In this study, the transition gestures were manually removed. A study on SIBI by [14] also employed RetinaNet [15] for object detection, achieving a WER of 4.1% and an SAcc of 78.99%. While these results demonstrate progress in handling complex backgrounds, there is still room for improvement. The performance metrics, such as mAP, WER, and SAcc, can potentially be enhanced by leveraging more advanced object detection methods like YOLOv7 and YOLOv8. These latest models offer improved accuracy, speed, and robustness in detecting and segmenting objects in complex environments.

The third challenge SLT faces, especially in BISINDO, is the lack of datasets. The latest research by [11] only used 40 sentences with 152 words, while other studies, such as [16]–[18], still utilized as letter datasets, not as sentences. This dataset limitation fundamentally constrains system development and performance validation, necessitating the expansion of the dataset to create an end-to-end BISINDO SLT.

Addressing the identified challenges in SLT systems, particularly for BISINDO, this study develops an end-to-end translation application that tackles three critical challenges: (1) Dataset limitations - expanding beyond previous constrained datasets to create a more comprehensive BISINDO representation; (2) Object detection optimization - implementing cutting-edge YOLOv7 [19] and YOLOv8 [20] architectures to enhance mAP and overcome complex background interference; and (3) Automated epenthesis removal - adapting the TCRF method [21] to BISINDO for eliminating transitional gestures, thereby improving model training and system accuracy. These integrated solutions enable fully automated operation while making three key contributions: first, establishing a robust BISINDO dataset; second, advancing detection capabilities through state-of-the-art YOLO models; and third, pioneering transitional gesture processing in BISINDO through automated epenthesis removal.

## 2. RELATED WORKS

Table 1 presents several studies related to SLT systems. in the context of Indonesian sign language and foreign sign languages, which have provided valuable theoretical foundations to support the upcoming research.

Table 1. Related works in the field of SLT systems

Author	Summary	Method	Dataset	Results (%)
[22]	DeepCNN for feature extraction on SIBI	Resnet50 and MobileNetv2	SIBI Sentences	Acc: 99.89
[23]	BISINDO letter level using YOLOv3	YOLOv3	BISINDO Sentences	Image: 100 f1-score; Video: 84.38 f1-score
[24]	Modified skeleton, detection head, and loss function of YOLOv5	YOLOv5 combined with the LSTM network and OpenPose	Mandarin Words	Acc: 98.87
[25]	Using Google MediaPipe to extract key points from the palm	LSTM	Bangla Sentences	Acc: 92.07
[26]	Sign Language Detection Using CNN-YOLOv8	CNN-YOLOv8	Indian Word	Acc: 98.9
[27]	Malaysian Sign Language Mobile Application Using Deep Learning	YOLOv5	Malaysian Sign Language	mAP: 87.23

TCRF [28] a framework that considers thresholds by adding labels for non-sign G patterns in the original CRF and utilizing the weights from the transition functions and existing features in the CRF. Therefore, TCRF includes labels  $S = \{Y1, \dots, Yl, G\}$ , where  $(1, \dots, l)$  refers to the number of CRF labels. The weight of the feature function for the non-sign label G,  $\mu_m$ , is calculated using Equation 1. Thus, determining this threshold value is crucial in optimizing system performance.

$$\mu_m(G) = \bar{\mu}_m + T\sqrt{\sigma_{\mu_m}} \quad (1)$$

the value  $\bar{\mu}_m = \frac{\sum_{k=1}^l \mu_m(Y_k)}{l}$ , where k is the number of CRF labels, and  $\sigma_{\mu_m}$  is the variance of weight m. Several studies that implement TCRF, such as [29] to categorize human actions, [30] to recognize facial expressions, and [31] to interpret human body movements for robot commands.

### 3. METHOD

Developing an end-to-end application for translating BISINDO sign language into text requires a comprehensive approach encompassing several crucial stages. These stages include dataset preparation, experimental design, and result evaluation, as illustrated in Figure 3.

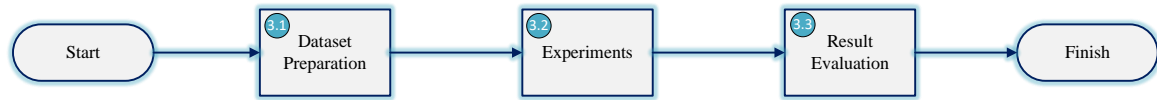


Figure 3. Methodology

In the dataset preparation stage, we will explain how the dataset for this research was collected, how the dataset was preprocessed, and what annotations were involved. In the experiment stage, we will explain the step-by-step process of addressing the problems targeted in this research. Finally, the evaluation metrics used to measure the performance of the end-to-end application for translating BISINDO SLT into text will be explained. A detailed discussion of each stage will be presented in the following sections.

#### 3.1. Dataset preparation

##### 3.1.1. Dataset collection

We extended the dataset initially employed by [11], which comprised 40 BISINDO sentences spanning 152 word classes, incorporating an additional 150 sentences to improve the system's lexical coverage and enhance its generalization capability, increasing the total number of word classes to 352. The integration of the original and supplementary datasets yielded a comprehensive corpus of 190 BISINDO sentences across 435 classes, with representative examples presented in Table 2. For the additional dataset, we collected primary data in collaboration with the Language Research Laboratory team at the Fakultas Ilmu Budaya, Universitas Indonesia (LRBI FIB UI). Two deaf signers demonstrated the 150 BISINDO sentences three times, resulting in 900 videos, while two interpreters translated their gestures (Figure 4). Figures 4(a) and 4(b) illustrate the dataset collection process, while Figures 4(c) and 4(d) depict the deaf signers involved.

Table 2. Example of 150 BISINDO sentences

No.	Sentences	BISINDO Gloss
1	<i>Dia Suka Mengenakan Baju Merah Muda</i> He/She likes to wear pink clothes.	<i>Dia - baju - merah muda - dia - suka - pakai.</i> He/She - Shirt - Pink - He/She - Likes - Wear.
2	<i>Hari Ini Adalah Hari Minggu</i> Today is Sunday.	<i>Hari - ini - minggu.</i> Day - This - Sunday.
3	<i>Saya Butuh Penggaris Untuk Mengukur Panjangnya.</i> I need a ruler to measure its length.	<i>Penggaris - ini - saya - butuh - untuk - apa - panjang - ukur.</i> Ruler - This - I - Need - For - What - Length - Measure.
4	<i>Saya Memiliki Saudara Kandung Yang Baik.</i> I have a kind sibling.	<i>Saya - punya - saudara kandung - dia - baik.</i> I - Have - Sibling - He/She - Kind.
5	<i>Jangan Lupa Makan Jambu Saat Sarapan.</i> Don't forget to eat guava at breakfast.	<i>Pagi - jambu - makan - lupa - jangan.</i> Morning - Guava - Eat - Forget - Don't.

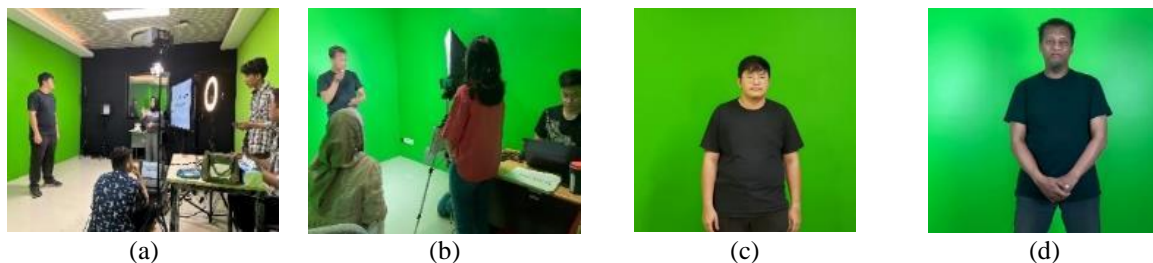


Figure 4. Dataset collection process, (a) dataset collection documentation for signer one, (b) dataset collection documentation for signer two, (c) signer one, and (d) signer two

The selection of words for the 150-sentence BISINDO dataset was based on the BISINDO dictionary, with the primary objective of this study being to incorporate all vocabulary entries from the dictionary to ensure comprehensive linguistic representation. A comparative analysis of the datasets used in

this study is presented in Table 3. We visualized the relationship between the two datasets using a Venn diagram (Figure 5), revealing 69 overlapping word classes.

Table 3. Detail dataset

Aspect	40-BISINDO sentences	150-BISINDO sentences
Signer	4	2
Sentences	40	150
Number of videos	420	900
World classes	152	352
Dictionary word cover (%)	13.02	30.16
Total frames	92,246	124,021

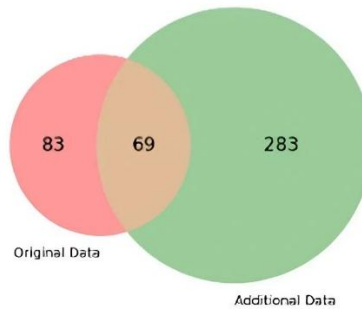


Figure 5. Venn diagram of the dataset

### 3.1.2. Dataset pre-processing

The BISINDO gestures were recorded using a smartphone at a resolution of 1920×1080 and a frame rate of 30 FPS, resulting in 124,021 frames. Each frame was resized to 224×224 for efficient training and processing. To simulate real-world conditions, backgrounds were modified to represent nine environments: bank, office, classroom, shopping mall, market, restaurant, hospital, school, and street. Figure 6 outlines the dataset preparation process.

### 3.1.3. Dataset annotation

The data annotation involved applying two types of annotations. First, each frame was labeled with the corresponding word based on the gesture, including transitional labels. Figure 7 illustrates the word annotation process, where the annotated sentence is "*Dia Lari Karena Takut*" ("He/She Runs Because Scared"). However, the original sentence order is "*Karena Takut, Dia Lari*" ("He runs because he is scared"). The labels for the sentence are 0-63-0-77-0-65-0-133, with label 0 indicating transition gestures at the beginning, between words, and at the end of gestures. Five annotators participated in this annotation process to minimize bias.

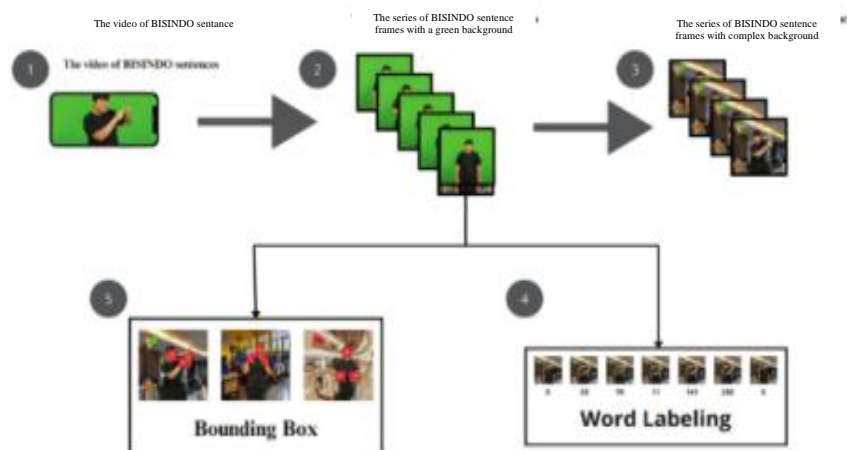


Figure 6. Dataset preparation steps

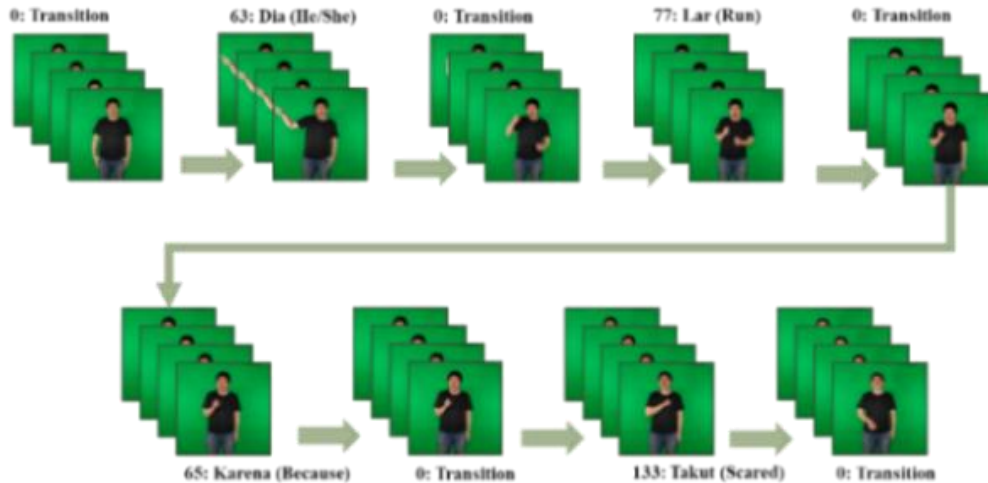


Figure 7. Word labeling annotation

The subsequent process involved assigning bounding box labels for object detection tasks and establishing ground truth for three categories: face, right hand, and left hand. Nine annotators performed the bounding box labeling. Given the enormous volume of frames requiring bounding boxes, the mean squared error (MSE) [32] and structural similarity index (SSIM) [33] methods were employed with a 94% threshold. This systematic approach bound box labels for 38,736 out of 124,021 frames. We use the open-source application labelling for this task.

### 3.2. Experiments

This research consists of three experiments, as illustrated in Figure 8. Experiment 1 aims to improve the object detection method used in the previous study by [11] with the latest object detection method. Experiment 2 aims to determine the most effective method for recognizing transition gestures. Experiment 3 aims to develop the optimal model for end-to-end BISINDO SLT.

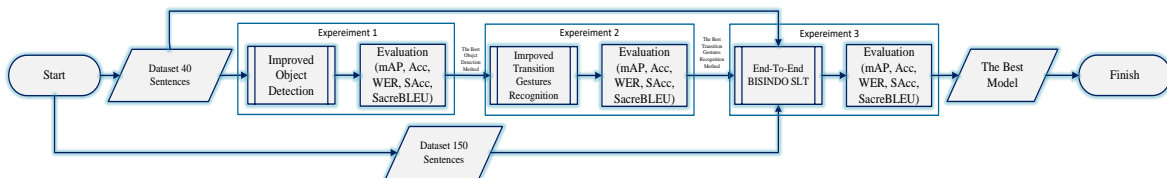


Figure 8. Experimental flow

#### 3.2.1. Baseline

A study conducted by [11] aim to develop an automated Bisindo recognition system in a complex background, utilizing 40 BISINDO with stages of object detection, skin color segmentation, feature extraction, and classification (Figure 9), which will serve as a baseline. The results of Faster R-CNN and YOLOv5 showed mAP scores of 71.3% and 74.1%, respectively, and WER, SAcc, and SacreBLEU were 26.81%, 35.00%, 45.86%, and 16.42%, 49.29%, 67.77%, respectively, which indicates that YOLOv5 outperforms Faster R-CNN, it should be noted that in this study, removing transitional gestures is done manually in the data preprocessing process (Figure 9, process 1). Based on these results, there are still gaps that can be improved, such as in object detection, which will impact the performance of the translation results and the implementation of automatic transition gesture removal, so that the system built reflects real-world conditions.

#### 3.2.2. Experiment 1 (Improved object detection)

This experiment aims to enhance system performance in object detection (Figure 9, Process 2) by utilizing the YOLOv7x and YOLOv8x models, as shown in Figure 10. Both models offer distinct advantages: YOLOv7 focuses on optimizing the E-ELAN architecture and scaling for efficiency, while



YOLOv8 emphasizes multi-level feature extraction through a combination of CSPDarknet, SPPF, and PAN. Although YOLOv8 is a newer version of YOLOv7, the original YOLOv8 paper does not explicitly compare the performance of the two YOLO versions. The experiment utilizes 40 BISINDO sentences, with the object detection method as the independent variable. Its success is measured by the mAP, WER, SAcc, and SacreBLEU scores.

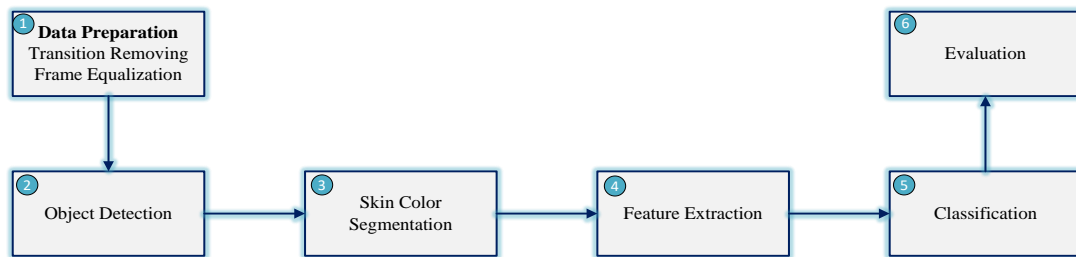


Figure 9. Baseline [11]

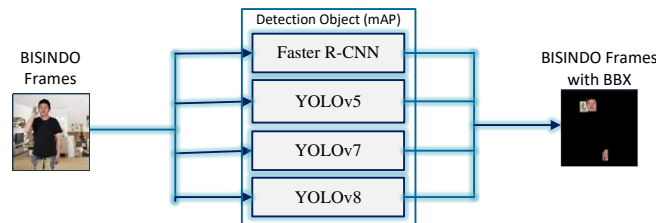


Figure 10. Improved detection objects

### 3.2.3. Experiment 2 (Improved transition gestures recognition)

This experiment addresses the problem of removing transitional gestures. In the study [11], the removal of transitional gestures was carried out at the data preprocessing stage (Figure 9, process 1), so in this study, a stage will be added between stages 4 and 5 (Figure 9, processes 4 and 5), transitional gestures recognition so that the system can remove them automatically. The ultimate goal is to build an end-to-end BISINDO SLT. In addition, we also added the sandwich majority voting and short word-frame sequence relabeling stages after the transitional gestures stage and before the automatic removal of transitional gestures, so that the resulting data are clean, as seen in Figure 11. Instead of using original labels to recognize transition gestures, we employ two approaches: first, MobileNet-predicted labels, and second, TCRF-predicted labels. The dataset used consists of 40 BISINDO sentences, with the transition gesture recognition method as the independent variable.

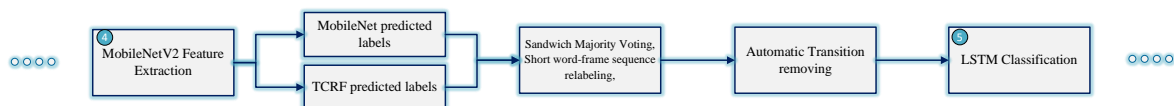


Figure 11. Improved transitional gesture removal

### 3.2.4. Experiment 3 (Proposed end-to-end BISINDO SLT)

This study proposes an end-to-end BISINDO SLT application, developed through Experiments 1 and 2. The system processes BISINDO video frames as input and generates corresponding BISINDO sentences as output, operating fully automatically without human intervention in real-world conditions (Figure 12). The stages involved are (1) object detection, (2) skin color segmentation, (3) feature extraction, (4) transition gestures recognition, (5) sandwich majority voting and short word-frame sequence relabeling, (6) transition removing, (7) frame equalization, and (8) classification. Evaluation employs three dataset configurations: 40 sentences (152 labels), 150 sentences (352 labels), and their combined 190-sentence dataset (435 labels), with sentence and label quantity as the independent variables for assessing the impact of system performance.

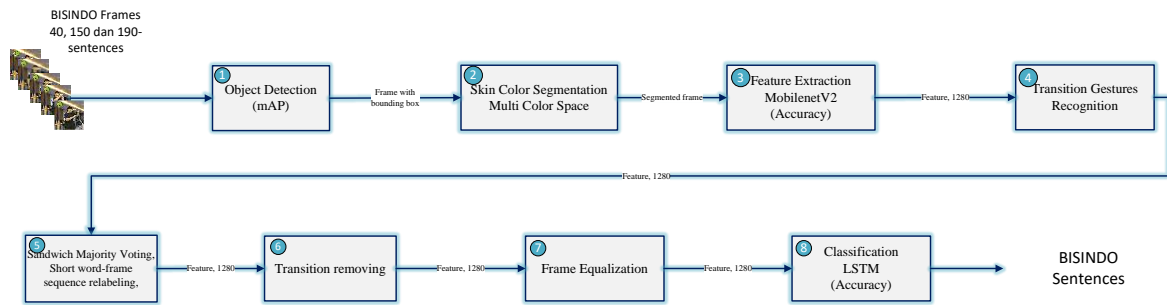


Figure 12. End-to-end of BISINDO SLT

Object detection is the initial stage of identifying the signer's face, right hand, and left hand in input images with complex backgrounds, producing isolated regions. The aim is to extract key features for further processing. The models were trained with a batch size of 16 for 150 epochs. Performance is measured using mAP on the test set, with the baseline set as the results from [11]'s study, which serves as the ground truth to ensure reliable detection quality.

The second stage is skin color segmentation. Images from the previous stage still contain bounding boxes, requiring an additional step to isolate and extract only the face and hands. This stage applies the method proposed by [34], by combining normalized RGB, HSV, and YCbCr color spaces.

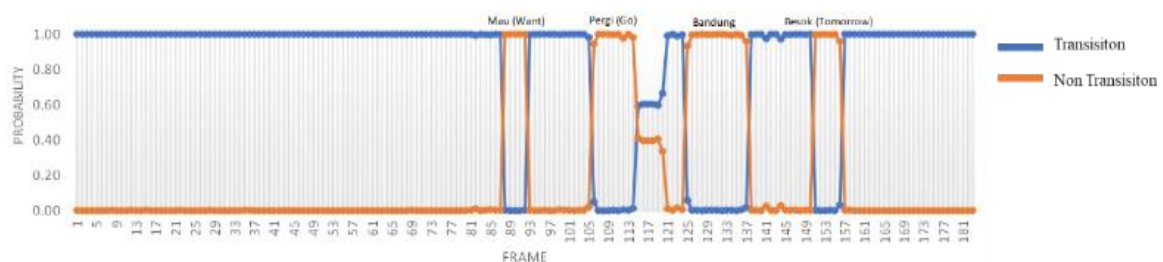
The third stage performs feature extraction using MobileNetV2 [35] to transform segmented hand and face regions into meaningful gesture representations. This stage processes segmented frames and their BISINDO labels through a model trained with a batch size of 16 for 100 epochs, generating 1280-dimensional feature vectors (stored as NumPy arrays) for each frame. The purpose is to create discriminative features that capture essential gesture characteristics while reducing computational complexity. Predicting accuracy measures performance, which validates how well the extracted features represent BISINDO gestures.

The fourth stage focuses on transition gesture recognition using MobileNetV2 and TCRF prediction labels. These two methods aim to determine the best robust method for recognizing transition gestures. Mobilenet plays a dual role; in addition to being an extraction feature, it is also tasked with making predictions based on the resulting labels. The output of this stage is a sequence of features for each BISINDO gesture, as shown in Figure 13.

Gloss	Transition	Mau (Want)	Transition	Pergi (Go)	Transition	Bandung	Transition	Besok (Tomorrow)	Transition
Frame Feature									
MobileNetV2 Label	0 0 238 0 0	0 238 238 238 0 0 0 0 0	0 41 41	0 0 0 0 0 0	0 25 25 0	0 0 0 0 0 0	6 0 0 6 6 0	251 0 0 0 0	
TCRF Label	0 0	0 0 0 238 238 238 238 0 0 0 0 0	297 297 297 297 0 0 0 0 0	25 25 25 25 25 25 0 0 0 0 0	54 54 54 54 54 54	0 0 0 0 0			

Figure 13. Illustration of transition gesture recognition

TCRF is trained with a batch size of 16 and 100 epochs, and the accuracy is measured for each threshold value. As an illustration, Figure 14 presents the probability graph for each class (transition or word gestures) based on the BISINDO sentence "*Mau Pergi Bandung Besok*" ("Want-Go-Bandung-Tomorrow"). The best-performing MobileNetV2 and TCRF models generated this graph. Word gestures appear in frames 88–92, 106–114, 126–137, and 151–156, while all other frames represent transitional gestures.

Figure 14. Probability graph for "*Mau-Pergi-Bandung-Besok*" ("Want-Go-Bandung-Tomorrow")



The fifth stage comprises two sequential steps: Sandwich minority voting. This step replaces transition labels sandwiched between word labels with the corresponding word labels. Short word-frame sequence relabeling: Word labels flanked by transition labels are reclassified as transition labels. In alignment with [10], this study applies a constraint whereby a maximum of two consecutive labels may be modified in each step, as illustrated in Figure 15.

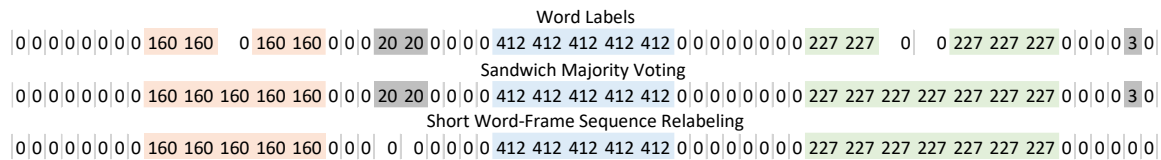


Figure 15. Post-transition gesture recognition

The sixth stage involves removing transitional gestures, where transitional gestures are identified and eliminated using MobileNetV2 and TCRF-predicted labels. Frames associated with a label value of zero and their corresponding labels are discarded, as depicted in Figure 16. This process ensures that only meaningful gesture segments are preserved for subsequent word prediction.

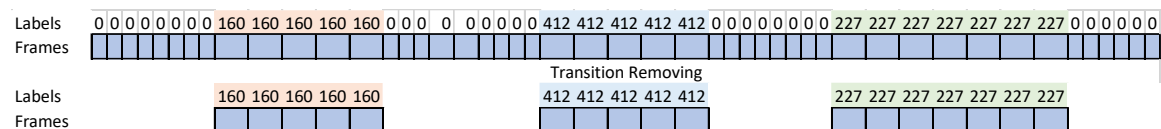


Figure 16. Transition removing

The seventh stage involves frame equalization, as shown in Figure 17, ensuring uniform input sequence lengths required by the LSTM-based seq2seq architecture. This standardization process begins by calculating the average frame length after removing the transition. For sequences exceeding this average length, the system calculates the MSE between adjacent frame pairs, ranks them accordingly, and removes the last  $n$  frames (where  $n$  equals the difference between current and average lengths) as shown in Figure 17(a). Conversely, shorter sequences are padded by duplicating their final frame until they reach the target length (Figure 17b). The target length of 15 frames per word was determined by dividing the total frame count across all words by the number of word classes. This equalization procedure is applied uniformly to both feature frames and their corresponding labels.

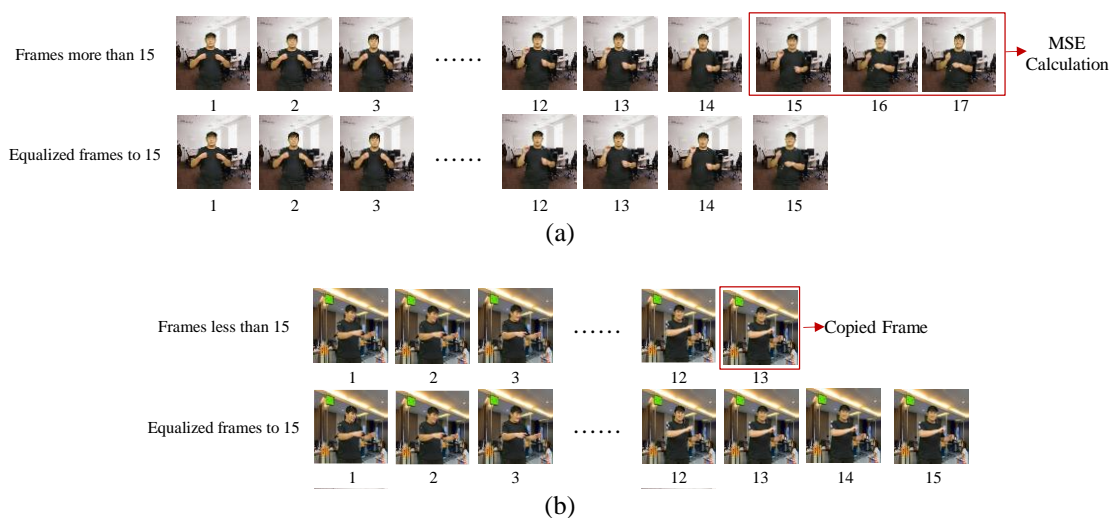


Figure 17. Frame equalization (a) Frame more than 15, (b) Frame less than 15

The final stage of the proposed methodology is classification, utilizing a 1-layer Bidirectional long short-term memory (LSTM) network [36]. This model was chosen based on prior research [11]. Configured with 256 hidden units, a batch size of 16, and trained for 300 epochs, the model ensures effective learning and precise classification of input data, optimizing performance for BISINDO gesture recognition.

All processes were conducted using the same machine as in the study by [11], an NVIDIA DGX-1 server with Linux OS (#48-Ubuntu SMP), an 80-core x86\_64 CPU, 512 GB DDR4 memory, and an NVIDIA Tesla V100-SXM2-32GB GPU. This allows the results to be compared with previous research.

### 3.3. Evaluation

Five metrics were used: mAP, Accuracy, WER, SacreBLEU, and SAcc. mAP assesses object detection by comparing predictions to ground truth using the intersection over union (IoU) metric. It is calculated using Equation 2.

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \quad (2)$$

Where N is the number of classes, and AP<sub>i</sub> is the AP for the i-th class.

Accuracy measures the percentage of correctly predicted data, as defined in (3). It is crucial for evaluating performance in feature extraction, transition gesture recognition, and classification stages.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Where true positive (TP) is the number of positive cases correctly predicted. True negative (TN) is the number of negative cases correctly predicted. False positive (FP) is the number of negative cases incorrectly predicted as positive. False negative (FN) is the number of positive cases incorrectly predicted as negative by the model.

WER measures the ratio of the number of incorrect words generated by the system to the total number of words in the given document or dataset. In (4) calculates WER.

$$WER = \frac{S+D+I}{N=H+S+D} \quad (4)$$

Where S is the number of substitution errors, D is the number of deletion errors, I is the number of insertion errors, and N is the total number of words in the analyzed sentence or document.

SacreBLEU assesses the quality of machine translations between languages. Its goal is to evaluate how effectively machine translations compare to human translations. SacreBLEU uses a formula modified based on n-gram precision to evaluate this. The equation can be explained as shown in (5).

$$SacreBLEU = BLEU \times BP \quad (5)$$

Where BLEU is the score calculated using n-gram precision-modified, BP is the brevity penalty factor, which addresses the issue when machine translation outputs are shorter than human reference translations.

SAcc measures the correctness of an entire sentence or sequence of predictions by comparing it to a reference sentence. SAcc ensures that the generated sentence matches the reference without errors. It is calculated using (6).

$$Sentence Accuracy (SAcc) = \frac{Number of Correctly Predicted Sentence}{Total Number of Sentences} \quad (6)$$

## 4. RESULTS AND DISCUSSION

### 4.1. Results of experiment 1 (Improved object detection)

Experiment 1 aims to improve object detection performance using YOLOv7 and YOLOv8, the latest object detection methods. The success of this experiment is measured by the increase in mAP on object detection, the improvement in accuracy on MobileNet extraction features, the enhancement in LSTM classification, and the reduction in computation time at each stage (Table 4).

As shown in Table 4, YOLOv8 significantly improves object detection performance, as indicated by its increasing mAP scores. This advancement is complemented by enhanced accuracy in MobileNet's feature extraction capability, with frame label classification showing substantial improvement when compared to

ground truth labels. The LSTM-based word gesture classification achieves higher accuracy, strengthening overall system performance. YOLOv8 also exhibits reduced computational time at each processing stage, indicating a more efficient operational capability; however, YOLOv7 remains slower than YOLOv5. This progress can be attributed to YOLOv8's advanced dual-path architecture, which combines a feature pyramid network (FPN) for multi-scale feature extraction and a path aggregation network (PAN) for effective feature fusion. YOLOv8's end-to-end translation capability—a metric seldom evaluated in prior studies—highlights its potential for real-time sign language interpretation. These findings align with the research by [11] confirming that higher mAP scores correlate with better translation outcomes.

Table 4. Results for each stage for improved object detection

Method		Object Detection (mAP 50 – 90)	Feature Extraction MobileNetV2 (Accuracy)	LSTM Classification (Accuracy)
YOLOv5 [11]	Metric (%)	74.10	76.42	83.58
	Time (seconds)	0.0138	0.0158	0.0188
YOLOv7	Metric (%)	76.30	79.13	85.39
	Time (seconds)	0.023	0.0166	0.026
YOLOv8	Metric (%)	88.00	82.04	88.60
	Time (seconds)	0.010	0.0145	0.0165

The gesture classification results from the LSTM model were quantitatively evaluated using WER, SAcc, and SacreBLEU metrics to measure their correspondence with the original sentences (Table 5). The analysis demonstrates a positive correlation between improved system components - object detection, feature extraction quality, LSTM classification performance, and reduced total computational time. To support the results in Table 5, Table 6 contains translation results for each object detection method.

As shown in Table 6, YOLOv8 outperforms YOLOv5 and YOLOv7 in BISINDO gesture translation. While YOLOv5 exhibits significant word-level classification errors (e.g., misidentifying "*besok*" [tomorrow] as "*pulang*" [go home] and "*makan-semangka*" [eat-watermelon] as "*terima*" [accept]), YOLOv7 shows inconsistent sentence-level predictions with repetitive errors such as the persistent misclassification of "*makan-semangka*" [eat-watermelon]. These comparative results demonstrate that YOLOv8 can accurately detect hand and face positions, enabling the system to recognize gestures shown by the signer.

Table 5. LSTM performance in the BISINDO translation system for improved object detection

Method	WER (%)	SAcc (%)	SacreBLEU	Computational Time (seconds)
YOLOv5 [11]	16.42	49.29	67.77	0.0458
YOLOv7	14.61	55.00	68.61	0.0656
YOLOv8	11.40	62.14	77.96	0.041

Table 6. Comparison of prediction results for each object detection method

Methode	Original Sentences	Prediction Sentences	WER (%)	SAcc (%)
YOLOv5	<i>Saya – pergi – bandung – besok</i> (I – go – Bandung – tomorrow)	<i>Saya – pergi – bandung – pulang</i> (I – go – Bandung – Go home)	25	0
	<i>Saya – bingung – karena – tidak-boleh – makan-semangka</i> (I – confused – because – not-allowed – eat-watermelon)	<i>Saya – bingung – karena – karena – terima</i> (I – confused – because – because – accept)	40	0
YOLOv7	<i>Saya – pergi – bandung – besok</i> (I – go – Bandung – tomorrow)	<i>Saya – pergi – bandung – besok</i> (I – go – Bandung – tomorrow)	0	100
	<i>Saya – bingung – karena – tidak-boleh – makan-semangka</i> (I – confused – because – not-allowed – eat-watermelon)	<i>Saya – makan-semangka – bingung – kenapa – makan-semangka</i> (I – eat-watermelon – confused – why – eat-watermelon)	40	0
YOLOv8	<i>Saya – pergi – bandung – besok</i> (I – go – Bandung – tomorrow)	<i>Saya – pergi – bandung – besok</i> (I – go – Bandung – tomorrow)	0	100
	<i>Saya – bingung – karena – tidak-boleh – makan-semangka</i> (I – confused – because – not-allowed – eat-watermelon)	<i>Saya – bingung – karena – tidak-boleh – makan-semangka</i> (I – confused – because – not-allowed – eat-watermelon)	0	100

The experiment highlights the significance of these advancements in real-time sign language interpretation, as evidenced by YOLOv8's higher translation accuracy, lower WER, better sentence-level predictions, and reduced total computational time compared to YOLOv5 and YOLOv7. However, unanswered questions remain regarding the model's generalization across diverse sign language datasets and

its robustness in the real world. Future research should explore the latest versions of object detection methods, such as transformer-based object detection or segment anything model (SAM).

#### 4.2. Result of experiment 2 (Improved transition gestures recognition)

Experiment 2 aims to develop a transition gesture recognition method for BISINDO to automatically remove it. Using two approaches, MobileNetV2 and TCRF prediction labels. For TCRF, we tested Threshold values on 1, 1.5, 2, 2.5, 3, 3.5, and 4, then calculated the accuracy. Accuracy was determined by comparing the predicted labels against ground truth labels (Figure 13) for both MobileNetV2 and TCRF approaches. The detailed accuracy results for each threshold value are presented in Table 7. Among all tested thresholds, a value of 1 demonstrated optimal performance, achieving an accuracy of 95.683%. In comparison, the MobileNetV2 achieved an accuracy of 82.04% (Table 4). These results demonstrate TCRF's superior performance in BISINDO transition gesture recognition.

Table 7. Accuracy comparison of threshold values for TCRF

1 (%)	1.5 (%)	2 (%)	2.5 (%)	3 (%)	3.5 (%)	4 (%)
95.683	95.204	95.204	94.964	94.964	94.964	94.964

The predicted labels generated by both MobileNet and TCRF undergo subsequent processing through two key stages: (1) Sandwich minority voting and (2) Short word-frame sequence relabeling (as illustrated in Figure 15). Following these refinement processes, all labels identified as transitional gestures (label = 0) are automatically filtered out (Figure 16). This preprocessing ensures that only meaningful word gesture labels proceed to the final LSTM classification stage. Table 8 shows that TCRF outperforms MobileNetV2 in all metrics (WER, SAcc, SacreBLEU, and computational time)

Table 8. Performance comparison of LSTM-based transition recognition between MobileNetV2 and TCRF

Methods	WER (%)	SAcc (%)	Sacre BLEU	Computational Time (seconds)
MobileNetV2	93.53	0.00	0.28	0.129
TCRF	6.89	71.69	81.23	0.032

MobileNetV2 demonstrates fundamental limitations for sequential SLT processing due to its lack of temporal modeling capabilities. While effective for static feature extraction, the architecture exhibits three critical flaws: (1) inability to capture gesture motion trajectories, (2) failure to maintain inter-frame dependencies, and (3) computational inefficiency. These deficiencies lead to the misclassification of transitional movements as distinct signs, which in turn results in poor performance.

In contrast, TCRF excels in sequential analysis, as demonstrated by its successful sentence predictions where MobileNetV2 failed (Table 9). Its inherent capacity to model contextual dependencies between frames enables more accurate, temporally consistent labeling. These results conclusively demonstrate that effective temporal modeling is crucial for BISINDO SLT systems, with TCRF's architecture proving significantly superior to MobileNetV2's static approach.

Table 9. Comparison of prediction results using MobileNet and TCRF labels

Original Sentence	MobileNetV2		TCRF	
	Labels Predictions	WER (%)	Labels Predictions	WER (%)
<i>Kakak – Beri</i> (Older – Sibling – Give)	<i>Kakak – Jalan-Kaki – Kakak – Jalan-Kaki</i> (Older-Sibling – Walk – Older-Sibling – Walk)	150	<i>Kakak – Beri</i> (Older – Sibling – Give)	0
<i>Motor – Saya – Mau – Jual</i> (Motorcycle – I – Want – To Sell)	<i>Menit – Motor – Jual – Ini</i> (Minute – Motorcycle – Sell – This)	100	<i>Motor – Saya – Mau – Jual</i> (Motorcycle – I – Want – To Sell)	0
<i>Buku – Bahasa – Inggris – Baca – Sudah – Pernah</i> (Book – Language – English – Read – Already – Ever)	<i>Inggris – Baca – Harus</i> (English – Read – Must)	.66.67	<i>Buku – Bahasa – Inggris – Baca – Sudah – Pernah</i> (Book – Language – English – Read – Already – Ever)	0

The experimental results validate the effectiveness of combining YOLOv8 for robust object detection with TCRF (threshold = 1) for automated transition gesture recognition, enabling the removal of

these objects automatically in BISINDO SLT. This integrated approach successfully addresses the critical challenges in developing a fully functional End-to-End BISINDO SLT system. The findings highlight the necessity of temporal modeling for accurate SLT and suggest TCRF as a promising solution for real-time SLT systems. The limitation is the number of words the model can translate, so adding a dataset to generalize the model is necessary. Future research could explore hybrid models, adaptive thresholding, and applications to other sign languages to further enhance robustness and generalizability.

#### 4.3. Result of experiment 3 (End-to-end BISINDO SLT)

This experiment aims to develop an end-to-end BISINDO SLT system by integrating YOLOv8 for real-time object detection and TCRF (threshold = 1) for automated transition gesture removal, based on Experiments 1 and 2. The system will be trained and evaluated on datasets comprising 40, 150, and 190 BISINDO sentences to assess scalability and performance. Using video frames as input, the pipeline processes visual data through YOLOv8 for spatial localization, followed by TCRF-based temporal modeling to filter transitional noise, ultimately generating complete and accurate BISINDO sentences as output. This approach addresses the critical need for robust, real-time SLT systems by combining state-of-the-art object detection with context-aware sequential modeling. The results are presented in Table 10.

Table 10 reveals a performance decline when testing the 150-sentence dataset compared to the 40-sentence dataset. This degradation occurs due to differences in model capacity between the datasets - while the 40-sentence dataset utilizes four signers, the 150-sentence dataset employs only two signers, resulting in reduced representation capability. However, performance improves when using the combined 190-sentence dataset due to two key factors: (1) the inclusion of 69 overlapping words enhances lexical distribution, thereby better capturing real-world sign variation; and (2) increased training diversity strengthens the model's pattern recognition capabilities. Regarding computational efficiency, all datasets exhibited consistent processing times during testing, demonstrating stable performance across varying dataset sizes.

A post-processing grouping algorithm was implemented following LSTM classification to enhance output quality. This method consolidates consecutive duplicate labels, achieving three key improvements: (1) reduction of redundant predictions, (2) elimination of interpretation ambiguities, and (3) generation of more coherent sign sequences. The quantitative benefits of this approach are detailed in Table 11.

Table 10. LSTM performance for end-to-end BISINDO SLT

Dataset (label)	WER (%)	SAcc (%)	Sacre BLEU	Computational Time testing (seconds)
40 sentences (152)	6.89	71.69	81.23	0.032
150 sentences (352)	23.18	63.76	57.87	0.032
190 sentences (435)	11.38	80.95	82.57	0.032

Table 11. LSTM performance for end-to-end BISINDO SLT with grouping

Dataset (label)	WER (%)	SAcc (%)	Sacre BLEU	computational time (seconds)
40 sentences (152)	6.89	71.69	81.23	0.032
150 sentences (352)	16.02	73.83	69.03	0.032
190 sentences (435)	8.31	84.13	87.08	0.032

Table 12. Example of translation results using TCRF's label with grouping

Dataset (label)	Without Grouping	WER (%)	With Grouping	WER (%)
150 sentences (352)	<i>Anak – main – main – sudah – harus</i> (child – play – play – already – bath – must)	40	<i>Anak – main – sudah – mandi – harus</i> (child – play – already – bath – must)	0
190 sentences (435)	<i>teman-teman – teman-teman – ketemu - nanti</i> (friends – friends – meet – later)	33.3	<i>teman-teman – ketemu – nanti</i> (friends – meet – later)	0

The experimental results (Tables 10 and 11) reveal distinct performance patterns across different dataset scales. While translation accuracy remained consistent for the 40-sentence dataset (suggesting minimal label repetition), we observed significant improvements for both the 150- and 190-sentence datasets. This graduated enhancement demonstrates the efficacy of our grouping method in addressing prediction redundancy, particularly for frequently occurring signs. Table 12 confirms that our approach eliminates duplicate outputs while preserving semantic content and improving translation fidelity (resulting in significantly reduced WER). These findings corroborate the framework established in [10], validating frequency-sensitive processing.

This study successfully developed an end-to-end BISINDO SLT by integrating YOLOv8 for real-time hand detection and TCRF (threshold = 1) for automated removal of transition gestures. The system was rigorously evaluated across three datasets (40, 150, and 190 sentences), demonstrating its ability to handle varying vocabulary sizes while maintaining consistent computational efficiency (0.032s). A key innovation was the post-processing grouping algorithm, which significantly improved output quality by eliminating redundant predictions, reducing WER in critical test cases. Performance analysis revealed important insights: while the 150-sentence dataset showed limitations due to having fewer signers (2 vs. 4), the 190-sentence dataset achieved optimal results by leveraging overlapping vocabulary and enhancing training diversity. These findings validate frequency-sensitive processing as an effective strategy for SLT systems. However, the study identified areas for future improvement, including handling regional sign variations, developing adaptive thresholds for different signing speeds, and incorporating NLP techniques for better semantic understanding. The results provide a strong foundation for building more robust and inclusive communication tools for the Deaf community, while highlighting the importance of dataset diversity in SLT development.

## 5. CONCLUSION

This study systematically progressed through three interconnected experiments to develop an end-to-end BISINDO SLT, where each phase built upon the previous findings to address distinct challenges. Experiment 1's demonstration of YOLOv8's superiority (88.00% mAP, 0.041s total computational time) in accurate and efficient object detection directly enabled Experiment 2's breakthrough in temporal modeling, where TCRF's optimized threshold=1 configuration achieved 95.68% transition gesture recognition accuracy, resolving MobileNetV2's critical limitations in capturing sequential dependencies. These foundational advancements then converged in Experiment 3's end-to-end system, where the integration of YOLOv8's spatial detection with TCRF's temporal processing, enhanced by a grouping algorithm, yielded unprecedented translation fidelity (reducing WER to 8.31% for 190-sentence datasets) while maintaining real-time performance. The research provides transformative insights for both technical and social domains: technically, it establishes hybrid spatial-temporal architectures and frequency-aware processing as essential paradigms in BISINDO SLT; socially, it delivers a robust framework for accessible communication tools, though the observed performance variance and vocabulary sizes highlights the need for expanded datasets encompassing regional dialects and signing styles. Future work must bridge these gaps through transformer-based detection for nuanced sign semantics, adaptive thresholding for personalized signing patterns, and multimodal NLP integration - ultimately translating these laboratory achievements into practical solutions that empower the deaf community through more inclusive technology.

## ACKNOWLEDGEMENTS

This research is supported by the Tokopedia-UI AI Center of Excellence and Funded by the Program for Innovation from the University of Indonesia 2024 - P3 No: PKS-93/UN2.INV/HKP.05/2024 and No: 0019/PKS/ES/2024. We extend our gratitude to the Lembaga Riset Bahasa Indonesia (LRBI), Fakultas Ilmu Budaya of Universitas Indonesia, for providing the dataset used in this study.

## FUNDING INFORMATION

The authors state no funding is involved.

## CONFLICT OF INTEREST STATEMENT

The authors state no conflict of interest.

## DATA AVAILABILITY

Data availability does not apply to this paper as no new data were created or analyzed in this study.

## REFERENCES

- [1] R. E. Mitchell and M. A. Karchmer, "Chasing the mythical ten percent: Parental hearing status of deaf and hard of hearing students in the United States," *Sign Language Studies*, vol. 4, no. 2, 2004, doi: 10.1353/sls.2004.0005.
- [2] W. C. Stokoe and M. Marschark, "Sign language structure: An outline of the visual communication systems of the american deaf," *Journal of Deaf Studies and Deaf Education*, vol. 10, no. 1, pp. 3–37, 2005, doi: 10.1093/deafed/eni001.






- [3] S. Kumar, P. Kumar, P. Mishra, and P. Tewari, "A robust sign language and hand gesture recognition system using convolutional neural networks," in *Proceedings - IEEE 2023 5th International Conference on Advances in Computing, Communication Control and Networking, ICAC3N 2023*, 2023, pp. 395–399, doi: 10.1109/ICAC3N60023.2023.10541471.
- [4] S. Sivamohan, S. Anslam Sibi, T. R. Divakar, and S. Jagan, "Hand gesture recognition and translation for international sign language communication using convolutional neural networks," in *Proceedings - 2nd International Conference on Advancement in Computation and Computer Technologies, InCACCT 2024*, 2024, pp. 635–640, doi: 10.1109/InCACCT61598.2024.10551000.
- [5] S. Chi, Y. Zhou, F. Liu, and J. Li, "Real time recognition methods for Chinese sign language based on detection model," in *2024 IEEE 4th International Conference on Electronic Technology, Communication and Information, ICETCI 2024*, 2024, pp. 88–93, doi: 10.1109/ICETCI61221.2024.10594623.
- [6] M. P. Geetha, S. Swetha, M. Subitsha, and K. V. Visnupriya, "Gesture based sign language recognition for specially challenged using Yolov5," 2024, doi: 10.1109/ICSTEM61137.2024.10560764.
- [7] P. Sonsare, D. Gupta, J. Sayyed, P. Nandha, and S. Laura, "Sign language to text conversion using deep learning techniques," 2024, doi: 10.1109/IC457434.2024.10486568.
- [8] N. Sarma, A. K. Talukdar, and K. K. Sarma, "Real-time indian sign language recognition system using YOLOv3 model," in *Proceedings of the IEEE International Conference Image Information Processing*, 2021, vol. 2021-Novem, pp. 445–449, doi: 10.1109/ICIIP53038.2021.9702611.
- [9] N. Palfreyman, "Budaya tuli Indonesia dan hak bahasa (Indonesian deaf culture and the linguistic rights)," in *Coference Paper, Seminar Tahunan Linguistik, Universitas Pendidikan Indonesia*, 2015, no. June 2015.
- [10] E. Rakun, I. G. B. H. Widhinugraha, and N. F. Putra Setyono, "Word recognition and automated epenthesis removal for Indonesian sign system sentence gestures," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 3, p. 1402, Jun. 2022, doi: 10.11591/ijeecs.v26.i3.pp1402-1414.
- [11] M. A. Saputra and E. Rakun, "Recognizing Indonesian sign language (Bisindo) gesture in complex backgrounds," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 36, no. 3, pp. 1583–1593, 2024, doi: 10.11591/ijeecs.v36.i3.pp1583-1593.
- [12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017, doi: 10.1109/TPAMI.2016.2577031.
- [13] T. F. Dima and M. E. Ahmed, "Using YOLOv5 algorithm to detect and recognize american sign language," *2021 International Conference on Information Technology, ICIT 2021 - Proceedings*, pp. 603–607, 2021, doi: 10.1109/ICIT52682.2021.9491672.
- [14] M. H. Nugraha and E. Rakun, "Solving complex background problem using retinanet for sign system for Indonesian language (SIBI) gesture-to-text translator," in *Proceedings - ICACISIS 2022: 14th International Conference on Advanced Computer Science and Information Systems*, 2022, pp. 45–52, doi: 10.1109/ICACISIS56558.2022.9923450.
- [15] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, vol. 2017-Octob, pp. 2999–3007, doi: 10.1109/ICCV.2017.324.
- [16] D. Joan, V. Vincent, K. J. Daniel, S. Achmad, and R. Sutoyo, "BISINDO hand-sign detection using transfer learning," 2023, doi: 10.1109/ICRAIE59459.2023.10468194.
- [17] R. Wiliam, C. Lufian, Meiliana, and A. Y. Zakiyyah, "Recognitions of bahasa isyarat Indonesia (BISINDO) alphabets using SVM and mediapipe," 2024, doi: 10.1109/ICORIS63540.2024.10903898.
- [18] R. A. Pranadesta and I. S. Suwardi, "Indonesian sign language (BISINDO) translation system with ORB for bilingual language," in *Proceeding - 2019 International Conference of Artificial Intelligence and Information Technology, ICAIIT 2019*, 2019, pp. 502–505, doi: 10.1109/ICAIIT.2019.8834677.
- [19] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2023, vol. 2023-June, pp. 7464–7475, doi: 10.1109/CVPR52729.2023.00721.
- [20] D. Reis, J. Kupec, J. Hong, and A. Daoudi, "Real time flying object detection with YOLOv8," 2024, [Online]. Available: <http://arxiv.org/abs/2305.09972>.
- [21] H. D. Yang, S. Sclaroff, and S. W. Lee, "Sign language spotting with a threshold model based on conditional random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 7, pp. 1264–1277, 2009, doi: 10.1109/TPAMI.2008.172.
- [22] N. F. P. Setyono and E. Rakun, "Recognizing word gesture in sign system for Indonesian language (SIBI) Sentences using DeepCNN and BiLSTM," in *2019 International Conference on Advanced Computer Science and Information Systems, ICACISIS 2019*, 2019, pp. 199–204, doi: 10.1109/ICACISIS47736.2019.8979772.
- [23] S. Daniels, N. Suciati, and C. Fathichah, "Indonesian sign language recognition using YOLO method," in *IOP Conference Series: Materials Science and Engineering*, 2021, vol. 1077, no. 1, p. 012029, doi: 10.1088/1757-899x/1077/1/012029.
- [24] X. He, Y. Lin, Z. Hu, X. Xu, R. Xu, and W. Xiang, "AI Chinese sign language recognition interactive system based on audio-visual integration," in *2023 IEEE International Conference on Electrical, Automation and Computer Engineering, ICEACE 2023*, 2023, pp. 962–968, doi: 10.1109/ICEACE60673.2023.10442295.
- [25] S. M. M. Mahin, M. R. Islam, and S. M. M. Ahsan, "Phrase level Bangla sign language recognition using keypoints from hand gesture video," 2023, doi: 10.1109/NCIM59001.2023.10212460.
- [26] C. Singh, D. Sharma, A. P. Dubey, and N. Tyagi, "Sign language detection using CNN-YOLOv8l," in *2023 International Conference on Advances in Computation, Communication and Information Technology, ICAICIT 2023*, 2023, pp. 12–17, doi: 10.1109/ICAICIT60255.2023.10465792.
- [27] A. A. A. Azahari and S. Nordin, "Malaysian sign language mobile application using deep learning," in *2023 4th International Conference on Artificial Intelligence and Data Sciences: Discovering Technological Advancement in Artificial Intelligence and Data Science, AiDAS 2023 - Proceedings*, 2023, pp. 361–365, doi: 10.1109/AiDAS60501.2023.10284708.
- [28] H. Chung and H.-D. Yang, "Conditional random field-based gesture recognition with depth information," *Optical Engineering*, vol. 52, no. 1, p. 017201, 2013, doi: 10.1117/1.oe.52.1.017201.
- [29] M. H. Siddiqi, M. Alruwaili, A. Ali, S. Alanazi, and F. Zeshan, "Human activity recognition using gaussian mixture hidden conditional random fields," *Computational Intelligence and Neuroscience*, vol. 2019, 2019, doi: 10.1155/2019/8590560.
- [30] B. Hasani and M. H. Mahoor, "Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields," *Proceedings - 12th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2017 - 1st International Workshop on Adaptive Shot Learning for Gesture Understanding and Production, ASLAGUP 2017, Biometrics in the Wild, Bwild 2017, Heteroge*, pp. 790–795, 2017, doi: 10.1109/FG.2017.99.
- [31] M. Hasanuzzaman, T. Zhang, V. Ampomaramveth, and H. Ueno, "Gesture-based human-robot interaction using a knowledge-based software platform," *Industrial Robot*, vol. 33, no. 1 SPEC. ISS., pp. 37–49, 2006, doi: 10.1108/01439910610638216.

- [32] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, pp. 01–11, 2015, doi: 10.5121/ijdkp.2015.5201.
- [33] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004, doi: 10.1109/TIP.2003.819861.
- [34] R. F. Rahmat, T. Chairunnisa, D. Gunawan, and O. S. Sitompul, "Skin color segmentation using multi-color space threshold," *2016 3rd International Conference on Computer and Information Sciences, ICCOINS 2016 - Proceedings*, pp. 391–396, 2016, doi: 10.1109/ICCOINS.2016.7783247.
- [35] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018, doi: 10.1109/CVPR.2018.00474.
- [36] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

## BIOGRAPHIES OF AUTHORS



**Satria Putra**    received his bachelor's degree in Computer Systems from Sriwijaya University in Palembang, Indonesia, in 2014. He is a master's student at the University of Indonesia's Faculty of Computer Science. His research interests are computer vision, artificial intelligence, and deep learning. From 2018 to the present, he has worked for the Indonesian government as an IT staff member in the Directorate General of Horticulture, Ministry of Agriculture. He can be contacted at email: satria.putra21@ui.ac.id.



**Erdefi Rakun**    received her bachelor's degree in Electrical Engineering from the University of Indonesia in Jakarta, Indonesia, in 1982. She received her M. Sc. in Computer Science from the University of Minnesota, USA, in 1988. She received her Ph.D. in Computer Science from the University of Indonesia in 2017. From 1986 to the present, she has been a full-time lecturer in the Faculty of Computer Science at the University of Indonesia, holding the Academic rank of Associate Professor. Her research interests include machine learning, deep learning, and image processing for Indonesian sign language recognition systems. She can be contacted at email: erdefi.rakun@cs.ui.ac.id.