CHAPTER V
CONCLUSION AND RECOMMENDATION

This chapter presents the conclusions and recommendations based on the results of the research conducted. The conclusions summarize the main findings regarding the implementation of the DETR model for BISINDO alphabet detection with on-the-fly data augmentation, while the recommendations are intended to guide future development.

5.1 Conclusion

Based on the results of the research conducted, the following conclusions can be drawn:

1. The Detection Transformer (DETR) model with a ResNet-50 backbone was successfully applied to detect 43 classes of BISINDO alphabets and words. The model was trained for 300 epochs using the Adam optimizer and the CosineAnnealingWarmRestarts scheduler across two scenarios: Scenario 1 without on-the-fly augmentation and Scenario 2 with on-the-fly augmentation, which included spatial transformations such as randomSizedBboxSafeCrop and horizontalFlip, as well as photometric transformations like colorJitter. Both models successfully converged with a significant reduction in loss from an initial value above 3.8 to below 0.6 at the end of training.
2. The application of on-the-fly data augmentation shows that Scenario 2 with augmentation outperforms Scenario 1 on nearly all metrics. AP@[IoU=0.50:0.95] increased from 0.6772 to 0.7188 (+6.1%), AP@0.75 increased from 0.8118 to 0.9192 (+13.2%), and AR@maxDets=100 increased from 0.7219 to 0.7949 (+10.1%). The most dramatic improvements occurred for medium-sized objects, with AP medium rising by 27.6% and AR medium by 31.7%. The only metric that was nearly equivalent was the AP@0.50, where both scenarios were close to perfect (0.9892 vs 0.9917). These results indicate that augmentation not only enhances robustness but also directly improves localization accuracy and model-scale generalization ability.
3. Scenario 1 shows rapid convergence early on but begins to exhibit signs of mild overfitting in the final phase, marked by a final train–val loss gap of 0.165 and

a stagnant validation loss while the training loss continues to decrease. Scenario 2 converges more slowly but more stably, with both curves continuing to decline together until epoch 300 and a final gap of only 0.061. The minimum validation loss value for Scenario 2 is also lower (0.447 vs 0.525), confirming better generalization ability.

4. Under normal conditions with the same subjects, Scenario 1 already experienced dropout, failing to detect anything at all, in 20 out of 43 classes (53.5% success rate), while Scenario 2 successfully detected all classes without any dropout. When lighting was varied using sunset LED lights that produced a dynamic color cast, Scenario 1 still experienced dropout in 17 classes, while Scenario 2 remained at 100%. Under the most challenging conditions, with different subjects and varying lighting, Scenario 1's success rate plummeted to just 32.6%, whereas Scenario 2 maintained 100% detection success. Furthermore, the failure of Scenario 1 is not merely dropout but also involves bounding boxes that are far off from the gesture with a high confidence score, a more dangerous form of failure because it is not immediately apparent as a failure.
5. Every weakness indicated by the PyCOCOTools metrics for Scenario 1, lowerAP@0.75, weak medium AR, and a large train–val gap, was proven to be real in real-time testing in the form of misaligned bounding boxes, failure to detect hands outside the memorized size template, and rapid performance degradation when conditions changed. The consistency between these metric findings and real-time observations reinforces the validity of the evaluation methodology used.
6. The Scenario 2 model, selected as the best model, can run in real-time at 30–55 FPS on a GPU and 5–9 FPS on a CPU, thereby meeting real-time response requirements. A web-based interface implementation using Streamlit was also successfully developed to support a more accessible demonstration without requiring special installation.

5.2 Recommendation

Based on the results of the research and analysis conducted, the following recommendations for future research are as follows:

1. The dataset in this study is still limited to a single subject in a relatively homogeneous environment. Future research is recommended to involve more subjects with varying physical characteristics (such as hand size, age, and skin color) as well as more diverse data collection conditions, including camera angle, distance, lighting, and background.
2. The augmentation techniques in this study are still limited to a few basic transformations. Further development could explore other methods such as mixup, cutmix, random erasing, or generative model-based approaches, as well as conducting ablation studies to determine the most optimal combination of augmentations.
3. To make this system truly useful in everyday communication, future development could integrate speech-to-text and text-to-speech modules to enable two-way communication: BISINDO users speak through hand gestures that are translated into text or speech, while the conversation partner can respond verbally, which is then visualized as text.
4. Performance of 5–9 FPS on a CPU indicates that the current model is still too resource-intensive to run without a GPU. Future research could explore lighter models such as RT-DETR, DETR-Lite, or knowledge distillation approaches to produce smaller yet accurate models, enabling the system to run effectively on mobile devices or embedded systems without requiring a dedicated GPU.
5. The evaluation in this study was conducted technically using quantitative metrics and robustness testing. For future research, it is recommended to conduct evaluations involving actual BISINDO users to obtain feedback on the system's accuracy in recognizing gestures performed by real users.