khazanah informatika

Jurnal Ilmu Komputer dan Informatika

Object Detection of BISINDO Sign Language Letters Using Residual Network

Maulidina Norick Eriyadi[1], Gunawan Abdillah[1], Ridwan Ilyas[1*], Asep Id Hadiana[2]

[1]Teknik Informatika
Universitas Jenderal Achmad Yani
Cimahi, Indonesia
[2]Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka
Melaka, Malaysia
*rdi@if.unjani.ac.id

Abstract-Indonesian Sign Language or BISINDO is an alternative language used by people who suffer from disabilities, especially those who have hearing impairments. This language grew and developed from the deaf community, so its use is based on the visual aspect. This research aims to apply Residual Networks to detect objects in the context of Bisindo Letter Sign Language, with the hope of increasing accuracy and efficiency in letter recognition. Object detection goes through 2 stages, namely feature extraction and model training. During image capture, sign language letters from A to Z, following BISINDO standards, are demonstrated. Additionally, hand movements for letters J and R are incorporated. ResNet is a type of Convolutional Neural Network (CNN) architecture that utilizes models that have been previously trained so that it can save the time required in the model development process. In this research, Residual Network (ResNet) was used for feature extraction to recognize important aspects in the Bisindo letter sign image, such as hand position, finger shape characteristics, and direction of movement. The research results show that the new dataset used as training data and test data has a fairly good ability to detect with a division of 70% train set, 20% valid set, and 10% test set with size 640x640 with 300 epochs for the training model.

Keywords: Object Detection; Sign Language; Bisindo Letters; Residual Networks; Single Frame.

Article info: *submitted December 21, 2023, revised April 2, 2024, accepted April 2, 2024*

1. Introduction

Communication is a process in which someone conveys messages, ideas, or concepts to others through oral or verbal means, and it can be understood by both parties. The situation is different for children with special needs who are deaf, as they face barriers to hearing and difficulties in capturing verbal communication, which can be classified based on its frequency and intensity [1]. Individuals who are deaf interact with fellow deaf individuals or with the general community through the use of sign language. The expression of sign language typically involves hand movements, facial expressions, and body gestures that form symbols to interpret specific letters or words [2]. One commonly used sign language method is Indonesian Sign Language (Bisindo). Bisindo is supported by the Indonesian Welfare Movement for the Deaf (Gerkatin) and developed by the deaf community, making it a practical and effective communication system for the deaf in Indonesia because Bisindo originates from the deaf community itself [3].

Many individuals do not understand the sign language of their conversation partner because sign language is rarely used by people without disabilities. This becomes a limitation in communication between individuals with disabilities and those without disabilities [4]. Based on these issues, research is conducted on the detection of Bisindo sign language letters using residual networks.

The introduction of letters in BISINDO is key to the development of technology that supports accessibility and effective communication for the deaf community. With its distinctive hand and body movement representations, Indonesian Sign Language provides a means to convey meaning and information beyond verbal mediums. Standardizing the letters in BISINDO becomes an essential guideline that refers to the rules of letter and word communication using hand signs and body movements. During image capture, sign language letters from A to Z, following BISINDO standards, are demonstrated. Additionally, hand movements for letters J and R are incorporated. Indonesian Sign Language is not just a communication tool but also a cultural heritage that binds the deaf community in the unity of its identity [5].

Research in the development of automatic Sign Language recognition systems reflects efforts to enhance inclusivity. This technology is not only about technological advancement but also about empowering individuals with hearing impairments to fully engage in everyday life, both in education and in the workforce. By integrating this technology into everyday devices, we open the door to more opportunities for integration and reduce communicative gaps between the deaf and hearing communities.

The choice of Residual Networks (ResNets) architecture in the development of deep learning models for sign language detection is based on several key reasons. ResNets utilize residual blocks that facilitate learning complex representations more efficiently. By leveraging residual blocks, the model can overcome challenges in object detection by performing convolutional layers only on the entire image, making it efficient for real-time object detection applications, especially in the context of sign language involving rapid movements [6] several tasks in computer vision have actively deployed CNN models for feature extraction. However, the conventional CNN models have a high computational cost and require high memory capacity, which is impractical and unaffordable for commercial applications such as real-Time on-road object detection on embedded boards or mobile platforms. To tackle this limitation of CNN models, this paper proposes a wide-residual-inception (WR-Inception.

The ability of ResNets to handle object detection problems is considered highly relevant. With the presence of shortcut connections in residual blocks, ResNets enable better gradient flow during training, improving the stability and accuracy of the model [7]. This advantage makes ResNets a suitable choice for developing sign language detection models that can be effectively integrated into mobile devices, providing easy and practical access for users. ResNets are also known for their ability to handle objects of various scales and sizes [8].

2. Methods

This research begins with capturing images using a camera to create a diverse dataset of BISINDO sign language letters. Emphasis is placed on image quality to support accurate object detection in later stages. After collecting the dataset, the next step involves annotating sign language letter areas in each image, crucial for training the object detection model. The Residual Network (ResNet) method is used for modeling, known for addressing the vanishing gradient problem. The final step includes data evaluation using metrics like IoU, precision, recall, and mAP to assess the model's performance in recognizing BISINDO sign language letters. mAP is the primary parameter for overall performance, while IoU, precision, and recall provide insights into detection accuracy and completeness.

a. Data Collection and Annotation

In the image capture process, assistance is provided through the demonstration of sign language letters according to the BISINDO standards for all letters from A to Z. Specifically, two letters, J and R, are added with hand movements during image capture. The approach applied in this research is object detection with a limitation to a single frame.

For each class, 7 images are captured with 4 different outfits (subtracting 1 image for 1 outfit) as shown in Figure 1, resulting in a total of 26 classes. The total number of images is 727, with a distribution of 70% (508 images) for the train set, 20% (147 images) for the validation set, and 10% (72 images) for the test set. The dataset is split into 70%, 20%, and 10% portions to prevent overfitting of the model and to enhance its accuracy.

The object detection data annotation process utilized the Roboflow platform. Roboflow is a web-based platform that provides various functionalities related to datasets. The use of Roboflow allows users to share datasets and efficiently process datasets. Some features implemented in this research include the ability to annotate or label objects to be detected using bounding boxes. Additionally, Roboflow can be used for dataset pre-processing, such as conversion to grayscale and automatic augmentation. In this case, auto-oriented with a size of 640x640 and automatic augmentation were employed.

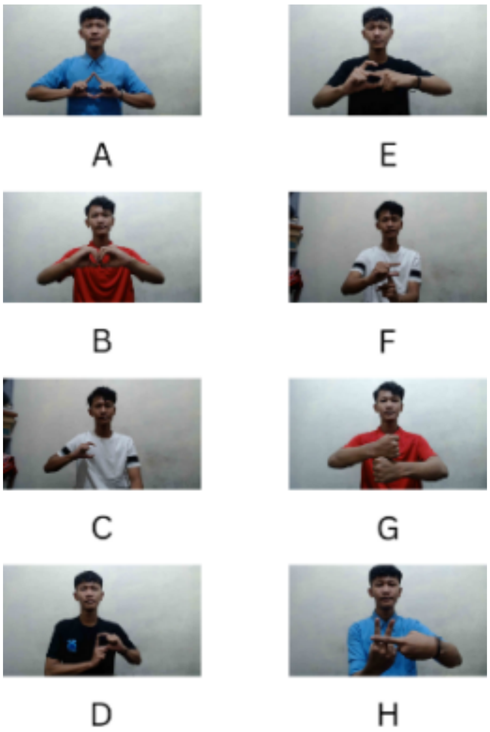

Figure 1. Example Image Data for Each Character

The entire collected image data is then annotated according to their respective classes. The process begins by selecting a folder and uploading it to Roboflow for annotation. The annotator then marks the areas of the hand shapes representing BISINDO sign language letters with the available bounding boxes. Subsequently, they are labeled according to their classes, such as ‘Letter A’ in Figure 2. The annotated data is then uploaded as a dataset, where Roboflow automatically separates it into the train set, validation set, and test set, as explained in point 1.

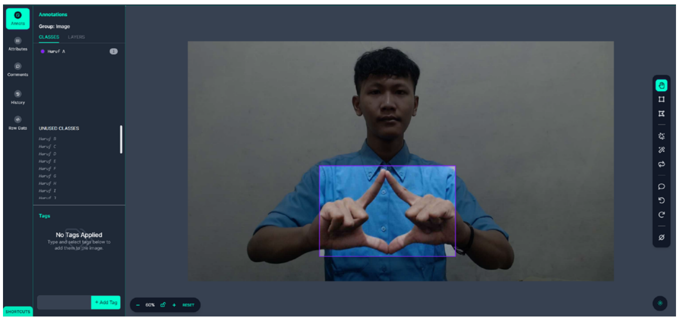

Figure 2. Example Annotation of Letter A Data

b. Modeling with ResNets and Evaluation

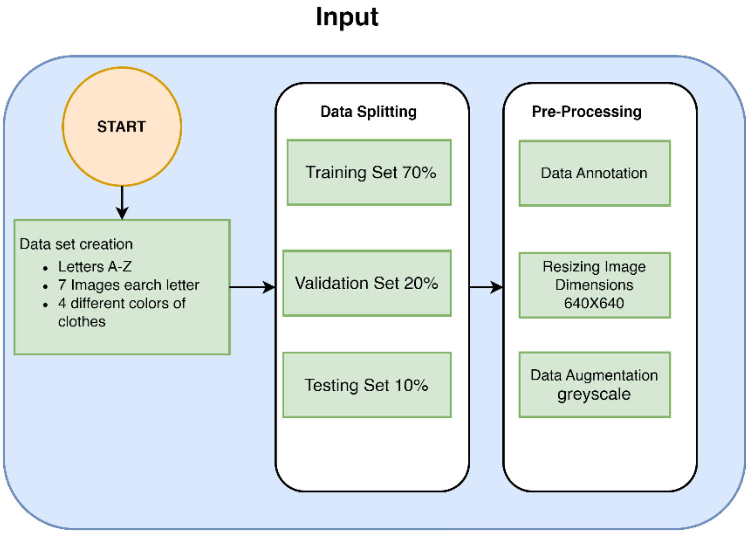

Figure 3a. Input Diagram Method of Sign Character Classification

This research aims to create a high-quality image dataset containing a series of hands forming letters in Indonesian Sign Language (BISINDO), following established standards. To achieve this goal, several tools and specifications are used meticulously. The tools utilized include a Logitech Webcam and a stabilizing Tripod, chosen to ensure optimal image quality.

In accordance with what has been written in Figure 3a, scenario for capturing images is based on a selfie process, where subjects direct their hands to form each desired BISINDO letter. It is important to note that the entire image capture process must adhere to BISINDO standards, including specific movements required for certain letters such as J and R. In this process, an object detection approach is applied within a single frame to ensure suitability and consistency in results. Each letter class is represented by seven images taken with four different clothing options, resulting in a total of 728 images. The dataset is divided meticulously, with 70% used as a train set, 20% as a validation set, and 10% as a test set. Factors such as adequate lighting and the correct distance between the camera and subject are also taken seriously. Sufficient and uniform lighting is key to ensuring the resulting images are clear and interpretable, while a consistent distance of 1.5 meters from the camera ensures consistency in perspective and object size. Thus, the generated dataset is expected to be a valuable resource for research and development in the field of Indonesian Sign Language.

Pre-processing is a vital initial step in creating the dataset, aiming to prepare the data for the machine learning model’s input requirements. Resizing the image dimensions to a standardized 640x640 pixels is essential for ensuring the model can efficiently process the images uniformly, enabling accurate analysis across the dataset.

Furthermore, data annotation is pivotal for associating the images with their corresponding letters, facilitating the model's learning process by establishing the correlation between images and labels.

Data augmentation, another crucial technique, serves to enhance the dataset's size and the model's generalization capabilities. Within this context, converting the images to grayscale is employed, a common method to reduce color complexity and focus on the letter's shape and texture. This approach aids the model in learning robust features, thereby improving its accuracy in recognizing letters amidst various colors and backgrounds. By leveraging data augmentation, the machine learning model can acquire a broader range of patterns and relationships, enhancing its proficiency in accurately classifying unseen data.

A Convolutional Neural Network (CNN) is a type of artificial neural network that is commonly used for image classification tasks. The CNN is made up of various layers that perform specific functions to extract features from the input data. Based on the Figure 3b, the process begins with a kernel, which is a small matrix that is used to perform a convolution operation on the input data. The input data is typically an image, and the output is a feature map that highlights the presence of certain features in the input data.

Hyperparameter tuning is the process of adjusting the hyperparameters of the CNN to optimize its performance. This is an important step in the CNN development process, as it can significantly impact the accuracy of the model. Pooling is another technique used to reduce the spatial dimensions of the feature maps, while retaining important information. This is done to reduce the computational complexity of the CNN and to prevent overfitting [9].

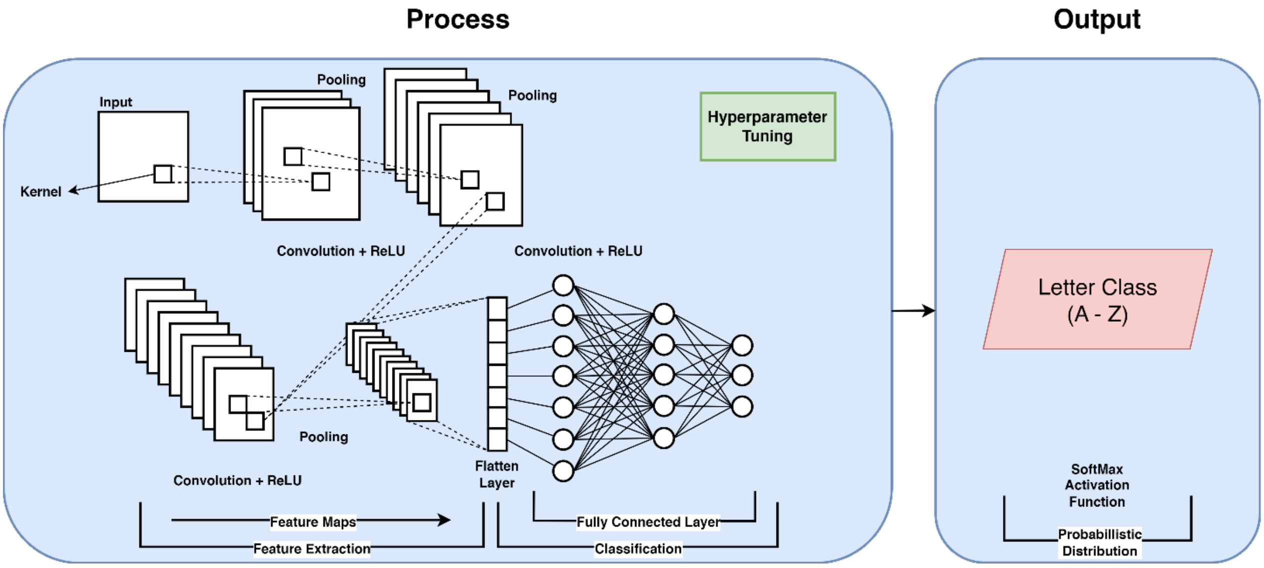

Figure 3b. Process and Output Diagram Method of Sign Character Classification

Convolution is a mathematical operation that combines the kernel and the input data to produce a feature map. The rectified linear unit (ReLU) is an activation function that is applied to the feature maps to introduce non-linearity. This is important because the input data is typically linear, and the CNN needs to be able to learn non-linear relationships.

Feature extraction is the process of extracting features from the input data using the CNN. The CNN extracts these features by applying a series of convolutions and pooling operations to the input data. These features are then used to classify the input data. The final layer of the CNN is the fully connected layer, which is responsible for producing the final classification of the input data. The fully connected layer uses the features extracted by the previous layers to determine the class of the input data.

The activation function is a mathematical function that is applied to the output of the CNN to introduce non-linearity. The SoftMax function is a type of activation function that is commonly used in the final layer of the CNN to produce a probabilistic distribution over the possible classes. This is important because it allows the CNN to make probabilistic predictions about the class of the input data.

Overfitting is when the CNN is too complex and learns patterns in the training data that do not generalize to new data. Underfitting is when the CNN is not complex enough to learn the patterns in the training data. Both of these issues can be addressed through hyperparameter tuning and regularization techniques. CNNs have a wide range of applications, including image and video recognition, natural language processing, and speech recognition. They are particularly well-suited for image and video recognition tasks because they are able to extract features from the input data that are invariant to translation, scaling, and rotation. In summary, the CNN is a powerful tool for image classification tasks, and it is made up of various layers that perform specific functions to extract features from the input data. Hyperparameter tuning and pooling are important techniques for optimizing the performance of the CNN, and the activation function is used to introduce non-linearity into the model.CNNs have a wide range of applications, and they are particularly well-suited for image and video recognition tasks [10].

Residual Networks (ResNets), introduced in the 2016 paper "Deep Residual Learning for Image Recognition" by Shaoqing Ren, Kaiming He, Jian Sun, and Xiangyu Zhang, have emerged as a prominent and successful deep learning model. Residual learning, a key component of ResNets, addresses challenges in training deep networks by tackling the vanishing gradient problem. Traditional networks struggle with this issue as gradients diminish during backward propagation through multiple layers, hindering effective weight updates in early layers. In residual learning, "skip connections" or "residual connections" are introduced to enable the network to learn the residual mapping between input and output directly. By focusing on learning differences instead of the complete mapping, this approach facilitates easier optimization of weights. These skip connections pass through one or more layers, allowing for more direct gradient flow during backpropagation [11].

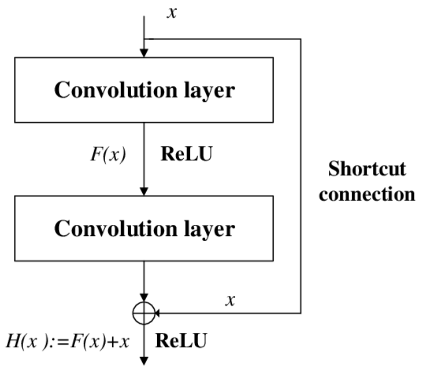

Figure 4. Residual Learning: Building Block

The study demonstrates that the residual learning architecture as shown Figure 4, effectively tackles the vanishing gradient problem in detecting Indonesian Sign Language (BISINDO) letters within a single frame. This success in image recognition tasks, particularly with deep residual networks (ResNets), is leveraged to optimize the performance of recognizing hand movements representing BISINDO letters. The implemented residual learning architecture in this research enhances accuracy and stability in single-frame object detection for BISINDO sign language letters. Key to its effectiveness are the skip connections in the residual block, aiding the model in discerning differences between hand representations and the desired BISINDO letter representations [12]. This approach proves to be a standard and effective choice for improving BISINDO sign language letter detection at the single-frame level. It significantly enhances model accuracy and stability by addressing variations in poses and shifts within a frame, which is crucial for interpreting hand movements that visually represent letters in sign language. The skip connections play a vital role in capturing differences between input and output representations [13], allowing the model to focus on detecting critical features in hand signals within a single frame. The inherent capability of residual networks to address the vanishing gradient [14] problem is pivotal for optimizing object detection in hands and sign language movements.

Intersection over Union (IoU) is a metric used to measure the extent to which two bounding boxes overlap. In accordance with what has been written in Figure 5, IoU is calculated by dividing the overlap area between the predicted bounding box and the ground-truth bounding box by the total combined area of both. IoU is often used to assess how accurate detection is in object detection algorithms. The higher the IoU value, the better the performance of the object detection. This metric helps determine whether detection is considered true or false in the evaluation of the algorithm [15].

Precision is the ability of a model to recognize only objects that are truly relevant or important. Precision is calculated by dividing the number of true positive object detections by the total number of positive detections made by the model. Precision measures the accuracy of the model in classifying objects as positive, meaning objects that are genuinely relevant or important [16]. The higher the precision value, the fewer objects misclassified as positive.

Recall is the ability of a model to detect all relevant or important objects. Recall calculation is done by dividing the number of true positive object detections by the total number of objects actually present in the image or video. Recall measures the effectiveness of the model in finding all relevant or important objects, regardless of whether the model classifies the objects correctly or incorrectly [17]. The higher the recall value, the more objects successfully found by the model.

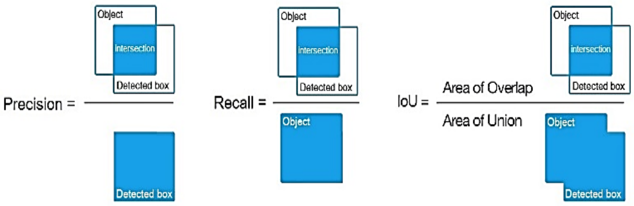

Figure 5. The Evaluation Methods

mAP (mean Average Precision), as written in Equation 1, is an evaluation metric that measures the accuracy level of an object detection model across all object classes in a dataset. mAP is calculated by taking the average precision value for each object class. Average precision itself is computed by measuring the area under the precision-recall curve for each object class. mAP provides a comprehensive overview of the object detection model's performance for all object classes in the dataset [18] but an efficient implementation is still absent. Current implementations can only count true positives (TP's. The higher the mAP value, the better the performance of the object detection model. mAP is often used as the primary evaluation metric in various object recognition contests.

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

3. Results

The object detection research for Indonesian Sign Language (BISINDO) letters in a single frame starts with the data training phase, utilizing the residual learning architecture to address the vanishing gradient problem. This phase focuses on learning distinctive hand-feature representations of BISINDO letters. Following the training, the resulting models are evaluated to select the best-performing one. Evaluation metrics include accuracy, precision, recall, and others relevant to single-frame object detection.

The data training process employs ResNet architecture to recognize hand characteristics representing BISINDO letters. ResNet, with its skip connections, enables effective learning of differences between input hand images and desired outputs, overcoming the vanishing gradient problem. The training data, consisting of visual representations of sign language gestures, is applied to the ResNet model through iterations. The model adjusts by minimizing differences between predicted and actual labels, enhancing its ability to recognize and classify BISINDO letters. The training aims to produce an optimal model capable of understanding sign language letters at the single-frame level.

a. Loss Classification Result

The Box Loss graph displays results from the Box Loss algorithm, aiming to minimize object detection errors below a specified threshold [19] they require image processing algorithms to inspect contents of images. This project compares 3 major image processing algorithms: Single Shot Detection (SSD. The analysis results from the figure 6, a lower Box Loss value signifies a better understanding of the dataset. Similarly, the Class Loss graph, utilizing the Class Loss algorithm, focuses on reducing classification errors, with a smaller value indicating improved dataset understanding. The Object Loss graph combines both object detection and classification errors, aiming for a minimized value to enhance overall understanding [20]. Analysis reveals decreasing trends in all graphs, signifying algorithmic performance improvement on the dataset. Consistent trends without significant points or changes imply clear visibility of differences between the training and testing phases. In summary, these results demonstrate the algorithm's successful learning from the dataset.

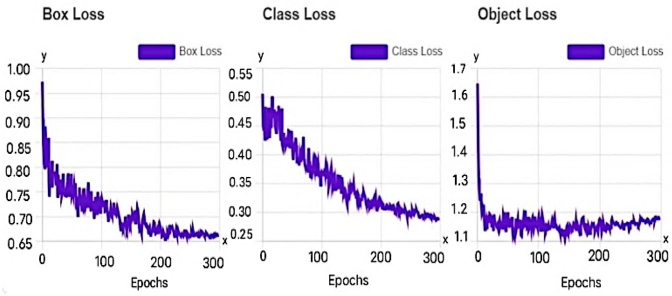

Figure 6. Evaluation Loss

b. mAP Result

The analysis results from the Figure 7 provide a clear understanding of the ResNet model's performance and influencing factors [21] systematic research on exploring the model performance and guiding the design of new convolutional neural network (CNN). An epoch in machine learning refers to the number of times the entire training dataset has been processed by the algorithm during training [22] algorithms that use deep learning to recognize faces have become more popular. The majority of them are predicated on extremely accurate but complicated Convolutional Neural Networks (CNNs). The model's performance, especially in terms of mAP and FPS, shows a steady improvement until reaching epoch 200. However, after reaching its peak, there is a noticeable decline in performance, indicating that the model may struggle to sustain learning and optimization beyond epoch 250. Then, changes in data distribution have changed between epoch 250 and the next epoch, causing worse model performance.

The mAP, crucial for evaluating object recognition models in tracking, exhibits a decline after epoch 200. mAP, covering the overall precision of predicted objects, shows a decrease, while munn, calculating the precision of outer objects, also experience a drop. On the other hand, the propony and Epochs metrics, used to assess ResNet model performance, show consistent improvement up to epoch 200 but then decline. It may indicate that the model begins to recognize objects with very small proportions or detects them incorrectly. Overall, the analysis results suggest that the ResNet model's performance decreases after reaching epoch 250. This could be interpreted as the model starting to experience "forgetting" and struggling to maintain the proportions of predicted objects.

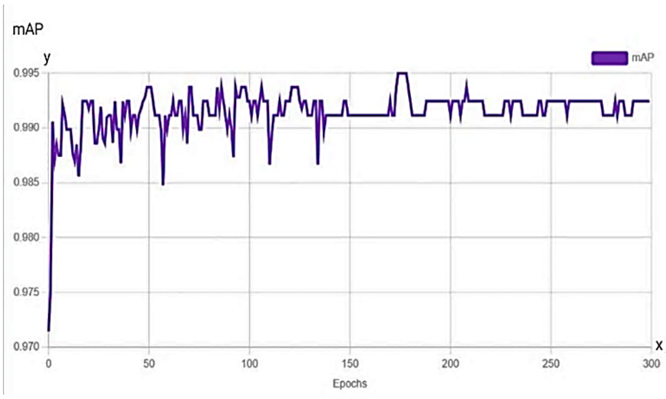

Figure 7. mAP Model

4. Discussion

The exceptional performance of the Residual Networks, even when applied to a different sign language, highlights the strength and versatility of the model given a comprehensive and representative dataset. The high level of accuracy, as reflected by the PbC and mAP metrics, indicates the model's ability to effectively identify and interpret various signs or letters within the new sign language. This has significant implications for the deaf community, as it signifies the potential for advanced technology to improve communication and accessibility in their everyday lives. A precise sign language recognition system can facilitate smoother interactions between deaf individuals and the broader community, ultimately enhancing access to information and services. The near-perfect AP values observed in certain classes, such as 'B', underscore the model's precision and reliability, which are essential for effective communication and understanding in sign language interactions. Consequently, these results not only demonstrate the technical capabilities of the algorithm but also its potential to positively impact the lives of deaf individuals by overcoming communication barriers.

5. Conclusion

The analysis of the Box Loss, Class Loss, and Object Loss graphs underscores the ResNet model's progressive learning from the dataset, with decreasing trends indicating algorithmic performance enhancement. However, performance peaks around epoch 200, followed by a decline suggesting potential struggles in sustaining learning beyond epoch 250, possibly due to changes in data distribution. This decline is reflected in metrics such as mAP and FPS, highlighting a potential challenge in maintaining object recognition precision and proportions. Moreover, the Precision by Class metric demonstrates high accuracy across various classes, with some achieving near-perfect scores, emphasizing the model's proficiency in multi-class data. Practically, these findings imply the need for strategies to mitigate performance decline post-epoch 200, possibly through continual dataset refinement or model adaptation techniques to prevent "forgetting" phenomena and maintain accuracy. Future directions could explore dynamic learning rate adjustments or novel architectures to address these challenges and sustain model performance over extended training periods.

6. Acknowledgement

This research received funding from the Directorate General of Higher Education (Dikti) Research Grant program in the category of Junior Lecturer Research for the 2023 period. The study was also conducted in collaboration with PT. Inovasi Disabilitas Indonesia (Indonesian Disability Innovation).

References

[1] M. Dewi, T. Wahyuningrum, and N. A. Prasetyo, "Pengenalan Kata Bahasa Isyarat Indonesia (BISINDO) Menggunakan Augmented Reality (AR)," *INISTA: Journal of Informatics, Information System, Software Engineering and Applications*, vol. 3, no. 2, pp. 53–60, 2021, [Online]. Available: https://journal.ittelkom-pwt.ac.id/index.php/inista/article/view/256

[2] J. Zheng, Y. Chen, C. Wu, X. Shi, and S. M. Kamal, "Enhancing Neural Sign Language Translation by highlighting the facial expression information," *Neurocomputing*, vol. 464, pp. 462–472, 2021, doi: 10.1016/j.neucom.2021.08.079.

[3] R. I. Borman and B. Priyopradono, "Implementasi Penerjemah Bahasa Isyarat Pada Bahasa Isyarat Indonesia (BISINDO) Dengan Metode Principal Component Analysis (PCA)," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 3, no. 1, pp. 103–108, 2018, doi: 10.30591/jpit.v3i1.631.

[4] S. Apendi, C. Setianingsih, and M. Paryasto, "Deteksi Bahasa Isyarat Sistem Isyarat Bahasa Indonesia Menggunakan Metode Single Shot Multibox Detector," *eProceedings of Engineering*, vol. 10, no. 1, pp. 249–255, 2023, [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/19322

[5] D. R. Kurnia and T. Slamet, "MENORMALKAN YANG DIANGGAP 'TIDAK NORMAL' (Studi Kasus Penertiban Bahasa Isyarat Tunarungu di Sekolah Luar Biasa [SLB] dan Perlawananya di Kota Malang)," *Ijds*, vol. 3, no. 1, pp. 34–43, 2016, [Online]. Available: http://ijds.ub.ac.id

[6] Y. Lee, H. Kim, E. Park, X. Cui, and H. Kim, "Wide-residual-inception networks for real-Time object detection," *IEEE Intelligent Vehicles Symposium, Proceedings*, pp. 758–764, 2017, doi: 10.1109/IVS.2017.7995808.

[7] D. Han, J. Kim, and J. Kim, "Deep pyramidal residual networks," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 6307–6315, 2017, doi: 10.1109/CVPR.2017.668.

[8] M. Effati and G. Nejat, "A Performance Study of CNN Architectures for the Autonomous Detection of COVID-19 Symptoms Using Cough and Breathing," *Computers*, vol. 12, no. 2, 2023, doi: 10.3390/computers12020044.

[9] H. Q. T. Ngo, "Design of automated system for online inspection using the convolutional neural network (CNN) technique in the image processing approach," *Results in Engineering*, vol. 19, no. August, 2023, doi: 10.1016/j.rineng.2023.101346.

[10] F. Wagner, A. Eltner, and H. G. Maas, "River water segmentation in surveillance camera images: A comparative study of offline and online augmentation using 32 CNNs," *International Journal of Applied Earth Observation and Geoinformation*, vol. 119, no. April, p. 103305, 2023, doi: 10.1016/j.jag.2023.103305.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.

[12] R. R. E. PRASETYO and M. ICHWAN, "Perbandingan Metode Deep Residual Network 50 dan Deep Residual Network 152 untuk Deteksi Penyakit Pneumonia pada Manusia," *MIND Journal*, vol. 6, no. 2, pp. 168–182, 2021, doi: 10.26760/mindjournal.v6i2.168-182.

[13] O. K. Oyedotun, K. Al Ismaeil, and D. Aouada, "Training very deep neural networks: Rethinking the role of skip connections," *Neurocomputing*, vol. 441, pp. 105–117, 2021, doi: 10.1016/j.neucom.2021.02.004.

[14] H. Wang, S. Xu, K. bin Fang, Z. S. Dai, G. Z. Wei, and L. F. Chen, "Contrast-enhanced magnetic resonance image segmentation based on improved U-Net and Inception-ResNet in the diagnosis of spinal metastases," *Journal of Bone Oncology*, vol. 42, p. 100498, 2023, doi: 10.1016/j.jbo.2023.100498.

[15] C. Caroline, A. Yogta, R. Thayeb, H. Hermawati, S. Dwijayanti, and B. Y. Suprapto, "Identifikasi Jalan Kampus Universitas Sriwijaya Berbasis Fully Convolutional Networks," *Jurnal Surya Energy*, vol. 4, no. 1, pp. 353–358, 2020, doi: 10.32502/jse.v4i1.2057.

[16] K. Oksuz, B. C. Cam, E. Akbas, and S. Kalkan, "Localization Recall Precision (LRP): A New Performance Metric for Object Detection" arXiv:1807.01696v2 [cs.CV] 5 Jul 2018, doi: https://arxiv.org/abs/1807.01696.

[17] R. Padilla, W. L. Passos, T. L. B. Dias, S. L. Netto, and E. A. B. Da Silva, "A comparative analysis of object detection metrics with a companion open-source toolkit," *Electronics (Switzerland)*, vol. 10, no. 3, pp. 1–28, 2021, doi: 10.3390/electronics10030279.

[18] B. Wang, "A Parallel Implementation of Computing Mean Average Precision," no. 2016, pp. 1–15, 2022, [Online]. Available: http://arxiv.org/abs/2206.09504

[19] S. Srivastava, A. V. Divekar, C. Anilkumar, I. Naik, V. Kulkarni, and V. Pattabiraman, "Comparative analysis of deep learning image detection algorithms," *Journal of Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00434-w.

[20] W. Chen, Y. Li, Z. Tian, and F. Zhang, "2D and 3D object detection algorithms from images: A Survey," *Array*, vol. 19, no. June, p. 100305, 2023, doi: 10.1016/j.array.2023.100305.

[21] F. Chen and J. Y. Tsou, "Assessing the effects of convolutional neural network architectural factors on model performance for remote sensing image classification: An in-depth investigation," *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, no. June, p. 102865, 2022, doi: 10.1016/j.jag.2022.102865.

[22] A. J and P. Suresh L, "A novel fast hybrid face recognition approach using convolutional Kernel extreme learning machine with HOG feature extractor," *Measurement: Sensors*, vol. 30, no. January, p. 100907, 2023, doi: 10.1016/j.measen.2023.100907.