한국컴퓨터그래픽스학회
Korea Computer Graphics Society Vol. 31, No. 3, P. 161~170

정적 손 제스처 인식을 위한 딥러닝 모델 비교 연구: DenseNet vs ViT with HaGRID

후세인 무하마드 아브라르[1] 김성기[1*]

[1]조선대학교 컴퓨터공학과

abrar@chosun.ac.kr, skkim@chosun.ac.kr

A Comparative Study of Deep Learning Models for Static Hand Gesture Recognition: DenseNet vs. ViT with HaGRID

Hussain Muhammad Abrar[1] Kim SeongKi[1*]

[1]Department of Computer Engineering, Chosun University

요 약

본 연구는 증강현실·가상현실 등 HCI 응용 분야에서 중요한 손 제스처 인식(HGR)을 위해 대규모 HaGRID v2 512px 데이터셋을 활용하여 DenseNet-121과 Vision Transformer (ViT-B/16) 모델을 비교 평가하였다. DenseNet-121은 95.32%, ViT-B/16은 95.45%의 테스트 정확도를 기록하였으며, 특히 ViT-B/16은 시각적으로 유사한 제스처 간 오분류를 줄이는 데 효과적이었다. 연구 결과는 트랜스포머 기반 모델이 그래픽 중심의 HCI 분야에서 실용적임을 보여준다.

Abstract

Hand Gesture Recognition (HGR) is essential for natural and intuitive interactions in fields like augmented reality, virtual reality, and mixed reality, significantly enhancing the user experience in human-computer interaction (HCI) applications. In this study, we present an extensive evaluation of static Hand Gesture Recognition models using the large-scale HaGRID v2 512px dataset, comprising 1,086,167 RGB images, covering 33 gesture classes along with a dedicated no_gesture category, from over 65,977 unique individuals. We systematically benchmark two state-of-the-art deep learning models: a lightweight convolutional neural network (DenseNet-121), trained from scratch, and a Vision Transformer (ViT-B/16) fine-tuned from ImageNet-21k pre-training. Under identical training conditions, DenseNet-121 achieves a validation accuracy of 94.98% and a test accuracy of 95.32%, whereas ViT-B/16 attains a validation accuracy of 94.71% and a test accuracy of 95.45%, demonstrating clear comparative performance. Additionally, ViT-B/16's global self-attention mechanism notably reduces misclassification errors, particularly in visually similar gesture classes. Our results highlight the viability and efficiency of transformer-based architectures for accurate, real-time gesture recognition in graphics-intensive HCI applications.

키워드: 손동작 인식, 비전 트랜스포머, DenseNet-121, HaGRID, 컴퓨터 그래픽스

Keywords: Hand Gesture Recognition, Vision Transformer, DenseNet-121, HaGRID, Computer Graphics

1. Introduction

Hand gesture recognition (HGR) has emerged as a pivotal technology for intuitive human-computer interactions (HCI), enabling natural and seamless user experiences across various application domains, including augmented reality (AR), virtual reality (VR), mixed reality (MR), automotive interfaces, and smart home environments. By interpreting human gestures, systems become more responsive and

*corresponding author: Kim SeongKi / Chosun University (skkim@chosun.ac.kr)

Received : 2025.06.13./ Review completed : 1st 2025.06.30. 2nd 2025.07.10./ Accepted : 2025.07.14.
DOI : 10.15701/kcgs.2025.31.3.161
ISSN : 1975-7883(Print)/2383-529X(Online)

immersive, significantly improving usability and accessibility. Recent advancements in deep learning have significantly advanced the field

Gestures	Percentage
Call	2.6
Dislike	2.9
Fist	2.9
Four	2.9
Grabbing	3.3
Grip	3.4
Hand Heart	2.7
Hand Heart2	2.9
Holy	3.6
Like	2.9
Little Finger	3.3
Middle Finger	3.5
Mute	3.0
No Gesture	0.2
Ok	2.9
One	2.9
Palm	2.9
Peace	2.9
Peace Inverted	2.7
Point	3.5
Rock	3.0
Stop	2.9
Stop Inverted	2.8
Take Picture	2.6
Three	2.8
Three Gun	3.5
Three2	2.7
Three3	3.7
Thumb Index	4.3
Thumb Index2	1.7
Timeout	2.7
Two Up	2.8
Two Up Inverted	2.8
Xsign	3.6

Table 1. Distribution of Gesture classes in HaGRID

of human-computer interaction (HCI), particularly through vision-based systems that enable more natural, intuitive interfaces. Convolutional neural networks (CNNs) remain a strong baseline for this task due to their proven ability to extract local spatial hierarchies. Among these, DenseNet-121 stands out for its dense connectivity

pattern, which facilitates feature reuse, mitigates vanishing gradients, and achieves parameter efficiency qualities that are particularly advantageous in recognizing fine-grained gesture differences.

Traditional input devices like keyboards and mice are increasingly being supplemented by gesture-based controls, allowing users to interact with digital environments using body movements. As noted by Sachdeva (2023) [1], the evolution of HCI is marked by a shift from basic graphical interfaces to intelligent systems capable of interpreting gestures, voice, and other non-verbal inputs, driving ongoing innovation in seamless human machine communication.

The proliferation of HGR technologies closely aligns with developments in computer graphics, particularly in rendering realistic, interactive, and responsive environments. Accurate gesture recognition directly enhances graphical user interfaces by facilitating real-time, gesture-driven interactions, thus expanding the horizons of immersive experiences in AR, VR, and MR systems. Advanced HGR can significantly improve user engagement, interaction fidelity, and overall graphical immersion by accurately capturing and interpreting fine-grained gestures in dynamic scenarios. Recent work by Padmakala (2024) [2] further reinforces this by demonstrating how hyperparameter-optimized deep convolutional networks can dramatically improve recognition performance on gesture-rich datasets like HaGRID, highlighting the growing importance of tailored architectures in high-fidelity interaction systems.

A. S. M. Miah (2023) [3] proposed a multi-culture sign language recognition framework using graph-based and deep learning models, emphasizing the importance of robust spatial-temporal feature representation. Their findings reinforce the need for flexible and scalable HGR models applicable across varied linguistic and cultural contexts.

Despite recent progress, substantial challenges persist. Variability in human gestures due to individual differences, environmental conditions, occlusions, and diverse lighting scenarios complicates accurate gesture recognition. Moreover, achieving real-time performance alongside high recognition accuracy remains challenging, particularly on resource-constrained platforms typical in mobile and

wearable devices (Kapitanov et al., 2024) [4]. Thus, there is an ongoing need for systematic evaluation and optimization of robust and efficient HGR models.

To address these challenges, large-scale, diverse, and annotated datasets are crucial. The HaGRID v2 512px dataset used in this study contains over 1 million images spanning 34 gesture classes, covering a wide range of scenarios and subject variability. Collected from more than 65,000 unique subjects and annotated at high resolution, it is among the most comprehensive publicly available datasets for static hand gesture classification. Table 1 illustrates the distribution of gesture classes in the HaGRID v2 512px dataset, highlighting both class diversity and the imbalance in sample counts. Figure 1 shows the sample gesture images from the HaGRID v2 dataset, representing a diverse range of static hand poses used for interaction in HCI applications. These include single-hand and two-hand gestures, as well as culturally specific and device-control gestures. Leveraging this dataset, researchers can more reliably benchmark advanced neural architectures, exploring deep learning techniques such as convolutional neural networks (CNNs) and Vision Transformers (ViT).

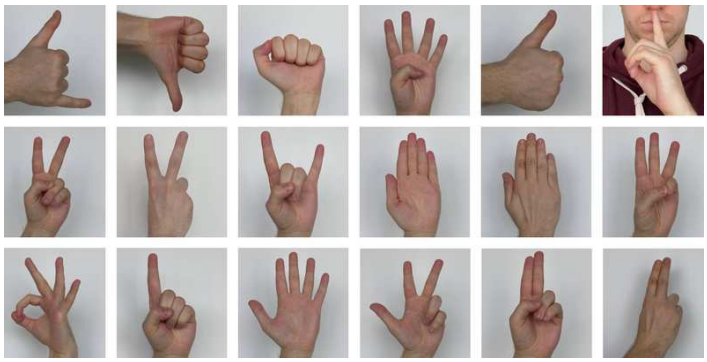

Figure 1. Sample Gestures of HaGRID

A key technical contribution of this work lies in its detailed interpretability analysis, which goes beyond standard accuracy reporting. By examining confusion matrices in depth, we identify and interpret gesture-level misclassifications, especially between semantically and visually similar gestures (e.g., peace vs peace_inverted, three2 vs three3). This analysis reveals model-specific strengths in fine-grained gesture discrimination, where DenseNet-121 showed better differentiation in certain closely related gestures despite having significantly fewer parameters than ViT-B/16.

In addition, we employed a custom training pipeline that includes learning rate scheduling and class weight balancing, which contributed to stable convergence and improved recognition performance for underrepresented gesture classes. These implementation choices enhanced both sensitivity and specificity, particularly in the presence of class imbalance a common challenge in real-world gesture datasets.

In this study, we present a comparative analysis of two deep learning models for hand gesture recognition: DenseNet-121 (trained from scratch) and ViT-B/16 (fine-tuned from pre-trained weights). Using the HaGRID v2 512px dataset, we evaluate their accuracy and efficiency for real time gesture based interaction in graphics intensive applications. Our findings support the development of gesture-driven interfaces in computer graphics and HCI, offering practical insights for future systems that require both visual and interaction realism.

2. Related Work

Hand gesture recognition (HGR) plays a critical role in advancing natural user interfaces within human-computer interaction (HCI), smart environments, and immersive systems such as augmented and virtual reality. With the increasing need for intuitive and contactless interaction methods, researchers have extensively explored deep learning approaches to improve recognition accuracy, real-time responsiveness, and robustness across diverse environments.

Sharma et al. (2021) [5] proposed a vision-based hand gesture recognition system utilizing deep learning for the interpretation of sign language. Their approach leverages a deep convolutional neural network to extract discriminative spatial features from hand images, enabling effective recognition of complex static sign gestures. The study highlights the capability of CNNs to address variability in hand pose and lighting conditions, supporting robust sign language translation in diverse environments.

Building CNN-based architectures for practical deployment, Sahoo et al. (2022) [6] presented a real-time hand gesture recognition framework using a fine-tuned convolutional neural network. Their method emphasizes rapid and accurate gesture classification, optimized for low-latency scenarios essential in real-time HCI applications. The experimental results confirm that, with suitable fine-tuning and efficient pipeline design, CNNs can deliver high recognition rates without significant computational overhead.

In recent years, transformer-based models have begun to reshape the landscape of gesture and sign language recognition. Hu et al. (2021) [7] introduced SignBERT, a pre-trained transformer model specifically designed to capture hand-model-aware representations for sign language recognition. By focusing on both spatial and temporal cues

from video sequences, SignBERT demonstrates superior performance in learning fine-grained gesture dynamics, offering a pathway to improved generalization across users and sign variations.

Montazerin et al. (2023) [8] further advanced transformer-based HGR by proposing a novel model that integrates instantaneous and fused neural decomposition of high-density electromyography (EMG) signals. Their Compact Transformer-based Hand Gesture Recognition (CT-HGR) framework efficiently captures both temporal and spatial dependencies in muscle activity data, outperforming conventional CNNs and classical machine learning methods, particularly for complex and subtle hand motions.

Smith et al. (2023) [9] employed a Deep Convolutional Neural Network (CNN) combined with a novel sterile data augmentation technique, using an FMCW mmWave radar dataset consisting of real and synthetic ("sterile") hand gesture data, achieving an accuracy of 95.4% on static hand gesture classification tasks.

Bristy Chanda(2024) [10] approached a combining semantic segmentation using the U-Net architecture and a score-level fusion of fine-tuned convolutional neural networks (ResNet50 and VGG16) was employed for static hand gesture recognition. The model was evaluated on the National University of Singapore (NUS) hand posture dataset II, which comprises 2000 images equally distributed among 10 gesture classes. Experimental results demonstrated that the proposed method achieved superior performance, reaching an accuracy of 99.92%, outperforming alternative CNN architectures such as VGG16 (99.75%), VGG19 (99.00%), ResNet50 (98.70%), and Inception V3 (96.75%).

Raju et al. (2025) [11] developed a CNN-based real-time static hand gesture recognition system trained on a publicly available static hand gesture image dataset (such as the ASL alphabet dataset), attaining an accuracy of 96.1% and demonstrating effective real-time processing capability at approximately 30 frames per second.

Our work contributes to the ongoing discussion in the field by benchmarking compact deep learning models that maintain reasonable accuracy while being suitable for deployment in resource-limited environments, such as mobile AR systems or embedded smart home controllers.

3. Proposed Method

Hand gesture recognition relies on selecting efficient deep learning architectures. This study compares DenseNet-121 and Vision Transformer (ViT-B/16) for multi-class gesture classification using uniformly preprocessed images resized to 224×224.

Figure 2 shows the architecture used for gesture recognition based on DenseNet-121, a densely connected convolutional neural network. DenseNet mitigates vanishing gradients and promotes better feature reuse by connecting each layer to every other layer in a feed-forward manner. Within a dense block, feature maps from all preceding layers are concatenated, enabling deeper supervision and richer representations.

The network begins with a convolutional layer of size 224×224×3, followed by the DenseNet-121 backbone, where feature extraction is performed. The resulting tensor is passed through a Global Average Pooling (GAP) layer to reduce the spatial dimensions. This is followed by a Dropout layer (rate=0.5) to reduce overfitting, then a Dense layer with 256 ReLU-activated neurons, and another Dropout layer (rate=0.3). Finally, the output is mapped to gesture classes using a Dense softmax layer that outputs class probabilities, mathematically

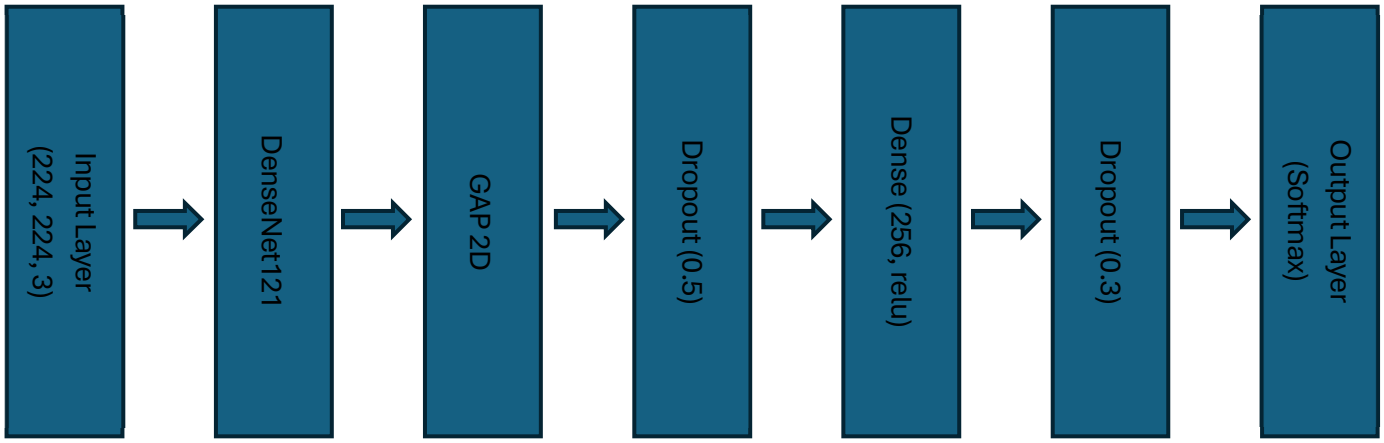

Figure 2. DenseNet-121 Gesture Recognition Architecture

described as:

$$x_1 = H_l([x_0, x_1, \dots, x_{l-1}]) \quad (1)$$

where x_l is the output of the $l-th$ layer, $[x_0, x_1, \dots, x_{l-1}]$ denotes the concatenation of feature maps from layers 0 to $l-1$, and $H_l(.)$ represents a composite function of batch normalization, ReLU activation, and convolution.

The output feature tensor after the final dense block is globally average pooled, resulting in a feature vector $z \in R^d$. A fully connected layer projects z onto a C dimensional output, where C is the number of gesture classes. The softmax activation computes the class probabilities:

$$p_i = \frac{\exp(w_i^T z + b_i)}{\sum_{j=1}^{c} \exp(w_i^T z + b_i)} \quad (2)$$

where w_i, b_i are the weights and bias for class i.

The model is trained using the categorical cross-entropy loss:

$$L = -\sum_{i=1}^{c} y_i \log(p_i) \quad (3)$$

where y_i is the true label (one-hot encoded) and p_i is the predicted probability for class i.

Optimization is performed with the Adam optimizer, and class weights are used to address imbalance in gesture categories.

Figure 3 depicts the ViT-B/16-based gesture recognition architecture. ViT replaces convolutional layers with self-attention mechanisms, allowing the model to capture both local and global dependencies across the input image.

The 224×224×3 input image is split into non-overlapping 16×16 patches, flattened into embeddings with a prepended class token and positional encodings. These are processed through Transformer Encoder blocks containing MHSA and MLP layers with layer normalization and residual connections. The class token output is then passed through a Dense layer (256, ReLU), a Dropout (0.3), and a final softmax layer for classification.

These patch embeddings are combined with a class token and positional encodings to form the input sequence:

$$z_0 = [x_{class}; x_p^1 E; x_p^2 E; \dots; x_p^N E] + E_{pos} \quad (4)$$

where x_{class} is a class token, x_p^i represents the i^{th} image patch, E is the embedding matrix, and E_{pos} is the positional encoding.

The sequence is processed through multiple transformer encoder blocks, each comprising a multi-head self-attention (MHSA) module and a multilayer perceptron (MLP), with residual connections and layer normalization:

$$z_l' = MHSA\left(LN(z_{l-1})\right) + z_{l-1} \quad (5)$$

$$z_l = MLP\left(LN(z_l')\right) + z_l' \quad (6)$$

where $MHSA$ denotes multi head self attention, LN denotes layer normalization, and MLP is a two-layer feed forward network.

After the final encoder layer, the output corresponding to the class token is passed through a classification head consisting of a dense layer with

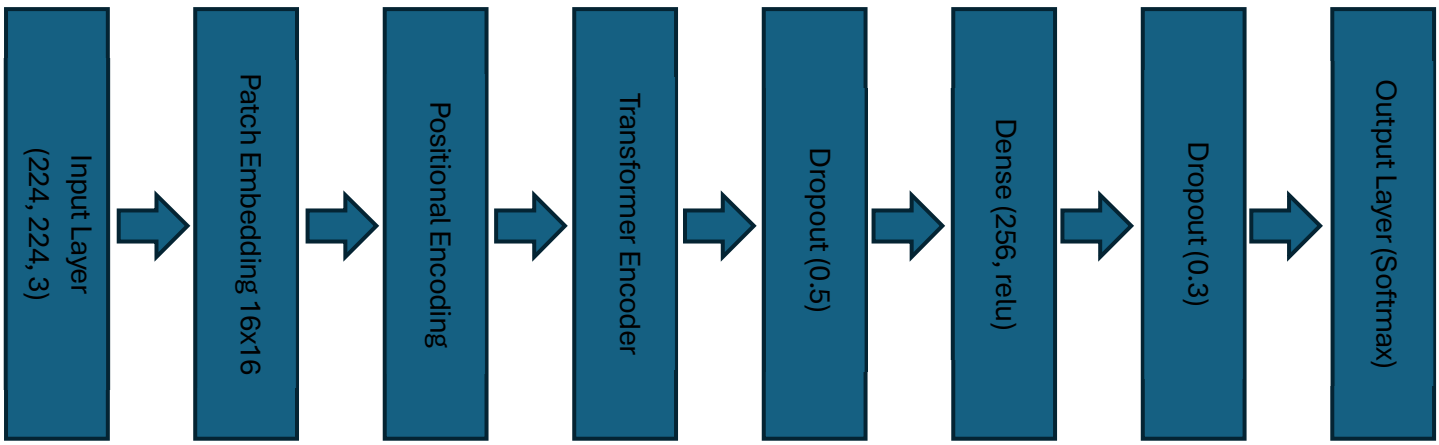

Figure 3. Vision Net Gesture Recognition Architecture

ReLU and dropout followed by a softmax layer to produce a probability distribution over the C gesture classes:

$$p_i = \frac{\exp\left(w_i^T z_L^{class} + b\right)}{\sum_{j=1}^{c} \exp\left(w_i^T z_L^{class} + b_i\right)} \quad (7)$$

where z_L^{class} is the class token output of the final transformer layer, and w_i, b_i are the weights and bias associated with class i.

The model is trained using the categorical cross-entropy loss,

$$L = -\sum_{i=1}^{c} y_i \log\left(p_i\right) \quad (8)$$

where y_i is the ground-truth label. Training is performed using the Adam optimizer with early stopping and learning rate scheduling. Identical data augmentation, preprocessing, and evaluation protocols are used for both the DenseNet-121 and ViT-B/16 models to ensure a fair and direct comparison.

The ViT-B/16 model leverages self-attention to capture long-range spatial dependencies, making it effective for distinguishing gestures with similar local features but different global structures. By comparing it with DenseNet-121's localized feature extraction, the study highlights the trade-offs and strengths of both architectures in static gesture recognition.

4. Training and Validation Process

To ensure a fair comparison, both DenseNet-121 and ViT-B/16 were trained under identical conditions on the HaGRID v2 512px dataset, split into 70% training and 30% testing, with 10% of the training data reserved for validation. Stratified sampling preserved class distributions. All images were resized to 224×224 and normalized to [0,1]. Identical augmentations horizontal flips, brightness adjustments, and zooming were applied to improve generalization.

Both models were trained using the Adam optimizer (learning rate 1×10 −4, batch size 32) with categorical cross-entropy loss, early stopping (patience=2), and learning rate reduction (factor = 0.5). Class imbalance was addressed through weighted losses. Experiments were conducted on an Intel Core i9 CPU with 64GB RAM and an NVIDIA RTX 4090 Graphics Card (24GB). DenseNet-121 trained from scratch for 5 epochs (≈5 hours), while ViT-B/16 was fine-tuned for 6 epochs (≈6.5 hours) from ImageNet-21k weights, with all transformer layers unfrozen to adapt fully to the gesture data. Both models employed dropout (0.5 and 0.3) to minimize overfitting. DenseNet-121 and ViT-B/16 have about 8 million and 86 million parameters, respectively, with complexities of roughly 2.8 and 17.5 GFLOPS for 224×224 inputs (Dosovitskiy et al., 2021) [12] [13].

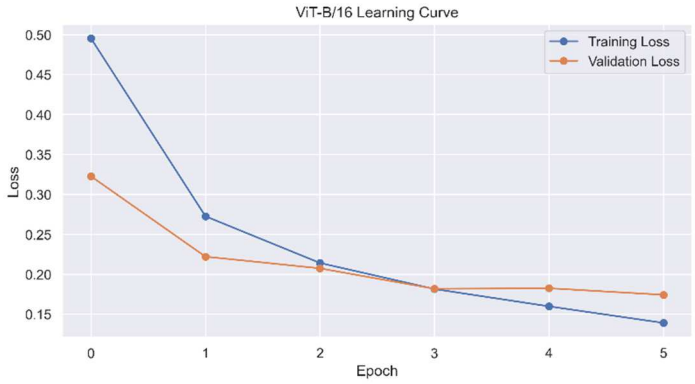

Figure 4. DenseNet-121 Training and Validation Loss over Epochs

Figures 4 and 5 show training and validation loss and accuracy curves.

DenseNet-121 converged rapidly, stabilizing by epoch three, while ViT-B/16 improved more gradually, reaching slightly better validation

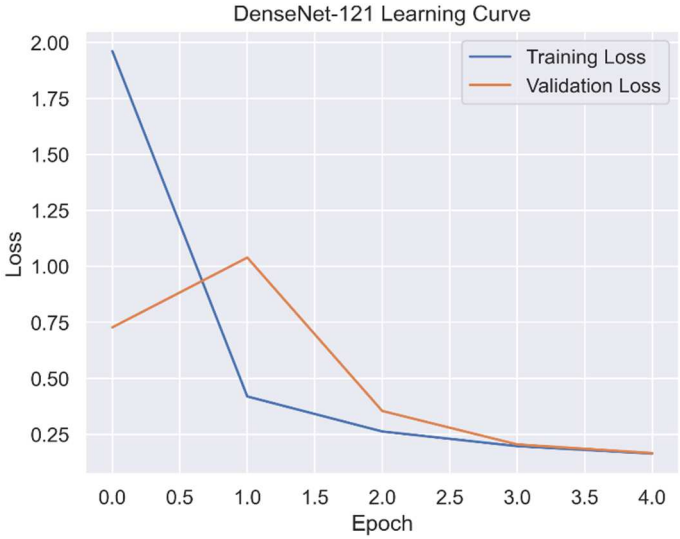

Figure 5. ViT-B/16 Training and Validation Loss over Epochs

performance with minimal overfitting. These trends confirm effective convergence and generalization under the shared setup.

5. Experiments and Results

The trained DenseNet-121 and Vision Transformer (ViT-B/16) models were evaluated on the test set from HaGRID v2 512px dataset under identical experimental conditions. Comprehensive performance metrics including classification accuracy, sensitivity, specificity, and

confusion patterns were analyzed to assess their comparative effectiveness for gesture recognition.

The Vision Transformer achieved a final test accuracy of 95.45% and validation accuracy of 94.71% after 6 epochs. In contrast, DenseNet-121 achieved a test accuracy of 95.32% and a higher validation accuracy of 94.98% after 5 epochs. While the overall performance of both models is comparable, ViT-B/16 exhibited slightly higher test accuracy, while DenseNet generalized slightly better on the validation set.

To better understand class-wise behavior, we computed sensitivity (recall) and specificity for each class. The Vision Transformer achieved a macro-average sensitivity of 0.945 and specificity of 0.998, whereas DenseNet-121 slightly outperformed with a macro-average sensitivity of 0.950 and the same specificity of 0.998. This suggests that while ViT-B/16 attained higher overall classification accuracy, DenseNet-121 was slightly more consistent in detecting true positives across all gesture classes.

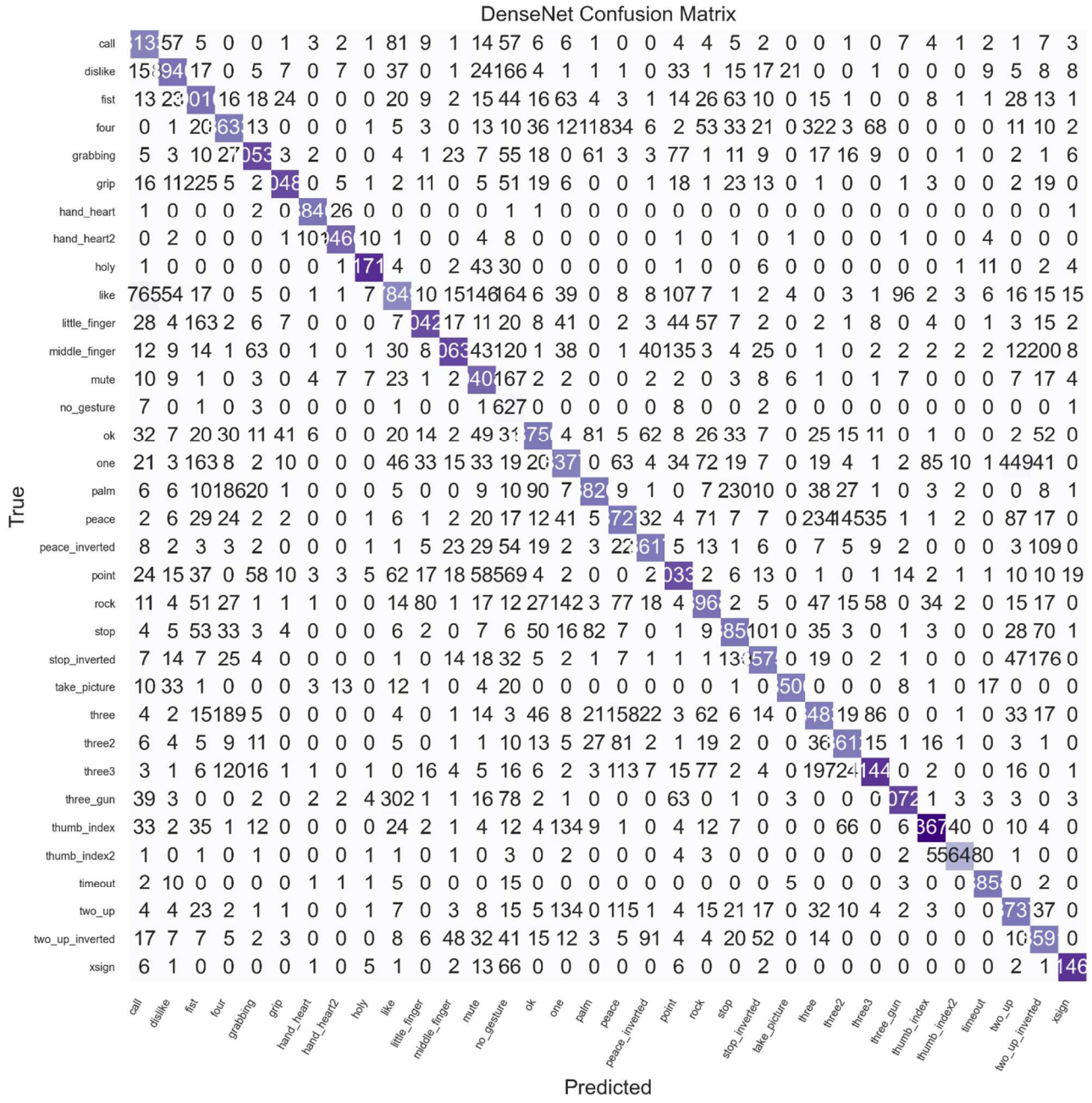

Figure 6. DenseNet-121 Confusion Matrix on the HaGRID

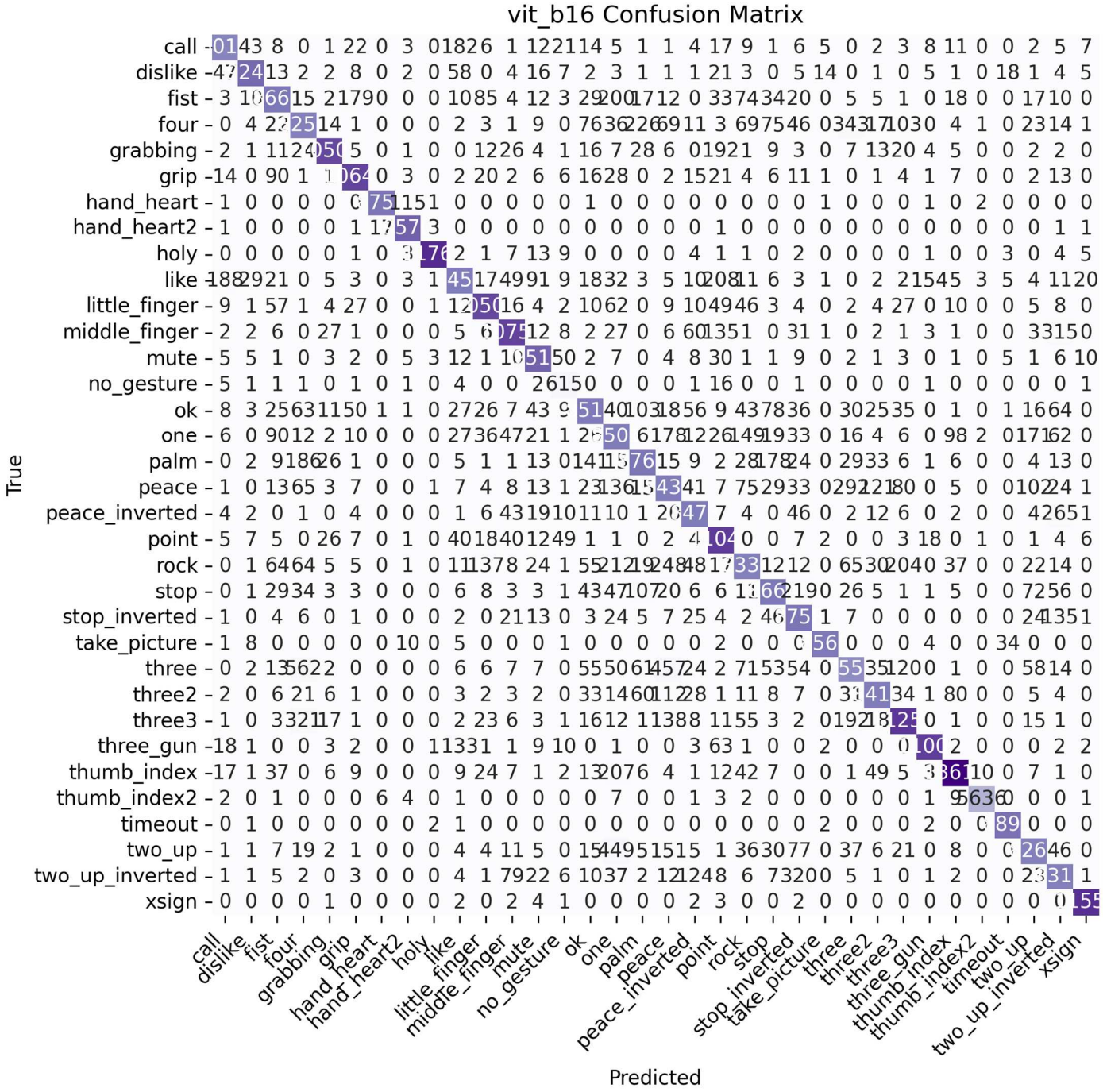

Figure 7. ViT-B/16 Confusion Matrix on the HaGRID

Figures 6 and 7 present the confusion matrices for DenseNet-121 and ViT-B/16. To improve interpretability, we highlight that most misclassifications occur between visually or semantically similar gestures. For instance, both models showed some confusion between peace and peace_inverted, as well as between three2 and three3, likely due to their similar hand configurations. However, the DenseNet model demonstrated slightly lower confusion between these gesture pairs compared to ViT-B/16, indicating better fine-grained discrimination despite ViT-B/16's marginally higher overall classification accuracy. Additionally, both models exhibit strong diagonal dominance across classes, signifying high per-class accuracy. This aligns with their reported high sensitivity (≥0.945) and specificity (~0.998), as true positives are concentrated along the diagonal and false positives are minimal across columns. The clear separation of class predictions in the confusion matrices supports the models' robustness in distinguishing gestures, with few ambiguous misclassifications overall.

6. Conclusion

In this study, we presented a comparative evaluation of two state-of-the-art deep learning architectures DenseNet-121 and Vision Transformer (ViT-B/16) for static hand gesture recognition using the large-scale HaGRID v2 dataset. Both models were trained under identical conditions, ensuring a fair assessment of their classification performance, generalization ability, and robustness across 34 gesture categories.

Our results demonstrated that the Vision Transformer achieved a slightly higher test accuracy of 95.45%, while DenseNet-121 maintained a better macro sensitivity of 0.950. Despite the performance similarity, the two models showed complementary strengths: ViT-B/16 exhibited superior capability in capturing global visual dependencies, while DenseNet-121 provided more balanced recognition across all gesture classes and required fewer computational resources.

The learning curves and confusion matrices validated the effectiveness of both models in minimizing overfitting and achieving stable convergence. Importantly, our evaluation highlights that high-performing hand gesture recognition can be achieved not only through advanced attention-based models but also through compact, efficiently structured CNNs like DenseNet when properly tuned and regularized.

In addition to reporting high accuracy, future work will incorporate statistical validation methods such as paired *t*-tests to evaluate whether performance differences between models are statistically significant. This would provide a more rigorous comparison and strengthen the credibility of the results beyond simple performance metrics.

This work contributes to the growing body of research at the intersection of gesture recognition and human-computer interaction, providing insights into model selection for real-time, vision-based interface systems. Future work may explore the integration of temporal modeling for dynamic gestures, multi-modal fusion with depth or skeletal data, and deployment optimization on mobile or embedded platforms.

Acknowledgements

This study was supported by research fund from Chosun University, 2025

References

[1] K. Sachdeva and R. Sachdeva, “A Novel Technique for Hand Gesture Recognition,” in *2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)*, IEEE, Nov. 2023, pp. 556–560. doi: 10.1109/ICAICCIT60255.2023.10466087.

[2] S. Padmakala, S. O. Husain, E. Poornima, P. Dutta, and M. Soni, “Hyperparameter Tuning of Deep Convolutional Neural Network for Hand Gesture Recognition,” in *2024 Second International Conference on Networks, Multimedia and Information Technology (NMITCON)*, IEEE, Aug. 2024, pp. 1–4. doi: 10.1109/NMITCON62075.2024.10698984.

[3] A. S. M. Miah, Md. A. M. Hasan, Y. Tomioka, and J. Shin, “Hand Gesture Recognition for Multi-Culture Sign Language Using Graph and General Deep Learning Network,” *IEEE Open Journal of the Computer Society*, vol. 5, pp. 144–155, 2024, doi: 10.1109/OJCS.2024.3370971.

[4] A. Kapitanov, K. Kvanchiani, A. Nagaev, R. Kraynov, and A. Makhliarchuk, “HaGRID - HAnd Gesture Recognition Image Dataset,” *Proceedings - 2024 IEEE Winter Conference on Applications of Computer Vision, WACV 2024*, pp. 4560–4569, Jan. 2024, doi: 10.1109/WACV57701.2024.00451.

[5] S. Sharma and S. Singh, “Vision-based hand gesture recognition using deep learning for the interpretation of sign language,” *Expert Syst Appl*, vol. 182, p. 115657, Nov. 2021, doi: 10.1016/J.ESWA.2021.115657.

[6] J. P. Sahoo, A. J. Prakash, P. Pławiak, and S. Samantray, “Real-Time Hand Gesture Recognition Using Fine-Tuned Convolutional Neural Network,” *Sensors 2022, Vol. 22, Page 706*, vol. 22, no. 3, p. 706, Jan. 2022, doi: 10.3390/S22030706.

[7] H. Hu, W. Zhao, W. Zhou, Y. Wang, and H. Li, “SignBERT: Pre-Training of Hand-Model-Aware Representation for Sign Language Recognition,” 2021.

[8] M. Montazerin, E. Rahimian, F. Naderkhani, S. F. Atashzar, S. Yanushkevich, and A. Mohammadi, “Transformer-based hand gesture recognition from instantaneous to fused neural decomposition of high-density EMG signals,” *Sci Rep*, vol. 13, no.

1, Dec. 2023, doi: 10.1038/S41598-023-36490-W,.
[9] J. W. Smith, S. Thiagarajan, R. Willis, Y. Makris, and M. Torlak, “Improved Static Hand Gesture Classification on Deep Convolutional Neural Networks Using Novel Sterile Training Technique,” *IEEE Access*, vol. 9, pp. 10893–10902, 2021, doi: 10.1109/ACCESS.2021.3051454.
[10] B. Chanda, “Improving Static Hand Gesture Recognition with Semantic Segmentation and Fine-Tuned Convolutional Neural Network,” pp. 1158–1163, Jun. 2025, doi: 10.1109/ICCIT64611.2024.11022040.
[11] G. Raju, K. Sushmitha, M. Priyanka, S. Leela, V. V. Chowdary, and M. Yashwantha, “Real Time Hand Gesture Recognition Using CNN,” *Global Journal of Engineering Innovations & Interdisciplinary Research GJEIIR*, vol. 5, no. 2, p. 44, 2025.
[12] A. Dosovitskiy *et al.*, “AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE,” 2021, Accessed: Jul. 16, 2025. [Online]. Available: https://github.com/
[13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 2261–2269, Aug. 2016, doi: 10.1109/CVPR.2017.243.

〈저자소개〉

Muhammad Abrar Hussain
• received the Bachelor’s degree in Computer Science from Mirpur University of Science and Technology (MUST), in December 2021. He is currently pursuing a Master’s degree in Computer Engineering at Chosun University, which he began in March 2025. His research interests include computer graphics and artificial intelligence.
• https://orcid.org/0009-0007-0863-3346

Seongki Kim
• received the Ph.D. degree in computer science and engineering from Seoul National University, in 2009. He is an Associate Professor with Chosun University. He researched and developed software for GPU, GPGPU, and dynamic voltage and frequency scaling (DVFS) at Samsung Electronics, from 2009 to 2014. He was also with Ewha Womans University, Keimyung University, Sangmyung University, and Chosun University, from 2014 to 2025. His current research interests include graphics/game algorithms, virtual/augmented reality, artificial intelligence, algorithm optimization through the GPU, and high-performance computing with CPU and GPU.
• https://orcid.org/0000-0002-2664-3632