

# Deep Learning Based Graph Convolutional Network Using Hand Skeletal Points For Vietnamese Sign Language Classification

Manh-Hung Ha<sup>\*(C,A)</sup>, Duc-Chinh Nguyen<sup>\*</sup>, Thai-Kim Dinh<sup>\*</sup>, Tran Tien Tam<sup>\*</sup>, Do Tien Thanh<sup>\*</sup>, and Oscar Tzyh-Chiang Chen<sup>\*,\*\*</sup>

**Abstract:** This paper develops a robust and efficient method for the classification of Vietnamese Sign Language gestures. The study focuses on leveraging deep learning techniques, specifically a Graph Convolutional Network (GCN), to analyze hand skeletal points for gesture recognition. The Vietnamese Sign Language custom dataset (ViSL) of 33 characters and numbers, conducting experiments to validate the model's performance, and comparing it with existing architectures. The proposed approach integrates multiple streams of GCN, based on the lightweight MobileNet architecture. The custom dataset is preprocessed to extract key skeletal points using Mediapipe, forming the input for the multiple GCN. Experiments were conducted to evaluate the proposed model's accuracy, comparing its performance with traditional architectures such as VGG and ViT. The experimental results highlight the proposed model superior performance, achieving an accuracy of 99.94% test on the custom ViSL dataset, reach accuracy of 0.993% and 0.994% on American Sign Language (ASL) and ASL MINST dataset, respectively. The multi-stream GCN approach significantly outperformed traditional architectures in terms of both accuracy and computational efficiency. This study demonstrates the effectiveness of using multi-stream GCNs based on MobileNet for ViSL recognition, showcasing their potential for real-world applications.

**Keywords:** Deep Learning, Graph Convolutional Network, Vietnamese Sign Language, Skeletal Join Points, Classification.

## 1 Introduction

SIGN language plays an essential role in both research and real-life applications, particularly for the hearing-impaired community. In practice, it serves as the primary communication tool that helps hearing-impaired individuals connect with society, break down language barriers, and promote inclusion in education, work, and

daily life. Beyond being a means of exchanging information, sign language also reflects the cultural identity and linguistic diversity of communities [1][2].

Sign language, with its rich diversity, presents a unique challenge in the field of automatic recognition due to the high demand for accuracy and contextual understanding [2]. This becomes even more crucial when considering Vietnamese Sign Language (VSL) [3][4], a complex system with unique characteristics that are not completely aligned with international sign languages like American Sign Language (ASL) [5][6]. Efforts to enhance the communication and interaction capabilities of the deaf community necessitate the development of a system capable of accurately recognizing VSL from visual data. We explore the synergy between Deep Learning based on Graph Convolutional Networks (GCN) to fully capture the complexities of sign language

Iranian Journal of Electrical & Electronic Engineering, 2026.

Paper first received 28 Jan 2025 and accepted 15 Sep 2025.

\* The authors are with the Faculty of Applied Sciences, International School, Vietnam National University, Hanoi 100000, Vietnam.

\*\* The author is with the Department of Electrical Engineering, National Chung Cheng University, Chiayi, 62102, Taiwan.

E-mail: [hunghm@vnu.edu.vn](mailto:hunghm@vnu.edu.vn)

Corresponding Author: Manh-Hung Ha

communication. Deep Learning, known for its ability to learn from large and complex data without explicit programming, has proven to be a powerful tool in solving this issue, particularly due to its proficiency in processing time series and spatial data [7]. This makes it an ideal choice for sign language recognition tasks that require accurate capture of continuous motion and positional changes. Meanwhile, GCN is chosen for its ability to handle graph-structured data, particularly useful when working with hand skeletal data represented as graphs. GCN understands and leverages spatial relationships between nodes (here, hand skeletal points), enabling it to capture hand structures and movements naturally and accurately. The paper [8] emphasizes the application of an enhanced Dynamic Graph Convolutional Neural Network (DGCNN) for grasp area detection on 3D objects. By leveraging GCN's strength in modeling spatial relationships, the method significantly improves feature extraction and accuracy in robotic grasping tasks, especially with complex object geometries. The proposed model in this study is based on their ability to process on their deep analysis of complex spatial relationships, making them particularly suited to the challenge of recognizing sign language from images. Through these models, we aim to achieve significant progress in accurately and automatically recognizing VSL, opening new avenues in sign language recognition technology research and development. Moreover, this study focuses on developing a new ViSL dataset comprising 33 letters and numbers, collected from 6 project participants. Each sign was recorded from various angles and under consistent lighting conditions to ensure diversity and minimize bias. The data collection process included cross-checking among members to maximize the accuracy of each sign.

The structure of this paper is organized as follows: Section II provides a comprehensive literature review, exploring prior methodologies for sign language recognition and the application of Deep Learning and GCN in processing skeletal data. Section III introduces our proposed model, detailing the dataset preparation, the deployment of various deep learning models for analysis, and the specific use of GCN for ViSL recognition. Experimental results and a thorough discussion on the findings, challenges encountered, and the evaluation of the GCN model's effectiveness are presented in Section IV. The development and integration of a web application to utilize the GCN model for sign language recognition services are also outlined. Section V concludes the paper with a summary of our contributions and insights for future research directions.

## 2 Literature Review

The chronicle of hand gesture recognition methodologies is a testament to the evolving interplay

between technological ingenuity and computational theory. Appearance-Based methods, with their utilization of classifiers and regressors to map image features onto hand poses, have historically been the bulwark of hand gesture language recognition. Such methodologies have played a pivotal role in the field's advancement, yet they are invariably tethered to voluminous datasets and exhibit inherent limitations in their capacity to generalize across diverse scenarios [9].

Distinguishing further between the sensor-based and vision-based approaches reveals a dichotomy in gesture recognition methodology. Sensor-based methods have the distinct advantage of precision, with instruments like recognition gloves providing exacting measurements of hand movements. Vision-based methods, on the other hand, present a more organic form of interaction by leveraging visual data, thus fostering a naturalistic and pliable user experience. This modality, while rich in potential, is particularly susceptible to the vicissitudes of environmental factors such as lighting and necessitates intricate computations to ensure reliable recognition [10].

The paper presents an activity recognition system using the MobileNet architecture, tailored for smartphones in home environments. MobileNet lightweight design, featuring inverted residuals and linear bottlenecks, ensures computational efficiency while maintaining high accuracy. The system achieves reliable recognition of daily activities with minimal energy consumption and low latency, making it ideal for mobile and real-time applications. This study highlights MobileNet potential for enabling AI-driven solutions in resource-constrained, real-world scenarios like smart homes [11].

In [12], discusses the VGGNet architecture, which uses deep convolutional layers with small 3x3 filters to improve image recognition accuracy. It emphasizes simplicity, replacing larger filters with deeper stacks of smaller ones. VGGNet achieved state-of-the-art performance on ImageNet, significantly influencing modern deep learning.

Vision Transformers (ViT) [13], which adapt transformer architectures from NLP for image recognition. By dividing images into fixed-size patches (16x16), ViT processes them as sequential inputs like words in text. The model achieves state-of-the-art results on image classification benchmarks, outperforming convolutional networks when trained on large-scale datasets, highlighting its potential in computer vision.

Within the deep learning milieu, Graph Convolutional Networks (GCNs) have emerged as a formidable architecture, distinct in their capacity to map the spatial relationships inherent in skeletal data. GCNs are adept at delineating the complex, hierarchical structures within

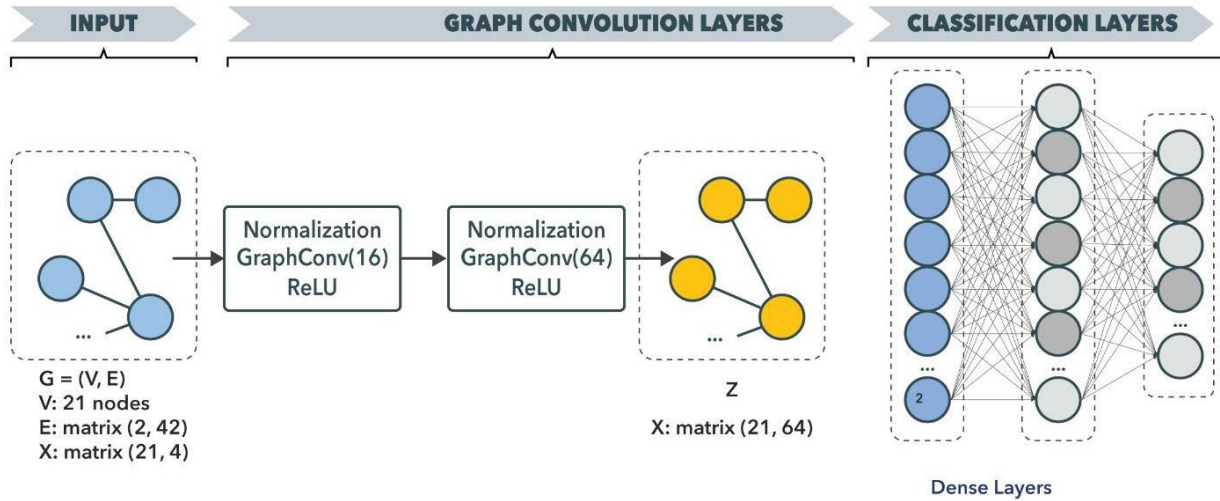


Fig 1. Architecture of a Graph Convolutional Network for Gesture Classification

and between frames of skeletal data, granting them a pronounced edge in the domain of dynamic sign language movement recognition. The theoretical underpinnings of GCNs are grounded in their ability to operate over graph structured data, harnessing the power of topological data analysis to perceive and process the intricate patterns of connectivity that define human gestures. By integrating the principles of graph theory and convolutional methodologies, GCNs can identify and extrapolate the relational features intrinsic to sign language, enabling a profound comprehension of its spatial-temporal dynamics [14].

The ramifications of GCNs' advancements in sign language recognition are multifold. Not only do they hold transformative potential for enhancing accessibility and facilitating real-time communication for the deaf and hard-of-hearing communities, but they also portend a breadth of applications across diverse sectors. These sectors range from augmented reality and human-computer interaction to assistive robotics and beyond, each standing to gain from the intricate gesture recognition capabilities afforded by GCNs [15].

The evolution of GCNs epitomizes the synergy between deep learning innovation and the quest for universal communication. As the nexus between computational architecture and linguistic expression, GCNs not only pave the way for a new era of inclusivity but also signify a watershed moment in the confluence of artificial intelligence and human language [16].

In this study, MediaPipe Hand developed by Google is to generate the hand key points, which ensures sufficient reliability and seems similar to the two methods mentioned above for estimating hand key points. At the core of MediaPipe Hand's functionality is a sophisticated machine learning model that can accurately interpret hand movements by identifying 21 distinct landmarks on

a single hand in real- time. This technology is highly advanced, capable of detecting not just the presence of a hand in the video frame, but the position and movement of individual fingers and the wrist [17][18].

### 3 Proposed Graph Convolutional Neural network

To deploy an efficient system capable of recognizing hand gestures, allowing flexible and accessible interaction for a broad user base in everyday life, we have chosen an image-based approach. Our initial idea was to develop a classification task based on graph convolutions, employing techniques for feature extraction, processing, and classification based on graphs. Broadly, our model consists of three main parts (as shown Figure 1):

- Feature extraction of the hand's skeletal framework
- Preprocessing, feature augmentation, and graph structure construction
- Classification model: implementation of models for graph classification tasks

#### 3.1 Proposed Graph Convolutional layers-Type1

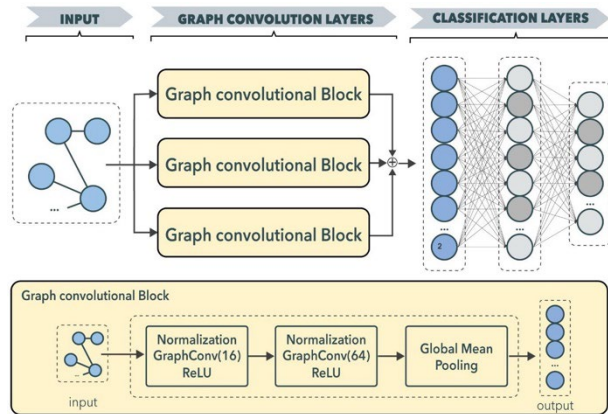
With the initial idea to deploy a classification task based on graph convolution, techniques for extraction, processing, and classification were used. In this study, Figure 1 shown an orientation to build a system based on graph classification methods, we start with a simple classification module that includes 2 hidden blocks of the graph convolutional network. Each Graph Convolutional Block (GCB) in the architecture consists of a normalization layer, a graph preconvolutional layer, and an activation function. As preprocessed in the feature extraction and preprocessing stages, the input is an undirected graph with 21 nodes interconnected through an adjacency matrix E (2, 42), with each node having 4 features. Through the use of graph convolution,

the number of features at each layer is incrementally expanded to 16 features at the first hidden block and 64 features at the second hidden block. After the feature aggregation and extraction process, the graph is synthesized through a global mean pooling layer before passing through two dense layers for classification.

Recognizing the exemplary performance of the model achieved with minimal computational resources, particularly its efficiency in parameter usage at just 2,500 parameters, our enthusiasm is buoyed towards further enhancements. We aim to bolster the model's performance while preserving its inherent compactness, an attribute critical for deployment in resource-constrained environments.

Embracing a multi-stream paradigm, the refined model we propose delineates input data into three distinct streams. This triadic configuration is engineered to scrutinize various dimensions of the graph data in parallel, akin to an ensemble of specialized analytical processes concurrently dissecting a complex dataset.

Each stream inherits the proven two-layer graph convolutional topology, ensuring that the robust feature extraction methodologies established by preceding models are perpetuated. This structural recursion is strategic, facilitating the isolation of salient features within the high-dimensional space.



**Fig 2.** Multi-Stream Graph Convolutional Network for Gesture Classification

Post-extraction, an integration of these diversified feature matrices ensues, synthesizing disparate data interpretations into a comprehensive feature representation. The ensuing matrix exhibits a rich tapestry of features, subsequently presented to the classification layers.

### 3.2 Proposed Multiple Graph Convolutional layers-Type2

In the proposed model, in the Graph Convolutional Layer phase, we use 3 parallel GCB blocks. Each GCB block contains a graph convolutional 16 and a graph

convolutional 64 connected to each other as shown in Figure 2. The architecture's classification mechanism is a composite of multilayer perceptrons, each neuron serving as an integrator of the extracted features. This multi-layered approach is not merely a nod to traditional neural networks but an articulate choice, enhancing the translational capacity of complex features into precise predictive outputs. Moreover, the integration of global mean pooling within each convolutional stream underscores the model's commitment to dimensional reduction. This pooling modality distills voluminous feature sets into their most expressive constituents, thereby elevating the most significant features for subsequent classification processes.

An emphasis on input normalization prior to convolutional operations ensures a homogenized scale for input data. Such standardization is imperative for gradient stability across training epochs, especially when engaging with heterogeneous datasets that may otherwise introduce scale variance and compromise learning efficacy.

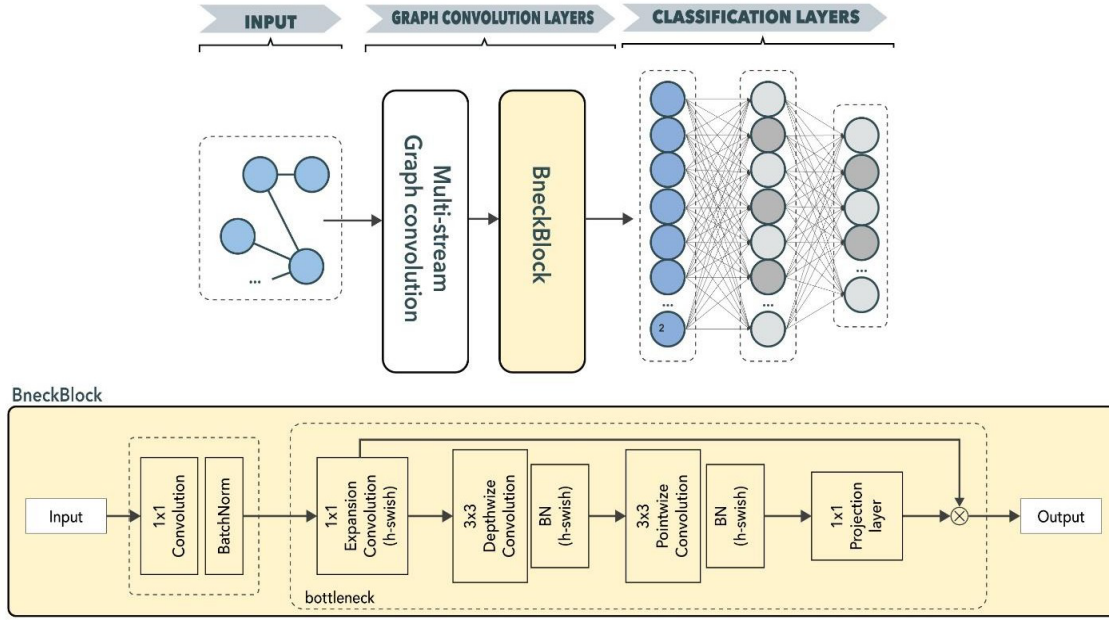
Building upon the foundational design, the proposed GCN architecture incorporates a novel module aiming to enhance the extraction of higher-level features from input data through advanced graph convolution processes. The strategic augmentation is premised on the model's imperative to remain dimensionally economical while striving for minimal impact on performance efficacy.

### 3.3 Proposed Multiple Stream Graph Convolutional layers based Bottleneck (BneckBlock)-Type3

The module draws inspiration from the architectural ingenuity of MobileNet bottleneck design, a paradigm of classification models recognized for their potent performance and judicious use of computational resources. The crux of MobileNet efficiency lies within its bottleneck structure, an elegantly orchestrated constriction of data channels that simultaneously engages with lightweight depthwise separable convolutions. This mechanism offers a dual benefit: it curtails the computational overhead while ensuring that essential information is retained and amplified for subsequent processing layers.

Adopting this principle, Figure 3 show the proposed model call Type 3 introduces a bottleneck block in the aftermath of the feature combination layer—a pivotal juncture where the feature maps from multiple graph convolution streams coalesce. This integration embodies a harmonization of distinct feature perspectives, each stream having parsed the graph data through its specialized convolutional lens. The bottleneck block's role is thus to distill this confluence of features, pruning the informational expanse into a dense, potent representation.





**Fig 3.** Proposed MobileNet-Inspired Bottleneck Structure in a Graph Convolutional Network for Gesture Classification

In operational terms, the bottleneck block applies a series of convolutions that initially compress the feature channels, deliberately reducing dimensionality. This compression is immediately succeeded by an expansion convolution, a maneuver that broadens the information channels once more, albeit with an acute focus on the most salient features as dictated by the training regimen. The expansion is tactical, reinforcing the model's capability to accentuate relevant patterns while precluding extraneous data.

The bottleneck's depthwise convolutions impart another layer of refinement, deploying a fine-grained filter across each channel, reinforcing the model's interpretative precision. A subsequent global pooling layer stands sentinel at the terminus of this block, its charge to aggregate the spatial information into a singular vector that encapsulates the essence of the input's topological characteristics.

Such a vector then traverses to a projection layer, whose purpose is to project the condensed feature array onto the output space. Here, the classification process culminates, with each feature vector undergoing a transformation into a probabilistic interpretation of class membership. The application of the softmax function at this final layer ensures a probabilistic distribution over potential classes, crystallizing the

network's deductions into definitive predictions.

In synthesis, the described enhancements to the GCN architecture encapsulate a meticulous balance between feature richness and computational frugality. Through the judicious application of a bottleneck module, the

architecture achieves a feat of maintaining a compact footprint while exhibiting formidable analytical prowess—a quintessential exemplification of efficiency in deep learning architectures.

To highlight the distinctions among the proposed models, we provide a clear comparison (Table 1). Type1 serves as a baseline, consisting of a simple two-layer GCN architecture. Type2 enhances feature extraction by introducing three parallel GCN branches, allowing the model to learn from multiple perspectives simultaneously. Type3 builds upon Type2 by integrating a bottleneck module inspired by MobileNet, aiming to reduce the number of parameters while preserving performance. The following table summarizes the key differences



**Fig. 4** Data collection example

## 4 Illustrations

### 4.1 Dataset

Vietnamese sign language with standard Vietnamese signs for the aim of being used by Vietnamese people. Unlike the English alphabet with 26 letters and 10 numerals, the Vietnamese character set comprises 33 characters, including 23 letters and 10 numerals. The dataset process is operated by recording video and extracting it into images to ensure standards of dataset characteristics and accurate detailed data collection. In order to get a sufficient label size and balanced data distribution for optimal machine learning performance, this dataset encompasses a total of 118,800 images, including 33 distinct labels: 23 static letters (from A to Y) and 10 numerals (from 0 to 9). Each label containing 3600 frames. Some examples shown in Figure 4.

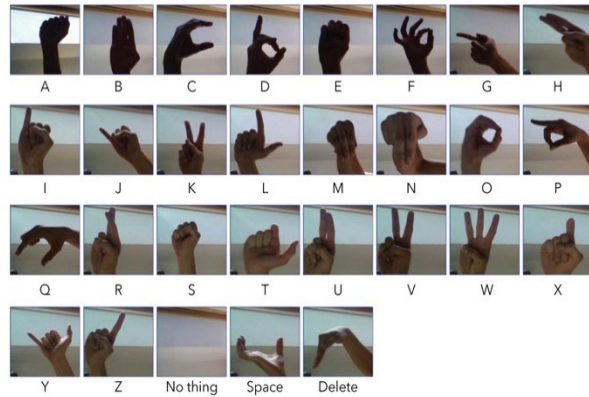


Fig 5. ASL dataset [19].

The American Sign Language (ASL) dataset used in this study [19] consisted of approximately 87,000 images, organized into 29 categories representing the letters A to Z in ASL, as shown in Figure 5. Each image was standardized to a resolution of 200x200 pixels, providing a large and varied sample for evaluating model accuracy and generalization.

The Sign Language MNIST (Figure 6) dataset employed in this study was the sign [20]. This dataset offers a unique resource for analyzing American Sign Language, particularly through its simplified format. It includes 24 static alphabet classes, omitting the letters J and Z due to their reliance on motion, and contains a total of 27,455 grayscale images with a resolution of 28x28 pixels. Modeled after the original MNIST digit dataset, this version substitutes handwritten numerals with images of hand gestures representing ASL letters. Utilizing this dataset enables comparative analysis of model performance on classification tasks involving visual hand sign inputs, thereby contributing to a deeper comprehension of ASL's visual complexity and variability.



Fig 6. ASL dataset [20]

### 4.2 Experimental environment

All experiments in this study were conducted on Google Colab, which provided sufficient computational resources for both training and evaluation. The environment was equipped with an Intel(R) Xeon(R) CPU @ 2.20GHz, featuring 1 core and 2 threads, with a cache size of 56320 KB. The proposed model, based on a Graph Convolutional Network, was implemented in PyTorch. For optimization, we used the Adam optimizer with a learning rate of 0.01, along with an exponential learning rate scheduler (gamma = 0.95) to gradually reduce the learning rate over time. The loss function was Binary Cross-Entropy with Logits. Training was carried out for 100 epochs with a batch size of 32.

### 4.3 Empirical results to the ViSL dataset

In our experimental section, we present a detailed comparative analysis that evaluates the performance of three distinct pre-trained neural network models alongside three variations of our primary Graph Convolutional Network (GCN) models. The pre-trained models under scrutiny are VGG16 [12] and Vision Transformer (ViT) [13], each fine-tuned on our Vietnamese Sign Language (VSL) dataset to ensure relevance and applicability to the task at hand.

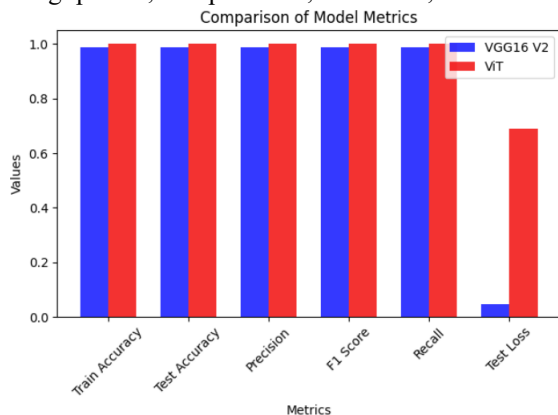
The GCN models, denoted as Type1, Type2, and Type3, have been meticulously designed and iterated upon to optimize sign language classification. Type1 serves as our baseline GCN, providing a foundation for comparison. Type2 incorporates advanced strategies such as multi-stream graph convolution blocks and a MobileNet-inspired bottleneck feature to enhance feature extraction while maintaining model efficiency. Type3 (Proposed) represents our most sophisticated architecture, potentially featuring further layers or complex configurations, as evidenced by its increased parameter count and lowest test loss as shown in Table 1

**Table 1** Comparison of the proposed GCN-based models (Type1, Type2, Type3)

Model	Structure	Branches	Fusion Method	Special Feature
Type1	2-layer GCN	1	Global mean pooling	Simple baseline
Type2	2-layer GCN $\times$ 3	3	Feature concatenation	Multi-stream architecture
Type3	2-layer GCN $\times$ 3 + Bottleneck	3	Bottleneck + pooling	Efficient fusion (MobileNet-inspired)

Through rigorous training and testing protocols, each model has been evaluated on metrics that include training accuracy, test accuracy, precision, F1 score, and test loss. The parameters count is also recorded to weigh the models' computational demands against their performance. This comparison not only demonstrates the potential of GCN models in sign language recognition tasks but also provides valuable insights into the benefits and trade-offs of employing complex pre-trained networks versus specialized graph-based architectures. The results of this comparative study are crucial in guiding future research directions and potential real-world implementations of sign language recognition systems.

The Figure 7 presents the performance metrics for two given model, VGG16 and ViT, which have been trained and tested on a dataset, likely related to image recognition tasks such as sign language detection. By proposed three distinct Graph Convolutional Network (GCN) models, which have been fine-tuned for the task of Vietnamese Sign Language recognition. Each model is evaluated based on its accuracy in both training and testing phases, its precision, F1 score, and test loss.



**Fig 7.** Performance Metrics Comparison between VGG16 V2 and ViT Models

Additionally, the table includes the number of parameters for each model, providing insight into the model's complexity.

**Table 2** Comparative performance metrics of three GCN Models

Models	Train Acc	Test Acc	Precision	F1 score	Test Loss	Params
VGG16	98.75	98.75	98.77	98.75	0.0477	-
ViT	99.94	99.92	99.94	99.94	0.0069	-
MobileNet	99.37	99.01	99.11	99.31	0.0016	-
Type1	98.14	99.15	99.15	99.15	0.0025	4441
Type2	99.40	99.60	99.60	99.60	0.0009	8665
Type3 (Proposed)	99.98	99.94	99.92	99.94	0.0008	7673

The Table 2, the ViT model, on the other hand, showcases extremely high testing accuracies with rounding up to 99.92%. This remarkable consistency demonstrates the model's ability to learn and generalize from the data exceptionally well. The precision, F1 score, and recall metrics all match the accuracy, indicating an almost perfect classification system with a balance between the sensitivity and precision of the model. Yet, the test loss for the ViT model is significantly higher than that of VGG16 at 0.0069, which is counterintuitive given the other metrics.

The discrepancy between the ViT model's test loss and its other performance metrics could indicate an issue with how the loss was calculated, reported, or perhaps a sign of an anomaly in the test dataset that did not impact accuracy. Typically, a high-test loss would correspond to lower accuracy metrics, so further investigation is needed to understand the reason behind this inconsistency.

In comparison, while the ViT shows superior accuracy, the VGG16 model may be more reliable in terms of loss metric consistency. This could suggest that in scenarios where interpretability of loss is crucial, VGG16 might offer a more comprehensible performance profile.

Starting with Type1, the transition from training to testing accuracy indicates a model that generalizes well, reflected in its significant precision and F1 score of 99.15%. The low-test loss further substantiates the model's ability to maintain its performance on unseen data, an essential quality for practical applications where predictability is crucial.

Type2 elevates these metrics, exhibiting a remarkable synergy between training and test accuracies, testing exceeding 99.4%. The precision and F1 score alignment suggest a high true positive rate and a balanced sensitivity and specificity—attributes that are critical in systems where misclassification could significantly impact the user experience. The reduced test loss implies

an enhancement in the model's ability to not just fit but truly understand the nuances of the ViSL data.

Type3 presents a conundrum: while it shows the highest training accuracy and an impressive test accuracy that nearly matches. This incongruity hints at a potential overfitting or an error in data reporting or processing that requires attention, as such a loss value contradicts the other performance indicators.

When contextualizing these models against the backdrop of pre-trained models such as VGG16 and ViT, an intriguing narrative unfolds. The pre-trained models, drawing on extensive and diverse visual datasets, demonstrate exceptional test accuracies with ViT standing out with almost perfect metrics across the board. These models, particularly ViT, have leveraged their vast exposure to various visual patterns to deliver exemplary performance on ViSL recognition—a task that requires discerning subtle differences in hand gestures.

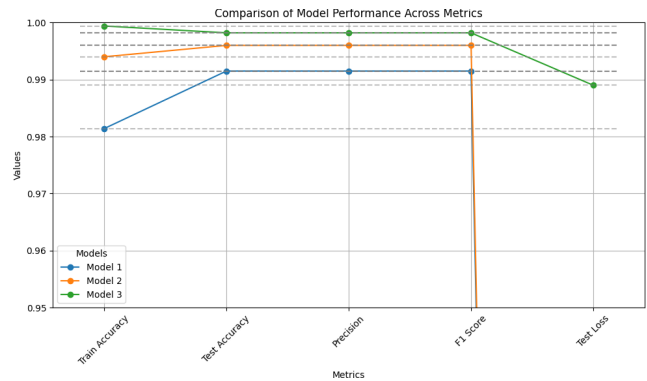
The performance metrics of the three GCN models tailored for Vietnamese Sign Language recognition highlight their strengths and potential areas for improvement. While MobileNet is a lightweight CNN optimized for general image classification on flat grids, it struggles to capture the crucial spatial dependencies of skeleton-based data. In contrast, the ViT model excels at modeling global contextual relationships through self-attention mechanisms, resulting in superior generalization compared to MobileNet.

The proposed GCN models are specifically designed to process graph-structured data, naturally capturing the relational and temporal dynamics of joint movements in Vietnamese Sign Language. This architectural alignment enables the GCNs to outperform fine-tuned pre-trained models such as VGG16 and ViT, which, although powerful, are not inherently tailored for skeleton data.

The conclusion drawn from juxtaposing the GCN models against the pre-trained architectures is multidimensional. Firstly, the performance of the pre-trained models affirms the power of transfer learning, especially in domains rich with visual nuances such as

sign language. Among these, ViT's performance is notably stellar, likely due to its ability to capture the sequential and relational aspects inherent in sign language through its self-attention mechanisms.

Secondly, the GCN models, with their specialized architecture designed to capture relational data, illustrate the potential for bespoke solutions in sign language recognition. Particularly, Type 2 model performance suggests that there is a sweet spot where a model can be both efficient and highly accurate without an overbearing parameter count.



**Fig 8.** Performance Metrics of GCN Models Across Key Evaluation Criteria

This exploration reveals a compelling argument for a hybrid approach: combining the broad learning capabilities of pre-trained models with the specialized, graph-based nuance understanding of GCNs. Such a hybrid could potentially capture the best of both worlds, achieving high performance with computational efficiency as shown in Figure 8.

Further investigation and experimentation could focus on reconciling the disparities in Type 3-test loss, integrating lessons learned from both the pre-trained and GCN models, and potentially exploring ensemble methods or novel architectures that marry the conceptual strengths of both approaches. The ultimate goal remains to develop an interpretable, reliable, and efficient model that can democratize communication for the ViSL community in diverse, real-world environments.

**Table 3** Performance comparison of three proposed models (Type 1, Type 2, and Proposed Type 3) on two public dataset

Model	ASL						ASL MNIST					
	Train Acc	Test Acc	Precision	F1 score	Test Loss	Para	Train Acc	Test Acc	Precision	F1 score	Test Loss	Para
Type 1	0.9873	0.9902	0.9881	0.9881	0.0046	3856	0.8578	0.9194	0.9254	0.919	0.0329	3986
Type 1	0.9945	0.9911	0.9892	0.9893	0.0046	6928	0.932	0.9528	0.9556	0.953	0.0209	7314
Type 3-Proposed	0.9975	0.9929	0.9908	0.9911	0.0025	7520	0.9791	0.9944	0.9948	0.9944	0.0138	7554



#### 4.4 The impact of the three proposed models on two types of public datasets (ASL and ASL MNIST)

The table 3 above shows the performance comparison of three proposed models (Type 1, Type 2, and Type 3) when trained and tested on two public datasets: ASL and ASL MNIST. Overall, Type 3 gives the best results on both datasets, with higher accuracy, recall, and F1 score than the other two models. For example, on the ASL dataset, Type 3 achieved F1 score of 0.991, with a high accuracy of 0.993. This suggests that Type 3 is better at learning the features of sign language images.

For the ASL MNIST dataset, all models performed well. Type 3 stood out with accuracy of 0.994 and an F1-score of 0.994, showing its strong accuracy in recognizing static hand signs. However, Type 3 also has more parameters than the other models, meaning it may require more computing resources. In summary, Type 3 is the best choice if the main goal is high accuracy in sign language classification.

#### 4.5 Visualization Performance Insights

In the accuracy chart, we see that the validation accuracy starts high and maintains a relatively stable trajectory, hovering around the 97-98% range. This demonstrates a good level of generalization from the outset. However, the training accuracy begins at a lower point and progressively climbs, suggesting that the model is learning and improving its performance on the training data over time. The gap between training and validation accuracy indicates that there might be some overfitting, as the model performs better on the training data than on the validation set by the end of the observed epochs.

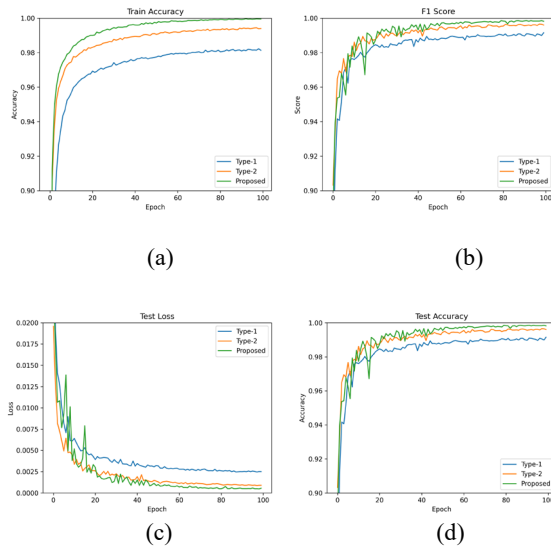


Fig 9. Illustration of three distinct GCN variants for ViSL Dataset

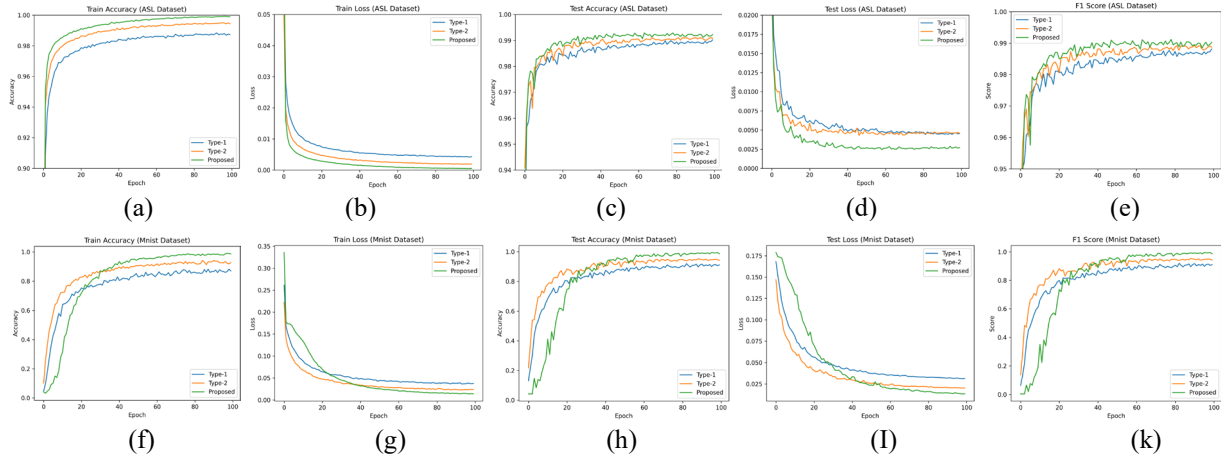
The loss chart complements this view. The training loss decreases sharply and then plateaus, which is

expected behavior as the model begins to fit the training data more closely. Conversely, the validation loss is low from the start and remains flat throughout the epochs. This is an encouraging sign, implying that the model is not just memorizing the training data but is also effectively generalizing to the validation data. However, one might expect the validation loss to decrease alongside the training loss if the model were learning effectively. The stability of the validation loss at a low value is positive, but the ideal scenario would be a convergence of training and validation loss, which is not observed here.

Building upon the foundation laid by the pre-trained models, we pivot to a focused analysis of the learning evolution within our tailored Graph Convolutional Network models. This visual assessment spans a comprehensive 100 epochs, capturing the nuanced progression of three distinct GCN variants, each meticulously calibrated for the Vietnamese Sign Language recognition task. We observe the following:

Figure 9(a), train Accuracy: While all models improve over time, the 'Proposed' model shows the highest train accuracy, which, coupled with its test performance, indicates an effective learning process with minimal overfitting. Figure 9(b), F1 Score: All models display an upward trend in F1 score, with the 'Proposed' model achieving and maintaining the highest score earlier in the training process. This reflects its superior balance between precision and recall, signifying a robust classification system capable of maintaining high accuracy consistently for both positive and negative classes. Figure 9(c), test Loss: The proposed Type 3 model shows a rapid decline in test loss, stabilizing at the lowest point compared to the 'Type-1' and 'Type-2' models. This suggests that the proposed model has the best generalization capability, indicating a precise understanding of the underlying patterns in the ViSL dataset. Figure 9(d). Test Accuracy: The 'Proposed' model similarly outperforms in terms of test accuracy, reaching peak performance rapidly and maintaining it throughout the remainder of the epochs. High test accuracy is crucial for the practical application of sign language recognition, as it translates to reliable communication in real-world scenarios.

Figure 10 shows the training and testing, and F1 score performance of three models (Type-1, Type-2, and Type 3-Proposed) on two datasets: ASL and ASL MNIST. Overall, the Proposed model (Type-3) performs the best in most curves. On the ASL dataset, it reaches the highest accuracy and F1 score, and has the lowest training and testing loss. This means the Proposed model can learn the features of sign language images more effectively and with more stability than the other two models.



**Fig 10.** Illustration of three distinct GCN variants for ASL and ASL MNIST Dataset. Fig 10 (a) and Fig 10 (f) the curve for training accuracy, Fig 10 (b) and Fig 10 (g) the curve for training loss, Fig 10 (c) and Fig 10 (h) the curve for testing accuracy, Fig 10 (d) and Fig 10 (i) the curve for testing loss, Fig 10 (e) and Fig 10 (k) the curve for F1 score.

For the **ASL MNIST dataset**, all models learn quickly, but the Proposed model starts slower in the early training steps. However, after about 40 epochs, it catches up and gives better accuracy and F1 scores than Type-1 and Type-2. Type 3 also keeps the lowest loss during testing. This shows that the Proposed model needs more time to learn in the beginning, but once trained, it becomes more accurate and consistent.

## 5 Conclusion

This study has successfully demonstrated the application of a Deep Learning-based Graph Convolutional Network utilizing hand skeletal points for the classification of Vietnamese Sign Language. By enhancing the GCN model with specific feature extractors, notably VGG16, we achieved a remarkable accuracy rate of 99,6% on the ViSL dataset. This represents a significant improvement over traditional CNN models, highlighting the effectiveness of integrating advanced deep learning architectures with graph-based analysis in recognizing complex gestures. The integration of deep learning techniques with skeletal data has provided a nuanced understanding of the spatial dynamics and intricate patterns inherent in sign language gestures. This approach not only enhances the accuracy of gesture recognition but also enriches our understanding of the subtleties involved in sign language communication. The ability of the GCN model to capture and analyze the structured data of hand movements effectively demonstrates its potential for advancing sign language recognition technologies. Furthermore, the results of this research showcase the promising capabilities of GCNs in the domain of gesture recognition. The adaptability and precision of the GCN model, supported by the robust feature extraction capabilities of VGG16, pave the way for future developments in assistive communication technologies.

This could significantly impact the way communication aids are developed for the hearing impaired, offering them more nuanced and accessible tools for interaction. As we continue to refine and improve upon these models, the potential applications of such technologies extend beyond the realm of sign language recognition, suggesting broader implications for the fields of human-computer interaction and automated behavioral analysis. Future research will focus on further enhancing the model's accuracy, reducing computational costs, and expanding the dataset to include more diverse representations of ViSL gestures, which will undoubtedly open new avenues for innovation in this vital field.

## Conflict of Interest

All financial, commercial or other relationships that might be perceived by the academic community as representing a potential conflict of interest must be disclosed. If no such relationship exists, authors will be asked to confirm the following statement:

*The authors declare no conflict of interest.*

## Author Contributions

The Author Contributions section is mandatory for all articles, including articles by sole authors. The Author Contributions statement must describe the contributions of individual authors referred to by their initials and, in doing so, all authors agree to be accountable for the content of the work.

## Funding

This research is funded by International School, Vietnam National University, Hanoi (VNU-IS) under code CS.2025-01.

## Acknowledgment




This research is funded by International School, Vietnam National University, Hanoi (VNU-IS).

## References

- [1] Pu, Junfu, Wengang Zhou, and Houqiang Li. "Iterative alignment network for continuous sign language recognition." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019.
- [2] Kheddar, Hamza, Mustapha Hemis, and Yassine Himeur. "Automatic speech recognition using advanced deep learning approaches: A survey," *Information Fusion*, 102422, 2024.
- [3] Vo, Anh H., Van-Huy Pham, and Bao T. Nguyen. "Deep learning for Vietnamese Sign Language recognition in video sequence." *International Journal of Machine Learning and Computing* 9.4: 440-445, 2019.
- [4] Vo, Duc-Hoang, et al. "Dynamic gesture classification for Vietnamese sign language recognition." *International Journal of Advanced Computer Science and Applications*, 8.3, 2017. doi:<https://doi.org/10.14569/ijacsa.2017.080357>.
- [5] Uthus, Dave, Garrett Tanzer, and Manfred Georg. "Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus." *Advances in Neural Information Processing Systems* 36, 2024.
- [6] Parikh, Harsh, et al. "American sign language classification using deep learning." *International Journal of Biometrics*, 16.6: 640-659, 2024.
- [7] Manh-Hung Ha and Osacl T C Chen, "Deep neural networks using capsule networks and skeleton-based attentions for action recognition," *IEEE Access*, vol. 9, pp. 6164–6178, January 2021
- [8] Merrikhi H, Ebrahimnezhad H. "Grasp Area Detection for 3D Object using Enhanced Dynamic Graph Convolutional Neural Network," *Iranian Journal of Electrical and Electronic Engineering*; 20 (4) :134-146, 2024
- [9] Hashi, Abdirahman Osman, Siti Zaiton Mohd Hashim, and Azurah Bte Asamah. "A Systematic Review of Hand Gesture Recognition: An Update From 2018 to 2024," *IEEE Access*, 2024.
- [10] Mohamed, Noraini, Mumtaz Begum Mustafa, and Nazean Jomhari. "A review of the hand gesture recognition system: Current progress and future directions," *IEEE access* 9: 157422-157436, 2021.
- [11] Chen, Oscar Tzyh-Chiang, Manh-Hung Ha, and Yi Lun Lee. "Computation-affordable recognition system for activity identification using a smart phone at home," *IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2020.
- [12] Boesch, Gaudenz. "VGG Very Deep Convolutional Networks (VGGNet) What you need to know." Read more at: <https://viso.ai/deep-learning/vgg-very-deep-convolutional-networks>, 2022.
- [13] Dosovitskiy, Alexey. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).
- [14] Nguyen, Duc-Chinh, et al. "RHM: Novel Graph Convolution Based on Non-Local Network for SQL Injection Identification." 2024 *IEEE Symposium on Industrial Electronics & Applications (ISIEA)*, 2024.
- [15] Y. Wang, L. Chen, J. Li and X. Zhang, "HandGCNFormer: A Novel Topology-Aware Transformer Network for 3D Hand Pose Estimation," 2023 *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, 2023, pp. 5664-5673, doi: 10.1109/WACV56688.2023.00563.
- [16] Ikram Kourbane, Yakup Genc, A hybrid classification-regression approach for 3D hand pose estimation using graph convolutional networks, *Signal Processing: Image Communication Volume* 101, February 2022, 116564
- [17] Google Developers. (n.d.). MediaPipe. [online] Available at: <https://developers.google.com/mediapipe>.
- [18] Bazarevsky, Valentin, and Fan Zhang. "On-device, real-time hand tracking with mediapipe." *Google AI Blog*, 2019.
- [19] Fostiropoulos, Iordanis, Jiaye Zhu, and Laurent Itti. "Batch model consolidation: A multi-task model consolidation framework." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.
- [20] Bayrak, S., Nabyev, V., & Atalar, C. American Sign Language Recognition Model Using Complex Zernike Moments and Complex-Valued Deep Neural Networks. *IEEE Access*, 2024.

## Biographies



**Manh-Hung Ha**    received the M.S. degrees in Information Communication Technology from University of Paris 13, France, and the Ph.D. degree with the Department of Electrical Engineering, National Chung Cheng University, Taiwan in 2014 and 2021, respectively. He was Lecturer with the Faculty of Electrical Engineering, Phenikaa University, Hanoi, Viet Nam, from September 2021 to July 2022. Since July 2022, he has been a Lecturer with the

Faculty of Applied Science, International School, Vietnam National

University, Hanoi. He is a visiting scholar with the Department of Computer Science and Engineering, KL University, India in 2024. He also served as the Head of the group leader in Cognitive Machine Intelligence (CoMI) and the Head of the Combined Bachelors and Masters Degree Program on Financial Technology and Digital Business, International School, Vietnam National University, Hanoi from 2024. His major research interests include multimedia/image/video analytics, computer vision, speech signal processing, and machine learning. He can be contacted at email: [hunghm@vnu.edu.vn](mailto:hunghm@vnu.edu.vn)



**Duc-Chinh Nguyen**     received the Degree of Engineer from the School of Information and Communications Technology at Hanoi University of Science and Technology, Vietnam in 2019. Since graduation, he has worked as a Web Development Engineer at Temona Inc, Tokyo, Japan. Currently, Duc-Chinh is pursuing a Master's degree in Master of Informatics and Computer Engineering (MICE) at the International School, Vietnam National University,

Hanoi. His research interests focus on computer vision, machine learning, and deep learning, particularly in graph-related architectures. He can be contacted at email: [22075057@vnu.edu.vn](mailto:22075057@vnu.edu.vn).



control, and robotics. He can be contacted at email: [thaikd@vnu.edu.vn](mailto:thaikd@vnu.edu.vn).





**Thai Dinh Kim**     was born in 1984. He received the M.S. degree in control engineering and automation from the Thai Nguyen University of Technology, Vietnam, in 2013, and the Ph.D. degree in electrical and communications engineering from Feng Chia University, Taichung, Taiwan, in 2020. He is currently a Lecturer with the International School, Vietnam National University, Hanoi. His research interests include computer vision, intelligent



information geometry, and fitting curves on manifolds. His email address is [tamtt@vnu.edu.vn](mailto:tamtt@vnu.edu.vn).

**Tien-Tam Tran** received M.S. degrees in Mathematics from Aix-Marseille University, Marseille, France and Paul Sabatier University, Toulouse, France, in 2016 and 2020, respectively, and a Ph.D. degree from LIMOS, University Clermont Auvergne, in 2023. Since July 2023, he has been a Lecturer with the Faculty of Applied Science at the International School, Vietnam National University, Hanoi. His major research interests include Gaussian processes,



**Thanh Tien Do**     has been awarded the M.A. degree from Swinburne University of Technology, Australia, since 2018. He is currently working toward a PhD degree at the International School – Vietnam National University Hanoi. His research interests include IT architectures, e-government models, machine learning and optical communication. He can be contacted at the email address [thanhdo@vnuis.edu.vn](mailto:thanhdo@vnuis.edu.vn).



**Oscal Tzyh-Chiang Chen** (Senior Member, IEEE) received the B.S. degree in electrical engineering from National Taiwan University, in 1987, and the M.S. and Ph.D. degrees in electrical engineering from the University of Southern California, Los Angeles, USA, in 1990 and 1994, respectively. He worked with the Computer Processor Architecture Department, Computer Communication & Research Laboratories (CCL), Industrial Technology Research Institute (ITRI), serving a System Design Engineer, Project Leader, and the Section Chief, from 1994 to 1995. Then, he was an Associate Professor with the Department of Electrical Engineering, National Chung Cheng University (NCCU), Chiayi, Taiwan, from September 1995 to August 2003. He also served as the Director of the Academic Development Division, Office of Research and Development, NCCU, from July 2001 to July 2004, and the Director of the Technology Transfer Center, NCCU, from July 2003 to July 2004. In August 2003, he became a Professor with the Department of Electrical Engineering, NCCU, where he currently serves as the Department Chair. He was a Visiting Scholar with the Department of Electrical and Computer Engineering, Carnegie Mellon University, USA, from December 2007 to May 2008, and the Department of Electrical and Computer Engineering, University of Wisconsin-Madison, USA, from February 2011 to July 2011. He has published more than 170 journal and conference papers and five book chapters, and holds 36 Taiwan patents, 22 U.S. patents, and one Chinese patent. His research interests include multimedia processing and understanding, neural networks, VLSI systems, and communication systems. He is a Life Member of Chinese Fuzzy Systems Association. In the technical society, he was an Associate Editor of IEEE Circuits & Devices Magazine, from August 2003 to December 2006, and a Founding Member of the Multimedia Systems and Applications Technical Committee of IEEE Circuits and Systems Society. He also participates in the technical program committee of many IEEE international conferences and symposiums.