Applying Swin Architecture to diverse Sign Language Datasets

Yulia Kumar *, Kuan Huang *, Chin-Chien Lin, Annaliese Watson, J. Jenny Li, Patricia Morreale, Justin Delgado

Posted Date: 27 February 2024

doi: 10.20944/preprints202402.1506.v1

Keywords: Swin Transformer; ASL detection; Deep Learning; The Unvoiced

Article

Applying Swin Architecture to Diverse Sign Language Datasets

Yulia Kumar *, Kuan Huang *, Chin-Chien Lin, Annaliese Watson, J. Jenny Li, Patricia Morreale and Justin Delgado

Kean University (Union, NJ, USA); wl02090177@gmail.com, [1], watsanna@kean.edu, juli@kean.edu, pmorreal@kean.edu, delgajus@kean.edu

* Correspondence: ykumar@kean.edu, khuang@kean.edu

Abstract: In the era of Artificial Intelligence (AI), comprehending and responding to non-verbal communication is increasingly vital. This research extends AI's reach in bridging communication gaps, notably benefiting American Sign Language (ASL) and Taiwan Sign Language (TSL) communities. It focuses on employing various AI models, especially the Hierarchical Vision Transformer with Shifted Windows (Swin), for recognizing diverse sign language datasets. The study assesses Swin architecture's adaptability to different sign languages, aiming to create a universal platform for Unvoiced communities. Utilizing deep learning and transformer technologies, hybrid application prototypes have been developed for ASL-to-English translations and vice versa, with plans to expand this to multiple sign languages. The Swin models, trained on varied dataset sizes, show considerable accuracy, indicating their flexibility and effectiveness. This research underscores major advancements in sign language recognition and underlines a commitment to inclusive communication in the digital era. Future work will focus on enhancing these models and broadening their scope to include more sign languages, integrating multimodality and Large Language Models (LLMs), thereby fostering global inclusivity.

Keywords: Swin Transformer; ASL detection; deep learning; the Unvoiced

1. Introduction

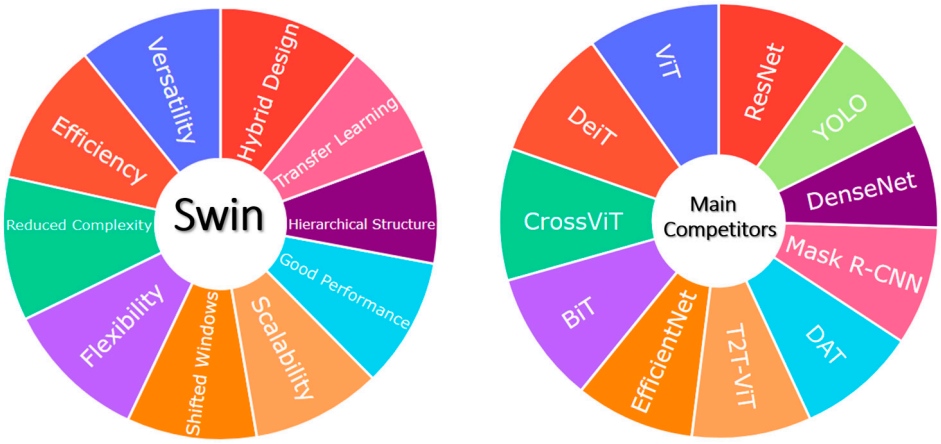
In the digital era, where rapid and error-free communication is paramount, a significant gap exists between verbal communication and the needs of individuals who rely on sign language, notably the deaf and hard-of-hearing communities. This research addresses this gap by leveraging Artificial Intelligence (AI) to bridge the communication divide. Focusing primarily on American Sign Language (ASL) while also incorporating Taiwan Sign Language (TSL) for comparative analysis, the study applies advanced AI techniques, particularly deep learning, and transformer-based neural networks. These technologies are employed to develop a comprehensive understanding of sign language data. The research includes the utilization of the Vision Transformer with Deformable Attention (DAT) and the Hierarchical Vision Transformer using Shifted Windows (Swin) models. These models are pivotal in recognizing ASL gestures, with a specific focus on the Swin model's expected superior performance. The study has achieved notable success, including 100% accuracy in ASL alphabet detection using a large Kaggle dataset. A key aspect of this research is the development of applications capable of real-time ASL-to-English translation and vice versa. These applications, designed for various user groups, integrate Text-To-Speech and Speech-To-Text functionalities. The research also emphasizes the crucial balance between accuracy, which requires significant computational resources, and the real-time responsiveness essential for practical usage. Through this study, researchers aim to not only provide tools for the Unvoiced community but also to educate others in sign language, thereby fostering a more inclusive digital environment.

 doi:10.20944/preprints202402.1506.v1

2. Understanding Swin Transformer

Swin Transformer, introduced in 2021 by Liu et. al. from Microsoft Research has become a subject of various projects [4–6] in both image and video processing [7–9]. It has gained significant popularity due to its unique approach and effectiveness in various vision tasks. Though, the approach of applying it to American Sign Language (ASL) and Taiwanese Sign Languages (TSL), showcasing the model's versatility, is novice.

The uniqueness of this transformer is in its hybrid design that combines Convolutional Neural Networks (CNNs) and Transformers with features like shifted windows and hierarchical structure. Those help it to handle the nuances of sign language across cultures.Most stand-out characteristics of Swin Transformer and its main competitors presented in Figure 1.



(**a**) (**b**)

Figure 1. (**a**) Swin Characteristics; (**b**) Main Competitors of Swin Transformer.

Figure 1 illustrates the numerous advantages of the Swin Transformer, a model renowned for its efficiency, scalability, and superior performance across various vision tasks. Its pioneering integration of Convolutional Neural Networks' (CNNs) local feature extraction and Transformers' global context awareness has redefined the standards in computer vision [10]. Concurrently, alternative models in this domain also exhibit unique strengths. This study compares the Swin Transformer's performance against other models such as VGG-16, Resnet-50, and DAT. A notable feature of the Swin Transformer is its capability to process both images and videos [4,6,10], a critical aspect for dynamic Sign Language recognition. The AI models depicted in Figure 1 vary in their suitability for different applications, depending on specific requirements like real-time processing, model size constraints, training data availability, and task complexity. The Swin Transformer distinguishes itself by providing an optimal balance between efficiency and accuracy, particularly in scenarios requiring a comprehensive understanding of both local and global image features [4,10]. It demonstrates exceptional ability in identifying and delineating multiple objects within images. Its application scope extends to semantic and instance segmentation, pose and depth estimation, transfer learning, and panoptic segmentation - a fusion of semantic and instance segmentation [4,11,12].

The Swin Transformer's unique shifted window scheme, which processes images in non-overlapping windows with self-attention, followed by layer-wise window shifting for cross-window connections, enables it to efficiently encompass broader contexts. Unlike traditional Transformers, Swin Transformers create feature maps at various resolutions, enhancing their applicability in diverse vision tasks [4,13,14]. Its versatility in processing images of different sizes makes it an advantageous tool for varied real-world applications, as demonstrated in this paper [4]. Achieving state-of-the-art results in benchmarks like ImageNet [4], the Swin Transformer is also adept at handling tasks such as tumor detection and organ segmentation in medical imaging [15].

 doi:10.20944/preprints202402.1506.v1

3. Diverse Datasets

The Unvoiced community in the US and Canada is remarkably diverse, encompassing numerous dialects including the predominant Black American Sign Language (BASL), as well as dialects from Bolivia, Burundi, Costa Rica, Ghana, Nigeria, Francophone regions, and Québec. Similarly, the United Kingdom and Australia have distinct versions of ASL: BSL and Auslan, respectively. These variations are analogous to regional differences in spoken American English, where accents and slang can create communication barriers even within the same language. In ASL, while there are no sound accents, variations manifest in the form of different signs and gestures. Figure 2 represent the training dataset from Kaggle (on top) [3] and dataset, collected especially for this study (at the bottom):

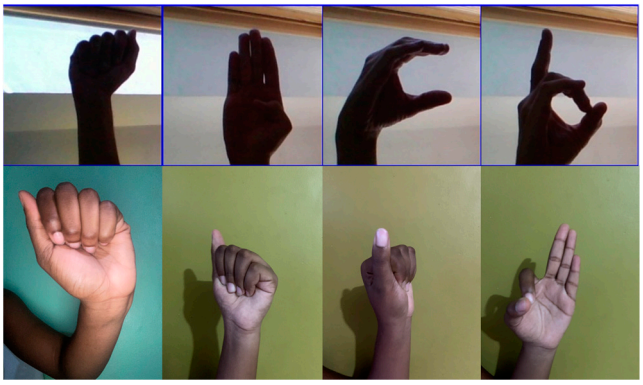
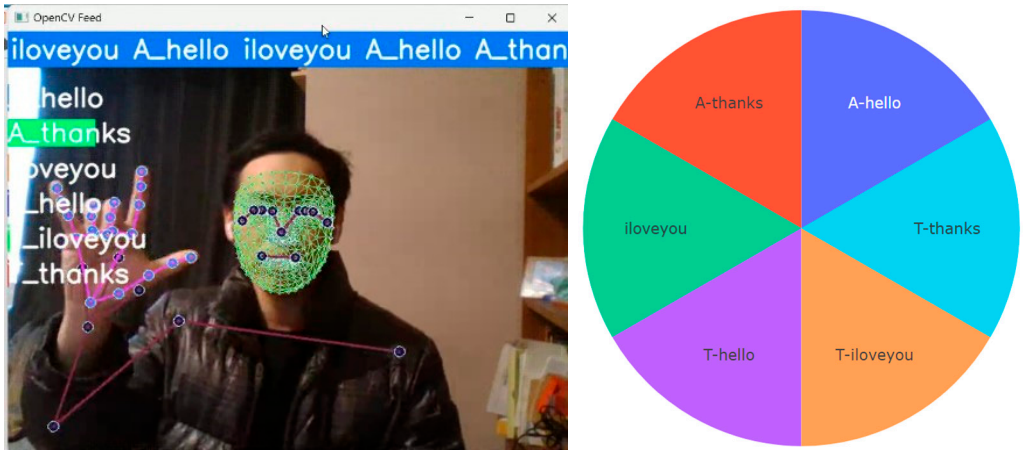


Figure 2. Examples of the training dataset (on top) vs testing dataset (at the bottom).

Regarding the project's ASL dataset, Figure 2 reveals that the initial training dataset predominantly features white male hands. To counter this limited representation, a supplementary dataset, curated by a female researcher proficient in ASL, introduces greater diversity. This step addresses a critical aspect of AI research in sign language recognition: the potential for algorithms to perpetuate biases. A diverse and representative dataset is imperative to ensure equitable recognition across different groups of signers, each with unique cultural backgrounds and signing styles. In sign language, communication is not just about hand movements; it also encompasses facial expressions and body language. The subtleties of these signs, influenced by different expressions and postures, are crucial for accurate interpretation and must be reflected in the training data to avoid biases. Figure 3 represents the Taiwanese dataset used in the study. This custom dataset is represented by three phrases trained with three same phrases in ASL, consisting of 6 classes.



(**a**) (**b**)

Figure 3. (**a**) Snapshot of Live TSL vs ASL Recognition; (**b**) Class Distribution in TSL vs ASL Dataset.

 doi:10.20944/preprints202402.1506.v1

At this moment the researchers are just experimenting with video recognition of ASL and TSL. In the future works recognition of other open-source datasets is planned. Among these potentially are:

- American Sign Language Lexicon Video Dataset, that consists of videos of more than 3,300 ASL signs in citation form, each produced by 1-6 native ASL signers [16].
- World Level American Sign Language Video Dataset on Kaggle: This dataset contains 12k processed videos of word-level ASL glossary performances [17].
- ASL Citizen by Microsoft Research: The first crowdsourced isolated sign language video dataset containing about 84k video recordings of 2.7k isolated signs from ASL [18].
- MS-ASL Dataset: A large-scale sign language dataset comprising over 25,000 annotated videos [19].
- OpenASL Dataset: A large-scale ASL - English dataset collected from online video sites, containing 288 hours of ASL videos in multiple domains from over 200 signers [20].
- How2Sign Dataset: A multimodal and Multiview continuous ASL dataset, consisting of a parallel corpus of more than 80 hours of sign language videos along with corresponding modalities including speech, English transcripts, and depth [21].
- YouTube-ASL Dataset: A large-scale, open-domain corpus of ASL videos and accompanying English captions drawn from YouTube, with about 1000 hours of videos [22].
- ASL video dataset - Boston University: a large and expanding public dataset containing video sequences of thousands of distinct ASL signs (produced by native signers of ASL), along with annotations of those sequences [23].

We experimented with various testing cases as well, some of which are presented in Figure 4.

Figure 4. Experiments with ASL testing scenarios.

Regarding Taiwanese Sign Language finding reliable dataset gets more challenging, as such, the Taiwanese Across Taiwan (TAT) corpus is a large-scale database of Native Taiwanese Article/Reading Speech collected across Taiwan. Although it's not a sign language dataset, it might be useful for voice recognition research [24]. A Survey of Sign Language in Taiwan by SIL International provides a comprehensive listing of the world's languages, including more than one hundred signed languages. It lists one sign language for Taiwan with two major dialects, Taipei, and Tainan [25].

4. Related Work

While working on the project, the team analyzed several products that currently produce ASL-To-English Translation. The Brazilian Hand Talk mobile app, that provides translation to the sign- language, is the most impressive [26]. The Hand Talk's app was developed by Ronaldo Tenório and used an animated character Hugo to produce a Portuguese into Libras translation. There are several other attempts as well as English-To-ASL converters [27]. Some researchers and developers prefer just to brainstorm and propose ideas about such an app, as its actual development requires time and joint efforts of several AI professionals as well as financial resources. One of these is posted by Ankit Jain on the Geeks for Geeks website [28]. There are several other ASL Apps [29–32].

 doi:10.20944/preprints202402.1506.v1

The recent advancements in Swin Transformers have led to significant developments across various fields. For instance, in paper [33] by Liang et al., the focus is on image restoration, demonstrating the Swin Transformer's effectiveness in enhancing image quality. Cao et al. [34] highlights its application in medical imaging, specifically in segmenting complex structures in medical images. In the realm of self-supervised learning, Xie et al.'s work [35] emphasizes the utility of Swin Transformers in learning without labeled data, a major step forward in machine learning. He et al. [36] further apply this technology to remote sensing images, improving semantic segmentation capabilities in geospatial analysis. Zu et al. [37] explore the use of Swin Transformers for classifying pollen images, showcasing their potential in environmental and botanical studies. Nguyen et al.'s research [38] in dynamic semantic communication demonstrates the model's efficiency in handling diverse computational requirements in communication systems. The versatility of Swin Transformers is further evidenced in MohanRajan et al.'s [39] study for land use and cover change detection, and Ekanayake et al.'s work in MRI reconstruction [40], showing its effectiveness in both environmental monitoring and medical imaging.

In the field of video analysis, Lu et al. [41] apply Swin Transformers for classifying earthwork activities, enhancing the accuracy of such tasks. Lin et al.'s CSwinDoubleU-Net model [42] combines convolution and Swin Transformer layers for improved medical image segmentation, particularly in detecting colorectal polyps. Moreover, Pan et al.'s study on renal incidentaloma detection [43] using a YOLOv4+ASFF framework with Swin Transformers marks an advancement in the detection and classification of medical conditions through imaging. Interesting results demonstrated paper by Kumar et. al. on applying Swin to Early Cancer Detection [44].

5. Applying Transformers to the ASL Dataset

To estimate which models are capable of accurately predicting ASL and can further be tuned for this purpose simulations were set up. Details can be seen in Table 1.

Table 1. Simulation Parameters.

Trial Parameter	Comments
Initial Dataset	87000 images
Trial Dataset	80% for training, 20% for testing at random
Classification	29 classes (A to Z, Space, Del, and Nothing)
Batch Size	16
Trial Dataset	256×256 (resized)
Optimizer used	SGD, learning rate 0.001
Number of Epochs	100
Pythorch version	1.12.1.

The trial parameters can be clearly seen from Table 1. 80% of the images were randomly selected for the training dataset and 20% of the images for testing. The input image was resized to 256×256. The optimizing method that worked best for appeared to be the stochastic gradient descent (SGD) method, with a learning rate of 0.001. The training took over 300 epochs on Ubuntu 20.04.5 Linux system with the following characteristics: AMD EPYC 7513 32-Core Processor 2.60GHz and 8 NVIDIA GeForce 3090 graphics cards, and each one has 24 Gigabyte memory.

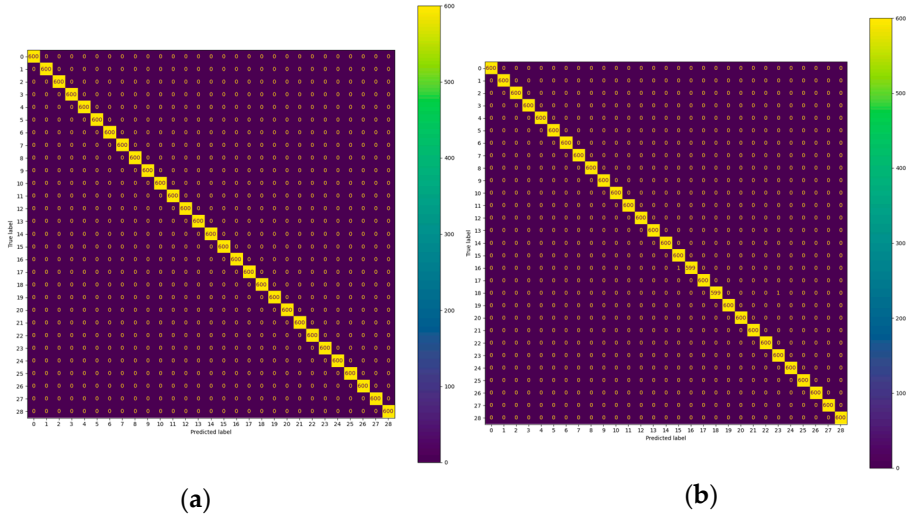
The simulation results can be seen in Table 2.

Table 2. Simulation Parameters and Accuracy achieved.

Trial Parameter	Number of Parameters	Accuracy
DAT Transformer	86,886,357	99.99%
VGG-16	165,845,085	**100%**
ResNet-50	23,567,453	**100%**
Swin Transformer	65,960,349	**100%**

 doi:10.20944/preprints202402.1506.v1

Table 2 presents a comparative analysis of deep learning models in terms of their parameters and accuracy in ASL to text translation. The table highlights ResNet-50 as the most parameter-efficient model, making it potentially more suitable for mobile applications. In contrast, VGG-16, with the highest parameter count due to its three fully connected layers, may be less optimal for such applications. Despite this, all models, including the DAT and Swin Transformers, achieve high accuracy, illustrating a balance between model complexity and performance in ASL translation tasks. Further insights into these models' performance are provided through training and testing loss curves in subsequent visualizations. It was found that the DAT transformer did not outperform the Swin Transformer in this project, which does not match the original paper of on the DAT transformer [45] that claimed that it should.



(**a**) (**b**)

Figure 5. (**a**) Confusion matrix for VGG-16, ResNet-50, and Swin Transformer; (**b**) Confusion matrix for DAT transformer.

Figure 6 represents Training and Testing Loss Curves for all models.

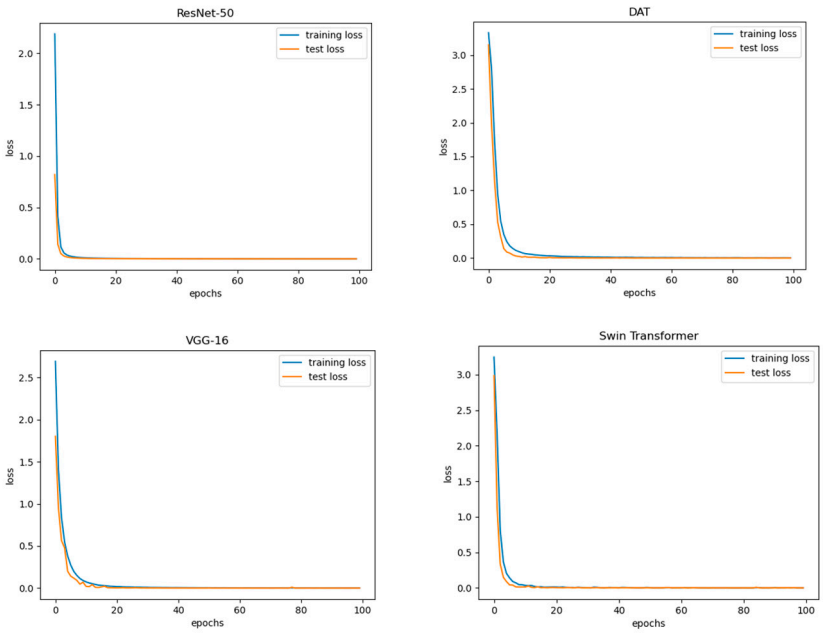
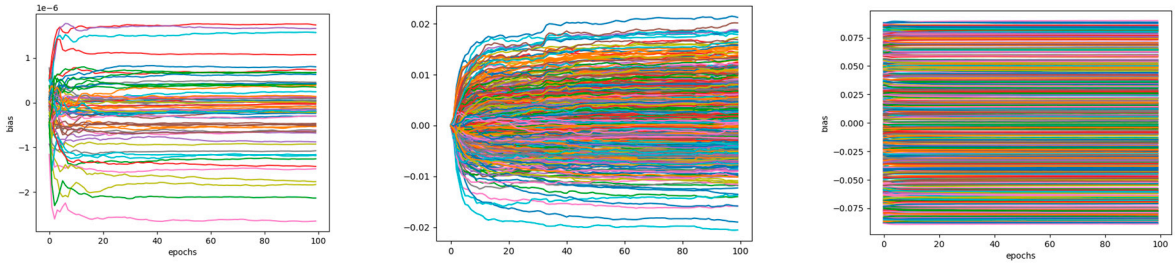


Figure 6. Training and Testing Loss Curves of the VGG-16 (bottom left), ResNet-50 (top left), DAT Transformer (bottom right), and Swin Transformer (bottom right).

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 February 2024 doi:10.20944/preprints202402.1506.v1

As can be seen from Figure 6, all four models demonstrated the biggest reduction in losses right away, around less than 10-15 epochs. To further understand the model, the researchers then proceeded with bias analyses. The analysis was built upon previously developed strategies [46,47].

The visualization of the biases discovered in the models can be seen in Figure 7:



(**a**) (**b**) (**c**)

Figure 7. Bias Visualizations: (**a**) The 1st convolutional layer of VGG-16; (**b**) The Multilayer perceptron of the 1st transformer stage of the Swin Transformer; (**c**) The Multilayer perceptron of the 1st transformer stage of the DAT Transformer.

The visualizations for VGG-16 demonstrate stable bias values after 15-20 epochs, within a range from -2.75 to 1.75 (1e-6). For the Swin and DAT Transformers, the focus is on the first transformer stage's Multilayer perceptron (MLP) biases [48]. The Swin Transformer shows a unique dome-like bias shape, suggesting a need for deeper analysis in terms of distribution, density, and outliers. Conversely, the DAT Transformer's biases converge around epochs 45-50, then stabilize, indicating less fluctuation post-convergence. This analysis aids in understanding the learning behaviors of these models.

ResNet-50 CNN model demonstrated no biases in its 1st convolutional layer. Parameters of the model can be seen from (1):

$$\textbf{self.}\textit{conv1 = nn.Conv2d(3, self.inplanes, kernel\_size=7, stride=2, padding=3, bias=False)} \quad (1)$$

The ResNet-50 model's first convolutional layer is designed without bias parameters to streamline the number of variables and potentially improve computational efficiency. This decision can be influenced by the fact that in deep learning models, especially in convolutional neural networks, bias terms are sometimes omitted. This is because batch normalization, often applied after convolutional layers, negates the effect of the bias by standardizing the output. Therefore, removing the bias parameter can reduce the model's complexity without significantly affecting its performance.

AI biases and Explainable AI are at the forefront of artificial intelligence research due to their importance in ensuring AI models are used responsibly in society. The intricacies of neural networks and transformers necessitate transparent AI decisions to foster public trust.

Further in this section, Class Activation Maps (CAMs) are employed to visually interpret the focus areas of deep learning models used in image classification, which is vital for understanding the decision-making process of AI [49]. CAMs produce heatmaps that identify critical regions influencing the classification decision, offering a comprehensive view when combined with bias data. This is particularly useful in explicating the opaque decision-making process in deep learning, enhancing user trust by demystifying AI classifications. The upcoming visualizations will compare the focus areas of different models, like CNN's ResNet-50 and the DAT transformer, highlighting their differing attention to image features which may affect accuracy.

 doi:10.20944/preprints202402.1506.v1

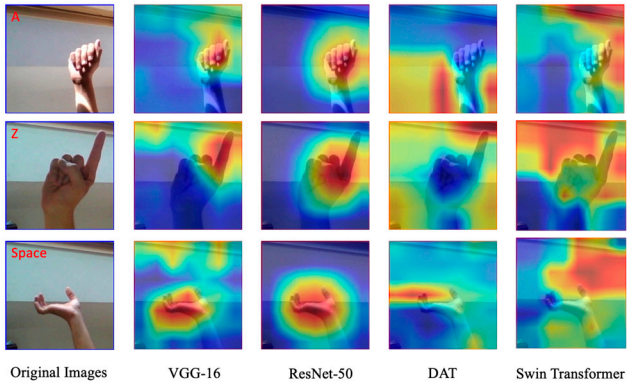
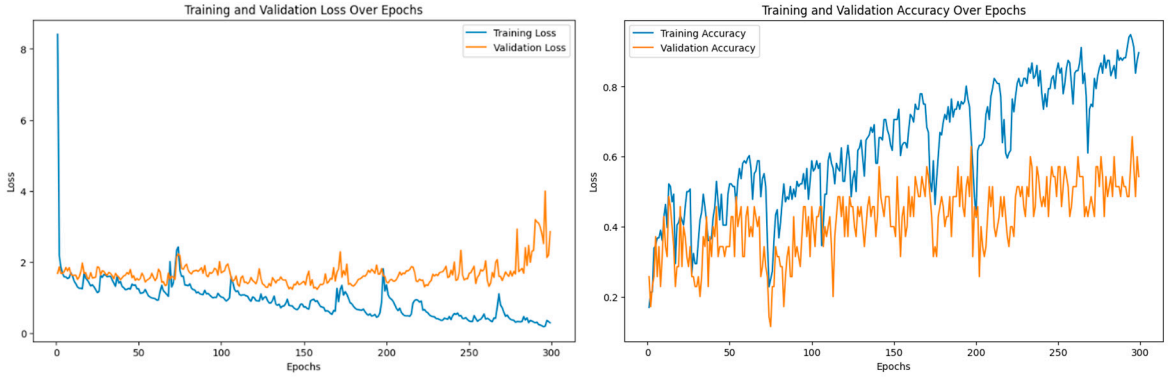


Figure 8. The CAMs of the Swin Transformer in comparison with other DL models.

Currently only preliminary results of ASL vs TSL video classification are available. Figure 9 demonstrates current losses and accuracy of 6 classes video classification mentioned above and demonstrated on Figure 3.



(**a**) (**b**)

Figure 9. (**a**) Training and Testing Losses of ASL vs TSL video classification; (**b**) Training and Testing Accuracy of ASL vs TSL video classification.

As can be seen training accuracy achieves 100% at 300 epochs what is well-performed considering a lightweight dataset of study.

6. Case Studies and Applications

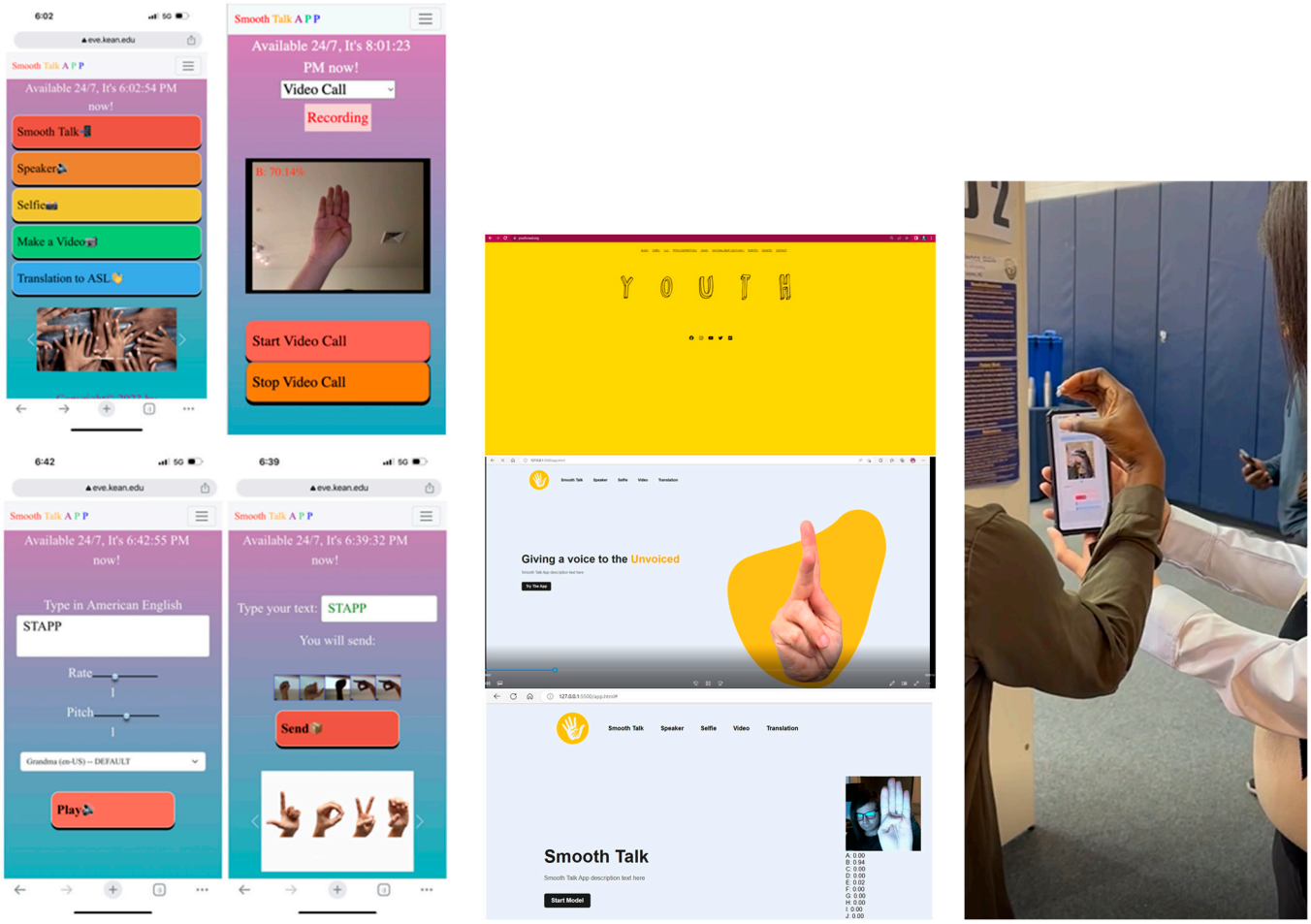
The research tackles the communication problem between those who know and use ASL every day and those who do not. In simple words, the researchers aim at creating a swift way for understanding by providing smooth communication for those involved. Case studies and associated applications consist of two types and associated cases:

(1) Develop a user-friendly interface for ASL translation, ensuring it is suitable for the intended users and use cases.
(2) Create interactive and engaging learning tools for ASL education.

To address the first case two different applications, the *Smooth Talk* app and *STApp* app were created. Both feature a very simple and user-friendly interface. The goal was originally to develop only one application and STApp stands for a Smooth Talk app as well but eventually two different apps were developed by the team of researchers. Both apps capture the ASL language live, translates it into English using python backend. The first yellowish version of the app was inspired by the Low-code AI TeachableMachines web tool [50], that was used to practice ASL language. Figure 10 (b, bottom) demonstrates the accuracy of 94% for the letter 'B' what was the top accuracy that could be

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 February 2024 doi:10.20944/preprints202402.1506.v1

achieved with the app as it was built to use a custom light ASL dataset, collected for the project. As can be seen from Figure 10 (a) another prototype demonstrates an accuracy of 70.14%. Figure 10 (c) demonstrates the live demo of the Smooth Talk app.



(**a**) (**b**) (**c**)

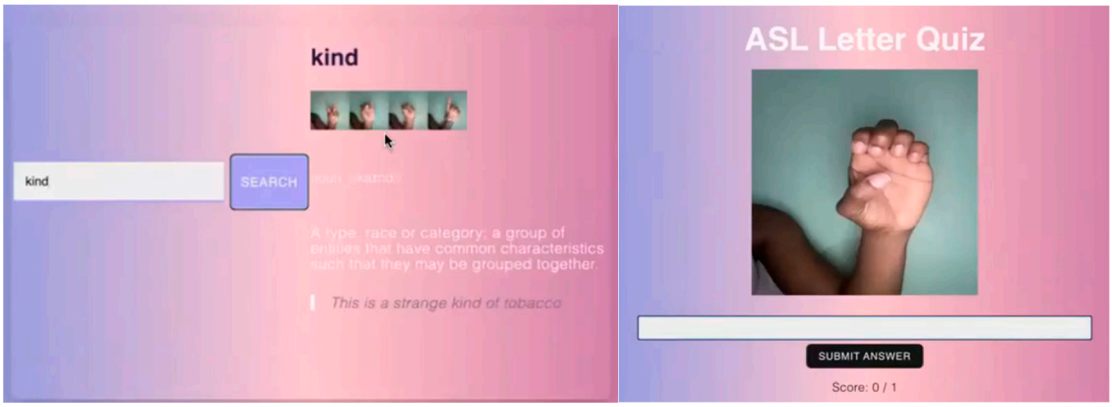
Figure 10. (**a**) Snapshot of the STApp app pages; (**b**) Home page of the NAD Youth Website [51] (top) vs the Smooth Talk Home Page (middle) vs ASL recognition page of the app(bottom); (**c**) Live demo of the app.

The mobile-first GUI of both apps relies on Bootstrap framework, CSS flex and other front-end technologies targeting responsive web design to further accommodate the users and potentially allow them to use the app from any place. They both can be considered currently hybrid apps working equally well both on a smartphone and in the web browser. The use JavaScript and its APIs, for example, to convert text to speech and deliver it to the hearing side of the conversation further enhances the prototype. The speech is converted to text using the same API. The text is translated into ASL letters using Map aka Dictionary Data Structure and displays the result to the *Unvoiced* person. The look and feel of the Small Talk app resemble the website of NAD Youth [51] - a project of the National Association of the Deaf [52]. It also inspired the STApp app GUI.

As can be seen from Figure 10 (a), the STApp app uses emoji as icons on its buttons what adds a uniqueness to this web design. The STApp app also provides the functionality of taking a selfie and prerecording a video, then it sends it to the other side of conversation accommodating asynchronous communication and making it aware of the translation accuracy.

To address the second educational case another app was developed (see Figure 11).

 doi:10.20944/preprints202402.1506.v1



(**a**) (**b**)

Figure 11. (**a**) Snapshot of the Educational app ASL Translation page; (**b**) Quiz page of the app.

At this point the app user can log in into the educational system, watch video recordings of the ASL language lesson and take the quiz to test their knowledge. The researchers are considering converting this quiz into an app game to add interactivity to it and attract a broader population. Obviously Swin Transformer and other AI models and tools can be used in ASL educational platforms for improved learning experiences and interactive applications, or games can further assist with that. Eventually the researchers plan on integrating into the platform for both ASL and TSL. Once the research project was launched the researchers themselves had to learn the basics of the ASL language to some degree.

7. Sign Language and LLMs

Integrating Large Language Models (LLMs) into this research can enhance interpretability and user interaction with the system. Such approaches as automated annotation with LLMs are becoming mainstream. LLMs can generate descriptions or labels for video clips or images based on predefined criteria, which can then be verified or refined by experts. As LLMs are constantly improving they will eventually be able to assist in identifying potential gaps or biases as it is critical to avoid biases in ASL recognition.

Latest LLMs such as ChatGPT-4-Vision and Gemini were tested on ASL dataset. Special AI assistant aka custom GPT *Sign Speak Guide* was created with the help of OpenAI API. It can be seen below in Figure 11:

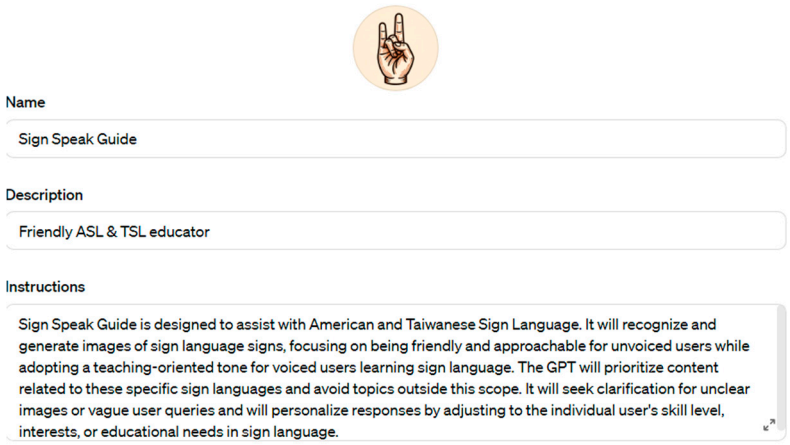


Figure 12. The configuration snapshot of the AI assistant *Sign Speak Guide*.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 February 2024 doi:10.20944/preprints202402.1506.v1

The assistant, that utilizes ChatGPT-4-Vision model can work with both ASL and TSL but accuracy of ASL is slightly better. Figure 13 demonstrates an image of the letter ‘A’ fed to the bot.

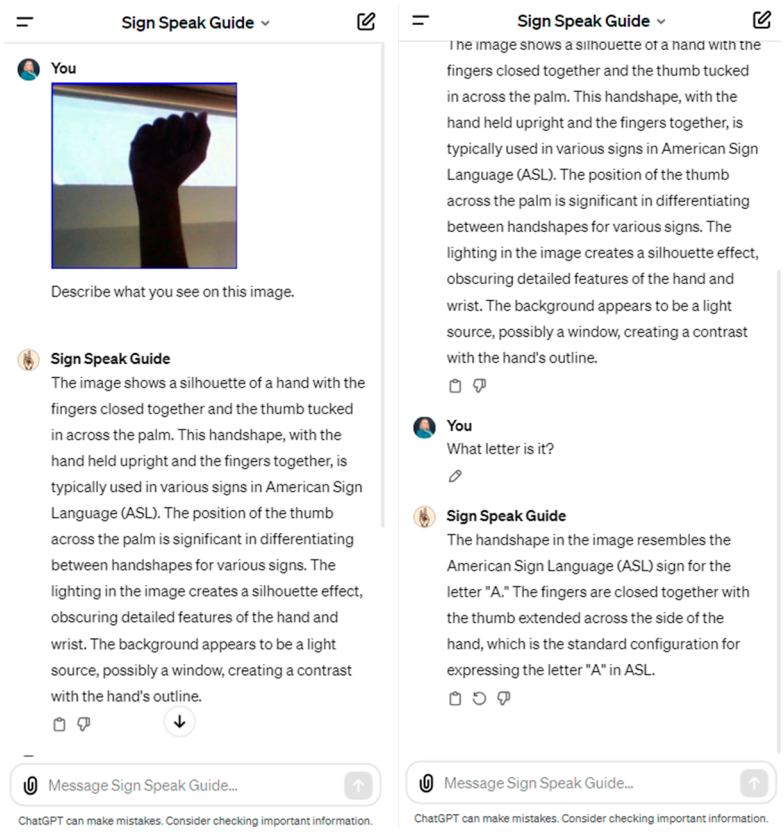


Figure 13. The snapshot of the AI assistant *Sign Speak Guide* in process of ASL recognition.

As can be seen from Figure 13 AI assistant correctly recognizes the letter ‘A’ provided.

8. Discussion

Sign language recognition and translation is a complex task that involves understanding subtle hand movements, facial expressions, and body language. It is also hard to find researchers or enthusiasts fluent in sign language and willing to devote their time and efforts to such a project. Sign language is both spatial (involving the positioning and movement of hands, fingers, facial expressions, etc.) and temporal (the meaning can depend on the sequence of movements). The researchers conclude that the Swin Transformer’s capabilities could potentially be very relevant to this dataset as the shifted window approach could efficiently handle the spatial aspect of sign language. Its ability to handle video data (spatial temporal) makes it suitable for interpreting sign language in a dynamic, continuous context. Integrating Large Language Models (LLMs) like GPT-4+ or similar technologies can further enhance the capabilities of the proposed apps. LLMs are expected to be particularly useful in ASL education and facilitate interactive language learning experiences. In data classification, they can assist in interpreting and summarizing classification results. LLMs can provide contextual translations of signs or offer cultural insights into sign language usage, help practicing sign language or explain concepts.

Integrating multimodality into the research on ASL dataset classification, especially when combined with Large Language Models (LLMs), can create a more comprehensive and effective system. Integrating not only images and text but audio and video will improve understanding and interaction. The top standard would be creating an interface that adapts to the user’s preferences and accessibility needs.

 doi:10.20944/preprints202402.1506.v1

9. Conclusions and Future Work

The research and developed app prototypes will facilitate communication between those who primarily communicate through Sign Language and those who do not. Our trials resulted with the following accurate outcomes: Swin Transformer achieved 100%, and CNN models achieved 100% as well. Future research will include exhaustive testing of the prototypes and LLMs in the field of ASL recognition. It is expected that validation will make a reliable ASL tool for all possible. The ethical scope of the problem we tackle is very sensitive, handling of personal data must be discussed. The integration of transformers and Large Language Models (LLMs) in American Sign Language (ASL) and Taiwan Sign Language (TSL) detection marks a significant advancement in the field of sign language recognition, challenging traditional methods and establishing new benchmarks.

The emergence of comprehensive datasets has been instrumental in the development and testing of advanced sign language recognition models. The Video Swin Transformer [3], with its potential in video-based sign language recognition, represents a new era in understanding and interpreting sign language through visual data.

Despite these advancements, challenges remain. One of the primary challenges is the creation of large, diverse, and high-quality datasets that accurately represent the complexity of ASL and TSL. Additionally, real-time processing capabilities are crucial for practical applications of these technologies. Future research should focus on tailoring transformer and LLM-based models to accommodate the specific requirements of sign language recognition more effectively. Bridging the gap between academic research and practical, real-world applications of ASL and TSL detection technologies is essential.

In conclusion, the integration of transformers and LLMs in ASL and TSL detection represents a significant advancement in the field. These technologies offer enhanced capabilities for interpreting sign language, leading to more accurate and efficient ASL and TSL detection systems. However, continuous efforts in data enhancement and model optimization are crucial to address existing challenges and further advance the field. The datasets and models discussed herein offer a glimpse into the current state of the field and its potential trajectory. Continuous advancements in this domain hold the promise of bridging communication gaps for the Deaf and Hard of Hearing communities globally, enhancing inclusivity and accessibility. Once future improvements in Swin Transformer and LLM approach the systems must be refined and updated to stay cutting edge apps.

Author Contributions: Conceptualization, Y.K. and K.H.; methodology, Y.K. and K.H.; software, J.D., C.-C.L. and A.W.; validation, C.-C.L., Y.K., K.H. and J.J.L.; formal analysis, J.D., C.-C.L. and A.W.; investigation, Y.K., J.D., C.-C.L., and K.H.; resources, Y.K.; data curation, P.M.; writing—original draft preparation, Y.K. and K.H.; writing—review and editing, J.J.L. and P.M.; visualization, Y.K. and K.H.; supervision, P.M.; project administration, Y.K.; funding acquisition, Y.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the NSF, grants awards 1834620 and 2137791, and Kean University's Students Partnering with Faculty 2023 Summer Research Program (SPF).

Data Availability Statement: Data available on request due to privacy restrictions (personal nature of ASL communication).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. GitHub repository of DAT transformer. Available online: https://github.com/LeapLabTHU/DAT (accessed on 2/24/2024).
2. GitHub repository of Swin transformer. Available online: https://github.com/microsoft/Swin-Transformer (accessed on 2/24/2024).
3. ASL Alphabet [Online], available at https://www.kaggle.com/datasets/grassknoted/asl-alphabet, last visited on 2/24/2024.

 doi:10.20944/preprints202402.1506.v1

4. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012-10022).
5. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., ... & Guo, B. (2022). Swin transformer v2: Scaling up capacity and resolution. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 12009-12019).
6. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., & Hu, H. (2022). Video Swin transformer. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3202-3211).
7. Lu, Y., You, K., Zhou, C., Chen, J., Wu, Z., Jiang, Y., & Huang, C. (2024). Video surveillance-based multi-task learning with Swin transformer for earthwork activity classification. Engineering Applications of Artificial Intelligence, 131, 107814.
8. Hu, X., Hampiholi, B., Neumann, H., & Lang, J. (2024). Temporal Context Enhanced Referring Video Object Segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 5574-5583).
9. Yu, Z., Guan, F., Lu, Y., Li, X., & Chen, Z. (2024). Video Quality Assessment Based on Swin TransformerV2 and Coarse to Fine Strategy. arXiv preprint arXiv:2401.08522.
10. Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).
11. He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
12. Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
13. Xie, Saining, et al. "Aggregated residual transformations for deep neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
14. Kirillov, Alexander, et al. "Panoptic segmentation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
15. Huang, Gao, et al. "Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation." arXiv preprint arXiv:2201.01266 (2022).Author 1, A.; Author 2, B. *Book Title*, 3rd ed.; Publisher: Publisher Location, Country, 2008; pp. 154–196.
16. Home page of ASLLVD (American Sign Language Lexicon Video Dataset). Available online: https://paperswithcode.com/dataset/asllvd (accessed on 2/24/2024).
17. WLASL Dataset on Kaggle. Available online: https://www.kaggle.com/datasets/grassknoted/asl-alphabet (accessed on 2/24/2024).
18. Microsoft Research ASL Citizen Dataset. Available online: https://www.microsoft.com/en-us/research/project/asl-citizen/ (accessed on 2/24/2024).
19. MS-ASL dataset. Available online: https://www.microsoft.com/en-us/research/project/ms-asl/ (accessed on 2/24/2024).
20. GitHub repository of OpenASL dataset. Available online: https://github.com/chevalierNoir/OpenASL (accessed on 2/24/2024).
21. GitHub repository of how2sign dataset. Available online: https://how2sign.github.io/ (accessed on 2/24/2024).
22. Uthus, D., Tanzer, G., & Georg, M. (2024). Youtube-asl: A large-scale, open-domain American sign language-English parallel corpus. Advances in Neural Information Processing Systems, 36.
23. Colarossi J. (2021) World's Largest American Sign Language Database Makes ASL Even More Accessible. Available online: https://www.bu.edu/articles/2021/worlds-largest-american-sign-language-database-makes-asl-even-more-accessible/ (accessed on 2/24/2024).
24. Home Page of TAT (Taiwanese Across Taiwan). Available online: https://paperswithcode.com/dataset/tat (accessed on 2/24/2024).
25. A survey of sign language in Taiwan. Available online: https://www.sil.org/resources/archives/9125 (accessed on 2/24/2024).
26. Sklar J. A mobile app gives deaf people a sign-language interpreter they can take anywhere. Available online: https://www.technologyreview.com/innovator/ronaldo-tenorio/ (accessed on 2/24/2024).
27. Ankit Jain. Project Idea | Audio to Sign Language Translator. Available online: https://www.geeksforgeeks.org/project-idea-audio-sign-language-translator/ (accessed on 2/24/2024).

28. English to Sign Language (ASL) Translator. Available online: https://wecapable.com/tools/text-to-sign-language-converter/, (accessed on 2/24/2024).
29. The ASL App (ASL for the People) on Google Play. Available online: https://theaslapp.com/about (accessed on 2/24/2024).
30. iASL App on speechie apps. Available online: https://speechieapps.wordpress.com/2012/03/26/iasl/ (accessed on 2/24/2024).
31. Sign 4 Me App. Available online: https://apps.microsoft.com/detail/9pn9qd80mblx?hl=en-us&gl=US (accessed on 2/24/2024).
32. ASL Dictionary App. Available online: https://play.google.com/store/apps/details?id=com.signtel&gl=US (accessed on 2/24/2024).
33. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., & Timofte, R. (2021). Swinir: Image restoration using swin transformer. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1833-1844).
34. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2022, October). Swin-unet: Unet-like pure transformer for medical image segmentation. In European conference on computer vision (pp. 205-218). Cham: Springer Nature Switzerland.
35. Xie, Z., Lin, Y., Yao, Z., Zhang, Z., Dai, Q., Cao, Y., & Hu, H. (2021). Self-supervised learning with Swin transformers. arXiv preprint arXiv:2105.04553.
36. He, X., Zhou, Y., Zhao, J., Zhang, D., Yao, R., & Xue, Y. (2022). Swin transformer embedding UNet for remote sensing image semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-15.
37. Zu, B., Cao, T., Li, Y., Li, J., Ju, F., & Wang, H. (2024). SwinT-SRNet: Swin transformer with image super-resolution reconstruction network for pollen images classification. Engineering Applications of Artificial Intelligence, 133, 108041.
38. Nguyen, L. X., Tun, Y. L., Tun, Y. K., Nguyen, M. N., Zhang, C., Han, Z., & Hong, C. S. (2024). Swin transformer-based dynamic semantic communication for multi-user with different computing capacity. IEEE Transactions on Vehicular Technology.
39. MohanRajan, S. N., Loganathan, A., Manoharan, P., & Alenizi, F. A. (2024). Fuzzy Swin transformer for Land Use/Land Cover change detection using LISS-III Satellite data. Earth Science Informatics, 1-20.
40. Ekanayake, M., Pawar, K., Harandi, M., Egan, G., & Chen, Z. (2024). McSTRA: A multi-branch cascaded Swin transformer for point spread function-guided robust MRI reconstruction. Computers in Biology and Medicine, 168, 107775.
41. Lu, Y., You, K., Zhou, C., Chen, J., Wu, Z., Jiang, Y., & Huang, C. (2024). Video surveillance-based multi-task learning with Swin transformer for earthwork activity classification. Engineering Applications of Artificial Intelligence, 131, 107814.
42. Lin, Y., Han, X., Chen, K., Zhang, W., & Liu, Q. (2024). CSwinDoubleU-Net: A double U-shaped network combined with convolution and Swin Transformer for colorectal polyp segmentation. Biomedical Signal Processing and Control, 89, 105749.
43. Pan, C., Chen, J., & Huang, R. (2024). Medical image detection and classification of renal incidentalomas based on YOLOv4+ ASFF swin transformer. Journal of Radiation Research and Applied Sciences, 17(2), 100845.
44. Kumar, Y., Huang, K., Gordon, Z., Castro, L., Okumu, E., Morreale, P., & Li, J. J. (2024). Transformers and LLMs as the New Benchmark in Early Cancer Detection. In ITM Web of Conferences (Vol. 60, p. 00004). EDP Sciences.
45. Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, Gao Huang (2022) Vision Transformer with Deformable Attention, https://doi.org/10.48550/arXiv.2201.00520.
46. Tellez, N., Serra, J., Kumar, Y., Li, J.J., Morreale, P. (2023). Gauging Biases in Various Deep Learning AI Models. In: Arai, K. (eds) Intelligent Systems and Applications. IntelliSys 2022. Lecture Notes in Networks and Systems, vol 544. Springer, Cham. https://doi.org/10.1007/978-3-031-16075-2_11.
47. J. Delgado, U. Ebreso, Y. Kumar, J. J. Li, and P. Morreale, "Preliminary Results of Applying Transformers to Geoscience and Earth Science Data," 2022 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2022, pp. 284-288, doi: 10.1109/CSCI58124.2022.00054.

 doi:10.20944/preprints202402.1506.v1

48. Hastie, Trevor. Tibshirani, Robert. Friedman, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, NY, 2009.
49. Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization." In Proceedings of the IEEE international conference on computer vision, pp. 618-626. 2017.
50. TeachableMachines web tool page. Available online: https://teachablemachine.withgoogle.com/models/TY21XA7_Q/ (accessed on 2/24/2024).
51. Home page of the NAD Youth. Available online: https://youth.nad.org/ (accessed on 2/24/2024).
52. Home page of the NAD. Available online: https://www.nad.org/resources/american-sign-language/learning-american-sign-language/ (accessed on 2/24/2024).