4-2024

TRANSFORMER-BASED DEEP LEARNING MODEL FOR SIGN LANGUAGE RECOGNITION

Ganzorig Batnasan

UAEU
جامعة الإمارات العربية المتحدة
United Arab Emirates University

MASTER THESIS NO. 2024: 62

College of Engineering

Department of Electrical and Communication Engineering

TRANSFORMER-BASED DEEP LEARNING MODEL FOR SIGN LANGUAGE RECOGNITION

Ganzorig Batnasan

April 2024

United Arab Emirates University

College of Engineering

Department of Electrical and Communication Engineering

TRANSFORMER-BASED DEEP LEARNING MODEL FOR SIGN LANGUAGE RECOGNITION

Ganzorig Batnasan

This thesis is submitted in partial fulfilment of the requirements for the degree of Master of Science in Electrical Engineering

April 2024

United Arab Emirates University Master Thesis
2024: 62

Cover: Transformer-Based Deep Learning for Sign Language Recognition

(Photo: By Ganzorig Batnasan)

Print: University Print Service, UAEU 2024

Declaration of Original Work

I, Ganzorig Batnasan, the undersigned, a graduate student at the United Arab Emirates University (UAEU), and the author of this thesis entitled "*Transformer-Based Deep Learning Model for Sign Language Recognition*" hereby solemnly declare that this is the original research work done by me under the supervision of Dr. Qurban Ali Andal Memon, in the College of Engineering at UAEU. This work has not previously formed the basis for the award of any academic degree, diploma, or similar title at this university or any other university. All materials borrowed from other sources (whether published or unpublished) and relied upon or included in my thesis have properly been cited and acknowledged in accordance with appropriate academic conventions. I further declare that there is no potential conflict of interest with respect to the research, data collection, authorship, presentation, and/or publication of this thesis.

Student's Signature: 

Date: May 18, 2024

Advisory Committee

1) Advisor: Dr. Qurban Ali Andal

Title: Associate Professor

Department of Electrical and Communication Engineering

College of Engineering

2) Co-advisor: Dr. Munkhjargal Gochoo

Title: Assistant Professor

Department of Computer Science and Software Engineering

College of Information Technology

Approval of the Master Thesis

This Master Thesis is approved by the following Examining Committee Members:

1) Advisor (Committee Chair): Dr. Qurban Ali Andal
Title: Associate Professor
Department of Electrical and Communication Engineering
College of Electrical Engineering, UAEU

Signature:  Date: 05-05-2024

2) Member: Dr. Munkhjargal Gochoo
Title: Assistant Professor
Department of Computer Science and Software Engineering
College of Information Technology, UAEU

Signature: Munkhjargal Gochoo Date: 05-05-2024

3) Member (Internal Examiner): Mahmoud Al Ahmed
Title: Professor
Department of Electrical and Communication Engineering
College of Engineering, UAEU

Signature:  Date: 05-05-202

4) Member (External Examiner): Muhammad Tahir Akhtar
Title: Associate Professor
Department of Electrical and Computer Engineering
Institution: Nazarbayev University, Kazakhstan

Signature: Muhammad Akhtar Date: 02-05-2024

This Master Thesis is accepted by:

Dean of the College of Engineering: Professor Mohammed Al-Marzouqi

Signature: 

Date: July 31, 2024

Dean of the College of Graduate Studies: Professor Ali Al-Marzouqi

Signature: 

Date: August 1, 2024

Abstract

Sign language recognition research aims to develop systems and tools that can interpret and translate sign language into text or spoken language. During the past two decades, the challenges faced in this domain are multifaceted. The first and foremost challenge is the complexity of sign language, which includes intricate hand gestures, facial expressions, and body movements. Recognizing and interpreting these components accurately is challenging. The second challenge is variability among different regions and communities, leading to variations in signs and gestures. This variability poses a challenge for developing universal recognition systems.

Limited data is another challenge which makes it difficult to train accurate recognition models. This scarcity of data hinders the performance and generalization of machine learning algorithms. Sign language communication often occurs in real-time, requiring recognition systems to process gestures quickly and accurately. Achieving real-time performance adds complexity to the design of recognition systems.

Some signs may have multiple meanings depending on context or subtle differences in execution. Disambiguating these signs accurately is crucial for reliable recognition. Furthermore, sign languages incorporate non-manual components such as facial expressions and body posture, which convey important linguistic information. Integrating these components into recognition systems poses additional challenges.

Sign language recognition systems may perform differently for different users based on factors such as signing speed, style, and proficiency. Developing systems that can adapt to individual users' signing characteristics is challenging. Additionally, deploying sign language recognition systems on hardware platforms with limited computational resources, such as mobile devices, presents challenges in achieving high performance while maintaining low latency.

This thesis on sign language recognition aims to address some challenges through various approaches. The foremost challenge addressed in this thesis is reduction in accuracy that uses transformer-based deep learning architecture in addition to preprocessing steps that include augmentations and transformations. The augmentations

and transformation helped increase the data size. Specifically, in-house signs have been generated using different persons for initial results. The video frames generated included facial expressions and both fingers, which were later stacked. Later, the model was validated using generic sign languages to address. For producing results, the model was trained and assessed on a set of frames. The comparisons with existing works are tabulated. Based on comparative results, it was found out that the accuracy of the proposed model assessed on WLASL2000, and ASL-Citizen datasets is higher than the state-of-the art models.

Keywords: Sign language, Sign language translation, Sign language recognition, Deep learning, Vision transformer.

Title and Abstract (in Arabic)

نموذج التعلم العميق القائم على المحولات للتعرف على لغة الإشارة

الملخص

تهدف أبحاث التعرف على لغة الإشارة إلى تطوير الأنظمة والأدوات التي يمكنها تفسير لغة الإشارة وترجمتها إلى نص أو لغة منطوقة. خلال العقدين الماضيين، كانت التحديات التي تمت مواجهتها في هذا المجال متعددة الأوجه. التحدي الأول والأهم هو تعقيد لغة الإشارة، والتي تتضمن إيماءات اليد المعقدة، وتعبيرات الوجه، وحركات الجسم. يعد التعرف على هذه المكونات وتفسيرها بدقة أمرًا صعبًا. التحدي الثاني هو التباين بين المناطق والمجتمعات المختلفة، مما يؤدي إلى اختلافات في الإشارات والإيماءات. يشكل هذا التباين تحديًا أمام تطوير أنظمة الاعتراف العالمية.

وتشكل البيانات المحدودة تحديًا آخر يجعل من الصعب تدريب نماذج التعرف الدقيقة. تعيق ندرة البيانات أداء وتعميم خوارزميات التعلم الآلي. غالبًا ما يحدث التواصل بلغة الإشارة في الوقت الفعلي، مما يتطلب أنظمة التعرف على معالجة الإيماءات بسرعة ودقة. يضيف تحقيق الأداء في الوقت الفعلي تعقيدًا إلى تصميم أنظمة التعرف.

قد يكون لبعض العلامات معاني متعددة اعتمادًا على السياق أو الاختلافات الدقيقة في التنفيذ. إن توضيح هذه العلامات بدقة أمر بالغ الأهمية للتعرف عليها بشكل موثوق. علاوة على ذلك، تتضمن لغات الإشارة مكونات غير يدوية مثل تعابير الوجه ووضعية الجسم، والتي تنقل معلومات لغوية مهمة. ويطرح دمج هذه المكونات في أنظمة التعرف تحديات إضافية.

قد تختلف أنظمة التعرف على لغة الإشارة باختلاف المستخدمين بناءً على عوامل مثل سرعة التوقيع والأسلوب والكفاءة. يعد تطوير الأنظمة التي يمكنها التكيف مع خصائص التوقيع الخاصة بالمستخدمين الفرديين أمرًا صعبًا. بالإضافة إلى ذلك، فإن نشر أنظمة التعرف على لغة الإشارة على منصات الأجهزة ذات الموارد الحسابية المحدودة، مثل الأجهزة المحمولة، يمثل تحديات في تحقيق الأداء العالي مع الحفاظ على زمن الوصول المنخفض.

تهدف هذه الأطروحة حول التعرف على لغة الإشارة إلى معالجة بعض التحديات من خلال أساليب مختلفة. التحدي الرئيسي الذي تتناوله هذه الأطروحة هو تقليل الدقة التي تستخدم بنية التعلم العميق القائمة على المحولات بالإضافة إلى خطوات المعالجة المسبقة التي تشمل التعزيزات والتحويلات. ساعدت التعزيزات والتحويلات على زيادة حجم البيانات. وعلى وجه التحديد، تم إنشاء علامات داخلية باستخدام أشخاص مختلفين للحصول على النتائج الأولية. تضمنت إطارات الفيديو التي تم إنشاؤها تعبيرات الوجه وكلا الأصابع، والتي تم تجميعها لاحقًا. وفي وقت لاحق، تم التحقق من صحة النموذج باستخدام لغات الإشارة العامة للتخاطب. للحصول على النتائج، تم تدريب النموذج واختباره على مجموعة من الإطارات. يتم جدولة المقارنات مع الأعمال الموجودة. وبناء على نتائج المقارنة، تبين أن دقة النموذج المقترح الذي تم اختباره على مجموعات بيانات WLASL2000 و ASL-Citizenأعلى من النماذج الحديثة.

مفاهيم البحث الرئيسية: لغة الإشارة، ترجمة لغة الإشارة، التعرف على لغة الإشارة، التعلم العميق، محول الرؤية

Acknowledgments

I extend my sincerest appreciation to Dr. Qurban Ali Andal Memon, who provided invaluable guidance and support throughout my master's thesis journey. I am also profoundly grateful to my parents for their unwavering presence and encouragement, and to my friends, who have been my pillar of strength and inspiration. Furthermore, I am thankful to my colleagues Munkh-Erdene Otgonbold and Erkhembayar Ganbold for their constant encouragement and support during this study.

Dedication

To my family

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ArSL	Arabic Sign Language
ASL	American Sign Language
BCI	Brain-Computer Interface
CNN	Convolutional Neural Network
CSL	Chinese Sign Language
CSLR	Continuous Sign Language Recognition
DBN	Dep Belief Network
GCN	Graph Convolutional Network
ISLR	Isolated Sign Language Recognition
KSL	Korean Sign Language
KSU-SSL	King Saud University Saudi Sign Language
LSTM	Long Short-Term Memory
NLP	Natural Language Processing
PSL	Pakistan Sign Language
ResNet	Residual Network
RNN	Recurrent Neural Network
SDCS	Saudi Deaf Companion System
SLP	Sign Language Processing
SLR	Sign Language Recognition
SLT	Sign Language Translation
STMC	Spatial-temporal Multi-cue
SVM	Support Vector Machine
TCN	Temporal Convolutional Network

Chapter 1: Introduction

1.1 Introduction to Sign Language

Sign Language serves as a crucial mode of communication primarily for individuals who are hard of hearing or deaf. This gesture-based language facilitates the expression of ideas and thoughts, overcoming barriers associated with hearing impairments. Historically, deaf communities have advocated for the recognition of signed languages as legitimate forms of communication. Despite this, signed languages have often struggled to gain acceptance, facing challenges such as misconceptions about their impact on speech development. The failure to acknowledge signed languages as fully-fledged natural language systems has had adverse consequences in the past. In today's increasingly digital world, it is imperative for natural language processing research to work towards ensuring that all individuals, including those who are deaf, have access to languages that resonate with their lived experiences.

Sign languages are complex visual forms of communication, employing manual gestures along with facial expressions and body motions to express ideas and concepts. They are crucial for communication among many deaf and hard-of-hearing individuals. Sign languages, like spoken languages, are structured by linguistic rules and evolve naturally over time. Despite similarities, they are not universally understood, and each has its own unique structure. For example, American Sign Language (ASL) is not simply a visual version of English but an independent language with distinct characteristics. Expertise in sign language linguistics is essential for incorporating its unique properties into model development.

Signed languages are highly sophisticated forms of communication, equivalent to spoken languages in linguistic and social capabilities. However, in societies that prioritize oral communication, deaf individuals are often encouraged to use spoken languages through methods like lip-reading or written text. The lack of inclusion of signed languages in modern language technologies exacerbates this preference for spoken languages, disregarding the strong preference of deaf communities to communicate using sign languages, both online and in-person. Making signed languages accessible is crucial to ensure inclusivity and accommodate the communication preferences of deaf individuals.

Deaf individuals have a strong preference for using sign languages, which are highly structured and naturally evolved communication systems governed by linguistic rules. However, most communication technologies are designed for spoken or written languages, excluding 300 sign languages used around the world [1]. This creates significant communication barriers for the 70 million deaf people worldwide who rely on sign languages. Recognizing, generating, and translating sign languages is an area of research with high potential impact in addressing these barriers and making communication more accessible for deaf communities.

The lack of widespread knowledge of sign languages poses a significant barrier to communication for the majority of the global population. Learning sign language requires considerable time and effort, deterring many from acquiring it. Sign language processing technologies have the potential to alleviate these barriers by making voice-activated services accessible to deaf sign language users, such as training personal assistants to respond to sign language. Additionally, these technologies could facilitate the adoption of text-based systems by converting signed content into written queries or automatically substituting displayed text with sign language videos. Other potential applications include automatic transcription of signed content for indexing and search, real-time interpreting, and various educational tools and applications.

Sign Language Processing (SLP) is an emerging field of artificial intelligence that deals with automatically processing and analyzing sign language content. It is a subfield of both Natural Language Processing (NLP) and Computer Vision (CV), although studies have primarily concentrated on the visual components of signed languages. Challenges to sign language processing encompass the translation of sign language video sequences into spoken language text, commonly referred to as sign language translation, translating spoken language text into sign language (sign language production), and sign language recognition for understanding sign language. Utilizing machine learning and image detection technologies presents a promising solution to the communication barriers faced by individuals using sign language. Implementing predictive models to automatically classify sign language symbols can enable real-time captioning for virtual conferences, such as Zoom meetings. This approach, when combined with voice-based captioning, can

create a two-way communication system online, significantly enhancing accessibility for people with hearing impairments.

Much of the research in SLP has predominantly focused on the visual aspects of signed languages, led by CV community, with limited involvement from NLP. This emphasis was logical given the historical lack of adequate CV tools for linguistic analyses of videos. However, signed languages, like spoken languages, are comprehensive linguistic systems with unique characteristics. Current SLP techniques inadequately manage the linguistic structure inherent in signed languages, which presents unique challenges for NLP because signed languages operate within a visual-gestural modality, involve simultaneity, rely on spatial coherence, and lack a written form. This absence of a written form makes traditional spoken language processing pipelines incompatible with signed languages, necessitating direct work on raw video signals for researchers.

SLP is intellectually engaging and holds significant potential to serve signing communities through various applications, including documenting endangered sign languages, educational tools, query and retrieval systems for signed language videos, responsive personal assistants, and real-time automatic interpretations. Collaboration with and guidance from deaf communities is essential, prioritizing their interests above all. Currently, most efforts to address this issue have focused on two main approaches: contact-based systems like sensor gloves, and vision-based systems using cameras. Vision-based systems, being more cost-effective and benefiting from advancements in deep learning, are gaining popularity. The sign language recognition (SLR) research have diversified due to different directions of sign language recognition and translational methods such as:

- Isolated sign language recognition (ISLR): This entails identifying individual signs or glosses within a specified segment of sign language video, often treated as a classification problem. Real-time applications require additional tasks like video segmentation.
- Continuous sign language recognition (CSLR): CSLR or sign language transcription, aims to predict all signs in a video sequence, suitable for real-world transcription. Depending on the approach, CSLR can extend ISLR.

- Continuous sign language translation: This entails translating a sequence of signs to a spoken language, typically an extension of CSLR.

Kaggle, a prominent platform for data science, offers numerous large training datasets for Sign Language [2]. One such dataset, "Sign Language MNIST" contains around 1,000 images representing each ASL letter. This dataset is public domain and free to use, providing pixel information for the sign language images, with the exclusion of J and Z as they are gesture-based signs.

1.2 Literature Review

Sign languages are essential for communication among the deaf and hard of hearing, employing visual-manual techniques. Sign language recognition comprises two main domains: isolated single sign recognition, which identifies individual signs conveying a single concept, and continuous sign recognition, which deals with recognizing flowing sequences of signs used in sentences or conveying multiple concepts. Recent progress in computer vision and its application such as [3-7] computer graphics, machine learning applications such as [8-12], and hardware has fueled research in sign languages, facilitating advancements in learning, communication, interpretation, translation, visualization, documentation, and skill development in this domain. The computer vision field has been dedicated to studying sign languages for the past 30 years, as evidenced by numerous studies [13,14]. The overarching goal of computational research in this area is to develop systems capable of translating and producing sign language [15]. These systems aim to convert sign language videos into spoken language sentences and vice versa, thereby enhancing the daily experiences of the Deaf community [16,17]. Most of the research conducted so far has focused on recognizing isolated signs and has utilized specific datasets tailored to particular applications, which limits the practical applicability of these technologies [18-20]. While recent studies have begun to address continuous sign language data [21-24], the transition from recognition to translation is still in its nascent stages [25].

Previous attempts at achieving sign language translation (SLT) were led by computational linguists. However, these studies focused solely on the text-to-text translation aspect and had limited scope, typically encompassing around 3000 total words

on average [26-28]. The notable sign language research review works are mentioned below.

The literature review conducted in [29] explores the use of wearable sensor devices systems to classify sign language gestures, aiming to address communication barriers faced by Deaf, deafened, hard of hearing, and non-verbal individuals. This review analyzes 72 research studies from 1991 to 2019 and identifies best practices, trends, and challenges in gesture recognition. The analysis and comparison entail exploring key aspects like variations in sign language, sensor setups, classification techniques, research methodologies, and evaluation criteria. The findings from this literature review offer insights that could form the creation of resilient and user-centered wearable sensor devices for sign language recognition, ultimately facilitating better communication for individuals relying on sign language.

The research paper [30] addresses the global need for effective local-level SLR tools, given the widespread presence of people with hearing impairments. Conducting a thorough review of machine learning methods and approaches used in SLR between 2014 and 2021, the study concludes that existing methods often need conceptual classification to interpret all data accurately. As a result, the paper shifts focus to common elements found in most SLR methodologies, discussing their respective strengths and weaknesses while proposing a general framework for researchers in the field. This study highlights the significance of input modalities in SLR, noting that the recognition based on combining sensor-based and vision data sources outperforms unimodal analysis. Recent advancements have enabled researchers to progress from merely recognizing sign language words and characters to translating continuous sign language communication with the least delay. While many models exhibit effectiveness across various tasks, none possesses the requisite generalization potential for commercial use at present. Nonetheless, the rapid advancement of research in the field is promising, and continued progress is anticipated with the resolution of specific challenge.

The importance of technology in improving the daily life for individuals with hearing impairment, who rely on sign language is also discussed in [31]. It aims to examine and assess articles involving sign language recognition experiments using sensor-based glove

systems to understand academic motivations, challenges, and recommendations in this field. The study searched major databases from 2017 to 2022 and identified key issues such as dataset size limitations and challenges in hand gesture recognition. Most research focused on static, single-handed, and isolated gestures, achieving recognition accuracy above 90%, but often with a limited number of gestures. The findings provide insights for future research and aim to raise awareness among researchers about sign language recognition.

The research paper [32] reviews technological progress in sign language recognition, synthesis, and visualization, aiming to address respective challenges. Employing the preferred reporting items for systematic reviews and meta-analysis (PRISMA) methodology, the study formulates research questions and conducts a systematic search for relevant articles published between 2010 and 2021 across various digital libraries. Utilizing Natural Language Processing tools streamlines initial steps, including duplicate removal and basic screening. Synthesizing existing knowledge, the study identifies notable progress in sign language recognition, synthesis, and visualization. Analysis of nearly 2000 papers reveals trends indicating that advancements in image processing and deep learning are propelling the development of new applications and tools, enhancing performance parameters in sign language recognition. The research identifies techniques, tools, and devices contributing to these advancements and identifies commonalities and gaps that could inspire novel avenues of research in the domain.

There have been numerous contributions, which discuss technological advancements to SLR. The most notable are discussed below.

The authors in [33] addresses the formidable task of CSLR in intelligent systems, which prompt responses in real-time alongside computationally demanding video analytics and sign language modeling. To address such challenges, the authors introduce SignBERT, a deep learning framework, combines Residual Neural Networks (ResNet) with Bidirectional Representations from Transformers (BERT). Its goal is to proficiently capture the structure and then extract spatial characteristics. Additionally, they suggest a multimodal adaptation of SignBERT that integrates hand images to improve feature alignment and minimize the discrepancy between the BERT model results and the hand

images. Simulation results demonstrate the efficacy of SignBERT compared to alternative CSLR approaches. SignBERT achieves higher accuracy and reduced word error rates across three continuous sign language datasets, highlighting its superiority in addressing the complexities of continuous sign language recognition tasks.

The research work [34] paper presents a network designed for isolated sign language recognition, addressing the need for effective communication among individuals with hearing impairments. The proposed network comprises three modules: the dynamic motion network, accumulative motion network, and sign recognition network. Key postures are extracted to handle variations in sign samples from different signers, aiding the DMN in learning spatiotemporal information. A novel technique fuses static and dynamic information into a single frame, enhancing feature representation. Evaluation on Arabic sign language datasets KArSL-190 and KArSL-502 shows a 15% improvement over other techniques in signer-independent mode. Furthermore, the proposed method outperforms well-known techniques on the LSA64 Argentinian sign language dataset, particularly excelling in recognizing static signs.

The considerable progress achieved by Transformer models in SLR and SLT owing to their ability to learn long-term dependencies in also discussed in [35]. However, shortcomings in the Transformer architecture hinder its effectiveness in sign language understanding. Specifically, the self-attention mechanism employed in transformers is tasked with learning sign video representation frame by frame, disregarding the temporal semantic structure of sign gestures. Additionally, the attention mechanism, coupled with absolute position encoding, lacks awareness of direction and distance, thus restricting its effectiveness. To tackle these issues, the paper introduces a novel approach featuring two key innovative convolution layers: content-aware and position-aware. These layers employ a unique content-aware neighborhood gathering approach to select pertinent features and aggregate them with position-informed temporal convolution layers to generate neighborhood-enhanced sign representations. Furthermore, the paper proposes the integration of relative position information into the attention mechanism, which is implemented within the encoder, decoder, and encoder-decoder cross attention modules, enhancing the understanding of sign language sequences by the model. The experimental findings reveal that the model consistently surpasses the vanilla transformer model across

extensive sign language benchmarks: PHOENIX-2014-T, PHOENIX-2014, and CSL. Moreover, comprehensive experimentation indicates that the proposed approach reaches state-of-the-art performance, with a notable improvement of +1.6 BLEU score.

The research contribution [36] introduces novel approaches for real-time handling of sign language recognition, translation, and production tasks. It utilizes the MediaPipe library and a hybrid model (CNN + Bi-LSTM) for extraction of pose details and generation of text to enhance accuracy. For sign gesture video production, it employs a hybrid model (NMT + MediaPipe + Dynamic GAN). The proposed model achieves over 95% classification accuracy and is evaluated across various phases, demonstrating significant improvements. Experimentation with multilingual benchmark sign corpora shows superior results in recognition accuracy and visual quality. Evaluation metrics include an improved average BLEU score, notable human evaluation scores, and other performance indicators such as FID, SSIM, Inception Score, PSNR, FID2vid, and TCM score, showing its effectiveness.

The paper [37] addresses the limitation of deep learning models in sign language understanding by proposing a spatial-temporal multi-cue (STMC) network that leverages multiple visual cues, such as hand shape, facial expression, and body posture, to learn implicit visual grammars in sign videos. The STMC network comprises two main modules: spatial multi-cue module that learns spatial representations of different cues using a pose estimation branch; and temporal multi-cue module, which models temporal corrections from intra-cue and inter-cue perspectives to explore the collaboration. Additionally, the paper introduces an optimization strategy and a segmented attention mechanism to effectively utilize multi-cue sources for sign language recognition and translation. Experimental validation on extensive sign language benchmarks such as PHOENIX-2014-T, PHOENIX-2014, and CSL, demonstrates that the employed approach reaches improved performance on all benchmarks, highlighting its effectiveness in video-based sign language understanding.

The research contribution described in [38] advocates for the transformer encoder's utility in sign language recognition. In addressing the recognition of static Indian signs, the authors have developed a vision transformer. This proposed methodology

demonstrates superior performance compared to other state-of-the-art convolution architectures in recognizing static Indian sign language. The approach involves segmenting the sign into positional embedding patches, which are then processed through a transformer block comprising four self-attention layers and a multilayer perceptron network. Experimental findings exhibit satisfactory gesture identification across various augmentation techniques. Additionally, the method achieves a remarkable 99.29 percent accuracy with minimal training epochs.

The paper [39] addresses the challenge of multilingual continuous sign language recognition, which is typically approached by building separate models for each language. However, the authors observe that different sign languages share common low-level visual patterns, suggesting that collaborative optimization could be beneficial. With this insight, the authors propose a unified framework for multilingual CSLR. The framework comprises three main components: a shared encoder for encoding visual information; multiple language-dependent modules designed to learn temporal dependencies specific to different languages; and a universal sequential module aimed at capturing commonalities across all languages. Additionally, the framework incorporates a language embedding mechanism to differentiate between languages within the shared temporal encoders. The authors also introduce a max-probability decoding approach to refine the alignment between sign videos and sign words for encoder improvement. The proposed approach is evaluated on continuous sign language recognition benchmarks: RWTH-PHOENIX-Weather, GSL-SD, and CSL. The simulation results demonstrate that the proposed approach outperforms individually trained models and obtains superior performance compared to well-known algorithms.

The paper [40] addresses the communication barrier faced by the speech and hearing-impaired community by proposing a hybrid vision-based deep neural network approach for recognizing Indian and Russian sign gestures. Existing sign language recognition systems often depend on expensive wearable sensors, limiting accessibility. Moreover, existing vision-based frameworks lack consideration of all spatial and temporal information crucial for accurate recognition. The proposed framework aims to overcome these challenges by developing a single framework capable of extracting and tracking multi-semantic properties, including non-manual components and manual co-

articulations. It employs a combination of techniques: spatial feature extraction that utilizes a 3D neural network with convolutions; sequential and temporal feature extraction that leverages attention-based Bidirectional Long Short-Term Memory (Bi-LSTM) networks; and abstract feature extraction that uses modified autoencoders to capture distinguished abstract features; and discriminative feature extraction that employs a hybrid attention module to differentiate sign gestures from unwanted transition gestures. Experimental validation conducted on a multi-signer Indo-Russian sign language dataset demonstrates the effectiveness of the model. The hybrid neural network-based sign language recognition framework outperforms existing state-of-the-art frameworks, indicating the possibility of improving the accessibility and communication for the hearing and speech-impaired community.

The paper [41] presents a spatial-temporal feature extraction network (STFE-Net) designed to address the challenge of (CSLT. This network optimally aggregates temporal and spatial features extracted by two sub-networks: the temporal feature extracting network (TFE-Net) and the spatial feature extracting network (SFE-Net). The SFE-Net conducts pose estimation for presenters appearing in videos, reducing the number of key points from 133 to 53 by leveraging characteristics specific to sign language. It employs high-resolution pose estimation on the hands to capture more detailed features. The features that are captured by SFE-Net, along with sign language words, are inputted into TFE-Net. This net utilizes transformer mechanism with position encoding for temporal feature extraction. The authors created a dataset for Chinese continuous sign language for performance assessment. STFE-Net obtains promising BLEU scores: 77.59 (BLEU-1), 75.62 (BLEU-2), 74.25 (BLEU-3), and 72.14 (BLEU-4). Moreover, its performance was assessed using public datasets: CLS and RWTH-Phoenix-Weather 2014T, achieving competitive BLEU scores on both. Overall, the simulation results demonstrate the promising performance of the STFE-Net in continuous sign language translation, showcasing its effectiveness in extracting discriminative spatial and temporal features for accurate translation.

The paper [42] introduces a novel approach called the mutual enhancement network (MEN) aimed at improving sign language recognition. The primary challenge addressed is the limited scale of training data in existing sign language recognition methods. The

proposed MEN comprises two main components: a sign language recognition system and a sign language education system. The sign language recognition system utilizes a spatial-temporal network to identify the semantic category of sign language videos. The sign language education system is designed to detect learner failure modes and provide guidance to help them sign correctly. A key contribution of the paper is formulating these two systems within an estimation-maximization (EM) framework, allowing them to iteratively enhance each other. The recognition system benefits from accurate training data collected by the education system, while the education system benefits from the hand shape analysis module of the recognition system, enabling more precise guidance to learners. The simulation findings on extensive sign language recognition datasets demonstrate the effectiveness and superior performance of the MEN framework.

The research survey [43] addresses the pressing need to integrate deaf-mute individuals into mainstream society in China by leveraging efficient sign language processing technologies. Sign language processing involves recognizing and translating sign language images and videos into text or speech systematically. The survey offers an overview of key research on Chinese sign language recognition, translation, classification, which includes explored features, available datasets, and future research trends.

The paper [44] introduces the application of deep belief networks (DBN) in wearable sensor device based Chinese Sign Language (CSL) recognition. Eight subjects participated in an experiment where they performed CSL recognition using a target word set consisting of 150 CSL sub words. Surface Electromyography (sEMG), accelerometer, and gyroscope signals were gathered during the experiment. The study explored three sensor fusion strategies: feature-level fusion, data-level fusion, and decision-level fusion. For feature-level fusion, two types of feature sources (hand-crafted and network-generated features), and two network structures (fully-connected net and DBN) were compared. The results indicate that feature-level fusion achieves the highest recognition accuracy among other fusion strategies, with the combination of network-generated features and DBN yielding the best performance. The study achieves a recognition accuracy of about 95% for user-dependent testing and about 88% for user-independent testing. This study is significant as it applies deep learning methods to wearable sensor-based CSL recognition, marking one of the first attempts to compare human-engineered features with network-

generated features in this field. The findings provide valuable insights into the use of network-generated features during sensor fusion and CSL recognition.

The paper [45] delves into limb motion decoding within the realm of Brain-Computer Interface (BCI) research, focusing specifically on the decoding of CSL from electroencephalograph signals during motor imagery and motor execution tasks. While previous studies often concentrate on decoding motor skills, such as ordinary motor imagery or simple upper limb movements, sign language presents a unique opportunity due to its rich semantic information and diverse executable commands. Twenty subjects were tasked with performing movement related execution and imagery based on Chinese sign language. Various classifiers were employed to classify EEG features, which included mean, sample entropy, power spectral density, and brain network connectivity, learned and selected through L1 regularization. The best average classification accuracy achieved was 89.90% for movement execution (83.40% for movement imagery), indicating the feasibility of decoding between different sign languages. Overall, the decoding strategy based on sign language yielded excellent classification results, offering valuable insights for future research on limb decoding utilizing sign language as a paradigm.

This paper [46] introduces an Arabic Sign Language (ArSL) recognition system utilizing 2D hands and key points of the body from video frames. The developed system is capable of recognizing recorded signs in signer-dependent and signer-independent modes by combining a 2D point convolution network with a 3D CNN skeleton network. To support this system, the authors developed a new ArSL video-based sign database containing 80 signs that are repeated by 40 persons. The signs encompass numbers, alphabet, and commonly used signs. To optimize the system for near real-time operation, the authors introduce the concept of efficiency score to determine the optimal number of successive frames required to make the decision for recognition. This tradeoff between speed and accuracy is important for avoiding delayed sign classification in an online sign recognition system. For signer-dependent mode, the system achieves an accuracy of about 98% for dynamic signs and over 88% for static signs. In signer-independent mode, the accuracy is over 96% for dynamic signs and about 86% for static signs. When the system

is trained with all signs (with static and dynamic signs mixed), accuracies about 88% and over 89% are achieved in signer-independent and signer-dependent modes, respectively.

The study [47] addresses the need for a cost-effective and user-friendly technique for Pakistan Sign Language (PSL) recognition, considering the reliance of deaf individuals on sign language for communication. While existing studies on PSL recognition have utilized colored-based hands or sensor-based approaches, these methods are often expensive and lack user-friendliness. In this study, the authors propose a technique for recognizing alphabets of PSL using only hands. They used sign language videos to build the dataset and then extract four vision-based features - local binary patterns, histograms (edge-oriented and gradient-oriented gradients) - to improve robustness. These features are classified using Support Vector Machines (SVM) with multiple kernels. A one-to-all scheme is employed to implement binary SVM into a multi-class classifier, and a voting scheme is used for recognition. The performance is assessed using recall, precision, accuracy, and F1-score. The experimental findings demonstrate excellent performance compared to current approaches, indicating the potential of the proposed technique for cost-effective and user-friendly PSL recognition.

The paper [48] proposes a solution to improve the recognition of dynamic continuous gestures in Chinese sign language by developing a complete recognition system. This system combines signals from Inertial Measurement Units (IMUs) and surface electromyography (sEMG). Following preprocessing and fusion of the sEMG and IMU signals, the system extracts gesture features such as acceleration, attitude quaternion, angular velocity, and sEMG information. Next, a model is built based on the BiLSTM framework, with temporal classification serving as the loss metric. This approach removes the limitations of imprecise pre-segmentation, enabling a complete dynamic continuous gesture recognition. Finally, the Chinese Sign Language Dataset is established that contains ten (10) types of Chinese sign language with 20,000 samples. The training and testing on this dataset demonstrate an average success rate of over 98%, confirming the effectiveness and superior performance of the proposed approach in continuous dynamic gesture recognition.

The solution proposed in [49] involves utilizing a two-stream deep learning network to effectively recognize dynamic signs, particularly focusing on Korean Sign Language (KSL). The first stream processes pose landmarks using Graph Convolutional Network (GCN) to extract graph-based features, which are then refined with a channel attention module and a general CNN model to enhance temporal context. Simultaneously, the second stream focuses on joint motion-based features using a similar method. The distinct features from both streams are aggregated and passed through classification for precise sign-word recognition. A significant aspect of this work is the development of a novel KSL video dataset, which addresses the lack of comprehensive data in this domain. The dataset includes skeletal data from 47 joint skeleton points as well as details from hands, body, and facial expressions, aiming to facilitate further research in KSL recognition. The ultimate goal of this approach is to contribute to KSL recognition, improving accessibility for the Korean hearing impaired community. Evaluation results on benchmark datasets demonstrate high recognition accuracies, surpassing current models on the basis of computational efficiency and accuracy.

The authors in [50] introduces the inaugural large-scale and annotated Qatari sign language database tailored for sign language processing. Focused on commonly encountered sentences and phrases in healthcare settings, the dataset comprises 6,300 records comprising 900 sentences. The collection process incorporates a diverse range of participants, including hearing-impaired persons and sign interpreters, to extract variations in signing speeds, styles, and linguistic nuances. Employing advanced technology such as true depth cameras, the data collection setup records signing movements comprehensively from various angles. The resulting dataset is rich in content, encompassing diverse signing variations and linguistic intricacies. Available publicly on IEEE Dataport, the paper also conducts an analysis of the captured data to discern trends and patterns. Given the escalating universal population with hearing impairments, there exists a critical demand for robust sign language recognition systems to close the communication chasm between the non-deaf and deaf communities. The introduction of the JUMLA-QSL-22 database marks an important advancement toward addressing this urgent requirement.

This paper [51] addresses the communication challenges faced by disabled individuals, particularly the deaf, and aims to promote their integration into society

through a sign translation companion called the Saudi Deaf Companion System (SDCS). The SDCS comprises three modules: the sign recognition module, the speech recognition and synthesis module for converting non-deaf individuals' speech to text, and an Avatar module for generating corresponding sign language gestures for the non-deaf speech. The system utilizes the King Saud University Saudi-SSL (KSU-SSL) dataset to perform 293 Saudi signs across various domains recommended by the Saudi Association for Hearing Impairment. These domains include healthcare, alphabets, verbs, common phrases, pronouns and adverbs, days, numbers, kings, family, and regions.

By integrating these modules, the SDCS facilitates communication between hearing-impaired individuals and the rest of society, aiming to transition them from the margins to active contributors in mainstream society. This contribution to the literature not only addresses communication barriers but also emphasizes the importance of inclusivity and accessibility in modern smart healthcare systems.

Regarding application-oriented sign language recognition systems, few research works were also found, which are mentioned below.

The research conducted in [52] introduces a cutting-edge dynamic sign language recognition system tailored for smart home applications. This innovative system consists of two key subsystems: a stochastic linear formal grammar module, and an image processing module. The image processing module is designed to recognize sign language words, employing a bag-of-features approach and a local part model for gesture recognition from videos. The module employs dense sampling to capture multiscale features of entire objects and their components, depicted through 3D histograms of gradient orientation descriptors. Clustering of visual words is accomplished using the k-means++ method, and gesture classification is performed with bag-of-features and nonlinear support vector machine techniques. Temporal context preservation is ensured through the adoption of a multiscale local part model. On the other hand, the SLFG module focuses on analyzing sign language sentences, assessing their syntactic validity, and predicting missing gestures to enhance recognition accuracy. By combining both modules, the DSLR system achieves an impressive accuracy rate of 98.65%, with the IP module alone achieving 97%, surpassing existing dynamic gesture recognition systems. The

utilization of syntactic pattern recognition in the SLFG module further enhances accuracy by about 2%, showcasing the effectiveness of the integrated approach in dynamic sign language recognition.

Only a small fraction of individuals who can hear are familiar with and utilize Mexican Sign Language (MSL), leading to a significant communication barrier between those with complete or partial hearing loss and those without. Addressing this issue, a study [53] introduces a real-time system capable of recognizing and animating a range of signs pertinent to general medical consultations in MSL. This innovation facilitates seamless and dynamic interaction between a hearing doctor and a deaf patient, enhancing communication without intrusiveness. The primary contribution of this study lies in the development of a bidirectional translation system tailored for MSL within the realm of primary healthcare services. This system not only includes fundamental signs for finger-spelling the alphabet and numbers but also incorporates additional features for conveying personal information like names and ages. Sign recognition is achieved through a Microsoft Kinect sensor, which captures sign images and relevant trajectories, subsequently processed using hidden Markov Models for real-time analysis. Simulation results demonstrate the successful recognition of 82 distinct signs by 22 users, yielding impressive F1 score and average accuracy of 88% and 99% respectively.

The study [54] introduces DeepSLR, a real-time end-to-end Sign Language Recognition (SLR) system aimed at facilitating communication between hearing-impaired individuals and others. Existing SLR systems face challenges in providing continuous recognition and accurately capturing finger and arm motions. DeepSLR addresses these issues by utilizing two armbands equipped sensors to capture both finger and arm movements. It proposes an attention-based encoder-decoder model with a multi-channel CNN to achieve precise and continuous SLR. This model is implemented on a smartphone and achieves an average word error rate of about 11% for continuous recognition and recognizes a sentence with four sign words in less than 1.1 seconds, demonstrating its efficiency and real-time capability in real-world scenarios.

1.3 Research Objectives

The objectives of the thesis are multi-fold. Below, we state each one of them.

- The existing methods are specific to one sign language only. In this thesis, the aim is to propose a novel approach to sign language recognition that includes a set of steps to be operated on image frames before training of the model starts. The steps include capturing head and both hand fingers of each sign. The objective is to enhance the accuracy of sign language recognition.
- Enhancing the dataset with different orientations is likely to increase the model performance. A pre-processing step is added to the captured frames to generate a set of frames for each sign frame to increase training data size to help improve accuracy of the model.
- The model is targeted to be applied to different sign languages to evaluate its performance for robustness.

1.4 Outline of the Thesis

In Chapter 2, well-known models applied to sign language recognition are discussed along with proposed approach. The weaknesses or shortcomings in different models are also stated. Further, different datasets that are commonly found in literature along with important benchmark metrics typically used for sign language recognition model performance are outlined in this chapter. Chapter 3 illustrates results from different experiments conducted using proposed method. In Chapter 4, conclusions are made, followed future directions.

Chapter 2: Existing Models and Proposed Approach

Sign language models are machine learning or deep learning models tailored to recognize and interpret movements and gestures in sign language. Their purpose is to comprehend and transcribe sign language into either spoken language or text, facilitating communication between sign language users and those who do not utilize sign language. The developed models can be trained on in-house or public domain large datasets of annotated sign language signs and gestures to learn to recognize a range of signs and gestures precisely. Some of the well-known models for sign language recognition are discussed below.

2.1 Video Recognition

Video recognition models designed for sign language seek to analyze and decipher the gestures and movements depicted in sign language videos. These models commonly utilize deep learning frameworks adept at processing spatiotemporal data. Several notable video recognition models for sign language encompass three-dimensional (3D) CNNs, transformer-based architectures, Long Short-Term Memory (LSTM) networks, two-stream networks, among others. Each of these models is discussed in further detail below.

2.1.1 3D Convolutional Neural Networks

CNNs are commonly employed for recognition tasks, including sign language recognition. They can learn spatial hierarchies of features from sign language images, making them effective for detecting and classifying hand gestures. Three dimensional (3D) CNNs extend the concept of CNNs to three-dimensional data, allowing them to extract both spatial and temporal features directly from videos. They are well-suited for SLR tasks involving video input.

Three primary challenges are present, encompassing the representation of spatial (image) features, temporal information representation, and the complexity of the model/computation. Carreira and Zisserman [55] demonstrated in their work that 3D CNNs, derived from 2D networks and pre-trained on ImageNet, offer a potential solution for learning spatial and temporal representations. However, it is noteworthy that 3D CNNs are considerably more resource-intensive than 2D CNNs and are prone to overfitting due

to their complexity. The study builds on the I3D approach and introduces variant architectures, Bottom-Heavy-I3D and Top-Heavy-I3D, to assess the necessity of 3D convolutions across different layers. Through meticulous exploration and "network surgery," the authors [61] not only question the need for 3D convolution but also propose a novel S3D-G model that significantly enhances accuracy over baseline methods in various video classification datasets. The architecture flow of this work is diagrammed in Figure 1. An alternative method to alleviate computational load involves substituting 3D convolutions with separable convolutions, wherein spatial convolution is initially performed in 2D followed by temporal convolution in 1D.

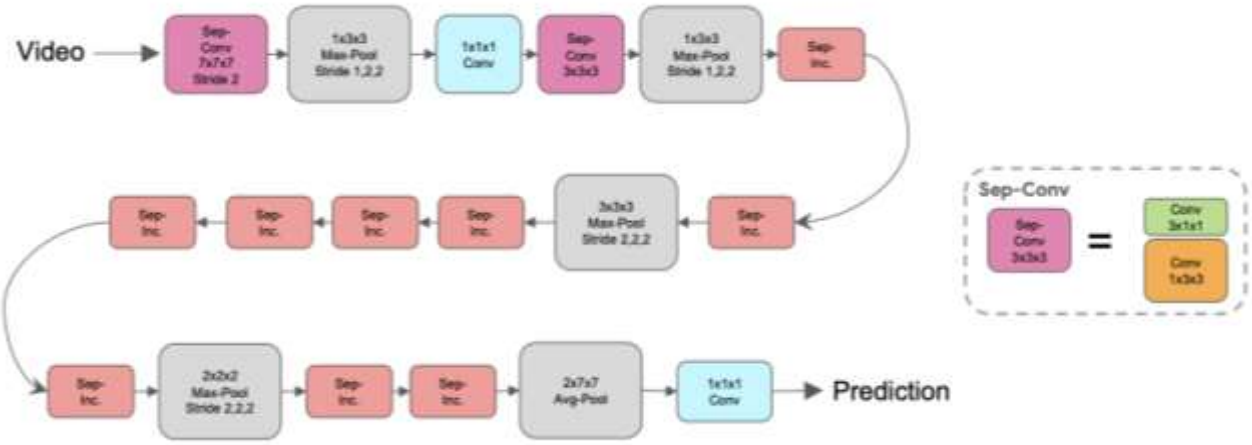

Figure 1: The architecture for learning spatial and temporal representations [71]

2.1.2 Long Short-Term Memory (LSTM) Networks

LSTM networks, belonging to the recurrent neural network (RNN) architecture, demonstrate notable efficacy in tasks involving sequence modeling, rendering them highly suitable for sign language recognition. Given that sign language recognition entails comprehending and deciphering sequences of hand gestures and movements over time, LSTM networks excel in capturing these temporal dependencies and long-range associations present within sequences. Each LSTM unit possesses an internal state capable of retaining information across extensive sequences, facilitating the preservation of context and the acquisition of patterns in sign language gestures. Furthermore, the memory cell within an LSTM unit exhibits the ability to selectively manage information flow via

gates such as the input gate, forget gate, and output gate. These gates regulate the influx, removal, and internal flow of information within the memory cell, enabling the LSTM to discern which information to retain and which to discard.

Sign language sequences often exhibit varying lengths, posing a challenge for conventional neural network architectures. However, LSTM networks demonstrate the ability to handle sequences of variable lengths by dynamically adjusting their internal state in response to the length of the input sequence. This inherent adaptability renders LSTMs well-fitted for the task of recognizing sign language movement and gestures in videos of differing durations. Moreover, LSTM networks can effectively extract features from raw video data. By sequentially processing frames of video input, LSTMs can learn to capture meaningful spatial and temporal features that encapsulate the subtleties of sign language gestures. These extracted features can then be propagated into subsequent layers of the network for classification or further analysis. Following training, LSTM networks are capable of categorizing sign language gestures into predefined classes. The network's final output can undergo processing through a softmax layer to compute the probability distribution across various sign language classes, thereby enabling the network to predict the most probable sign language gesture based on the input sequence.

In summary, LSTM networks provide a robust framework for sign language recognition by adeptly modeling temporal dependencies in sequential data, accommodating sequences of variable lengths, and extracting significant features from raw video input. Their capacity to capture extensive dependencies renders them especially suitable for recognizing intricate sign language gestures evolving over time.

2.1.3 Transformers-based Architectures

Transformers brought about a paradigm shift in sequence modeling by abolishing sequential processing, introducing self-attention mechanisms to grasp global dependencies, and efficiently parallelizing computations. In the realm of sign language recognition, transformer-based networks present distinctive benefits. They demonstrate exceptional proficiency in capturing complex temporal dependencies inherent in sign language sequences, which are vital for comprehending the subtle motions and gestures over time.

Through the utilization of self-attention mechanisms, transformers are able to dynamically assess the importance of various segments within the input sequence, effectively capturing both short and long-range dependencies inherent in sign language gestures. The self-attention mechanism stands as a cornerstone of the transformer architecture, facilitating each element in the sequence to attend to every other element, thereby capturing intricate relationships and dependencies irrespective of their relative positions. Essentially, self-attention computes a sum of all input elements, with the weights being calculated by the similarity between the current element and every other element in the sequence. This similarity, termed as attention score, is computed using learned parameters – Query-Q, Key-K, and Value-V – associated with each input element via dot-product attention. This involves taking the dot product of query, key, and value vectors linked with each input element. Subsequently, the attention scores undergo softmax transformation to acquire a probability distribution across all elements in the sequence, representing the significance or weight assigned to each element in relation to the current element. Finally, the sum of the weighted value vectors is computed, representing the attended representation of the input sequence, effectively encapsulating the dependencies and connections between elements. Mathematically, this process is represented as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{sqrt(d_k))}) * V \quad (2.1)$$

where d represents the dimension of scaling the dot products. Moreover, transformers have the capacity to acquire hierarchical representations of sign language gestures, encompassing both intricate details and broader semantic concepts. This hierarchical depiction empowers transformers to proficiently categorize sign language gestures into predetermined classes, even when faced with subtle variations or noise within the input data.

Utilizing vision transformers (ViTs) for sign language recognition entails adapting well-known architecture originally devised for image classification tasks to accommodate the distinctive characteristics of sign language data. Prior to feeding images into a vision transformer, they require resizing, normalization, and potentially augmentation, ensuring

they are represented in a format conducive to enhancing the model's robustness. Unlike traditional CNNs, vision transformers process the entire image simultaneously through self-attention mechanisms. This capability allows ViTs to capture both local and global spatial dependencies within the sign language images, which are crucial for comprehending the intricate hand movements and gestures. Typically, a classification head is appended to predict the sign language gesture depicted in the input sequence of images. This classification head may comprise a simple fully connected layer followed by an activation function to compute the probability distribution across various sign language classes. An illustration of a vison transformer model for hand gesture recognition is shown in Figure 2 [56].

To summarize, transformer-based networks present a compelling method for sign language recognition, capitalizing on their self-attention mechanisms, parallel processing prowess, and scalability to proficiently capture the temporal dependencies within sign language sequences. This approach attains state-of-the-art performance in addressing the complexities of this demanding task.

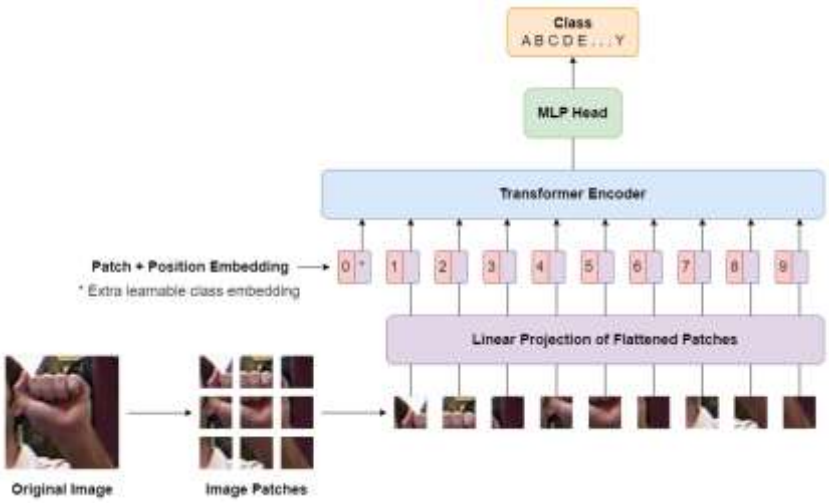

Figure 2: Vision transformer architecture for hand gesture recognition [56]

2.1.4 Two-Stream Networks

In sign language recognition, two-stream networks are frequently employed by incorporating two distinct streams of data: one dedicated to spatial features and the other to temporal features. These networks capitalize on the synergistic relationship between

spatial and temporal information present in sign language videos, aiming to improve recognition accuracy by effectively capturing both static hand shapes and the fluidity of hand movements characteristic of sign language gestures.

The spatial stream is geared towards extracting spatial characteristics from individual frames such as such as hand shapes, facial expressions, and positions within sign language videos or images. It commonly employs CNNs to discern visual patterns, shapes, and hand configurations evident in each frame. Operating independently on each frame, the spatial stream isolates spatial information. Meanwhile, the temporal stream focuses on capturing the dynamics and motion information such as movement and gestures over time across successive frames. It often utilizes RNNs or temporal convolutional networks (TCNs) to delineate the temporal relationships between frames. This stream processes sequences of frames to extract temporal information. Subsequently, after the separate extraction of spatial and temporal features, they are typically integrated with techniques such as concatenation, element-wise addition, or the utilization of learned attention mechanisms. The amalgamated features then proceed through additional layers, such as fully connected layers or classifiers, to execute sign language recognition tasks, including classification or sequence labeling. An illustration of a two-stream network for human action recognition is displayed in Figure 3 [57].

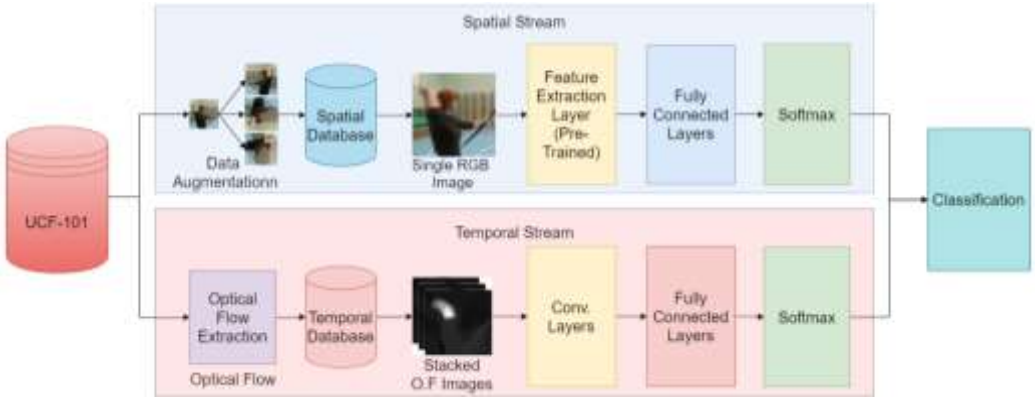

Figure 3: Human action recognition model using two-stream networks [57]

2.2 Self-Supervised Visual Representation Learning

Self-supervised learning offers a method to pre-train deep neural networks, enabling them to acquire significant representations from sign language videos or images, all

without the necessity of labeled data. Approaches like contrastive learning, as well as pretext tasks such as rotation prediction, jigsaw puzzle solving, or inpainting of missing segments, can be utilized to train the model effectively, enabling it to grasp pertinent features inherent in sign language data.

To capture temporal relationships within sign language videos, techniques such as temporal-contrastive learning can be employed. These methods aim to acquire representations that maintain temporal consistency while distinguishing between various segments of the videos. Additionally, self-supervised learning proves beneficial for domain adaptation by acquiring robust representations against alterations in lighting conditions, backgrounds, and camera perspectives. Through pre-training on extensive datasets using self-supervised methods, the model can acquire universal features transferable to the target sign language recognition task. Furthermore, self-supervised learning serves as a means of data augmentation. By generating extra training instances via pretext tasks like rotation prediction, colorization, or filling in missing components, the model can be exposed to a broader array of data, enhancing its ability to generalize. In situations where acquiring labeled data for sign language recognition proves scarce or costly, self-supervised learning emerges as a valuable avenue for unsupervised or weakly supervised learning. Through the utilization of extensive quantities of unlabeled sign language data, self-supervised techniques facilitate the training of resilient models even with minimal labeled data.

Numerous self-supervised visual representation learning models have been tailored for sign language recognition. Among them, simple contrastive learning of visual representations (SimCLR) [58] stands out as a prominent framework. It operates by discerning discriminative features from sign language videos or images, emphasizing agreement across differently augmented perspectives of the same image. Another noteworthy model is Momentum Contrast (MoCo) [59], a contrasting learning framework that utilizes a momentum encoder to generate a dynamic dictionary of representations. Demonstrating promising outcomes across diverse computer vision tasks, MoCo can be adjusted for sign language recognition to glean robust features from video sequences.

The temporal contrastive learning approach [60] concentrates on acquiring representations that maintain temporal consistency within videos. Through contrasting representations extracted from various segments of a video, models employing temporal contrastive learning can effectively grasp significant temporal dependencies evident in sign language gestures and movements. Additionally, rotation prediction models [61], which are trained to anticipate the rotation applied to an image, have been employed as pretext tasks within self-supervised learning. By mastering rotation prediction, these models can encapsulate invariant features essential for discerning sign language gestures from varying viewpoints.

Jigsaw puzzle solving [62] represents another prevalent pretext task employed in self-supervised learning. Here, models are trained to forecast the accurate spatial configuration of image patches rearranged akin to puzzle pieces. Such an approach aids in acquiring representations that remain unaltered by spatial transformations, a valuable asset for sign language recognition endeavors.

These models stand as potent instruments for acquiring representations from unlabeled sign language data, subsequently amendable for fine-tuning or transfer to downstream sign language recognition tasks, especially when labeled data is scarce. Through the utilization of self-supervised learning techniques, researchers can enhance the resilience and generalization capabilities of sign language recognition systems, ultimately advancing accessibility for individuals dependent on sign language communication. A self-supervised pre-trainable framework namely BEST [63], which leverages the success of BERT with the specific design to the sign language domain, is shown in Figure 4.

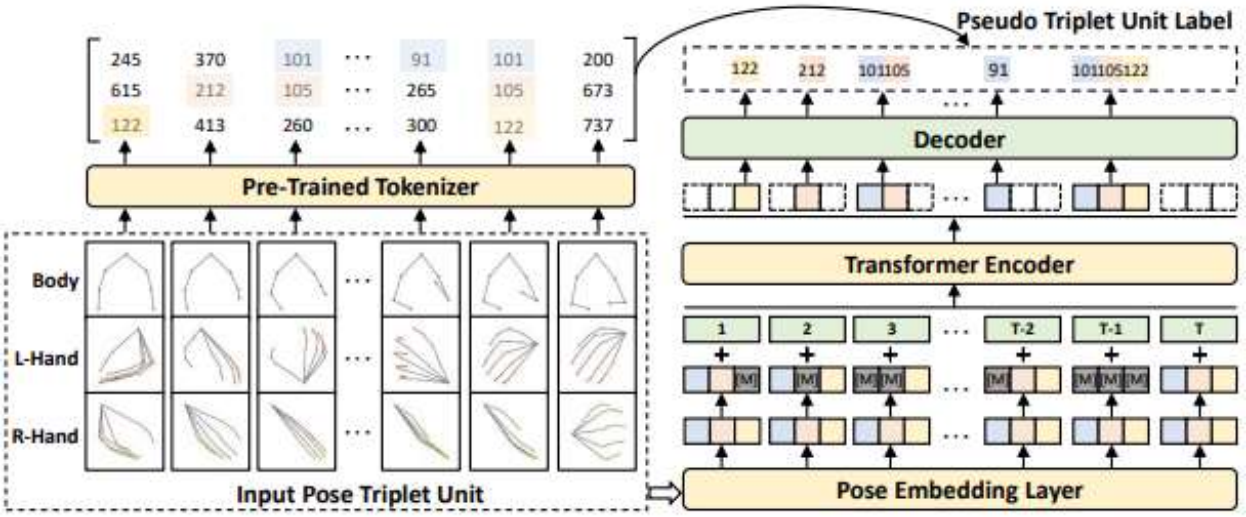

Figure 4: Self-learning framework with pre-training and task fine-tuning [63]

Continuous exploration of novel techniques and architectures is underway to bolster the efficacy and quality of self-supervised representation learning. Keeping abreast of the latest research literature remains crucial for identifying the most recent advancements in this domain.

2.3 Vision-Language Pre-Training

Vision-Language Pre-Training (VLP) for sign language involves pre-training a model on an extensive dataset containing paired visual and textual information pertaining to sign language. This method capitalizes on both visual and linguistic cues to develop unified representations that encapsulate the connections between signs and their associated textual descriptions or annotations.

In sign language, VLP entails training a model on sets of sign language videos or images paired with their respective textual descriptions, glosses, or annotations. Throughout the pre-training phase, the model establishes connections between visual attributes and linguistic representations, facilitating comprehension of the semantic nuances inherent in sign language gestures and expressions as they correspond to textual descriptions. Through pre-training on a diverse array of sign language data, the model develops a comprehensive grasp of both the visual and linguistic elements of sign language communication. This foundational knowledge can subsequently be refined through fine-tuning on targeted downstream tasks, even when labeled data is scarce.

The Contrastive Language-Image Pre-training (CLIP) model is a multimodal architecture capable of concurrently learning representations of both images and text [54]. Through training, CLIP develops associations between images and their respective textual descriptions, fostering an understanding of the interplay between visual and textual data. This enables the CLIP model to undertake different tasks that include image classification, zero-shot image classification, and image retrieval, among others, without necessitating task-specific training data. For illustration, the CLIP model is shown in Figure 5.

Adapting CLIP for sign language tasks entails training it on a dataset containing sign language video sequences or images paired with their corresponding textual descriptions or labels. Through fine-tuning, the model acquires the ability to correlate the visual characteristics of sign language gestures with their respective textual descriptions or labels, thereby enhancing its proficiency in recognizing and comprehending sign language.

In essence, VLP for sign language offers a promising avenue for narrowing the divide between visual and linguistic modalities in sign language processing, thereby enhancing the efficacy of diverse sign language-related tasks via unified representation learning.

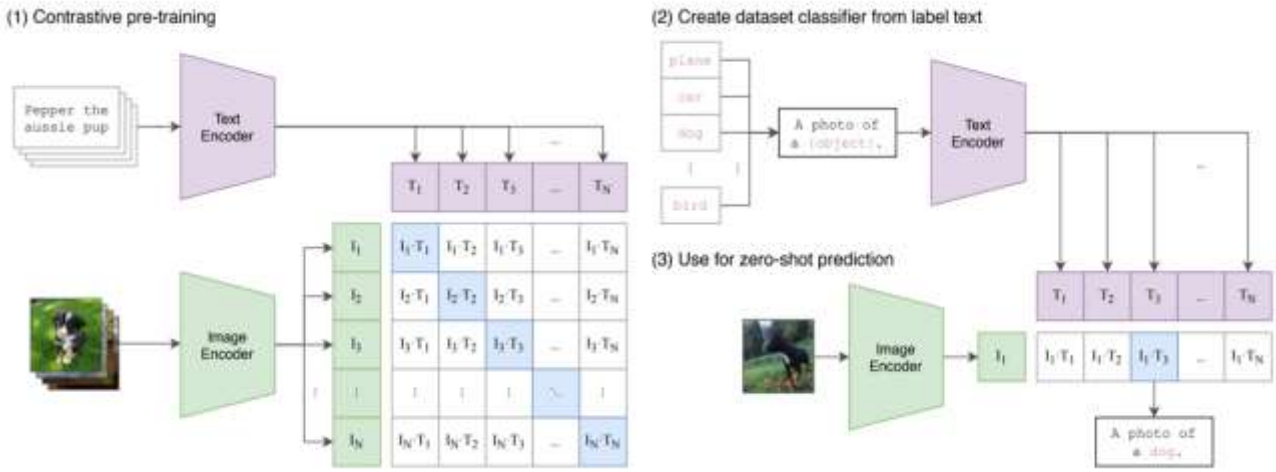

Figure 5: Approach adopted in CLIP model [64]

2.4 Weaknesses in Current Models

Current deep learning models applied to sign language recognition exhibit several weaknesses, including:

- Data Diversity and Scarcity: Sign language datasets commonly exhibit smaller sizes and less diversity compared to datasets in other domains. This limitation can result in overfitting and diminished generalization of models, especially when confronted with diverse signing styles, varying lighting conditions, and backgrounds.
- Limited Robustness to Variability: Deep learning models may encounter challenges in accommodating the inherent variability present in sign language, including variations in signing speed, handshapes, and movements. This variability poses obstacles to the model's capacity to precisely recognize signs across various contexts and users.
- Lack of Contextual Understanding: Deep learning models may encounter difficulties in comprehending the contextual significance of signs within sentences or conversations. Grasping the sequential structure of sign language and its underlying grammatical principles presents a notable challenge for existing models.
- Ambiguity and Similarity of Signs: Certain signs in sign language might exhibit visual resemblance or contextual ambiguity, rendering them challenging for models to differentiate accurately. This ambiguity can result in confusion and recognition errors, particularly in scenarios where signs share similar visual characteristics.
- Real-Time Processing Constraints: Deep learning models frequently demand substantial computational resources, which can constrain their practicality for real-time sign language recognition applications, including sign language interpretation systems or assistive technologies.
- Ethnic and Cultural Variations: Sign languages exhibit variability across diverse regions and cultures, encompassing unique vocabularies, grammatical structures, and signing styles. Deep learning models trained on one sign language might encounter difficulties in generalizing to others, thereby requiring extensive adaptation or retraining to accommodate varying linguistic contexts.

To overcome these limitations, advancements in model architectures, training methodologies, and dataset curation are essential, specifically tailored to the intricacies of sign language recognition. Additionally, integrating linguistic expertise and context-sensitive features into deep learning models has the potential to enhance their efficacy and resilience in handling sign language processing tasks.

2.5 Proposed Approach

In this thesis, a vision transformer-based model is proposed and implemented for Arabic and general American sign language recognition. The illustration of the proposed head and hands extraction method is shown in Figure 6, and resulting proposed model is shown in Figure 7. Sign languages rely on visual communication techniques such as handshape, facial expression, and body movement to convey information. In order to improve the modeling of sign languages, our proposal involves extracting head and hand regions from raw RGB frames to enhance the of visual representations. Specifically, given a temporally cropped video $V\epsilon\ \mathbb{R}^{T\times 3\times H_V\times W_V}$ with $T = 32$ frames and a spatial resolution of $H_V = W_V$= 256, we utilize the Mediapipe [65] library to estimate face, left hand, and right hand key points for each frame. From each set of estimated key points, we calculate the bounding box coordinates $(x_{min}, y_{\min})$ for the top-left corner and $(x_{\max}, y_{\max})$ for the bottom-right corner of the frame. We then crop and resize this bounding box region to fit ¼ of the original frame area. The original frame is then resized to $H = W = 128$ and stacked with the resized head, left and right-hand crops, maintaining the original frame size. In cases where no hand is detected in the frame, we fill the region with 0 values.

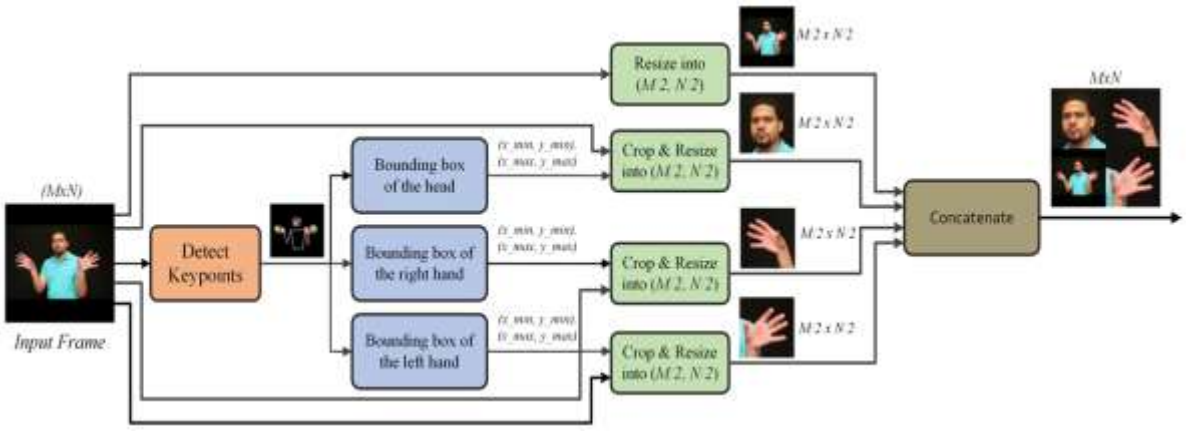

Figure 6: Proposed head and hands region extraction method

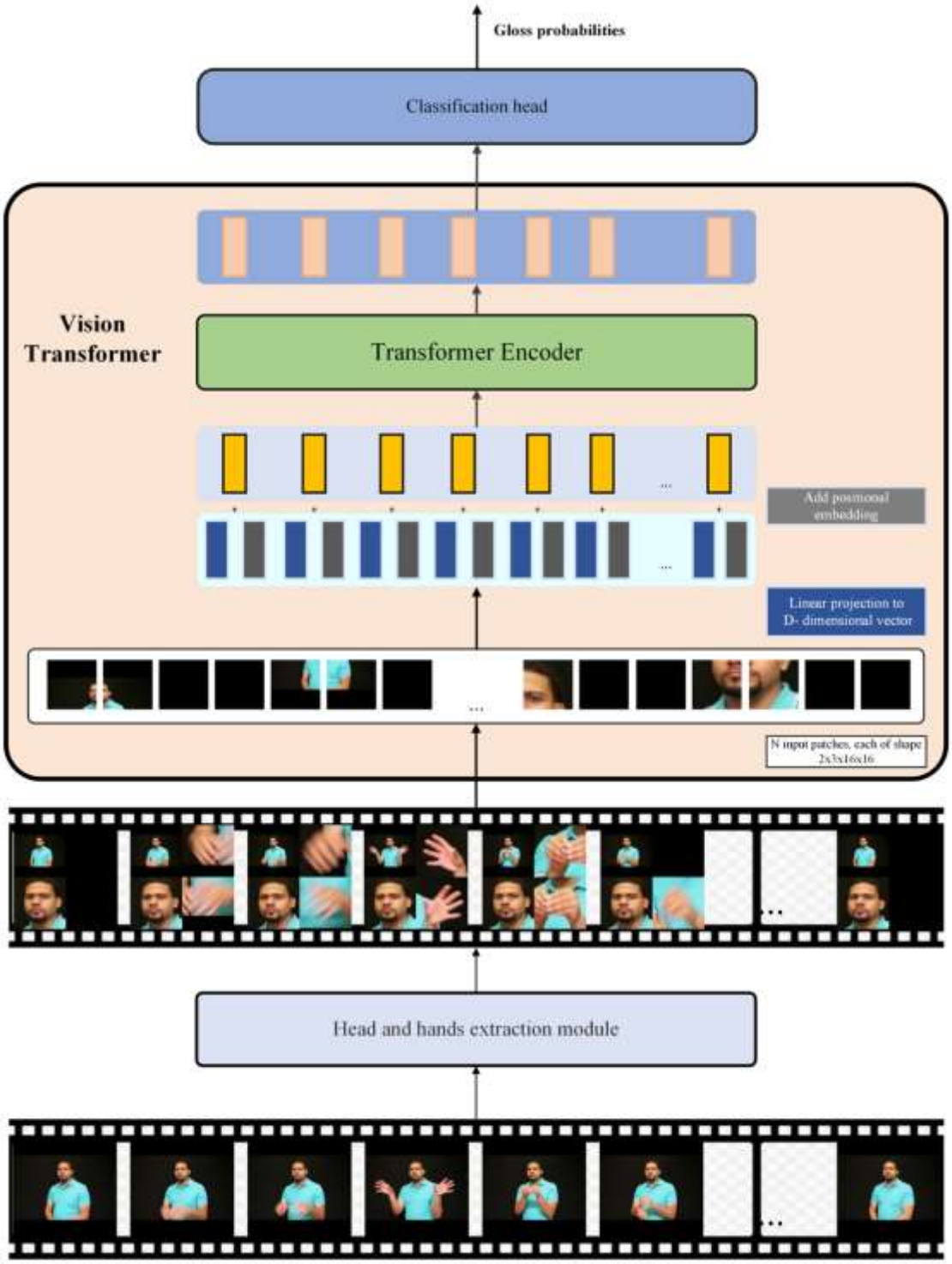

Figure 7: Proposed approach for sign video recognition

We employed ViT pre-trained by VideoMAEv2 [2] as our video feature extraction model. Our pre-trained VideoMAEv2 has three main components: cube embedding, position encoding and encoder. It processes sampled $I \epsilon \mathbb{R}^{C \times T \times H \times W}$ from a clip and uses the cube embedding to transform the frames into sequence of tokens. Cube embedding Φ_{emb}

encodes the local spatiotemporal features and builds the token list: $T = \Phi_{emb}$, where $T = \{T_i\}_{i=1}^{N}$ is the token sequence, T_i is the token produced by the embedding layer and then added with positional embedding, and N is the total token number. Encoder is a Vanilla Transformer Encoder [13] with 12 Transformer Blocks.

The Vision Transformer model consists of several consecutive transformer blocks, whose mathematical formulation is illustrated in equation. Below, we delve into the functionality of each component within the proposed transformer architecture.

$$z^l = \text{Norm}(\, x^l + \text{MultiHeadAttention}\,(\, x^l)) \tag{2.2}$$

$$\mathrm{x}^{\mathrm{l}+1} = \text{Norm}(\mathrm{z}^{\mathrm{l}} + \text{FeedForward}\,(\mathrm{z}^{\mathrm{l}}))$$

2.5.1 Position Embedding

Position embeddings in the proposed transformer architecture play a vital role during encoding the positional information of sequences, aiding the model in understanding the sequential order of input data. While the original transformer model [38] utilized sinusoidal positional encodings, integrating sine and cosine functions of varying frequencies into token embeddings to represent token positions accurately, the proposed transformer model opts for learnable position embeddings instead of predefined ones, diverging from the approach in the original paper. In the pretrained VideoMAEv2 [66], sinusoidal positional encodings are used, enabling the model to process videos with varying numbers of frames. Below are the sinusoidal embedding equations:

$$\mathrm{PE}_{(\mathrm{pos},2\mathrm{i})} = \sin\left(\frac{\mathrm{pos}}{10000^{\frac{2\mathrm{i}}{\mathrm{d}_{\mathrm{model}}}}}\right) \tag{2.3}$$

$$\mathrm{PE}_{(\mathrm{pos},2\mathrm{i}+1)} = \cos\left(\frac{\mathrm{pos}}{10000^{\frac{2\mathrm{i}}{\mathrm{d}_{\mathrm{model}}}}}\right)$$

2.5.2 Multi-Head Self-Attention

The multi-head attention layer within a transformer enables it to attend to different parts of the input sequence concurrently, extracting diverse patterns and relationships. This is formulated using the following set of steps:

• Input: $\{x_1, x_2, \dots, x_n\}$, $x_i \in \mathbb{R}^D$

In each head, we do the following process to get the output.

• Keys: $k_i = W_k x_i$, $k_i \in \mathbb{R}^{\frac{d}{n\_heads}}$

• Values: $v_i = W_v x_i$, $v_i \in \mathbb{R}^{\frac{d}{n\_heads}}$

• Query: $q_i = W_q x_i$, $q_i \in \mathbb{R}^{\frac{d}{n\_heads}}$

• Scores: $s_{ji} = \frac{k_j^T * q_i}{\sqrt{d}}$, $s \in \mathbb{R}^{nxn}$

• Attention weights: $a_{ji} = \text{Softmax}(s_i) = \frac{e^{s_{ji}}}{\sum_l e^{s_{li}}}$

• Outputs: $o_i = Attn(q_i, \{k_1, k_2, \dots, k_n\}, \{v_1, v_2, \dots, v_n\}) = \sum_l a_{li} v_i$. $o_i \in \mathbb{R}^{\frac{d}{n\_heads}}$

After getting the output from all heads we concatenate them to keep the dimension of the token.

2.5.3 Feed - Forward layer

The feedforward layer within the transformer block operates on each token individually, extracting information from each output token. This function is distinct from the attention layer, which is responsible for capturing relationships between tokens. This mechanism enables the model to capture specific spatial details from the input tokens, which is critical for precise recognition of sign language gestures.

2.5.4 Skip Connection

A skip connection layer is utilized after both the multi-head self-attention and feedforward layers by linking earlier layers with subsequent layers in the network to enable smooth gradient propagation and unrestricted flow of information. These

connections play a crucial role in addressing the issue of vanishing gradients, ensuring that the model can extract both low-level and high-level features from the input tokens. This mechanism enhances the model's capability to discern intricate patterns in sign language gestures.

2.5.5 Layer Normalization

Layer normalization in the proposed transformer model standardizes the activations across each layer using following equations, ensuring stable gradients during training, and aiding in faster convergence. This normalization method addresses challenges like internal covariate shift, as well as gradient explosions or vanishing, thus enhancing the overall training stability and efficacy of the vision transformer model for sign language recognition tasks. Following equations are typically used:

$$\mu = \frac{1}{d} \sum_{j=1}^{d} x_j \tag{2.4}$$

$$\sigma = \sqrt{\frac{1}{d} \sum_{i=1}^{d} (x_i - \mu)^2}$$

$$x := \frac{x - \mu}{\sigma}$$

Datasets play a crucial role in enhancing the accessibility of sign language translation and recognition technology by aiding in the creation of more precise and dependable recognition systems. Similar to various machine learning models, sign language recognition systems necessitate extensive labeled data for effective training. These datasets furnish the indispensable training material, comprising videos or images depicting sign language gestures alongside their corresponding annotations.

2.6 Datasets

Datasets serve as the foundation for evaluating the efficacy of sign language recognition models. Through the segmentation of datasets into training, validation, and testing subsets, researchers can train their models on a portion of the data, fine-tune model parameters using the validation set, and subsequently gauge the model's performance on

unseen data using the test set. Additionally, datasets function as benchmarks for assessing the performance of diverse sign language recognition algorithms and methodologies. Leveraging standardized datasets enables researchers to conduct equitable comparisons among various techniques, thereby facilitating the identification of the most efficacious approaches.

High quality datasets enhance model generalization to novel sign language gestures or variations in signing styles. Through the exposure of models to a broad spectrum of gestures from diverse signers, datasets contribute to the accurate recognition of signs from varied sources. A significant hurdle in vision-based SLT approach is the limited availability of useful datasets. Augmenting datasets can address this issue by bolstering their size and diversity, thereby enhancing the resilience and generalization capabilities of sign language recognition models. Techniques like flipping, rotation, scaling, and introducing noise to images or videos enable the generation of additional training instances without the necessity for new data collection.

Curating and annotating continuous sign language video sequences alongside spoken language translations is a labor-intensive endeavor. Datasets might encompass annotations or metadata that encapsulate particular hurdles in sign language recognition, such as occlusions, fluctuations in lighting conditions, or signer-independent recognition. Confronting these obstacles necessitates datasets that mirror real-world situations, empowering researchers to craft more resilient algorithms.

Several renowned datasets are utilized in sign language recognition research, each presenting distinctive characteristics and complexities. Here are a few prominent examples:

RWTH-PHOENIX-Weather 2014T: This dataset comprises video sequences of German sign language (Deutsche Gebärdensprache, DGS) depicting weather forecast. It encompasses recordings of signers delivering weather-related information, alongside annotations associating signs with spoken German. Similarly, the RWTH-PHOENIX-Weather 2014T Corpus dataset also features videos of DGS signers presenting weather forecasts. This dataset is annotated with glosses, providing written representations of signs, alongside English translations.

Phoenix-2014-T Dataset: This dataset comprises recordings of DGS signers delivering texts from diverse domains. It incorporates annotations that connect signs to spoken German, along with glosses and English translations.

American Sign Language Lexicon Video Dataset: This dataset is developed at the Center for Research in Computer Vision at the University of Central Florida. This dataset features videos showcasing individual ASL signs performed by various signers. It serves as a valuable asset for constructing ASL recognition systems. The latest addition to ASL datasets, WLASL, boasts an expanded vocabulary of 2,000 signs. This dataset encompasses 14,289, 3,916, and 2,878 image samples in the training, development, and testing sets, respectively. The Sem-Lex Benchmark offers 91,148 isolated sign videos representing a lexicon of 3,149 American Sign Language signs. These videos were contributed by 41 deaf participants and meticulously aligned with ASL-LEX 2.0 and ASL SignBank by a team of linguists. When combined with ASL Citizen, the Benchmark aggregates to a total of 174k videos.

CSL Dataset: This dataset, curated by researchers at Beijing University of Posts and Telecommunications, features videos showcasing CSL signs enacted by various signers. Its objective is to support advancements in CSL recognition and translation. The NMFs-CSL [30] dataset poses a significant challenge in CSL research due to its inclusion of numerous non-manual features. It encompasses 25,608 samples for training and 6,402 samples for testing, featuring a vocabulary size of 1,067. Nevertheless, the creators of the dataset furnish label indexes rather than glosses.

The Indian Sign Language (ISL) Recognition Dataset, crafted by researchers from the Indian Institute of Technology, showcases videos portraying ISL signs alongside their corresponding glosses and English translations. Its purpose is to bolster research endeavors in ISL recognition and translation systems.

Ankara University Turkish Sign Language Dataset (AUTSL): This dataset is an extensive, collection comprising isolated Turkish sign videos. It encompasses 226 distinct signs executed by 43 unique signers, totaling 38,336 video samples. These samples are captured utilizing Microsoft Kinect v2 technology, capturing data in RGB, depth, and skeleton formats. To ensure consistency, clipping and resizing operations to the RGB and

depth data is provided, resulting in a 512×512 resolution. The skeleton data provides spatial coordinates (x, y) for 25 junction points on the body, precisely aligned with the 512×512 data resolution.

Arabic Sign Language dataset (KArSL): KArSL stands as the most extensive video dataset for Word-Level Arabic Sign Language (ArSL). This database comprises 502 isolated sign words captured utilizing Microsoft Kinect V2 technology. Each sign within the dataset is executed by three proficient signers. To ensure comprehensive coverage, each of the signers repetitively performed each sign 50 times, culminating in a total of 75,300 samples across the entire database (502 signs × 3 signers × 50 repetitions).

Sign Language MNIST: Drawing inspiration from the classical MNIST dataset, this compilation showcases grayscale images depicting individual ASL gestures, each corresponding to letters and digits. It serves as a valuable data resource for training and evaluating models in fundamental sign language recognition endeavors.

Table 1: Different Sign languages and their representative features

Dataset	Number of Signs	Number of Videos	Source	Participants
AUTSL [63]	226	38336	Curated	Nonprofessional
MSASL [66]	1000	25513	Scraped	Unknown
WLASL [57]	2000	21083	Scraped	Unknown
ASL-Citizen [64]	2731	83399	Curated	Deaf
Sem-Lex Benchmark [65]	3149	91148	Curated	Deaf
NMF- CSL [67]	1067	26010	Curated	Deaf
Sign Language MNIST	26 letters	28000 images	Curated	Unknown
DGS	450	7300	Curated	Deaf
ISL	260	1289	Curated	Deaf
ArSL	502	7300	Curated	Deaf

These datasets exhibit diversity in the sign languages they portray, the range of gestures or signs incorporated, the intricacy of signing contexts, and the presence of

annotations. The ensuing Table 1 presents a compilation of notable sign languages along with their distinctive attributes.

2.7 Evaluation Metric

Various evaluation metrics are applicable to sign language recognition (SLR) tasks, contingent upon the particular goals and necessities of the application. The selection of evaluation metrics is influenced by factors like the nature of the SLR task, available data, and specific objectives. Employing multiple metrics is customary to furnish a thorough evaluation of SLR systems. The confusion matrix encompasses measurements such as true negatives (TN), true positives (TP), false negatives (FN), and false positives (FP). Below, the most commonly known evaluation metrics for SLR are presented, though accuracy is typically found sufficient in literature, unless speed is required for real timeliness:

Accuracy: Accuracy quantifies the ratio of accurately identified signs relative to the total number of signs within the dataset. This metric stands as one of the simplest metrics and is frequently employed as a primary evaluation criterion for SLR systems and is computed to signify the overarching model performance.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{2.5}$$

Precision and Recall: Precision quantifies the ratio of accurately identified signs among the signs predicted by the system, whereas recall gauges the ratio of accurately recognized signs among all the actual signs in the dataset. Precision and recall offer insights into the system's capacity to minimize false positives and false negatives, respectively. Precision assesses the precision of detection outcomes, computed by dividing the number of true positive (TP) detections by the total number of detections (as shown in equation 4.2). Conversely, equation (4.3) delineates recall, which assesses TP relative to the total possible detections.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{2.6}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{2.7}$$

F1 Score: This metric provides a balanced evaluation of a system's performance by representing the harmonic mean of precision and recall. It proves especially valuable in cases where there exists an imbalance between the number of positive and negative instances within the dataset. Its calculation is determined by the following equation:

$$\text{F1 Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{2.8}$$

Mean Average Precision (mAP): mAP, often employed in object detection endeavors, can be tailored for SLR applications as well. It quantifies the average precision across various classes or signs, proving beneficial in assessing SLR systems encompassing multiple classes. Calculating mAP involves averaging the average precision of each class, denoted by equation (4.1), which also serves as a measure for determining the accuracy of machine learning algorithms.

$$\text{mAP} = \sum_{q=1}^{Q} \frac{AveP(q)}{Q} \tag{2.9}$$

In this context, '*q*' denotes a query, 'Q' represents the total number of queries, and AveP(*q*) represents the average precision for a specific query.

Word Error Rate (WER): WER quantifies the ratio of inaccurately recognized signs or words in the output relative to the reference transcripts. This metric finds frequent application in tasks involving sequences of signs or words, such as sign language translation or captioning. WER is computed by dividing the total number of errors (including insertions, deletions, and substitutions) between the reference and hypothesis transcriptions by the total number of words in the reference transcription.

$$\text{WER} = \frac{(\text{Insertions} + \text{Deletions} + \text{Substituions})}{(\text{Number of words in ground truth})} \tag{2.10}$$

BLEU Score: BLEU is a metric frequently employed in machine translation endeavors. It can likewise be adjusted for SLR tasks where the output constitutes a sequence of signs or words, achieved by contrasting the predicted sequence with one or multiple reference sequences. The ultimate BLEU score is derived by calculating the geometric mean of all the adjusted precisions up to N, subsequently multiplied by BP. Here, N denotes the n-gram order utilized for computation, typically set at 4, encompassing uni, bi, tri, and tetra grams for assessment.

$$\text{BLEU} = \text{BP} \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \text{ where BP} = \begin{cases} 1 & \text{if } c > r \\ e^{\left(1-\frac{r}{c}\right)} & \text{if } c \leq r \end{cases} \quad (2.11)$$

where p_n is modified precision, and *w* is the weight. The parameter '*c*' signifies the length of the closest matching n-grams between the generated output and reference translations, while '*r*' indicates the length of the reference translation. These parameters are instrumental in calculating precision, which is an integral component of the overall BLEU score.

Chapter 3: Experimental Results and Discussions

This chapter presents a benchmarking evaluation of sign language recognition model proposed in chapter 4. The various experiments conducted are discussed below with results.

3.1. In-house Arabic Sign Language Recognition

An in-house dataset comprising forty-nine (49) signs of Emirates Sign Language (ESL) was compiled from recordings of 7 volunteers performing gestures in three distinct environments. All volunteers provided consent for the dataset to be utilized in the research. The signs performed by both hands were selected, resulting in the following classes (Table 2). Each sign was performed by each volunteer.

Table 2: List of different signs collected in-house

Gloss			
A	Astronaut	M	Mother's sister
	Ambassador		Mother's brother
B	Burj Khalifa	N	Nervous
	Burj Al Arab	P	Photograph
	Birth Certificate		People of determination card
C	Citizenship		Pharmacy
	Colors		Pray
	Chief		Prevent
D	Delay		Pregnant
	Death Certificate		Play
	Divorce Certificate	R	Read
F	Fire fighter	S	Salary
	First Aid		Salary Certificate
	Family		Son
G	Good Conduct Certificate	T	Translator
	Global Village		Time
	Guard	U	University Certificate
	Generous	V	Volunteer
H	Health Card	W	Work
	Happy		Wednesday
I	Imam		Week
J	Journalist	X	X-Ray
	Lawyer	Z	Zoo
	Membership Card		Zayed Mosque
	Marriage Certificate		

The original videos have 1920x1080 pixels resolution, a 30fps frame rate, and typically last about 5 seconds. Figure 8 shows nine sample frames extracted from videos recorded in-house. Each sign was performed twice by two volunteers wearing different clothing and against different backgrounds. The remaining seven volunteers performed each sign once. For each class, one volunteer's video was designated for testing, while the others were allocated to the training. The padding of zeros was done to each side of the frame that was shorter to achieve a square shape. Subsequently, the videos were resized to 224 x 224 pixels to input them into the model. During training, sixteen (16) frames were randomly sampled from the original video, while during testing, input frames were uniformly sampled over the video.

Figure 8: Nine sample sign frames

We conducted experiments using three types of augmentations on our video samples. Figure 9 showcases the results of different combinations of augmentation for one frame. Additionally, Figure 10 displays ten results obtained from a combination of shear and rotate transformations.

Figure 9: Augmentations applied to one frame

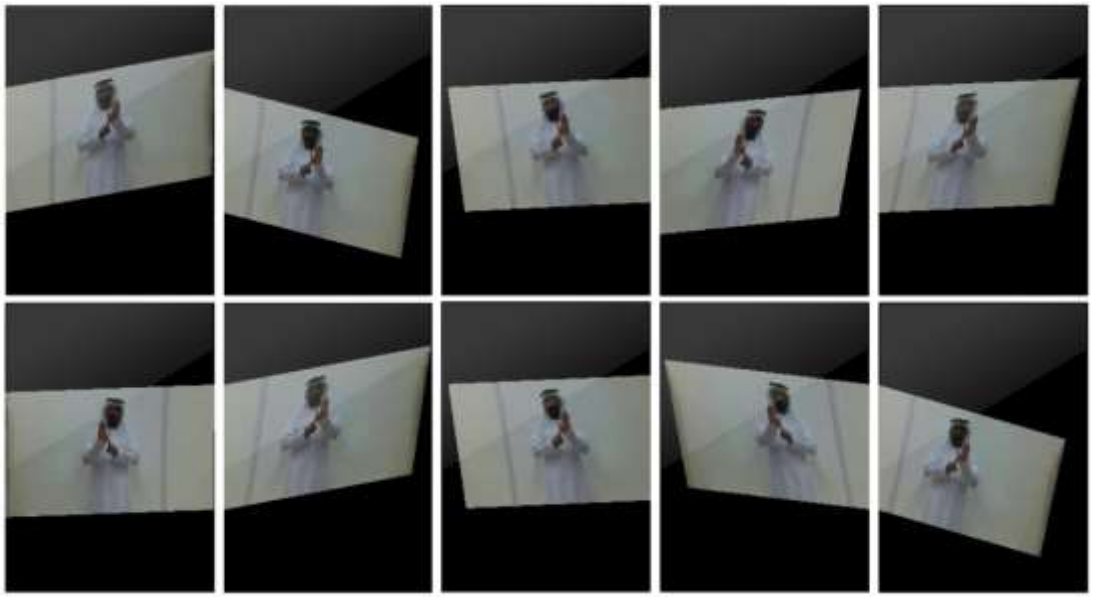

Figure 10: Combinations of shear and rotate transformations applied to one frame

Table 3: Model accuracy on in-house dataset with different augmentations

Shear	Translate	Rotate	Top-1 Accuracy
	✓		69%
✓		✓	85%
		✓	75%
✓	✓	✓	79%
✓			75%
		✓	77%

The results of testing on the dataset with various types of augmentations are summarized in Table 3. For each class, each video was augmented 10 times, resulting in 80 videos. It turned out that the model, which was trained on the dataset with shear and

rotate transformations achieved the highest accuracy, reaching 85% on the testing. Conversely, the dataset that was created solely from translational transformations exhibited the lowest performance, reaching an accuracy of 69%.

In another experiment, the numbers of augmentations were varied on the model, and the testing results are displayed in Table 4. The model that was trained on the dataset augmented up to 80 videos per class obtained the highest performance, reaching 85% accuracy on the testing. Conversely, the model that was augmented the least with about 16 videos per class exhibited the lowest accuracy, achieving 38%.

Table 4: Accuracy results for different number of samples per class

No. of samples per class	Top–1 Accuracy
80	85%
56	69%
40	71%
24	73%
16	38%

3.2 General Sign Language Recognition

Initially, the models are trained using the WLASL2000 [67] dataset. This initial version of the dataset contained 14 hours of isolated sign language videos in ASL, with a total of 2000 labels. However, upon a recent review of the dataset, significant shortcomings and discrepancies were discovered due to reliance on English translations for sign names. Consequently, we opted to use the manually corrected version of the WLASL2000 dataset. Based on the updated dataset labels, we created two benchmark dataset versions. The first version consists of 1518 classes, each with at least one sample in the train, test, and validation sets. The second version includes 1573 classes, ensuring each class has at least one sample in the train and test sets. Table 5 illustrates the testing accuracy results observed on ViT models that were trained on datasets with varying numbers of augmented videos per class. The table clearly shows that all models converged

to specific values without displaying signs of overfitting. Table 6 shows testing accuracy observed during training of models on revised WLASL2000 dataset.

Table 5: Different models performance on WLASL2000 dataset

Models	Top-1 Accuracy	Top-5 Accuracy
S3D – 64 frames [71]	51.15%	83.43%
VK Net 64 [69]	57.19%	88.29%
X-CLIP (ViT-B/64) [70]	41.69%	76.14%
X-CLIP (ViT-B/32) [70]	39.15%	72.53%
X-CLIP (ViT-B/16) [70]	39.63%	74.13%
VK-Net 64 (NLA-SLR) [69]	61.05%	91.45%
ViT-B 16 (Videomaev2) [66]	48.75%	83.32%
ViT-B 16 (VideoMAEv2) – Proposed	46.32%	81.28%
ViT-B 32 (Videomaev2, pretrained on ASL Citizen)	60.27%	90.17%
ViT-B 32 (Videomaev2, pretrained on ASL-Citizen)-Proposed	61.46%	91.22%

Table 6: Different models performance on revised WLASL dataset

Models	WLASL1573 (Accuracy)		WLASL1518 (Accuracy)	
	Top-1	Top-5	Top-1	Top-5
S3D – 64 [71]	72.33%	90.6%	72.25%	92.04%
VK-Net 64 [69]	77.63%	93.98%	78.87%	94.01%
X-CLIP (ViT-B/64) [70]	63.93%	87.04%	-	-
ViT-B 16 [66]	70.6%	91.06%	69.51%	90.94%
ViT-B 16 (pretrained on ASL-Citizen)	-	-	79.72%	95.28%
ViT-B 32 (pretrained on ASL-Citizen)	-	-	79.78%	95.35%
ViT-B 32 (pretrained on ASL-Citizen), Proposed	-	-	81.26%	95.81%

In another experiment, ASL- Citizen dataset was employed to examine models based on accuracy for comparison purposes. Table 7 shows comparative results.

Table 7: Different models performance applied on ASL-Citizen dataset

Model	Top-1 (Accuracy)	Top-5 (Accuracy)
I3D [55]	63.1%	86.09%
ST-GCN [72]	59.52%	82.68%
S3D – 64 [71]	61.63%	88.93%
ViT-B 16 (videomaev2) [66]	83.58%	98.19%
ViT-B 32 (videomaev2) - Proposed	88.25%	99.12%

For each experiment, we used 4 NVIDIA A100 – 40GB GPUs and trained for 50 epochs with a total batch size of 16. Our proposed method performed SOTA top-1 accuracy with 61.46% on the original WLASL dataset using only one modality, whereas VKNet [71] uses 2 modalities and 2 receptive field.

Chapter 4: Discussions and Conclusions

4.1 Transformer-based Video Extraction Model vs 3DCNN Model

Transformer-based video extraction models and 3D CNN models are commonly employed methods for video analysis tasks, each possessing its own set of advantages and drawbacks. Their performance is contingent upon the particular task and dataset at hand. The selection of the appropriate model is influenced by factors such as requirements of the task, the data availability, computational resources, and the desired level of interpretability. They differ mainly in the following ways:

Architecture: Transformer-based models adapt the Transformer architecture, initially designed for natural language processing, to process video data. They utilize self-attention mechanisms and feedforward neural networks operating across multiple frames. On the other hand, 3D CNNs extend traditional 2D convolutional neural networks by incorporating spatiotemporal convolutions, convolving across three dimensions (height, width, and time).

Temporal understanding: Transformers inherently capture long-range dependencies in sequences via self-attention mechanisms, effectively modeling temporal relationships in videos without explicit convolutions across time. Conversely, 3DCNNs process spatiotemporal information directly through 3D convolutions, enabling them to capture motion patterns and temporal dependencies. However, they may face challenges in capturing long-range dependencies effectively.

Parameter efficiency: Transformers are recognized for their scalability and parameter efficiency in comparison to traditional CNN architectures. They adeptly handle variable-length sequences and achieve promising results with relatively fewer parameters. Conversely, 3D CNNs typically demand more parameters due to the incorporation of an additional dimension (time) in the convolutional kernels. Consequently, this can result in escalated computational costs and memory requirements, particularly for large-scale video datasets.

Data efficiency: Transformers excel in leveraging pretraining on extensive text and image datasets, utilizing techniques like pretraining on large-scale language models or

self-supervised learning. This capability enables effective transfer learning to video tasks with limited annotated data. Conversely, 3DCNNs often demand substantial amounts of annotated video data for training, particularly when starting from scratch. While techniques such as transfer learning from pre-trained 3D models can assist, they may not match the effectiveness of pretraining on text or image data for certain tasks.

Robustness to Input Variations: Transformers demonstrate inherent flexibility in handling variable-length sequences, effectively capturing temporal dynamics across different time scales. This potentially enhances their robustness to variations in video duration and frame rate. On the contrary, 3DCNNs may encounter challenges with variable-length sequences, typically requiring fixed-length inputs. Preprocessing techniques such as sampling or padding may be necessary to manage input length variations effectively.

Interpretability: Transformers provide enhanced interpretability through attention mechanisms, enabling users to visualize which parts of the input are being attended to at each layer. This facilitates understanding of the model's decision-making process. In contrast, interpretability in 3DCNNs is generally lower as they primarily rely on convolutions without explicit attention mechanisms. Comprehending the learned representations may necessitate additional techniques or analysis.

4.2 Different Vision Transformers on the Sign Language Dataset

The ViT model, initially trained on the Kinetics dataset, showed promising transfer learning capabilities across diverse small-scale datasets using up to ten (10) augmented samples per category, as evidenced by the outcomes. Nevertheless, the study conducted in-house had a narrow focus, examining solely 49 two-handed signs, which constitute a fraction of the complete ArSL lexicon, and did not assess the ViT model against other cutting-edge deep learning models for action/video recognition. The results indicated that enhancing the dataset with additional samples might enhance the model's efficacy, yet further investigations are necessary to validate this assertion.

The augmentation techniques yielded results where all combinations achieved accuracies exceeding 75%. Additionally, it was noted that the combination of shear and

rotate methods resulted in enhanced accuracy compared to incorporating all three methods simultaneously.

Regarding dataset preparation, a random extraction approach was employed to select 16 frames from each video, potentially missing crucial attention points for particular gestures. Implementing smarter adaptive sampling methods for frame extraction could potentially enhance the model's effectiveness. The ViT model's input size is 224 x 224, subsequently divided into 16x16 cells for additional feature extraction. Considering sign language videos often require detailed features of both hand gestures and facial expressions, experimenting with enlarging the ViT model's input size could lead to improved performance.

The first phase of our research concentrated on fine-tuning the ViT model using an augmented small-scale in-house dataset for ArSL recognition. Our primary aim was to achieve satisfactory outcomes despite resource constraints and a limited dataset, involving only 10 volunteers and a maximum of 10 augmented samples per original video. Throughout the training phase, the training set exclusively comprised augmented videos, while the test set comprised original videos from a single volunteer. Analysis of the experimental results revealed that the combining rotation and shear surpassed other methods, attaining a top-1 accuracy of 85% on the test dataset. Although our study focused solely on 49 two-handed signs as a case study, we contend that it offers valuable insights into data-centric AI solutions targeting small-scale dataset-based SLR tasks and action/video recognition in general. The future endeavors entail further exploration of augmentation and intelligent extraction techniques for our samples, alongside an exploration of fundamental deep learning models. Additionally, we aspire to expand our existing in-house dataset to encompass a broader spectrum of ArSL vocabulary and make it publicly accessible for research purposes.

For the second part of the study, the proposed method managed to boost the state of the art top-1 accuracy from 61.05% to 61.46% using only raw RGB frames. This improvement hinges on three key enhancements integrated into our methodology. First, we used large-scale unsupervised pre-trained model, which helps the model better understand spatiotemporal representation of the video modality. Second, we have

developed a module specifically designed to extract head and hand movements, refining the model's ability to interpret sign language gestures accurately. Third, we have employed supervised pre-training using the ASL-Citizen dataset, which provides additional training data for the model. Notably, without this supervised pre-training and our proposed head and hands extraction module, the performance of the ViT model on the WLASL2000 dataset dropped to 48.75%. AddThe key improvement consists of 3 key elements: Large-scale unsupervised pre-training, head and hands extraction module and supervised pre-training on ASL-Citizen dataset. Without the supervised pre training on ASL-Citizen dataset and our proposed head and hands extraction module, ViT model performance on WLASL2000 dataset was 48.75%. Additionally, we explored the ASL-Citizen dataset, which contains a substantial 83,399 samples. Interestingly, the 3D CNN-based S3D model performed worse than the ViT model on this dataset, contrary to its performance on the WLASL-2000 dataset. This highlights the importance of dataset nuances and underscores the need for tailored approaches to achieve optimal results.

4.3 Future Work

The research conducted in this thesis offers opportunities for further enhancement in several aspects. Firstly, alternative attention mechanisms could be explored to better suit sign language recognition, enabling more effective capture of spatial-temporal relationships. Secondly, there is potential for integrating multiple modalities, including video, text, and audio, to enrich contextual information and thereby enhance accuracy. Thirdly, investigating models optimized for real-time performance could be beneficial for applications like assistive technologies for the deaf community. Additionally, exploring cross-lingual recognition systems to facilitate communication across different sign languages presents another avenue for research. Lastly, leveraging transfer learning to adapt pre-trained transformer models for sign language tasks could offer promising avenues for performance improvement.

References

[1] J. Jepsen, G. De Clerck, et al., *Sign Languages of the World: A Comparative Handbook*, De Gruyter Mouton Publication; 1st edition, October 16, 2015.

[2] Sign Language MNIST, Available: https://www.kaggle.com/datasets/datamunge/sign-language-mnist [Accessed: February 29, 2024].

[3] Q. Memon, N. Valappil, "On Multi-class Aerial Image Classification using Learning Machines," In *Computer Vision and Recognition Systems Using Machine and Deep Learning Approaches: Fundamentals, technologies, and applications*," Chap. 15, pp. 351-384, DOI: 10.1049/PBPC042E_ch15, IET Digital Library, https://digitallibrary.theiet.org/content/books/10.1049/pbpc042e_ch15, 2021.

[4] N. Valappil, Q. Memon, "Vehicle Detection in UAV Videos Using CNN-SVM," In Proceedings of the 12th International Conference on Soft Computing and Pattern Recognition, Advances in Intelligent Systems and Computing, Vol 1383, 2022, 221-232.

[5] A. Domyati, Q. Memon, "Machine Learning Based Improved Heart Disease Detection with Confidence," *International Journal of Online and Biomedical Engineering*, vol. 19, no. 08, pp. 130–143, 2023.

[6] N. Bachir, Q. Memon, "Investigating YOLOv5 for Search and Rescue Operations Involving UAVs," In Proceedings of ACM International Conference on Control and Computer Vision, 2022, pp. 200 – 204.

[7] N. Bachir, Q. Memon, "Benchmarking YOLOv5 models for improved human detection in search and rescue missions," *Journal of Electronic Science and Technology*, vol. 22, no. 1, 100243, 2024.

[8] M. Alameri, Q. Memon, "Experimental analysis of accelerometer data for human activity recognition," In Proceedings of SPIE International Conference on Mathematical and Statistical Physics, Computational Science, Education and Communication, vol. 12936, 129361F, 2023.

[9] Q. Memon, B. Wodajo, A. Alshamsi, S. Alshamsi, A. Alshebli, N. Alderei, "Experimental Analysis of Robust Forest Fire Detection for Sustainability," In Proceedings of the 9th International Conference on Computer Technology Applications, 2023, pp. 91–96.

[10] Q. Memon, M. Hassan, "Cluster Analysis of Patients' Clinical Information for Medical Practitioners and Insurance Companies," *International Journal of Online and Biomedical Engineering*, vol. 16, no. 04, pp. 128–138, 2020.

[11] B. Asadi, Q. Memon, "Efficient breast cancer detection via cascade deep learning network," *International Journal of Intelligent Networks*, vol. 4, pp. 46-52, 2023.

[12] Q. Memon, B. Wodajo, S. Tekleab, and E. Alshehi, "Detection of Static and Moving Objects behind Walls and Surfaces – An Experimental Investigation," In Proceedings of the 9th International Conference on Computing and Artificial Intelligence, 2023, pp. 8–12.

[13] S. Tamura, S. Kawasaki, "Recognition of Sign Language Motion Images," *Pattern Recognition*, vol. 21, no. 4, 1988.

[14] T. Starner, J. Weaver, and A. Pentland, "Real-time American Sign Language Recognition using Desk and Wearable Computer Based Video," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 20, no. 12, 1998.

[15] K. Cormier, N. Fox, B. Woll, A. Zisserman, N. Cihan Camgoz, and R. Bowden, "ExTOL: Automatic recognition of British Sign Language using the BSL Corpus," In Proceedings of the Sign Language Translation and Avatar Technology, 2019.

[16] H. Cooper, B. Holt, and R. Bowden, "Sign Language Recognition," In *Visual Analysis of Humans*, J. Spudich, and B. Satir, Eds. Springer, 2011.

[17] D. Bragg, et. al., "Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective," In Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility, 2019.

[18] N. Cihan Camgoz, A. Kindiroglu, S. Karabuklu, M. Kelepir, A. Ozsoy, and L. Akarun, "Bosphorus Sign: A Turkish Sign Language Recognition Corpus in Health and Finance Domains," In Proceedings of the International Conference on Language Resources and Evaluation, 2016.

[19] H. Wang, X. Chai, X. Chen, "Sparse Observation (SO) Alignment for Sign Language Recognition," *Neurocomputing*, vol. 175, 2016.

[20] S. Ebling, et al., "SMILE: Swiss German Sign Language Dataset," In Proceedings of the International Conference on Language Resources and Evaluation, 2018.

[21] O. Koller, S. Zargaran, H. Ney, R. Bowden, "Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition," In Proceedings of the British Machine Vision Conference, 2016.

[22] J. Huang, W. Zhou, O. Zhang, H. Li, and W. Li, "Video-based Sign Language Recognition without Temporal Segmentation," In Proceedings of the AAAI Conference on Artificial Intelligence, 2018.

[23] R. Cui, H. Liu, and C. Zhang, “Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition by Staged Optimization,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[24] R. Cui, H. Hu Liu, and C. Zhang, “A Deep Neural Framework for Continuous Sign Language Recognition by Iterative Training,” *IEEE Transactions on Multimedia*, vol. 21, no. 7, pp. 1880-1891, July 2019.

[25] N. Cihan Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, “Neural Sign Language Translation,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[26] S. Morrissey, H. Somers, R. Smith, S. Gilchrist, and S. Dandapat, “Building a Sign Language Corpus for Use in Machine Translation,” In Proceedings of the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, 2010.

[27] D. Stein, C. Schmidt, H. Ney, “Analysis, Preparation, and Optimization of Statistical Sign Language Machine Translation,” *Machine Translation*, vol. 26, no. 4, 2012.

[28] C. Schmidt, O., Koller, H. Ney, T. Hoyoux, J. Piater, “Using Viseme Recognition to Improve a Sign Language Translation System,” In Proceedings of the International Workshop on Spoken Language Translation, 2013.

[29] K. Kudrinko, E. Flavin, X. Zhu and Q. Li, "Wearable Sensor-Based Sign Language Recognition: A Comprehensive Review," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 82-97, 2021.

[30] M. Al-Qurishi, T. Khalid and R. Souissi, "Deep Learning for Sign Language Recognition: Current Techniques, Benchmarks, and Open Issues," *IEEE Access*, vol. 9, pp. 126917-126951, 2021.

[31] Z. Saeed, Z. Zainol, B. Zaidan and A. Alamoodi, "A Systematic Review on Systems-Based Sensory Gloves for Sign Language Pattern Recognition: An Update From 2017 to 2022," *IEEE Access*, vol. 10, pp. 123358-123377, 2022.

[32] B. Joksimoski et al., "Technological Solutions for Sign Language Recognition: A Scoping Review of Research Trends, Challenges, and Opportunities," *IEEE Access*, vol. 10, pp. 40979-40998, 2022.

[33] Z. Zhou, V. Tam and E. Lam, "SignBERT: A BERT-Based Deep Learning Framework for Continuous Sign Language Recognition," *IEEE Access*, vol. 9, pp. 161669-161682, 2021.

[34] H. Luqman, "An Efficient Two-Stream Network for Isolated Sign Language Recognition Using Accumulative Video Motion," *IEEE Access*, vol. 10, pp. 93785-93798, 2022.

[35] P. Xie, M. Zhao and X. Hu, "PiSLTRc: Position-Informed Sign Language Transformer With Content-Aware Convolution," *IEEE Transactions on Multimedia*, vol. 24, pp. 3908-3919, 2022.

[36] B. Natarajan et al., "Development of an End-to-End Deep Learning Framework for Sign Language Recognition, Translation, and Video Generation," *IEEE Access*, vol. 10, pp. 104358-104374, 2022.

[37] H. Zhou, W. Zhou, Y. Zhou, and H. Li, "Spatial-Temporal Multi-Cue Network for Sign Language Recognition and Translation," *IEEE Transactions on Multimedia*, vol. 24, pp. 768-779, 2022.

[38] D. Kothadiya, C. Bhatt, T. Saba, A. Rehman, and S. Bahaj, "SIGNFORMER: DeepVision Transformer for Sign Language Recognition," *IEEE Access*, vol. 11, pp. 4730-4739, 2023.

[39] H. Hu, J. Pu, W. Zhou, and H. Li, "Collaborative Multilingual Continuous Sign Language Recognition: A Unified Framework," *IEEE Transactions on Multimedia*, vol. 25, pp. 7559-7570, 2023.

[40] E. Rajalakshmi et al., "Multi-Semantic Discriminative Feature Learning for Sign Gesture Recognition Using Hybrid Deep Neural Architecture," *IEEE Access*, vol. 11, pp. 2226-2238, 2023.

[41] J. Hu, Y. Liu, K. -M. Lam and P. Lou, "STFE-Net: A Spatial-Temporal Feature Extraction Network for Continuous Sign Language Translation," IEEE Access, vol. 11, pp. 46204-46217, 2023.

[42] Z. Liu, L. Pang, and X. Qi, "MEN: Mutual Enhancement Networks for Sign Language Recognition and Education," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 1, pp. 311-325, Jan. 2024.

[43] S. M. Kamal, Y. Chen, S. Li, X. Shi, and J. Zheng, "Technical Approaches to Chinese Sign Language Processing: A Review," *IEEE Access*, vol. 7, pp. 96926-96935, 2019.

[44] Y. Yu, X. Chen, S. Cao, X. Zhang, and X. Chen, "Exploration of Chinese Sign Language Recognition Using Wearable Sensors Based on Deep Belief Net," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 5, pp. 1310-1320, May 2020.

[45] P. Wang, Y. Zhou, Z. Li, S. Huang, and D. Zhang, "Neural Decoding of Chinese Sign Language With Machine Learning for Brain–Computer Interfaces," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 2721-2732, 2021.

[46] M. A. Bencherif et al., "Arabic Sign Language Recognition System Using 2D Hands and Body Skeleton Data," *IEEE Access*, vol. 9, pp. 59612-59627, 2021.

[47] F. Shah, M. Shah, W. Akram, A. Manzoor, R. Mahmoud, and D. Abdelminaam, "Sign Language Recognition Using Multiple Kernel Learning: A Case Study of Pakistan Sign Language," *IEEE Access*, vol. 9, pp. 67548-67558, 2021.

[48] J. Li, J. Meng, H. Gong, and Z. Fan, "Research on Continuous Dynamic Gesture Recognition of Chinese Sign Language Based on Multi-Mode Fusion," *IEEE Access*, vol. 10, pp. 106946-106957, 2022.

[49] J. Shin, A. Miah, K. Suzuki, K. Hirooka and M. Hasan, "Dynamic Korean Sign Language Recognition Using Pose Estimation Based and Attention-Based Neural Network," *IEEE Access*, vol. 11, pp. 143501-143513, 2023.

[50] O. El Ghoul, M. Aziz, and A. Othman, "JUMLA-QSL-22: A Novel Qatari Sign Language Continuous Dataset," *IEEE Access*, vol. 11, pp. 112639-112649, 2023.

[51] M. Faisal et al., "Enabling Two-Way Communication of Deaf Using Saudi Sign Language," *IEEE Access*, vol. 11, pp. 135423-135434, 2023.

[52] M. R. Abid, E. M. Petriu and E. Amjadian, "Dynamic Sign Language Recognition for Smart Home Interactive Application Using Stochastic Linear Formal Grammar," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 3, pp. 596-605, 2015.

[53] C. O. Sosa-Jiménez, H. V. Ríos-Figueroa, and A. L. Solís-González-Cosío, "A Prototype for Mexican Sign Language Recognition and Synthesis in Support of a Primary Care Physician," *IEEE Access*, vol. 10, pp. 127620-127635, 2022.

[54] Z. Wang *et al*., "Hear Sign Language: A Real-Time End-to-End Sign Language Recognition System," *IEEE Transactions on Mobile Computing*, vol. 21, no. 7, pp. 2398-2410, 1 July 2022.

[55] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," In Proceedings of I EEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017 pp. 4724-4733.

[56] C. Tan, K. Lim, R. Chang, C. Lee, A. Alqahtani, "HGR-ViT: Hand Gesture Recognition with Vision Transformer," *Sensors*, vol. 23, 5555, 2023.

[57] S. Khan, A. Hassan, F. Hussain, A. Perwaiz, F. Riaz, M. Alsabaan, and W. Abdul, “Enhanced Spatial Stream of Two-Stream Network Using Optical Flow for Human Action Recognition,” *Applied Sciences*, vol. 13, 8003, 2023.

[58] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” In Proceedings of the 37th International Conference on Machine Learning, vol. 119, 2020.

[59] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," In Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020, pp. 9726-9735.

[60] I. Dave, R. Gupta, M. Rizve, M. Shah, “TCLR: Temporal contrastive learning for video representation,” *Computer Vision and Image Understanding*, vol. 219, 2022, 103406.

[61] S. Yamaguchi, S. Kanai, T. Shioda, and S. Takeda, "Image Enhanced Rotation Prediction for Self-Supervised Learning," In Proceedings of IEEE International Conference on Image Processing, Anchorage, USA, 2021, pp. 489-493.

[62] Y. Chen, X. Shen, Y. Liu, Q. Tao, J. Suykens, “Jigsaw-ViT: Learning jigsaw puzzles in vision transformer,” *Pattern Recognition Letters*, vol. 166, pp. 53-60, 2023.

[63] W. Zhao, H. Hu, W. Zhou, J. Shi, H., Li, “BEST: BERT pre-training for sign language recognition with coupling tokenization,” In Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence, vol. 37, Article 401, 2023, pp. 3597–3605.

[64] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning Transferable Visual Models From Natural Language Supervision,” In Proceedings of the 38th International Conference on Machine Learning, 18-24 July 2021, 8748-8763.

[65] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C. Chang, M. Yong, J. Lee, W. Chang, W. Hua, M. Georg and M. Grundmann, “Mediapipe: A Framework for Building Perception Pipelines,” In proceedings of IEEE Computer Vision and Pattern Recognition Workshop, 2019.

[66] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, Y. Qiao, VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking (Version 2), 2023, arXiv. https://doi.org/10.48550/ARXIV.2303.16727

[67] D. Li, C. R. Opazo, X. Yu and H. Li, "Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison," In proceedings

of IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass, CO, USA, 2020, pp. 1448-1458.

[68] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need (Version 7)," 2017, arXiv. https://doi.org/10.48550/ARXIV.1706.03762

[69] R. Zuo, F. Wei, and B. Mak, "Natural Language-Assisted Sign Language Recognition," in Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 2023, pp. 14890-14900.

[70] B. Ni, *et al.*, "Expanding Language-Image Pretrained Models for General Video Recognition," In *Lecture Notes in Computer Science*, vol 13664. Springer, Cham, 2022.

[71] S. Xie, C., Sun, J., Huang, Z., Tu, K., Murphy, "Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification," In V. Ferrari, M., Hebert, C., Sminchisescu, Y., Weiss (eds), ECCV 2018, vol 11219. Springer, Cham.

[72] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the 32nd Conference on Artificial Intelligence and 30th Innovative Applications of Artificial Intelligence, 2018, 7444–7452.

[73] O. Sincan, H. Keles, "AUTSL: A Large Scale Multi-modal Turkish Sign Language Dataset and Baseline Methods." *IEEE Access*, vol. 8, pp. 181340-181355, 2020

[74] A. Desai, L. Berger, F. Minakov, V. Milan, C. Singh, K. Pumphrey, R. Ladner, H. Daume III, A. Lu, N. aselli, D. Bragg, D., "ASL Citizen: A Community-Sourced Dataset for Advancing Isolated Sign Language Recognition," 2023 arXiv preprint arXiv:2304.05934.

[75] L. Kezar, J. Thomason, N. Caselli, Z. Sehyr, and E. Pontecorvo, "The Sem-Lex Benchmark: Modeling ASL Signs and their Phonemes," In Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility, 2023, NY, Article 34, pp. 1–10.

[76] H. Joze, O. Koller, "MS-ASL: A large-scale data set and benchmark for understanding American sign language," 2018, doi:10.48550/ARXIV.1812.01053

[77] H. Hu, W. Zhou, J. Pu, and H. Li., "Global-Local Enhancement Network for NMF-Aware Sign Language Recognition," *ACM Transactions on Multimedia*, Computer Communication Applications, vol. 17, no. 3, Article 80, 2021.

List of Publications

[1] M. Gochoo G. Batnasan, et al., "Fine-Tuning Vision Transformer for Arabic Sign Language Video Recognition on Augmented Small-Scale Dataset," In proceedings of IEEE International Conference on Systems, Man, and Cybernetics, Honolulu, USA, 2023, pp. 2880-2885, https://ieeexplore.ieee.org/document/10394501.

[2] G. Batnasan, Q. Memon, “Transformer Based Deep Learning Model for Improved Sign Language Recognition,” In *Networks and Machine Learning: Models, Algorithms, and Applications*, CRC Press, 2024

UAEU

جامعة الإمارات العربية المتحدة
United Arab Emirates University

UAEU MASTER THESIS NO. 2024:62

This thesis on sign language recognition aims to improve recognition that include augmentations and transformations. The augmentations and transformation helped increase the data size. Specifically, in-house signs were generated using different persons for initial results. The video frames generated included facial expressions and both fingers, which were later stacked. Later, the model was validated using generic sign languages.

Ganzorig Batnasan received his Master of Science in Electrical Engineering from the Department of Electrical and Communication Engineering, College of Engineering at the United Arab Emirates University, UAE. His research interest is in computer vision and machine learning.

www.uaeu.ac.ae