Universidade Estadual de Campinas
Instituto de Computação

Wladimir Arturo Garces Carrillo

Gesture recognition in Brazilian Sign Language (Libras) using Vision Transformer

Reconhecimento de gestos em Língua Brasileira de Sinais (Libras) utilizando Transformadores Visuais

CAMPINAS
2024

Wladimir Arturo Garces Carrillo

Reconhecimento de gestos em Língua Brasileira de Sinais (Libras) utilizando Transformadores Visuais

Gesture recognition in Brazilian Sign Language (Libras) using Vision Transformer

Dissertação apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Dissertation presented to the Institute of Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Computer Science.

Supervisor/Orientador: Prof. Dr. Marcelo da Silva Reis
Co-supervisor/Coorientadora: Dra. Emely Pujólli da Silva

Este exemplar corresponde à versão final da Dissertação defendida por Wladimir Arturo Garces Carrillo e orientada pelo Prof. Dr. Marcelo da Silva Reis.

CAMPINAS
2024

Ficha catalográfica
Universidade Estadual de Campinas (UNICAMP)
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

G164g

Garcés Carrillo, Wladimir Arturo, 1987-
Gesture recognition in Brazilian Sign Language (Libras) using Vision Transformer (ViT) / Wladimir Arturo Garcés Carrillo. – Campinas, SP : [s.n.], 2024.

Orientador(es): Marcelo da Silva Reis.
Coorientador(es): Emely Pujólli da Silva.
Dissertação (mestrado) – Universidade Estadual de Campinas (UNICAMP), Instituto de Computação.

1. Língua brasileira de sinais . 2. Redes neurais profundas. 3. Visão por computador. 4. Interação humano-máquina. 5. Reconhecimento de padrões. 6. Processamento de imagens. 7. Inteligência artificial. 8. Tecnologia assistiva. I. Reis, Marcelo da Silva, 1979-. II. Silva, Emely Pujólli da, 1990-. III. Universidade Estadual de Campinas (UNICAMP). Instituto de Computação. IV. Título.

Informações complementares

Título em outro idioma: Reconhecimento de gestos em Língua Brasileira de Sinais (Libras) utilizando Vision Transformer (ViT)
Palavras-chave em inglês:
Brazilian sign language
Deep neural networks
Computer vision
Human-machine interaction
Pattern recognition
Image processing
Artificial intelligence.
Assistive technology
Área de concentração: Ciência da Computação
Titulação: Mestre em Ciência da Computação
Banca examinadora:
Marcelo da Silva Reis [Orientador]
Ivani Rodrigues Silva
Hélio Pedrini
Data de defesa: 23-09-2024
Programa de Pós-Graduação: Ciência da Computação

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0009-0001-7904-2996
- Currículo Lattes do autor: http://lattes.cnpq.br/4301845793447809

Universidade Estadual de Campinas
Instituto de Computação

Wladimir Arturo Garces Carrillo

Gesture recognition in Brazilian Sign Language (Libras) using Vision Transformer

Reconhecimento de gestos em Língua Brasileira de Sinais (Libras) utilizando Transformadores Visuais

Banca Examinadora:

- Prof. Dr. Marcelo da Silva Reis
  Instituto de Computação/Universidade Estadual de Campinas
- Profa. Dra. Ivani Rodrigues Silva
  Faculdade de Ciências Médicas/Universidade Estadual de Campinas
- Prof. Dr. Hélio Pedrini
  Instituto de Computação/Universidade Estadual de Campinas

A ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Campinas, 23 de setembro de 2024

Dedicatória

Dedico este trabajo a mis dos madres: mi abuela Tilda Dolores Torreglosa y mi madre Margot Carrillo Torreglosa, cuyas sabias palabras, oraciones, apoyo y motivación me han permitido avanzar por los caminos de la vida. A mis sobrinos, quienes me ven como un ejemplo a seguir, aunque a veces yo mismo no lo crea. A mis hermanos, con quienes he compartido sonrisas y lágrimas, pero siempre hemos estado ahí para darnos aliento mutuo. También a Francesca Alvarado Florez, quien con su amor y paciencia me ha acompañado durante este largo período lejos de casa.

Dedico este trabajo a las comunidades sordas de Brasil y América Latina. No estamos solos; nuestros esfuerzos se unen para hacer de nuestro mundo un lugar mejor.

Finalmente, dedico este trabajo a mi orientador, el Profesor Doctor Marcelo Reis, y a mi coorientadora, la Doctora Emely da Silva, por su sabiduría y guía a lo largo de este proceso.

La caja de pandora

Siento una llama aquí dentro
Quema lo que se conoce por todo.
No siento frío ni calor,
Caliento el aire alrededor.

Por primera vez
El planeta me siente.
Por no percibirme antes,
Siente curiosidad.

Pájaros vienen a mi ventana
En las mañanas de frío,
Por tan ignea presencia.
Gatos e perros
Se acercan en las tardes
De mucho calor.

Siento una llama aquí adentro
Cómo si fuera un motor.
Aquel que calienta el aire de un globo,
Y lo eleva por los cielos.

Y es que...

Elevado voy
Cuando camino.
Elevado me siento,
Y es que no me puedo ni sentar.

Las piedras,
Mis amigas las piedras,
Me permiten caminar.

Una piedra en el zapato,
Tal vez dos o tres
Que no me permitan levitar.

Siento una llama aquí dentro,
Abre cajas.
Cajas que encienden llamas.
Llamas para no sentir frío ni calor.
Calor que globos elevan.
Elevado voy cuando camino
Camino con piedras en los zapatos.
Zapatos para que?...

...nunca más volverá a ser igual.

(WAGC)

Agradecimentos

Esta pesquisa foi financiada pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) com o Código de Financiamento 001, concedido pelo Instituto de Computação da Universidade de Campinas. Agradecemos o apoio fundamental da CAPES, que foi essencial para a realização desta proposta.

Agradecemos também à Samsung Eletrônica da Amazônia LTDA pelo apoio financeiro durante o período de cumprimento dos créditos das disciplinas, à *Diretoria do Instituto de Computação da Universidade Estadual de Campinas* por ter financiado três dos últimos cinco meses deste mestrado com a *Bolsa Alumni.* Da mesma forma, agradeço ao *Fundo de Apoio ao Ensino, Pesquisa e Extensão – FAEPEX* pelo financiamento dos últimos dois meses.

Agradeço aos meus colegas do Recod.ai, em especial a Gabriel, Paula, Manuel, Daniel, João, José e Diego, que sempre tiveram uma resposta para minhas dúvidas.

Resumo

Esta dissertação aborda o desafio da escassez de dados no reconhecimento de sinais da Língua Brasileira de Sinais (Libras) utilizando técnicas de aprendizado profundo, em particular o *Vision Transformer* (ViT). A falta de dados amplamente disponíveis e anotados corretamente dificulta o desenvolvimento de tecnologias de Reconhecimento Automático de Línguas de Sinais (ASLR), especialmente para Libras. Para enfrentar esse problema, propomos explorar diferentes abordagens de aumentação de dados sendo eles o padrão, e *modelos de difusão*. Além disso, utiliza-se o aprendizado por transferência, visando melhorar o desempenho em múltiplos conjuntos de dados de Libras. O trabalho também introduz uma abordagem inovadora de análise de movimento por meio de Imagens de Energia de Marcha Coloridas (CGEI), que permite capturar informações detalhadas sobre os sinais, contribuindo para o aumento da precisão dos modelos. Os modelos ViT e ResNet50 foram avaliados com base em dados de Libras, comparando o desempenho dessas arquiteturas em protocolos de divisão de dados estratificados e controlados. Os resultados demonstram que o uso de técnicas avançadas de aumento de dados, juntamente com os modelos de aprendizado profundo, se mostrou como uma solução promissora para lidar com a escassez de dados em Libras. Os modelos ViT, em particular, mostraram melhor desempenho em relação a outras abordagens previamente utilizadas para o problema. Espera-se que esta pesquisa contribua para o desenvolvimento de tecnologias mais precisas e acessíveis, promovendo a inclusão da comunidade surda no Brasil.

Abstract

This dissertation addresses the challenge of data scarcity in recognizing Brazilian Sign Language (Libras) signals using deep learning techniques, particularly *Vision Transformer (ViT)* . The lack of widely available and correctly annotated data hinders the development of Automatic Sign Language Recognition (ASLR) technologies, especially for Brazilian Sign Language (Libras). To tackle this issue, we propose exploring different data augmentation approaches, including standard methods and *diffusion models*. Additionally, transfer learning is utilized to enhance performance across multiple Libras datasets. The work also introduces an innovative approach to motion analysis through Color Gait Energy Image (CGEI), which allows for capturing detailed information about the signals, contributing to increased model accuracy. Vision Transformer (ViT) and Residual Network (ResNet50) models were evaluated based on Libras data, comparing the performance of these architectures in stratified and controlled data splitting protocols. The results demonstrate that the use of advanced data augmentation techniques, combined with deep learning models, proves to be a promising solution for addressing data scarcity in Libras. ViT models, in particular, showed better performance compared to other previously used approaches for this problem. This research is expected to contribute to the development of more accurate and accessible technologies, promoting the inclusion of the deaf community in Brazil.

Acronyms

3DCNN	3D Convolutional Neural Network	62, 64, 67
AFER	Automatic Facial Expression Recognition	28
ARSL	Arabian Sign Language	62
ASL	American Sign Language	33, 62, 63, 67
ASLLVD	American Sign Language Lexicon Video Dataset	26
ASLR	Automatic Sign Language Recognition	9, 26–28, 57, 60, 63, 65
BLSTM	Bidirectional Long Short-Term Memory	67
CEFET-Libras	Libras Database developed by CEFET-Rio de Janeiro	27, 28, 56, 57
CGEI	Color Gait Energy Image	9, 15, 73, 81, 93–97, 102–104, 107, 109, 110, 112–115, 123
CNN	Convolutional Neural Network	28, 46, 60, 62, 64–67, 76
CSL	Chinese Sign Language	62, 66, 67
CSS	Controlled Stratified Split	14, 15, 17, 87–89, 91, 92, 94–97, 102–104, 113, 114, 123

CTC	Connectionist Temporal Classification	67
CUST	Controlled User Split	15, 87, 97–102, 104, 113, 114, 123, 124
DB	database	28
DDPM	Denoising Diffusion Probabilistic Models	63, 66, 67
DL	Deep Learning	60, 62, 66
Elias Dataset	Libras Database	15, 27, 29, 56, 57, 71, 75, 84, 85, 90–97, 100–102, 106–114
FACS	Facial Action Coding System	58
FE	Facial Expression	28, 60
GAN	Generative Adversarial Network	64, 66, 67
GEI	Gait Energy Image	15, 37, 74, 75, 81, 93–96, 103, 105, 106, 108–111, 113–115
GSL	German Sign Language	62, 67
HCM	Histogram of Cumulative Magnitudes	62
HCR	Hand Configuration Recognition	60
HOG	Histogram of Oriented Gradients	62
ISDA	Implicit Semantic Data Augmentation	64
KNN	k-Nearest Neighbors	28, 60, 62, 64
LDA	Linear Discriminant Analysis	62
Libras	Língua Brasileira de Sinais	1, 2, 4, 8

Libras	Brazilian Sign Language	1, 2, 4, 9, 22, 27–29, 31–35, 55, 57, 58, 60–62, 66, 74, 123
LIBRAS-10	Libras Database, Libras-10 Dataset, Extension of Libras-34	27, 56, 57
LIBRAS-34	Libras Database, Libras-34 Dataset (Kinect v1)	27, 56, 57
LIBRAS-UFOP-ISO	Libras Database developed at the Federal University of Ouro Preto	28, 56, 58
LightGBM	Light Gradient Boosting Model	28
LSA	from Spanish, *Lenguaje de Señas Argentina*	62
LSTM	Long Short-Term Memory	62
MINDS-Libras	Libras Database developed by the Machine Intelligence and Data Science Laboratory - MINDS at the Federal University of Minas Gerais	15, 17, 28, 29, 56, 58, 71, 75, 85, 90–97, 100, 101, 103, 104, 113–115
NDA	Negative Data Augmentation	64
NMT	Neural Machine Traslator	67
PB	Brazilian Portuguese	27
PDF	Probability Density Function	48, 50, 52, 53
PHOENIX-2014	Continuous Sign Language Benchmark dataset	26, 67
ResNet50	Residual Network	9, 14, 15, 46, 75, 76, 84, 88, 89, 91–99, 101, 103, 113, 114, 123, 124
RF	Random Forest	28
RQ1	Research Question 1	29, 83, 84, 123

RQ2	Research Question 2	29, 83, 84, 87, 124
RSL	Russian Sign Language	62
SGD	Stochastic Gradient Descent	62
SGR	Static Gesture Recognition	60
SILFA	LIBRAS database developed by University of Campinas (UNICAMP) in partnership with Federal University of ABC (UFABC), called “Sign Language Facial Action Database”	28, 56, 58
SL	Sign Language	26, 58
SMOTE	Synthetic Minority Oversampling Technique (SMOTE)	28, 29, 64
SVM	Support Vector Machine	28, 62, 64
TF	Transformer	62, 67
UTA-ASL	University of Texas Arlington American Sign Language Dataset	26
ViT	Vision Transformer	9, 14, 15, 29, 46, 60, 62, 63, 67, 75–78, 80, 81, 83–86, 88–96, 98–104, 106–109, 111–114, 123, 124
XGBoost	eXtreme Gradient Boosting	28
XmDA	Cross Modality Data Augmentation	64, 65, 67
ZIM	Zernike Invariant Moment	62

List of Figures

List of Tables

Contents

Prefacio (Spanish Version)

Durante mis años de vida, he aprendido y soy fiel a la idea de que la comunicación es la principal herramienta para la construcción de sociedades. Esta no se limita solo a las sociedades humanas, ni tampoco a las animales. Si entendemos la comunicación como cualquier interacción en la naturaleza, encontraremos infinitas conexiones: entre plantas y animales, entre el comportamiento de los ríos o entre la proliferación de peces. Para muchas de estas interacciones, se han establecido sistemas de comunicación, algunos más complejos que otros, pero todos nacen de la necesidad de interactuar.

Reflexionando sobre esto, sin hacer ninguna discriminación entre especies o complejidades de sistemas, podemos identificar diversos modos de comunicación que son fundamentales para la interacción entre los seres vivos. La comunicación verbal, por ejemplo, se manifiesta a través del uso de palabras habladas o escritas, facilitando diálogos y conversaciones. En contraste, la comunicación no verbal abarca gestos que expresan ideas o emociones, como los movimientos de las manos y el cuerpo, así como expresiones faciales que reflejan cambios en el rostro para indicar emociones o reacciones. La postura corporal también desempeña un papel crucial, ya que la forma en que sostenemos nuestro cuerpo puede transmitir mensajes sobre nuestro estado emocional o nuestras intenciones.

Asimismo, la comunicación visual juega un papel importante, utilizando colores o luces para transmitir información. También se manifiesta en el dibujo y el arte, donde se crean imágenes para comunicar ideas o sentimientos. Por otro lado, la comunicación sonora incluye vocalizaciones, como gritos, cantos o murmullos, así como la música, que combina sonidos organizados para transmitir emociones o ideas.

Además, en el ámbito olfativo, encontramos feromonas que los animales utilizan para comunicarse, especialmente en contextos de apareamiento o marcaje de territorio, así como olores que transmiten información sobre identidad, estado de ánimo o ubicación. La comunicación táctil, que se da a través de toques, expresa afecto, dominio o advertencia, mientras que la danza utiliza movimientos corporales individuales o grupales (como en las danzas de las abejas) para transmitir información.

Estos modos de comunicación reflejan la diversidad de interacciones en el mundo natural. A partir de ellos, y considerando las diferencias técnicas y de complejidad que existen, las especies vivas han desarrollado distintos lenguajes, todos con la finalidad de interactuar con el entorno y expresar intención motivos, intereses e ideas de diversos tipos y niveles. En el caso de los humanos, esto también incluye la expresión de la gran creatividad que nos caracteriza, una necesidad intrínseca que se manifiesta en la "poíesis", el acto de crear que va más allá de la mera supervivencia y que se convierte en una forma de dar sentido a nuestra existencia. Esta capacidad de crear, de generar significado a través del lenguaje y otras formas de expresión, nos permite conectar profundamente con nuestro entorno y con los demás. El lenguaje se transforma así en una interfaz que nos ayuda a resolver el complejo problema de la vida misma.

Desde un punto de vista más antropocéntrico, el lenguaje natural forma parte de la vida; existe por sí mismo y está vivo. Varía según la ubicación geográfica, así como entre individuos, grupos etarios, grupos étnicos o pequeños grupos de personas que se identifican con variaciones de una lengua y las modifican coloquialmente, añadiendo matices que merecen ser respetados y valorados. La lengua evoluciona en relación a muchos factores, entre ellos la cultura y la sociedad, ofreciéndonos una ventana a las costumbres y tradiciones que se han tejido a lo largo del tiempo. Me gusta afirmar que el lenguaje existe como una entidad independiente y compleja que habita en un mundo separado de nosotros.

El lenguaje natural, en su riqueza y complejidad, ha sido uno de los problemas más estudiados y desafiantes para la inteligencia artificial. Comprender y emular la capacidad humana para procesar, interpretar y generar lenguaje implica enfrentarse a aspectos como la ambigüedad, el contexto y las variaciones culturales que se reflejan en cada lengua. Desde los primeros intentos de traducir oraciones simples hasta los modelos modernos capaces de mantener conversaciones complejas y contextuales, la IA ha intentado desentrañar las reglas explícitas e implícitas del lenguaje, desarrollando herramientas y algoritmos que emulan en parte la capacidad humana de comunicación.

El estudio del lenguaje en la IA ha permitido avances significativos en tecnologías como el procesamiento de lenguaje natural (NLP), que se aplica en asistentes virtuales, traductores automáticos y modelos de generación de texto. Sin embargo, esta tarea sigue siendo un reto, ya que el lenguaje humano es dinámico y evoluciona continuamente, reflejando no solo información objetiva, sino también emociones, pensamientos complejos e incluso aspectos de la identidad individual y colectiva. La inteligencia artificial, al intentar comprender este fenómeno, busca no solo responder preguntas o traducir frases, sino también capturar la esencia del lenguaje como medio de expresión y vínculo social. En este sentido, el lenguaje no solo se estudia como un problema técnico, sino como una ventana hacia las complejidades y sutilezas de la experiencia humana.

A pesar de los grandes avances, existen desafíos específicos cuando se trata de lenguas de señas, ya que estas no solo dependen de gestos manuales, sino también de expresiones faciales, movimientos corporales y el contexto en que se realizan. Las lenguas de señas son sistemas visuales-espaciales únicos que, a diferencia de los idiomas orales, requieren una interpretación precisa de elementos visuales en movimiento. Esto añade una capa de complejidad en los modelos de inteligencia artificial, que deben captar tanto la dinámica espacial como la sincronización de múltiples señales visuales en tiempo real para lograr una interpretación adecuada. El desafío reside en desarrollar algoritmos que puedan entender y generar estas lenguas de manera fluida, lo que representa un área de investigación activa y esencial para garantizar accesibilidad e inclusión de las comunidades sordas en el ámbito social.

La investigación en inteligencia artificial ha alcanzado un punto crucial en el cual la interacción entre humanos y máquinas puede beneficiarse enormemente de herramientas avanzadas de reconocimiento gestual, especialmente en el ámbito de las lenguas de señas. Como investigador en Ciencias de la Computación y apasionado por la innovación en IA, encontré en el problema de la clasificación de señales gestuales un desafío estimulante que me ha permitido conjugar mi interés por la visión por computadora, el aprendizaje profundo y el diseño de soluciones que acerquen la tecnología a las necesidades humanas.

Mi motivación por este proyecto no solo proviene de la curiosidad técnica, sino de experiencias personales con personas cercanas que enfrentan dificultades auditivas. Estos casos me han sensibilizado sobre la importancia de la accesibilidad en la comunicación y me han llevado a ver la inteligencia artificial como una herramienta poderosa para resolver problemas complejos de nuestra sociedad. Desde esta perspectiva, este trabajo pretende contribuir a mejorar el reconocimiento de señales en Libras, enfrentando el desafío de la escasez de datos representativos y explorando el potencial de modelos avanzados, como Vision Transformers, para optimizar la precisión y adaptabilidad en la clasificación de gestos.

Este estudio se basa en dos enfoques: el uso de modelos de aprendizaje por transferencia y el aumento de datos mediante técnicas de difusión, aplicados a datasets de Libras como Elias y MindsLibras. La metodología, los resultados y conclusiones reflejan el compromiso con la mejora de modelos en entornos desafiantes. Además, los Objetivos de Desarrollo Sostenible (ODS) han guiado este proyecto, especialmente en el objetivo de lograr una educación inclusiva y accesible, y fomentar la innovación en tecnología para la inclusión social.

Este proyecto ha sido un viaje de descubrimiento y superación, en el cual he contado con el apoyo invaluable de mi orientador, família, colegas y colaboradores. Presentar estos hallazgos en la Reunión Internacional de Inteligencia Artificial y sus Aplicaciones (RIIAA) ha sido una experiencia que reafirma la relevancia de esta investigación. Espero que este trabajo inspire a otros a continuar innovando y contribuyendo al desarrollo de tecnologías inclusivas y accesibles para todos.

Preface (English Version)

Throughout my life, I have learned and firmly believe that communication is the primary tool for building societies. This concept extends beyond human societies and even beyond the animal kingdom. If we understand communication as any form of interaction in nature, we discover infinite connections: between plants and animals, the behavior of rivers, or the proliferation of fish. Many of these interactions have established communication systems, some more complex than others, but all are born from the need to interact.

Reflecting on this, and without discriminating between species or the complexities of systems, we can identify various forms of communication that are fundamental to interaction among living beings. Verbal communication, for instance, manifests through spoken or written words, facilitating dialogues and conversations. In contrast, non-verbal communication includes gestures that express ideas or emotions, such as hand and body movements, as well as facial expressions that reflect changes in the face to indicate emotions or reactions. Body posture also plays a crucial role, as the way we hold our bodies can convey messages about our emotional state or intentions.

Similarly, visual communication plays an important role, using colors or lights to convey information. It also manifests in drawing and art, where images are created to communicate ideas or feelings. On the other hand, auditory communication includes vocalizations, such as shouts, songs, or murmurs, as well as music, which combines organized sounds to convey emotions or ideas.

Additionally, in the realm of olfactory communication, we find pheromones that animals use to communicate, especially in contexts of mating or territory marking, as well as scents that convey information about identity, mood, or location. Tactile communication, which occurs through touch, expresses affection, dominance, or warning, while dance uses individual or group body movements (such as the dances of bees) to transmit information.

These modes of communication reflect the diversity of interactions in the natural world. From them, and considering the technical and complexity differences that exist, living species have developed distinct languages, all aimed at interacting with their environment and expressing intentions, motives, interests, and various types and levels of ideas. In the case of humans, this also includes the expression of the great creativity that characterizes us, an intrinsic need that manifests in "poíesis", the act of creating that goes beyond mere survival and becomes a way of giving meaning to our existence. This capacity to create, to generate meaning through language and other forms of expression, allows us to connect deeply with our environment and with others. Language thus transforms into an interface that helps us solve the complex problem of life itself.

From a more anthropocentric perspective, natural language is part of life; it exists in itself and is alive. It varies according to geographic location, as well as among individuals, age groups, ethnic groups, or small groups of people who identify with variations of a language and modify it colloquially, adding nuances that deserve to be respected and valued. Language evolves in relation to many factors, including culture and society, offering us a window into the customs and traditions that have been woven over time. I like to assert that language exists as an independent and complex entity that inhabits a world separate from us.

Natural language, in its richness and complexity, has been one of the most studied and challenging problems for artificial intelligence. Understanding and emulating the human ability to process, interpret, and generate language involves confronting aspects such as ambiguity, context, and cultural variations that are reflected in each language. From the earliest attempts to translate simple sentences to modern models capable of maintaining complex and contextual conversations, AI has tried to unravel the explicit and implicit rules of language, developing tools and algorithms that partly emulate the human capacity for communication.

The study of language in AI has led to significant advances in technologies such as natural language processing (NLP), which is applied in virtual assistants, machine translators, and text generation models. However, this task remains a challenge, as human language is dynamic and continuously evolves, reflecting not only objective information but also emotions, complex thoughts, and even aspects of individual and collective identity. Artificial intelligence, in trying to understand this phenomenon, seeks not only to answer questions or translate phrases but also to capture the essence of language as a means of expression and social bond. In this sense, language is studied not only as a technical problem but also as a window into the complexities and subtleties of human experience.

Despite significant advances, there are specific challenges when it comes to sign languages, as they rely not only on hand gestures but also on facial expressions, body movements, and the context in which they are performed. Sign languages are unique visual-spatial systems that, unlike spoken languages, require precise interpretation of visual elements in motion. This adds a layer of complexity to artificial intelligence models, which must capture both the spatial dynamics and the timing of multiple visual signals in real-time to achieve adequate interpretation. The challenge lies in developing algorithms that can understand and generate these languages fluently, representing an active and essential area of research to ensure the accessibility and inclusion of deaf communities in the social realm.

Research in artificial intelligence has reached a crucial point where the interaction between humans and machines can greatly benefit from advanced gesture recognition tools, especially in the realm of sign languages. As a researcher in Computer Science and passionate about innovation in AI, I found the challenge of gesture signal classification to be a stimulating problem that has allowed me to combine my interest in computer vision, deep learning, and the design of solutions that bring technology closer to human needs.

My motivation for this project stems not only from technical curiosity but also from personal experiences with close individuals facing hearing difficulties. These cases have sensitized me to the importance of accessibility in communication and led me to see artificial intelligence as a powerful tool for solving complex problems in our society. From this perspective, this work aims to contribute to improving sign recognition in Brazilian Sign Language (Libras), tackling the challenge of the scarcity of representative data and exploring the potential of advanced models, such as Vision Transformers, to optimize accuracy and adaptability in gesture classification.

This study is based on two approaches: the use of transfer learning models and data augmentation through diffusion techniques, applied to Libras datasets such as Elias and

MindsLibras. The methodology, results, and conclusions reflect a commitment to enhancing models in challenging environments. Furthermore, the Sustainable Development Goals (SDGs) have guided this project, particularly the goal of achieving inclusive and accessible education and promoting innovation in technology for social inclusion.

This project has been a journey of discovery and overcoming challenges, during which I have had the invaluable support of my advisor, family, colleagues, and collaborators. Presenting these findings at the International Meeting on Artificial Intelligence and Its Applications (RIIAA) has been an experience that reaffirms the relevance of this research. I hope that this work inspires others to continue innovating and contributing to the development of inclusive and accessible technologies for all.

Chapter 1

Introduction

In this chapter, we start by providing background and motivation of this research, emphasizing the importance of improving gesture recognition in sign language. We then define the specific problem addressed in this work, followed by our proposed solution, which leverages advanced classification models for gesture recognition. Next, we present the research questions that guide this study and the main objectives we aim to achieve. Finally, we outline the structure of the dissertation, giving the reader an overview of its organization and flow.

1.1 Background and Motivation

The lack of accessibility for deaf communities remains one of the main challenges currently faced. Although there have been significant advances in public policies and technology, the deaf community still struggles to communicate and access basic services [5, 30, 54, 93]. Additionally, deaf individuals often face stigmas and prejudices that impact their social inclusion. It is important to emphasize that hearing impairment does not limit intellectual ability. However, negative stereotypes and a lack of understanding from employers often result in barriers to employment and restrict their access to the job market [5].

These situations highlight the importance of promoting and implementing measures that ensure inclusion and equal opportunities for deaf individuals. One of the strategies adopted by the scientific community is the preservation and documentation of the language through inventories that collect the various signs present in a given territory [3, 20, 65]. These collections form the corpora of a Sign Language and are continually maintained by the scientific community [6, 37, 78]. Different countries have been working to gather as much data as possible related to their country's specific sign language [3, 4, 11, 20, 37, 78, 97].

Some solutions have been explored to improve accessibility and inclusion within the deaf community, especially for Automatic Sign Language Recognition (ASLR) [8, 46, 47]. However, the available corpora are not ready-to-use datasets for these technologies. They are scarce and have a limited number of samples. Although there are larger datasets, such as American Sign Language Lexicon Video Dataset (ASLLVD) [13], University of Texas Arlington American Sign Language Dataset (UTA-ASL) [27], and Continuous Sign Language Benchmark dataset (PHOENIX-2014) [70], they still contain only a small number of samples for each sign.

In Brazil, the Brazilian Sign Language (Libras) is the chosen language of the deaf community for communication and is spoken by approximately 5% of the Brazilian population. It is recognized as a language and has regulation associated with Law 10.436/2002 [1], including Decree 5.626/2005 [2]. Libras is a linguistic system with all the characteristics found in spoken languages and is constructed through signs composed of the following parameters: configuration, point of articulation, and hand movement; orientation/direction of the palm; facial and body expressions [30], which can be seen in Figure 1.1. Lack of awareness of these laws, coupled with limited contact with Libras, can be considered limiting factors for communication and social inclusion of the deaf. Furthermore, the teaching of Libras in schools is not comprehensive, and there are few bilingual school options. This contributes to deaf individuals having fewer opportunities for education, employment, and participation in society.

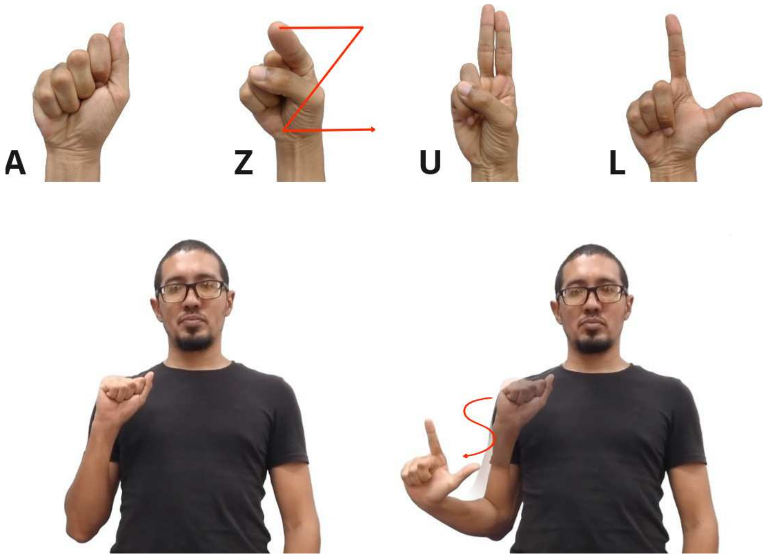


Figure 1.1: Image of the fingerspelling of AZUL and the sign "AZUL" in Libras.

Both academic studies and government actions have focused on improving accessibility for the deaf community. The Brazilian government invests in technologies that have facilitated deaf individuals' access to information. Notably among these projects is VLibras [54], which consists of an automatic translator from Libras to Brazilian Portuguese (PB) incorporated into most of the country's web pages. Additionally, several research groups have been dedicated to creating databases that drive ASLR technologies. Some examples are:

- Libras Database, Libras-34 Dataset (Kinect v1) (LIBRAS-34) [10];
- Elias Dataset [107];
- Libras Database, Libras-10 Dataset, Extension of Libras-34 (LIBRAS-10) [9];
- Libras Database developed by CEFET-Rio de Janeiro (CEFET-Libras) [53];

- MINDS-Libras [92];
- Libras Database developed at the Federal University of Ouro Preto (LIBRAS-UFOP-ISO) [23];
- SILFA [98].

These databases are essential for the development of increasingly advanced and accurate technologies for communication between deaf and hearing individuals. All sign language databases represent significant progress in this regard and should be valued and encouraged.

Within ASLR, hand tracking and hand configuration recognition are considered static processes [26]. In this sense, several studies have been developed to improve these processes. Some works propose recognition methods based on k-Nearest Neighbors (KNN) models [28]. Other studies have developed frameworks for real-time hand tracking using a single RGB camera, based on the hybridization of two models: a palm detector and a landmark model to predict the hand's skeleton [122]. In Libras, projects like HandArch [33] are capable of real-time hand pose recognition in videos to accelerate the development of ASLR applications.

Dynamic processes in ASLR involve recognizing isolated signs and sentences. In this context, one of the main challenges is the complexity of movement and shape patterns. The work of Gameiro et al. [53] proposes KNN and Random Forest (RF) algorithms for sign recognition based on computer vision, achieving an average accuracy of 65.81% on their own database (DB) of CEFET-Libras videos. Algorithms like Support Vector Machine (SVM), Light Gradient Boosting Model (LightGBM), or eXtreme Gradient Boosting (XGBoost) have also been tested and compared, incorporating various preprocessing techniques, achieving higher global accuracies depending on the DB used [83]. Other projects go a step further, focusing on methods for Automatic Facial Expression Recognition (AFER) in Libras through Convolutional Neural Network (CNN) combined with feature extraction approaches for Facial Expression (FE) [30].

The linguistic and cultural diversity of the country, coupled with the scarcity of resources for promoting accessibility and inclusion, makes the construction of comprehensive and accurate databases complex and challenging. Furthermore, the lack of standardization in sign languages used in different contexts and regions can lead to variations in message interpretation, further complicating the task of building reliable databases, which also represents a significant challenge in the development of ASLR technologies. These situations create a landscape of scarcity of annotated and standardized Libras video data.

1.2 Problem Definition

The scarcity of Libras data represents a significant challenge for the development of automatic sign language recognition technologies. To overcome this challenge, one possibility is to resort to data augmentation. One approach used by Passos et al. [83] is the application of augmentation techniques in feature space. The work involves evaluating three augmentation techniques in this space, called Synthetic Minority Oversampling Technique, Borderline-SMOTE, and Borderline-SMOTE SVM, which are commonly employed for data with class imbalance issues. Additionally, there are basic transformation techniques that perform augmentation in sample space. These transformations include: spatial transformations, such as translation, horizontal and vertical flipping, resizing, among others; and temporal transformations, such as scaling the duration of the sequence or warping in the time domain.

Zanon de Castro et al. [120] highlight the importance of these techniques in the quest to increase the volume of available data. Another possibility is the usage of transfer learning techniques, where resources learned from one task are used to improve performance in a related task, can be a promising approach. Indeed, the exploration of both possibilities is still an open problem.

1.3 Our Proposal

To tackle the aforementioned problem of data scarcity in the field of Libras, we propose a methodology based on two branches: in the first one, we make use of transfer learning through the usage of state-of-the-art vision transformer models; in the second one, we develop a data augmentation process based on diffusion models, which is trained using different Libras datasets such as Elias Dataset and MINDS-Libras.

1.4 Research Questions

Having defined the problem to be tackled and also the proposed approaches, the research questions (RQs) we aimed to answer were:

> **Research Question 1 (RQ1)**
> For a given Libras dataset (augmented or not), how does a Vision Transformer (ViT) performance compare to the machine learning models previously used for this problem?
>
> **Research Question 2 (RQ2)**
> Is it possible to augment data from a Libras database using data from other Libras datasets?

Those research questions guided the work developed in this Master's, whose details are presented in the following chapters. The answers we obtained to these research questions are provided in the dissertation's conclusions (Chapter 6).

1.5 Main Objectives

We believe that this work significantly contribute to the development of more precise and efficient solutions in automatic Libras recognition, since the combination of data augmentation techniques and ViT can help overcome challenges related to data scarcity, advancing the development of assistive technologies; thus, it is expected to improve accessibility for deaf communities.

1.6 Outline of the Dissertation

This Master's dissertation is organized as follows: in Chapter 2, we introduce fundamental concepts for reading this dissertation, including a presentation of Libras, moviment analysis techniques, transformers and difusion models; in Chapter 3, we present a review of related works, including recent works on sign language datasets, sign language classification and data augmentation for sign language; in Chapter 4, we describe the methodology developed in this dissertation, with dataset choice, data preprocessing and split, generative and classification models, and computational protocol of the experiments; in Chapter 5, we show and discuss the main results obtained in the computational experiments; finally, in Chapter 6, we recall the contents of the dissertation and make the final remarks on the classification and data augmentation experiments.

Chapter 2

Fundamental Concepts

In this chapter, we explore the fundamental concepts that underpin this work. We begin with a historical overview of Libras, highlighting its main linguistic and cultural features. Next, we focus on the motion analysis technique employed, which is crucial for understanding gestures. We then introduce concepts related to the attention model, covering both the transformer and its variant for computer vision, the Vision Transformer, which is the core model of our study. Finally, we discuss the principles of diffusion models, which also play an important role in this work.

2.1 Brazilian Sign Language - Libras

Brazilian Sign Language (Libras) is a visual-gestural-spatial language with its own grammar and linguistic structure, distinct from spoken Portuguese. It is a complete linguistic system that encompasses all the qualities present in spoken languages, making it a crucial means of communication for the deaf community in Brazil. This language allows deaf individuals to communicate effectively, expressing their ideas and feelings in a rich and complex manner [37, 38, 50, 72, 76, 101].

Libras is the preferred and officially recognized language for communication within the deaf community [1, 2, 72], representing about 5% of the Brazilian population. Federal Law No. 10.436/2002 [1] recognizes it, and Decree 5.626/2005 [2] made it legal, highlighting its importance as a linguistic system for conveying ideas and facts within the Brazilian deaf community [38]. This legal recognition was a significant milestone for the empowerment and inclusion of the deaf population, reaffirming their linguistic and cultural rights. Libras became the first sign language in Brazil to be included in the National Inventory of Linguistic Diversity, according to De Quadros et al. [38].

Formally, the education and integration of the deaf began in the 19th century, with the creation of the Imperial Institute of Deaf-Mutes in 1857, under the leadership of Ernest Huet [67, 35, 94]. The enactment of Federal Law No. 10.436 in 2002, which formally recognized Libras as the language of the deaf and gave it legal status in Brazil, marked an important milestone for the inclusion and rights of the deaf. Currently, Libras is being valued and accepted as an important language for the social processes of the deaf in Brazil, being officially recognized as a heritage of the deaf communities [1, 2, 38, 67, 72, 94].

However, deaf communities continue to face many barriers and prejudices. Despite official recognition, many deaf people are still denied the right to quality education, employment, and public services, and generally a dignified existence. Persistent prejudices include the belief that deaf people lack intellectual and cognitive capacities, which limits their participation in social and economic dynamics. These barriers hinder the full exercise of their rights and their integration into society [5, 67, 93].

Additionally, the teaching of Libras in schools is not comprehensive, and there are only a few options for bilingual schools. Many educational institutions do not include Libras as part of the curriculum, resulting in a lack of familiarity with the language among teachers and hearing students. The lack of Libras interpreters in educational institutions and other essential services is a critical issue. This can prevents the deaf from having equal access to education and other services, affecting their academic performance and social inclusion. Furthermore, the shortage of qualified interpreters means that many deaf people face significant difficulties in accessing health, justice, and other public services [50].

Socially, there is still a lack of awareness about the needs and rights of deaf people, perpetuating exclusion. Prejudices and negative stereotypes are common, and many deaf people are erroneously perceived as having inferior intellectual capacities. This leads to marginalization and discrimination in various areas, including the job market, where employment opportunities for the deaf are often limited.

In the technological field, the scarcity of data on Libras - including insufficient, unlabeled, non-standardized data - and the visual complexity of Libras phonology make it difficult to develop viable technologies for this community. Tools such as automatic translators, sign recognition software, and other technological resources face challenges due to the lack of high-quality data and the complex nature of the signs. Moreover, research and development in these areas are often underfunded, limiting progress in creating technological solutions that could significantly improve the quality of life for deaf people.

These problems highlight the need for more effective public policies and greater investment in education, interpreter training, and technological development to ensure that deaf people have the same opportunities and rights as the rest of the population.

Libras is now evolving and thriving as a vibrant and expressive language, exemplifying the rich diversity and tenacity of the Brazilian deaf community. It serves as a powerful symbol of identity and pride for the deaf, connecting them to their cultural history and empowering them to participate in all aspects of Brazilian culture actively.

2.1.1 Linguistic Characteristics

Libras is a complex and rich mode of communication that combines visual, gestural, and spatial elements. Communication occurs through gestures, facial expressions, and bodily movements, resulting in a fully autonomous linguistic system. Contrary to popular belief, Libras is more than just a translation of spoken Portuguese into gestures; it is a rich and expressive linguistic system capable of conveying nuances and complexities of meaning [18, 30, 36, 39, 101].

Libras' unusual grammar is intriguing. For example, word order in a phrase may differ from that in Portuguese, and specific resources exist to convey tense, aspect, mood, and negation. These grammatical features are required for comprehending and creating messages in Libras, exhibiting its richness and complexity as a fully evolved language [36].

Furthermore, it is crucial to stress Libras' phonological qualities. The phonology of sign languages, including Libras, is critical for the understanding of the structure and creation of signs. Each Libras sign consists of a variety of phonological aspects, including hand configuration, position, movement, palm orientation, and facial expressions, according to Stokoe [101], Da Silva et al. [30], De Quadros et al. [36], De Matos et al. [35] and Kumada [72]. These characteristics contribute to the variety of Libra signs and help to distinguish meanings and understand given signals, allowing users to communicate clearly and effectively. As a result, when analyzing the linguistic qualities of Libras, it is critical to acknowledge its complexity and originality. Libras is more than just another mode of communication for the deaf; it is a complete and alive language that embodies the culture and identity of the Brazilian deaf community. Understanding and honoring Libras is critical for encouraging inclusivity and guaranteeing equitable communication opportunities for everybody.

2.1.1.1 Hand Configuration (HC)

One of the distinguishing characteristics of sign languages is the range of hand configurations used to produce signs. The number of signs varies according to the author. According to Ferreira-Brito [51, 52], this system includes 46 hand configurations (HCs), similar to what is found in American Sign Language (ASL), although each sign language may have its own set of HCs. These signs are grouped based on their similarity, although they have not yet been classified as basic or variant. Therefore, these HCs only refer to surface manifestations, i.e., the phonetic level found in Libras. However, according to Felipe [49], there exist 64 HCs, including variations; moreover, recent recent studies report about 111 signs.

Each hand configuration is made up of a set of features, including finger form, hand position, and movement direction. These aspects combine to provide distinct meanings within the context of sign language. For example, extending or curling the fingers, as well as the palm orientation can radically change the meaning of a sign. Some hand configurations can also be used as icons to graphically depict objects, animals, or abstract notions. Precision in executing these hand configurations is critical for clarity and understanding in sign language communication. Signers must have fine motor skills as well as a comprehensive awareness of gestural nuances in order to effectively use varied hand configurations in verbal expressions. Therefore, signers must have fine motor skills as well as a comprehensive awareness of gestural nuances in order to effectively use varied hand configurations in verbal expressions.

2.1.1.2 Localization (L)

In sign linguistics, location refers to the area of space surrounding the body where signs are made. This quality is critical for comprehending and interpreting the meanings communicated by Libras and other sign languages. The particular areas used for sign production may differ depending on the thoughts or grammatical categories being represented [35, 36, 101].

For example, some animal signs may be generated close to the body (e.g., lion, cat, dog), whereas signs denoting distant items or locations may be done further out from the body (e.g., plate, pot, place, house, school, university). This spatial distinction not only helps Libras communicate clearly and effectively, but it also mirrors how deaf people perceive and organize their surroundings.

2.1.1.3 Moviment (M)

Sign language phonology relies heavily on hand and body movements. These movements are extremely dynamic and can vary in several ways, including direction, speed, amplitude, and fluidity. These differences are critical for communicating nuances of meaning and grammatical categories in Libras and other sign languages around the world [18, 51].

For example, the direction of movement might represent the direction of an object or action, whereas speed can suggest the intensity or urgency of the action. The amplitude of movement can emphasize the size of a region or the amount of something, whereas fluidity might show the smoothness or continuity of an action. All of these combined properties enable elaborate and exact expression in sign language communication [51].

In Libras, one significant example is the employment of quick hand movements to signify plurality or intensity in some signs. This quick succession of gestures might be read as several instances of the same item or action, giving depth and complexity to the gestural language.

2.1.1.4 Palm Orientation (PO)

The palm orientation plays a crucial role in the phonology of Libras, where several distinct orientations are recognized. This element refers to the direction in which the palm of the hand is facing when producing a sign and is essential for conveying specific meanings and linguistic nuances. In Libras, six palm orientations are identified: upward, downward, toward the body, forward, to the right, and to the left [51].

For example, when the palm of the hand is facing toward the signer's body, it generally indicates a reference to something related to the signer themselves, such as self-description or an action performed by them. On the other hand, when the palm of the hand is facing forward, it often suggests a reference to something external to the signer, such as an object, person, or event in the surrounding environment (as observed in the sign "NAME", where the palm facing the body means "My name" and facing forward means "Your name"). Additionally, the different palm orientations in Libras can be combined with other linguistic elements, such as movement, facial expressions, and specific locations, to create an even broader range of meanings and communicative contexts. This variation in palm orientation is fundamental to enriching the language [51].

Therefore, understanding the role of palm orientation in Libras is essential and contributes to a rich and meaningful linguistic interaction within the deaf community and beyond [51].

2.1.1.5 Non-gestual Expresions

The role of non-manual expressions in sign languages goes beyond simply accompanying hand movements. They play a multifaceted role in communication, contributing to the syntactic and semantic understanding of sentences. Facial expressions, including eye, mouth, and eyebrow movements, along with body movements such as head and torso gestures, are fundamental for conveying nuances and details in Libras [30, 42, 91].

Syntactically, non-manual expressions are used to mark different linguistic constructions. For example, they can indicate whether a sentence is a yes-no or a WH-question. Additionally, these expressions can be employed to highlight specific elements in the sentence, such as the topic of conversation or negation [30, 42, 91].

At the lexical level, non-manual expressions have the ability to modify the meaning of signs. They can mark specific references, such as pointing to a person or object in the environment, or indicate pronominal reference in a conversation. Moreover, these expressions can function as grammatical markers, such as the negative particle or adverb of manner.

It is important to note that several non-manual expressions can occur simultaneously, providing additional layers of meaning and complexity to communication in Libras. The dynamic interaction between hand gestures, facial expressions, and body movements allows users of sign language to express themselves in a rich and precise manner, thereby enriching communication and understanding within the deaf community Guimaraes and Maestri [55], dos Santos Paiva et al. [42], Rezende et al. [91].

2.1.2 Importance of Libras in Brazilian Society

The significance of Brazilian Sign Language (Libras) transcends mere linguistic utility; it embodies the fundamental principles of inclusivity, cultural preservation, and social equity within Brazilian society. Libras serves as a vital conduit for the expression of thought, emotion, and identity among the deaf community, fostering a sense of belonging and empowerment.

At its core, Libras is a catalyst for communication equality, breaking down barriers that hinder meaningful interaction between deaf and hearing individuals. By providing a robust means of expression for the deaf, Libras facilitates access to education, employment opportunities, healthcare services, and legal proceedings, thereby safeguarding their fundamental rights and enhancing their quality of life.

Moreover, Libras plays a pivotal role in preserving and promoting the rich cultural heritage of the deaf community in Brazil. Through its distinct linguistic features and expressive capabilities, Libras embodies a unique cultural identity, fostering pride and solidarity among its users. This cultural significance extends beyond mere communication; it fosters a sense of community, strengthens social bonds, and fosters inter-generational transmission of values, traditions, and narratives.

In a broader societal context, the recognition and promotion of Libras represent a commitment to diversity, inclusion, and social justice. By acknowledging Libras as an official language and investing in its dissemination and accessibility, Brazilian society affirms its dedication to upholding the rights and dignity of all its citizens, regardless of linguistic or sensory differences.

Furthermore, the integration of Libras into various sectors of society, including education, media, and public services, not only enhances accessibility but also enriches the cultural tapestry of the nation. Embracing linguistic diversity fosters a more inclusive and equitable society, where individuals of all backgrounds can fully participate, contribute, and thrive.

In conclusion, the importance of Libras in Brazilian society embodies the principles of equality, cultural identity, and social cohesion. By recognizing, supporting, and celebrating Libras, Brazil reaffirms its commitment to building a more inclusive, diverse, and harmonious society for present and future generations.

2.1.3 Challenges and Future Perspectives

Despite significant advances in gaining recognition and promoting inclusion through Brazilian Sign Language (Libras), there are still substantial challenges that need to be addressed to ensure its widespread acceptance and accessibility:

- **Limited Access to Education and Resources:** One of the main challenges faced by Libras users is the limited availability of educational resources and qualified instructors proficient in the language. Many deaf individuals struggle to access quality education in Libras, hindering their academic and professional development.

- **Linguistic Discrimination and Stigmatization:** Despite its official recognition, Libras still faces linguistic discrimination and stigmatization in Brazilian society. Negative attitudes towards sign languages persist, leading to social exclusion and barriers to employment, healthcare, and other essential services for deaf individuals.

- **Digital Divide and Technological Accessibility:** The digital divide poses a significant obstacle for deaf individuals, as many online resources and services are not accessible in Libras. Improving technological accessibility and promoting the development of digital content in Libras are essential steps to ensure equal opportunities for the deaf in the digital age.

- **Legal and Policy Implementation:** Although the legal framework for protecting the rights of the deaf, including the recognition of Libras, is in place, effective implementation and enforcement of these laws are often lacking. There is a need for greater advocacy and policy initiatives to ensure that the rights of the deaf are respected in practice.

- **Empowerment and Representation:** Empowering the deaf to become active participants in decision-making processes and advocating for their rights is crucial for advancing the status of Libras in Brazilian society. Increasing the representation of deaf individuals in political, cultural, and professional spheres can help challenge stereotypes and promote greater inclusivity.

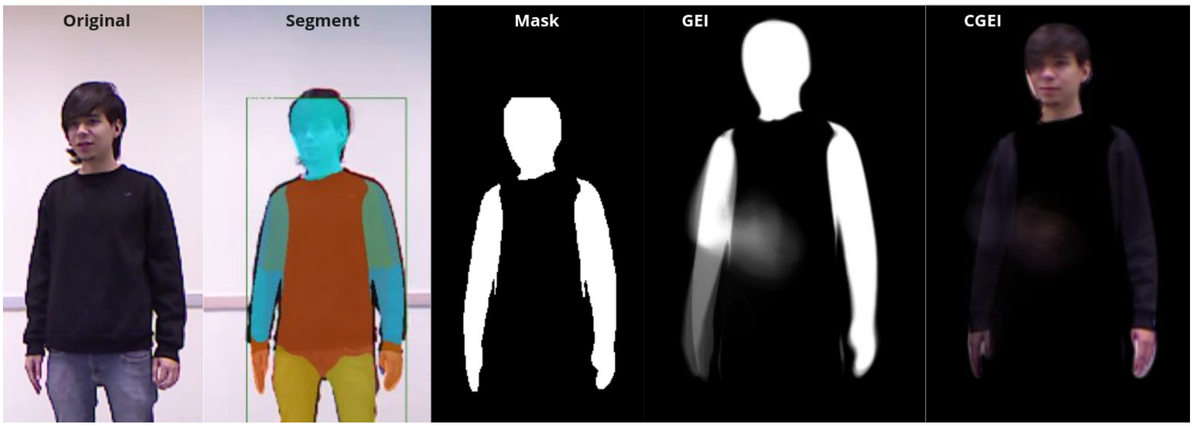
Despite these challenges, there are promising prospects for the future development and promotion of Libras in Brazil. Collaboration between government agencies, educational institutions, advocacy groups, and the deaf community is essential to address these challenges and promote a more inclusive society where the rights and linguistic diversity of all individuals are respected and preserved.

2.2 Movement Analysis techniques

2.2.1 Gait Energy Image (GEI)

Gait Energy Image (GEI) is a visual representation that captures crucial information about an individual's gait pattern. This concept arose from the need to quantify and objectively analyze human gait, a fundamental aspect of physical functionality and an important indicator of overall health. The GEI is generated from video sequences or sensor data that record body movements during walking. This technique utilizes image processing algorithms to extract significant features of the movement pattern and create a compact and informative visual representation (Figure 2.1.

The analysis of the Gait Energy Image has various applications in fields such as biomechanics, physical rehabilitation, elderly monitoring, and diagnosis of neurological disorders. By analyzing an individual's GEI, researchers and healthcare professionals can identify abnormal gait patterns that may indicate health issues, musculoskeletal injuries, or neurological impairments. Additionally, the GEI can be used to evaluate the effectiveness of therapeutic interventions and rehabilitation programs, providing objective feedback on the patient's progress over time.



(a) Original (b) Segment (c) Mask (d) GEI (e) CGEI

Figure 2.1: Different stages to calculate GEI and its variants. a) The original image (Source: Vidalón and Martino [107]), b) The image with the overlapping segments, c) Mask, d) GEI representation, e) GEI Color Representation. Source: Images taken from the dataset.

The GEI has also found application in sign language identification and recognition. Although initially developed to analyze walking patterns, its ability to capture distinctive movement characteristics can also be leveraged to interpret gestures and signs used in sign language communication. GEI analysis can extract important information about the dynamics and shape of gestures, enabling the development of sign language recognition systems based on computer vision. These systems have the potential to assist in communication between deaf individuals and hearing individuals, as well as in the automatic interpretation of gestures in environments where sign language is used.

The mathematical representation of the GEI is simple and can be interpreted as the average of each pixel in each frame of the sequence of images. The equation representing the GEI is:

$$G(x, y) = \frac{1}{N} \sum_{t=1}^{N} B_t(x, y), \tag{2.2.1}$$

where B_t and G represent, respectively, the mask sequences and the final GEI representation. x and y are the coordinates of each pixel in each sequence image, and N is the total number of frames in the video [16, 58, 83].

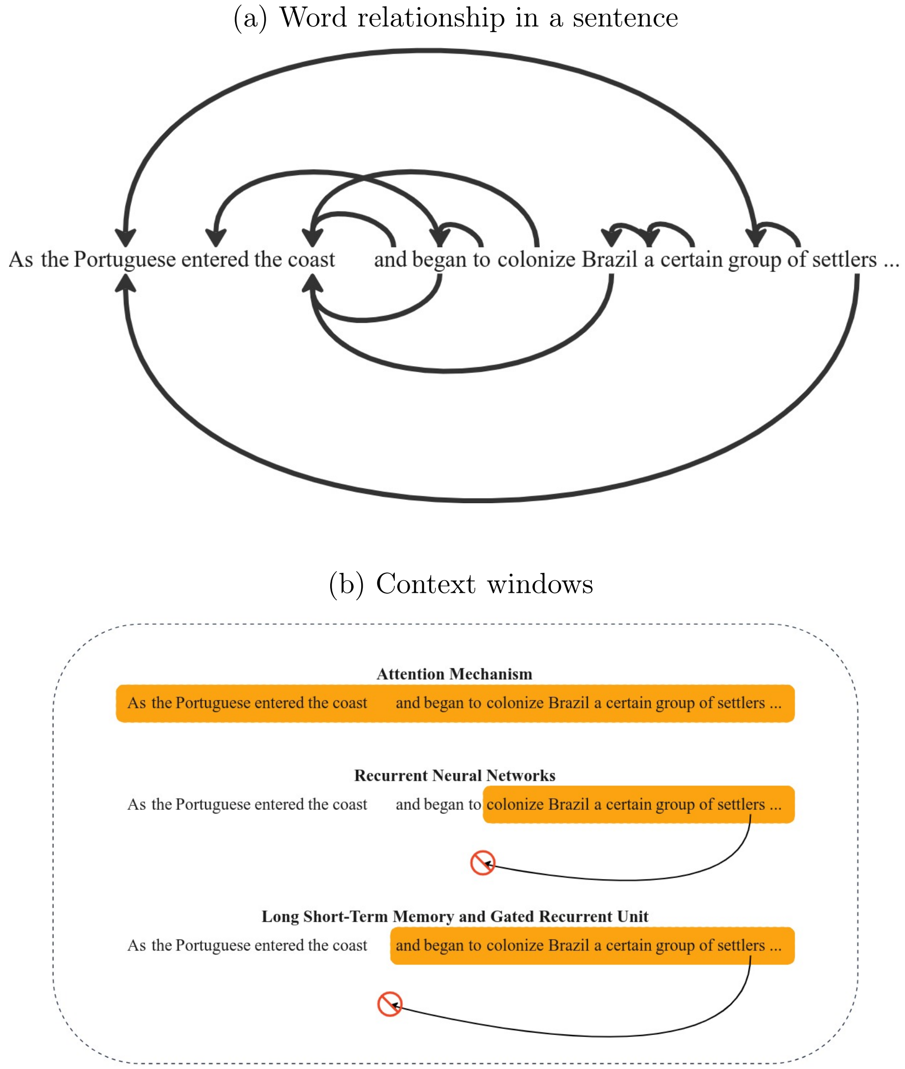
2.3 Transformers

Transformers represent a powerful class of machine learning models that have revolutionized various fields, such as natural language processing and computer vision. These models introduced an innovative approach to sequence modeling, moving away from conventional neural network architectures like Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), or Gated Recurrent Unit (GRU). Instead of relying on fixed network structures to capture sequential dependencies, transformers use attention mechanisms that allow for parallel processing of data sequences and capture complex long-range relationships.

A crucial aspect of transformers is their ability to handle long sequences and capture non-linear relationships between tokens. For example, consider an input sequence like “As the Portuguese arrived at the coast...” and the expected output sequence as “...and began to colonize Brazil a certain group of settlers...”. The model must be able to adapt and generate a coherent output sequence based on the input sequence of fixed size (also known as “context window”) that relates each word in the output to the corresponding words in the input. Figure 2.2a illustrates how different words in the output are related to words in the input sequence, showing that some words in the output sequence can be related to words very far back at the beginning of the input sequence, thus overcoming limitations of fixed context windows.

While networks like RNNs and even LSTMs have a limited context window, the attention mechanism of transformers allows for a context window size limited only by the available computational capacity. This means that to predict a word or token in a sequence, the model does not need to rely only on the immediately preceding token but can relate distant information within the input sequence. This feature is illustrated in Figure 2.2b, which compares the context windows for different models.

In practice, this translates into an improved ability to maintain cohesion and relevance of context in large volumes of text. While traditional networks may struggle to maintain textual cohesion in long texts due to their context limitations, Transformers are designed


(a) Word relationship in a sentence

(b) Context windows

As the Portuguese entered the coast and began to colonize Brazil a certain group of settlers ...
Attention Mechanism
As the Portuguese entered the coast and began to colonize Brazil a certain group of settlers ...
Recurrent Neural Networks
As the Portuguese entered the coast and began to colonize Brazil a certain group of settlers ...
Long Short-Term Memory and Gated Recurrent Unit
As the Portuguese entered the coast and began to colonize Brazil a certain group of settlers ...

Figure 2.2: a) Word relationship in a sentence. b) Context windows for different architecture: Attention mechanism has reference windows as large as the available computational resources; RNNs have short reference windows; GRUs and LSTMs improve RNN's reference window to some extent.

to effectively handle this complexity, resulting in better performance in tasks such as machine translation, text generation, and named entity recognition.

In this session, we will explore the fundamentals of transformers, including their architecture, internal mechanisms, and adaptations for vision tasks. Transformers excel in efficiently processing data sequences and capturing complex long-range relationships. Compared to traditional architectures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), transformers have shown significant improvements in many tasks, thanks to their attention capabilities and efficient parallelism. Their effective sequence modeling capability has made them the preferred choice in a variety of applications, including machine translation, text generation, text summarization, and named entity recognition.

2.3.1 Transformer Architecture

Understanding the architecture of the transformer (Figure 2.3) is crucial to comprehend how these models operate in sequence processing tasks. It consists of an encoder and a decoder, each one composed of n layers; those layers in turn have sublayers such as multiple self-attention blocks, fully connected layers, and layer normalization. Here is a detailed explanation of each component:

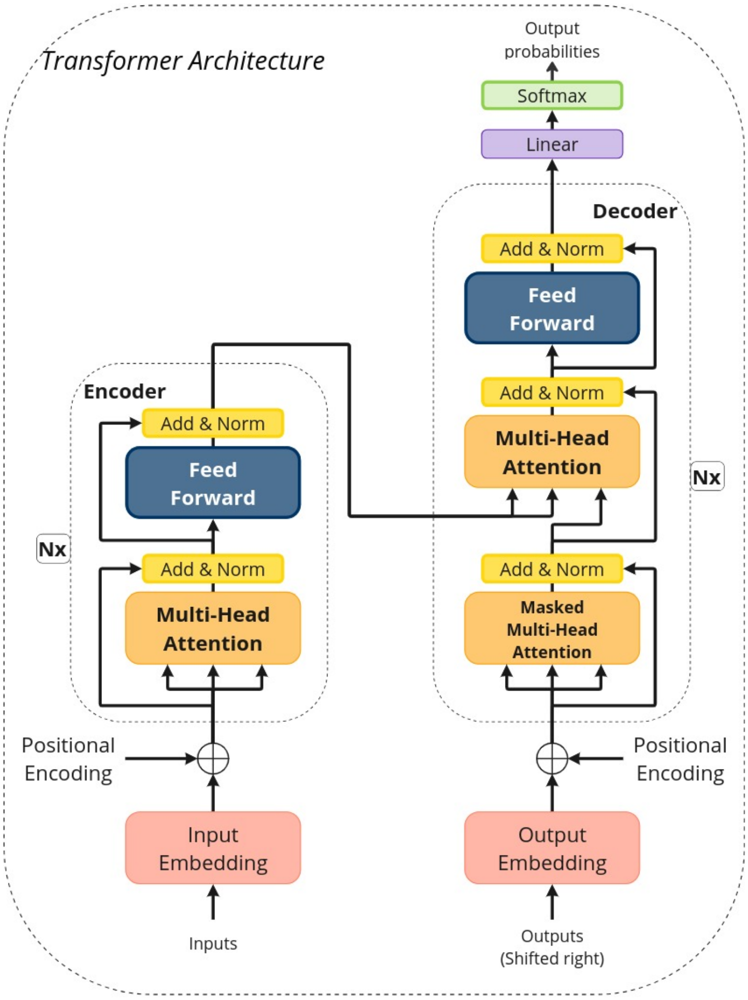


Figure 2.3: The Transformer model architecture. Adapted from: Vaswani et al. [106].

Encoder and Decoder

Encoder: The encoder in a transformer takes an input sequence and produces a contextualized representation of each token in the sequence. Each encoder layer consists of a self-attention block followed by a fully connected layer.

Decoder: The decoder is responsible for generating an output sequence based on both the contextualized representation of the input sequence and a context sequence, commonly referred to as the *"partial output"*, which consists of tokens already generated in previous steps. This enables the decoder to sequentially build the output, token by token.

To achieve this, each decoder layer incorporates two main components:

1. A self-attention block, which focuses on the partial output sequence. This block uses a masking mechanism to ensure each token only attends to previous tokens, preserving the causality needed in autoregressive tasks.
2. A cross-attention block, which attends to the encoded representation of the input sequence, allowing the decoder to integrate information from the input at every step.

Together, these components allow the decoder to generate each token based on both the previously generated sequence and the input context, ensuring consistency and relevance to the original input.

Self-Attention Block

- **Self-Attention:** The self-attention block is the fundamental unit of the Transformer. It computes relationships between all tokens in the input sequence, allowing the model to assign different attention weights to different tokens based on their contextual relevance. This mechanism is at the core of the transformer's ability to capture long-range relationships in sequential data.

- **Attention Weight Calculation:** During the calculation of attention weights, each token in the input sequence contributes to the attention of all other tokens, with the importance of each contribution dynamically determined by the calculated attention weights.

Fully Connected Layers and Layer Normalization

- **Fully Connected Layers:** After the application of the self-attention block, the representation of each token is passed through a fully connected layer, which applies a linear transformation followed by an activation function, such as ReLU (Rectified Linear Unit).

- **Layer Normalization:** Layer normalization is applied after each fully connected layer to stabilize training. This is achieved by normalizing activations within each layer, keeping them within a consistent range, and helping to prevent shifts in activation scale during training.

The processing flow of a transformer can be divided into six steps, which will be explained in the next sections.

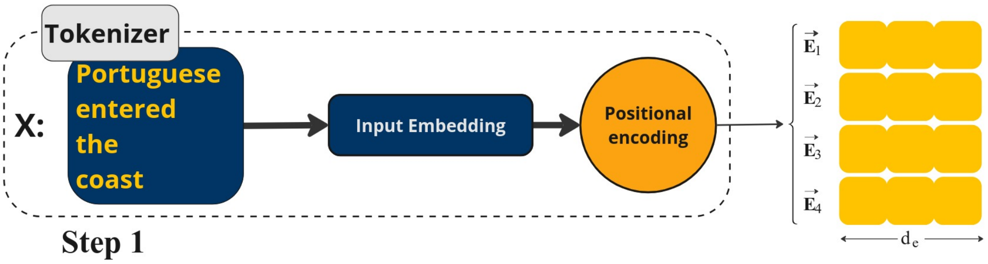


Figure 2.4: Text processing for transformer input, from left to right: *"X"* is the text input, following the word tokenizer. The tokens are inserted as input embeddings, subsequently, the positions are embedded, and finally, the embeddings representing the input text are output. d_e is the dimension of the embedding matrix.

2.3.2 Inputs in the Transformers Model

As a first step, Transformers uniquely process sequential data, using embedding vectors and positional information to represent input tokens (See ***Step 1*** in Figure 2.4). This subsection explores how input data is prepared and transformed before being fed into the model.

Tokens Representation

The input data, whether they are words in a sentence or patches of an image, need to be transformed into high-dimensional vectors d_e that the Transformer can process. This transformation is done through embeddings.

> **Input Embeddings:** Each token (word or patch) is mapped to a high-dimensional embedding vector. In NLP tasks, words are typically mapped using pre-trained embeddings, such as Word2Vec or GloVe, or learned during model training. In Vision Transformers (ViT), image patches are linearly projected into high-dimensional embeddings.

Positional Embeddings

Since transformers do not have an intrinsic order structure, it is necessary to add positional information to the token embeddings to preserve the sequential order of the data.

> **Positional Encoding:** Positional embeddings are added to the token embeddings to incorporate order information. These positional embeddings can be learned during training or generated using sine and cosine functions. They allow the transformer to differentiate tokens in different positions in the sequence.

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}}) \tag{2.3.1}$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}}) \tag{2.3.2}$$

2.3.3 Attention Mechanisms

Attention mechanisms in transformers are fundamental to the model's ability to capture complex relationships among elements of a sequence. They operate based on three main matrices: Q (of query vectors), K (of key vectors), and V (of value vectors), and employ the technique of scaled dot-product attention to calculate the importance of each token relative to others in the sequence.

Query, Key, and Value

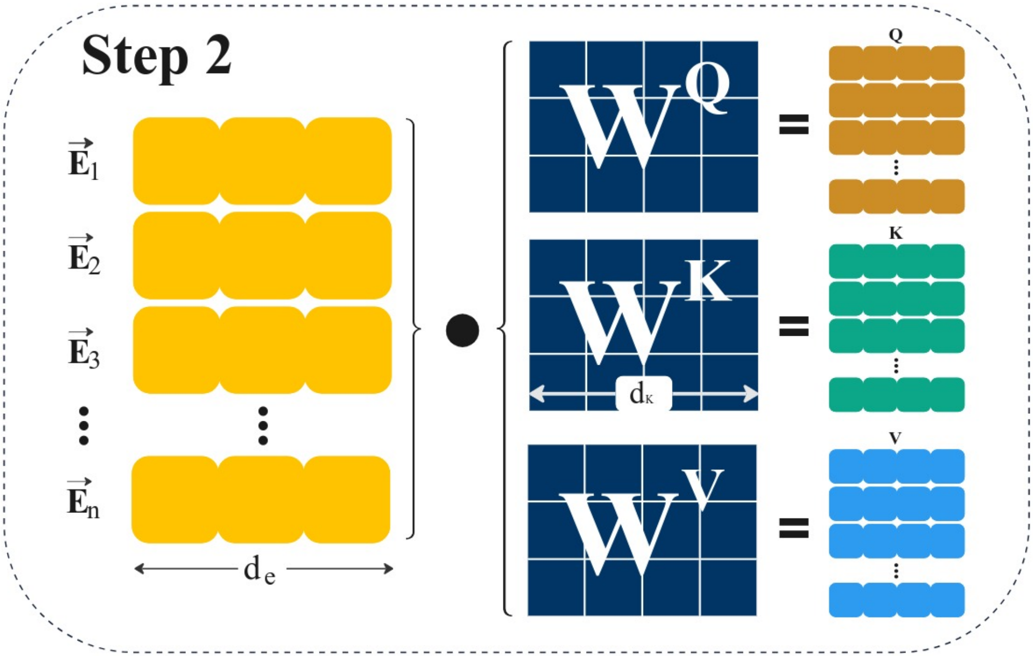


Figure 2.5: Calculation of *Query* and *Key-Value* pair matrices.

Attention mechanisms in transformers rely on three main elements: the query, the key, and the value. These elements are essential for calculating attention weights, which determine the importance of each token relative to other tokens in the sequence.

- **Query:** A query vector represents the token as a query, attempting to capture the purpose and likelihood of the token in context. In simple terms, we can think of the query as the question we are asking the model about a particular part of the sequence. For example, if we are trying to translate a sentence, the query can be the word at the current position that we are trying to translate into the target language. Each token in the input sequence has its own associated query. The query vector is projected from the token's representation in the previous layer of the transformer.

- **Key:** A key vector is used to calculate the relevance of each token in relation to the overall context, for a given query token. It encodes information about the context and content of each token in the sequence. For example, in translating a sentence, the key can represent the neighboring words of the current word we are trying to translate.

- **Value:** A value vector contains information about the content of each token in the sequence. It provides the context associated with each token and is used to calculate the attention weight assigned to each token relative to the query. Continuing with the example of translating a sentence, the value can represent the semantic meaning of each word in the original sequence.

The combination of query, key, and value vectors is crucial for calculating attention weights in an attention mechanism. The similarity between the query and the key of each token determines how much attention should be given to that token, while the weighted values are aggregated to compute the output of the attention layer. This process allows the Transformer to capture complex and long-range relationships between tokens in the sequence, resulting in rich and informative contextual representations, see ***Step 2*** in Figure 2.5.

Scaled Dot-Product Attention

The scaled dot-product attention is given by:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{2.3.3}$$

where d_k is the dimensionality of the key vector k and query vector q. It is a widely used technique for calculating the importance of each token relative to others in the sequence. It consists of three main steps:

- **Calculation of Similarity Scores:** For each pair of query and key, the similarity score is computed as the dot product between them. This measures the relevance of the key relative to the query and is known as unscaled attention, see ***Step 3*** in Figure 2.6.

- **Scaling of Scores:** To prevent the similarity score values from becoming too large, they are scaled by the square root of the dimension of the query and key vectors. This helps smooth the gradient during training and stabilize the attention process, see ***Step 4*** in Figure 2.6.

- **Softmax and Weighting of Values:** The scaled scores are passed through the softmax function to obtain normalized attention weights. These weights are then applied to the corresponding values to compute a weighted representation for each token in the sequence, see ***Step 5*** in Figure 2.6.

Multi-head Attention

Finally, the Multi-Head Attention mechanism in transformers is designed to capture multiple aspects of the relationships between tokens in a sequence. Instead of relying on a single "head" of attention to capture all relevant information, Multi-Head Attention uses

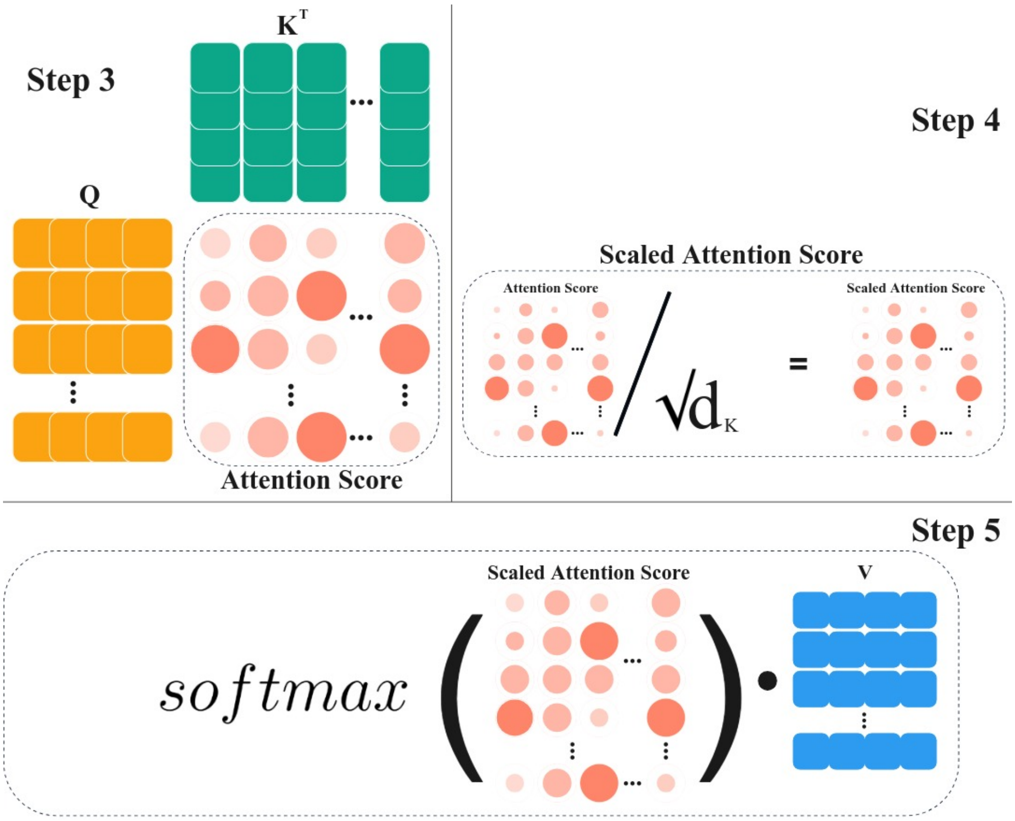


Figure 2.6: Explicit Scaled Dot-Product Attention process.

multiple attention heads simultaneously to obtain richer and more diverse representations. This process is represented by the equations:

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O, \tag{2.3.4}$$

where:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V). \tag{2.3.5}$$

Its ability to capture different aspects of token relationships allows models to capture complex and long-range contextual information, improving performance in tasks such as machine translation, text summarization, and text classification.

Masked Multi-Head Attention in the Decoder

In the Transformer architecture, the decoder generates each token in an output sequence based on the input representation and previously generated tokens. To prevent the decoder from accessing "future" tokens, it uses **masked multi-head attention**.

In each layer of the decoder, a *mask* is applied to restrict tokens from attending to any subsequent tokens. This mask is represented by a triangular matrix, allowing each token to attend only to itself and preceding tokens. This mechanism enforces a sequential generation process, ensuring the model does not "peek" ahead and maintains the causality needed for sequence generation tasks.

During training, masked multi-head attention ensures that each token prediction is based solely on prior tokens and the input, preventing information leakage from future tokens in the output sequence. This is essential for autoregressive tasks like machine translation and text generation, where the model must rely only on previous tokens to produce coherent and logically ordered outputs.

2.3.4 Vision Transformer (ViT)

In recent years, computer vision has been significantly driven by advances in deep learning models. Convolutional Neural Networks (CNNs) emerged as a powerful tool for image analysis, thanks to their ability to extract hierarchical features through local convolutions. Models such as LeNet, AlexNet, and VGGNet demonstrated the effectiveness of CNNs in image classification tasks. With the introduction of Residual Networks (ResNet50s) [59], there was a substantial leap in performance, allowing the construction of much deeper networks without the problems of gradient degradation. ResNet50 uses residual connections to facilitate the learning of deeper layers, further improving accuracy in computer vision benchmarks. More recently, Transformers, originally designed for natural language processing tasks, have been adapted for computer vision, resulting in the Vision Transformer (ViT) [43]. The ViT adopts an attention-based approach, dividing images into patches and processing them as sequences, which allows for capturing global dependencies more efficiently than traditional CNNs. In Figure 2.7, it is depicted the ViT architecture, whose main components will be presented in the following.

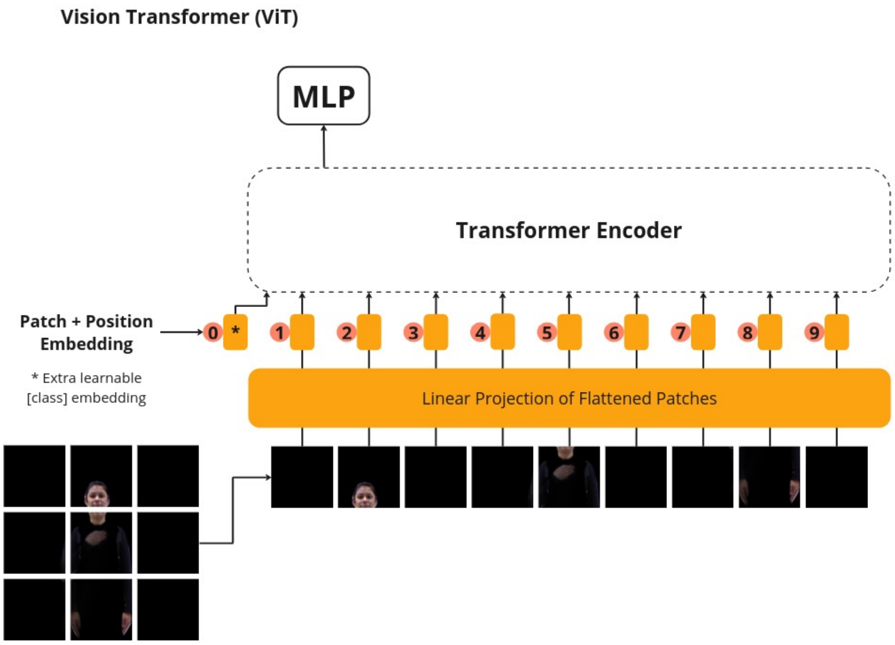

Vision Transformer (ViT)
MLP
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

Figure 2.7: Vision Transformer (ViT) model scheme. Adapted from: Dosovitskiy et al. [43].

Patch Segmentation

In ViT, patch segmentation helps to reduce the variable space of an image. By dividing the image into small patches, whether overlapping or not, it is possible to transform the two-dimensional image into a one-dimensional sequence of patches. This segmentation allows the model to capture spatial and contextual information from the image. Patch segmentation enables a more granular representation of the image, allowing the model to analyze specific regions in detail, which can be beneficial for tasks requiring a detailed understanding of visual features, such as object detection or semantic segmentation.

Token Projection

After segmenting the image into patches, each patch is linearly projected into an embedding space, generating a representation vector for each patch. These patch representation vectors are then treated as input tokens for the Transformer. Similar to natural language models, these patch embeddings pass through transformation layers in the Transformer encoder, where they are processed to capture contextual and relational information between patches.

Position and Position Embeddings

To preserve positional information in the image after patch segmentation, Vision Transformers include position embeddings that are added to the patch embeddings. These position embeddings encode the relative position information of the patches in the image. By incorporating position embeddings, ViTs ensure that the model can distinguish the relative location of patches in the image, allowing it to recognize spatial patterns important for scene understanding.

Classification

Finally, ViTs use an MLP head where classification will be performed. This head maps the patch final representations to the desired output classes. Each patch contributes to the final class decision, and the aggregation of these contributions results in the final class prediction for the image. The inclusion of the MLP heat in ViT allows the model to be trained end-to-end for image classification tasks, leveraging the features learned in various layers of the Transformer. In addition to image classification, the same approach can be applied to other vision tasks, where the MLP head is adapted according to the specific task, such as object detection, semantic segmentation, or object localization. This modular structure of ViT, with a separate classification layer, offers flexibility and adaptability for a variety of vision tasks, allowing the model to be easily fine-tuned and customized for different datasets and application requirements.

2.4 Diffusion Models

2.4.1 Introduction

Diffusion models represent an advanced family of probabilistic models that provide flexible structure, exact sampling, efficient handling of distributions, and computational economy in the calculation and evaluation of log probabilities and individual states, thus facilitating the modeling of complex datasets. These models are grounded in the statistical physics of non-equilibrium, characterized by the gradual and systematic deconstruction of the structure in a data distribution through an iterative process.

Within the field of generative Deep Learning, these models arrive as an innovative alternative to other traditional generative approaches. Diffusion models offer several comparative advantages that distinguish them from models such as normalized flows, autoregressive models [32], variational autoencoders [109], energy-based models, and generative adversarial networks (GANs).

First, unlike normalized flows [90], diffusion models do not require specifying an invertible transformation with constant volume, which allows for greater flexibility in modeling complex distributions. Normalized flows may be limited in terms of representational capacity due to the need to maintain a computable Probability Density Function (PDF), while diffusion models do not face this restriction, allowing for greater modeling capacity.

Compared to autoregressive models, diffusion models avoid the need to decompose the joint probability distribution into a series of one-dimensional conditional distributions. This can simplify the training process and reduce sample generation time. While autoregressive models are effective in generating sequences like text and music, they may be less efficient when generating high-resolution images.

In relation to variational autoencoders, diffusion models do not rely on formulating a variational lower bound (ELBO), which can introduce bias in the estimation of the target distribution. Additionally, variational autoencoders often face the problem of "posterior collapse," where the encoder disregards latent information. Diffusion models, by not depending on an explicit encoder-decoder, avoid this issue.

When compared to energy-based models, diffusion models allow for explicit and efficient probability evaluation. Energy-based models can be challenging to train due to the need to sample efficiently from an intractable distribution, a problem that diffusion models mitigate through their iterative and controlled diffusion and denoising process.

Finally, compared to generative adversarial networks, diffusion models avoid common issues associated with training GANs, such as generator-discriminator mismatch, mode collapse, and convergence difficulties. GANs require a delicate balance in training two competing networks, which can be unstable and difficult to manage. In contrast, diffusion models rely on a more direct and stable optimization process that does not require this adversarial balance.

Diffusion models stand out for their robustness, flexibility, and stability in modeling complex distributions, presenting themselves as a promising alternative to other generative approaches within the realm of Deep Learning. Their ability to handle high-dimensional distributions and ease of evaluating and sampling efficiently from these distributions positions them as a powerful tool in synthetic data generation and modeling complex phenomena.

2.4.2 Forward Diffusion Process (FDP)

The forward diffusion process, also known as forward diffusion, is a method in which noise is iteratively added to a sample from a real data distribution $\mathbf{X}_0 \sim p(\mathbf{x})$ (in our

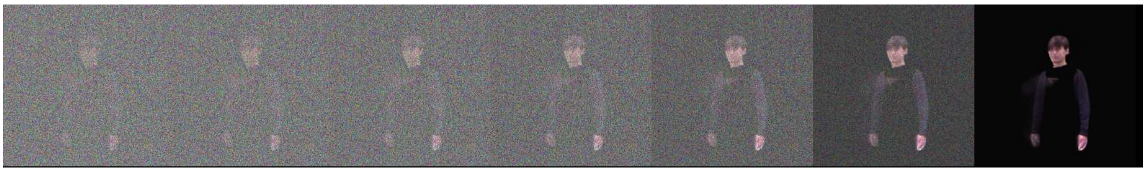


Figure 2.8: Noise addition process.

case, images), progressively disrupting the information. This process is fundamental for understanding diffusion models and can be mathematically formalized.

The process begins with an initial data distribution $p(\mathbf{x})$, from which a natural sample $\mathbf{X}_0$ is drawn. This distribution can be a representation of the original data or a simpler distribution (e.g., a Gaussian). Over T time steps, a small amount of Gaussian noise is added to the sample at each iteration, resulting in a progressive deterioration of the information.

Although this diffusion process is intuitively simple, it is crucial to examine its statistical consistency to ensure that the information is being disrupted in a controlled and mathematical manner.

Transition in Diffusion Process

To formalize this process, consider $\mathbf{x}_t$, where $t = 0, 1, ..., T$, representing the state of the data at step t of the diffusion process. As it is a probabilistic process, it is modeled with a probability density function (PDF):

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{2.4.1}$$

where β_t is a noise coefficient specific to step t, controlling the magnitude of the added noise. Choosing β_t appropriately is crucial to ensure the stability of the diffusion process. Small values of β_t ($10^{-4} < \beta_t < 10^{-2}$ for $1 < t < T$) allow for precise control over the amount of noise added, increasing gradually as the sample becomes noisier ($\beta_1 < \beta_2 < ... < \beta_t$).

This equation can be reparametrized using the reparametrization trick described by Weng [109], allowing the random variable $\mathbf{x}_t$ to be expressed as:

$$\mathbf{x}_t = \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{\epsilon}_{t-1}, \tag{2.4.2}$$

where $\boldsymbol{\epsilon}_{t-1} \sim \mathcal{N}(0, \boldsymbol{I})$ is standard Gaussian noise.

This formula shows that the transition between two consecutive states $\mathbf{x}_{t-1}$ and $\mathbf{x}_t$ is a weighted average between the less noisy sample and the Gaussian noise added at that step. This process is iteratively repeated, accumulating noise at each step.

A process of successive steps can be represented with a joint probability density function (PDF) over $[1:T]$ conditioned on the natural sample $\mathbf{X}_0$,

$$q(\mathbf{x}_{1:T} \mid \mathbf{X}_0) := \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1}), \tag{2.4.3}$$

where the natural sample $\mathbf{X}_0$ with probability $q(\mathbf{X}_0)$, whose analytical interpretation we do not know, can be described as the random choice of a sample from our training set.

Important Properties of the Forward Process

The forward diffusion process has three important properties:

1. **Fully Joint Probability Density Function (PDF) $q(\mathbf{x}_{0:T})$:**
Since this is a probabilistic process, there exists a set of possible trajectories that lead the natural sample $\mathbf{X}_0$ to a pure noise sample. This trajectory can be represented by a fully joint PDF:

$$q(\mathbf{x}_{0:T}) = q(\mathbf{X}_0)q(\mathbf{x}_{1:T} \mid \mathbf{X}_0). \tag{2.4.4}$$

Here, $q(\mathbf{x}_{1:T} \mid \mathbf{X}_0)$ is the joint PDF of the diffusion process over successive steps, which can be calculated using equation Equation (2.4.3).

2. **Marginal Distribution $q(\mathbf{x}_t \mid \mathbf{X}_0)$:**
The marginal distribution property of the forward diffusion process ensures that, for any step t in the interval $[0, T]$, the variable $\mathbf{x}_t$ follows a Gaussian distribution [62]:

$$\mathbf{x}_t \sim \mathcal{N}(\sqrt{\alpha_t}\mathbf{x}_0, (1 - \alpha_t)\mathbf{I}), \tag{2.4.5}$$

where $\alpha_t := 1 - \beta_t$. This representation implies that the random variables $\mathbf{x}_t$ in the interval $[0, T]$ can be sampled arbitrarily from their corresponding Gaussian distributions [110]. This is possible due to the Gaussian nature of the diffusion process, where each $\mathbf{x}_t$ can be expressed as a linear combination of the natural sample $\mathbf{X}_0$ and the *accumulated Gaussian noise* up to step t, $\boldsymbol{\epsilon}_t$. The distribution of the fusion of Gaussians with different variances is calculated as

$$\mathcal{N}(0, (\sum_{t=1}^{T}(1 - \alpha_t))\mathbf{I}),$$

and the combination of their standard deviation is:

$$\sqrt{1 - \alpha_1\alpha_2 \dots \alpha_t},$$

thus:

$$\bar{\alpha}_t := \alpha_1\alpha_2 \dots \alpha_t = \prod_{s=1}^{t} \alpha_s. \tag{2.4.6}$$

This logic ensures the elimination of dependence between intermediate random variables, preserving only the dependence with the natural variable $\mathbf{X}_0$. By reinterpreting Equation (2.4.1) using Equation (2.4.6), we obtain:

$$q(\mathbf{x}_t \mid \mathbf{X}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{X}_0, (1-\bar{\alpha}_t)I), \tag{2.4.7}$$

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{X}_0 + (1-\bar{\alpha}_t)\boldsymbol{\epsilon}. \tag{2.4.8}$$

Now, each $\mathbf{x}_t$ in the diffusion process depends only on the natural variable $\mathbf{X}_0$. We can also redefine the joint PDF of Equation (2.4.3) as:

$$q(\mathbf{x}_{1:T} \mid \mathbf{X}_0) \coloneqq \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{X}_0), \tag{2.4.9}$$

where $q(\mathbf{x}_t \mid \mathbf{X}_0)$ is described in Equation 2.4.7. Both Equation 2.4.3 and Equation 2.4.9 are equivalent, as they describe, for the same set of random variables, the same joint probability distribution.

Given that the parameter t can take any continuous value within the interval $[0, T]$ and the function $\bar{\alpha}_t$ is well-defined and continuous over this interval, any $\mathbf{x}_t$ generated will be consistent with the described diffusion model. This property ensures that the sampling process is robust and that any variable $\mathbf{x}_t$ obtained using this method will have a valid Gaussian distribution.

Therefore, this marginal distribution property not only allows us to sample intermediate variables efficiently but also provides a deep understanding of how the information from the original sample dissipates into Gaussian noise throughout the diffusion process.

3. **The Inverse Of The Forward Process** $q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{X}_0)$**:**
One of the most important features of the diffusion process is its reversibility. As will be explained later in Section 2.4.3, this process allows reconstructing a natural sample $\mathbf{X}_0$ from a pure noise sample by reversing the diffusion process. This reverse process is crucial for data generation in diffusion models.

2.4.3 Inverse Diffusion Process (IDP)

The inverse diffusion process is the method by which the original structure of the data is restored from a noisy distribution obtained through the *Forward Diffusion Process* (See Section 2.4.2). This process involves learning a model that can reverse the addition of noise step by step, thus reconstructing the original distribution.

As it is a reverse process, we start with a noisy distribution $q(\mathbf{x}_t) \sim \mathcal{N}(0, \mathbf{I})$ from which we sample a noisy $\mathbf{x}_t$. At each step of the reverse process, a small amount of noise is removed from the current state of the data. This process is performed iteratively in reverse from $t = T$ to $t = 0$. However, to understand the IDP, one must understand the *Inverse of the Forward Diffusion Process (IOFDP).*

Inverse of the Forward Diffusion Process (IOFDP)

In the previous section, we discussed two of the three important properties of the forward diffusion process: the Fully Joint Probability Density Function and the Marginal Distribution (See Section 2.4.2). However, perhaps the most important property of this process is the ability to reverse it.

To understand this, let's examine the PDF of the random variable $\mathbf{x}_t$ in Equation 2.4.1. By adding a redundant conditioned random variable $\mathbf{X}_0$, we can rewrite it as $q(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{X}_0)$. This expression does not affect the probability density for $\mathbf{x}_t$ because, by definition, given $\mathbf{x}_{t-1}$, the random variable $\mathbf{x}_t$ does not depend on any other random variable. Applying Bayes' rule to Equation 2.4.10, we get:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{X}_0) = q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{X}_0) \cdot \frac{q(\mathbf{x}_t \mid \mathbf{X}_0)}{q(\mathbf{x}_{t-1} \mid \mathbf{X}_0)}. \tag{2.4.10}$$

Note that the random variables $q(\mathbf{x}_t \mid \mathbf{X}_0)$ and $q(\mathbf{x}_{t-1} \mid \mathbf{X}_0)$ are known from the marginal distribution property of the *Forward Diffusion Process*. Rearranging Equation 2.4.10, the term in red becomes:

$$q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{X}_0) = q(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{X}_0) \cdot \frac{q(\mathbf{x}_{t-1} \mid \mathbf{X}_0)}{q(\mathbf{x}_t \mid \mathbf{X}_0)}. \tag{2.4.11}$$

Note that when $t = 1$, $q(\mathbf{X}_0 \mid \mathbf{x}_1, \mathbf{X}_0) \equiv 1$ due to the lack of uncertainty about it, which makes the first analytical loss function occur at $t = 2$.

Equation 2.4.2 describes a process that takes a noisy image $\mathbf{x}_t$ and transforms it into a less noisy image $\mathbf{x}_{t-1}$. Therefore, it describes the process in the opposite direction to the **Forward Process** and will be referred to as *The Inverse of the Forward Process*. This process is defined by the probability density of the forward process (in green) multiplied by a scaling factor (in blue). The three elements in Equation 2.4.2 were defined earlier, so simplifying it yields a multivariate Gaussian distribution:

$$q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{X}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_t(\mathbf{x}_t, \mathbf{X}_0), \tilde{\beta}_t I), \tag{2.4.12}$$

where:

$$\mu_t(\mathbf{x}_t, \mathbf{X}_0) = \frac{\sqrt{\bar{\alpha}_{t-1}\beta_t}}{1 - \bar{\alpha}_t}\mathbf{X}_0 + \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \alpha_t}\mathbf{x}_t, \tag{2.4.13}$$

and:

$$\tilde{\beta}_t = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}\beta_t. \tag{2.4.14}$$

Note that this process is conditioned on the initial image $\mathbf{X}_0$. This conditioning arises from the way we interpret the noise, as we are not starting from pure Gaussian noise but rather from noise close to pure Gaussian noise. Thus, in the IOFDP, we can obtain an image very close to $\mathbf{X}_0$. Additionally, this conditioning is unavoidable as it results from applying Bayes' rule to the forward process, which includes $\mathbf{X}_0$ when $t = 1$.

Inverse Process

An initial question we might suggest is: Why do we need an inverse process $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ when we already have the IOFDP $q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{X}_0)$? Can't we simply sample directly from that distribution, which we already know? The answer is that, although it is possible to do that, the IOFDP depends on the natural sample $\mathbf{X}_0$, the initial image, to define the behavior of the noise. This means that the IOFDP would return an image very close to the known $\mathbf{X}_0$. However, what we want is to be able to produce/generate new natural images independently of $\mathbf{X}_0$.

As in the IOFDP, the starting point of the inverse process is a pure noise sample, modeled by the standard multivariate Gaussian distribution $p_\theta(\mathbf{X}_T) \sim \mathcal{N}(\mathbf{X}_T; 0, \boldsymbol{I})$. In contrast to the IOFDP, the inverse process is not conditioned on any initial random variable, which allows generating new images from the noise:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_p(\mathbf{x}_t, t), \boldsymbol{\Sigma}_p(\mathbf{x}_t, t)). \tag{2.4.15}$$

In this formula, $\boldsymbol{\mu}_p$ is the mean vector and $\boldsymbol{\Sigma}_p$ is the covariance matrix. These are actually two deep neural networks that predict the *d-dimensional* mean and the $d \times d$ covariance matrix. Both networks receive two inputs: $\mathbf{x}_t$ and t, the noisy image and the time step, respectively. The latter is used to encode the diffusion process position in which we are. It is especially useful for providing additional positional context to the noise for the networks.

This makes sense if we consider that in statistical modeling, once we choose the distribution family, the hardest task is to define the values that completely specify that distribution. Since we choose a multivariate Gaussian distribution in this case, we need to predict the mean vector $\boldsymbol{\mu}_p$ and the covariance matrix $\boldsymbol{\Sigma}_p$ to describe the spectrum of images that are less noisy versions of $\mathbf{x}_t$.

These prediction functions are complex and it is impractical to use simple functions to model them. Therefore, it becomes necessary to use deep neural networks with millions of parameters that can handle these very difficult functions. For example, in Ho et al. [61], Nichol and Dhariwal [82], very high-dimensional architectures are described that predict the mean and covariance adaptively for each stage of the reverse process, allowing the model to gradually remove noise from the images.

In this context, in the inverse process, we can also establish a joint PDF that allows us to abbreviate the inverse process $p_\theta(\mathbf{X}_0, \mathbf{x}_1, \cdots, \mathbf{X}_T)$:

$$p_\theta(\mathbf{x}_{0:T}) := p(\mathbf{X}_T) \cdot \prod_{t=1}^{T} p(\mathbf{x}_{t-1} \mid \mathbf{x}_t). \tag{2.4.16}$$

The equation represents a joint PDF for $T+1$ random variables; it is, in itself, the product of all the recursive terms in Eq. (2.4.15) and $p_\theta(\mathbf{X}_T)$.

Finally, the model parameters, which are the weights of the neural networks, are adjusted through optimization. The goal is to find the appropriate values for these parameters so that, starting from a noise sample $p_\theta(\mathbf{X}_T)$, we can iteratively find $\mathbf{x}_{t-1}$ from the distribution $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$. By the end of the process, when we reach $\mathbf{X}_0$, we expect that sample to be a "denoised" version of the natural image.

In this way, we have defined the reverse diffusion process such that we understand its fundamentals and statistical coherence. We will not address the training process of the networks in this section, as we only intend to provide a fundamental overview of its operation.

Chapter 3

Literature Review

In this chapter, we will discuss the current state of sign language datasets, starting with variations across different languages, moving on to the types of data we can find, and concluding with a focus on the databases available for Libras. We will also cover the state of the art in sign language recognition technologies across various languages and end the chapter with an analysis of generative AI data augmentation techniques available to date.

3.1 Sign Language Datasets

Research in sign language recognition and classification has expanded significantly in recent decades, driven by the need for more inclusive communication systems for the deaf community. The core of this research lies in the diverse and detailed datasets that enable the training of machine learning models. These datasets vary not only in the sign languages they represent but also in the types of data they include. In Table 3.1, we can see the datasets reviewed in this literature.

3.1.1 Diversity of Sign Languages

Sign languages are not universal; each deaf community has developed its own language, reflecting its unique culture. These cultural and linguistic differences result in a rich diversity of sign languages worldwide. For example, American Sign Language (ASL) and British Sign Language (BSL) are significantly different, despite the countries sharing the same spoken language. This diversity is evidenced by the existence of multiple datasets for different sign languages, each reflecting the unique characteristics of its respective deaf community.

Here are some examples of datasets that reflect this diversity:

- **ASLLVD:** One of the largest datasets found in the literature, comprising high-quality video data useful for research. It contains recordings from multiple camera angles, with 3,314 signs from the Gallaudet Dictionary of American Sign Language.

Table 3.1: General dataset characteristics.

Libras Datasets

Datasets	Data provided			Mode		Body Part	Signers	Vocab.	Samples	Avail.
	RGB	Depth	Pose	Vid.	Img.					
CEFET-Libras [53]	✓			✓		Full body	20	24	547	No
MINDS-Libras [92]	✓	✓	✓	✓		Full body	12	20	1100	Yes
LIBRAS-10 [9]	✓	✓	✓	✓		Hand	10	10	100	Yes
SILFA [98]	✓		✓	✓		Facial	10	23	230	Yes
LIBRAS-34 [10]	✓	✓	✓	✓		Hand	5	34	170	Yes
LIBRAS-UFOP-ISO [23]	✓	✓	✓	✓		Full body	5	56	3040	Yes
V-Librasil [93]	✓			✓		Full body	3	1364	4089	Yes
Elias Dataset [107]	✓	✓	✓	✓		Full body	2	26	166	Yes

Others

Datasets	Data provided			Mode		Body Part	Signers	Vocab.	Samples	Avail.
	RGB	Depth	Pose	Vid.	Img.					
BSL Corpus [20]	✓			✓		Full body	249	25k	∼	Yes
ASL-LEX 2.0 [95]	✓			✓		Full body	129	2K7	∼	Yes
ArASL [73]	✓				✓	Hand	40	32	54K	Yes
ISL-CSLTR [87]	✓			✓	✓	∼	6+7	∼	700/18K	Yes
How2Sign [45]	✓	✓	✓	✓		Full body	11	16K	∼	Yes
RWTH-PHOENIX [70]	✓			✓		Full body	9	1081/7K	∼	Yes
ASLLVD [13]	✓			✓		Full body	6	3314	9794	Yes
UTA-ASL [27]	✓	✓		✓		Full body	2	1113	1313	No
ISL-translate [66]	✓			✓		Full body	∼	11K	31K	∼

Each video includes annotations on hand shapes in the initial and final frames, hand and face positions, and a general label with the "approximate English translation". This dataset consists of 9,800 samples performed by six native signers [13].

- **BSL Corpus:** A dataset composed of video recordings of hundreds of BSL users, collected from different regions of the United Kingdom [20].

- **RWTH-PHOENIX:** This dataset is based on the GSL (German Sign Language), and contains real videos from a weather channel, including a corpus of 1,081 words and 7,000 sentences from 9 signers. It stands out by addressing five main focuses: tracking, features, signer dependency, visual modeling, and language modeling. It also combines synthetic data for comparisons [70].

- **ArASL:** Arabic Alphabets Sign Language Dataset is a fully labeled dataset of Arabic Sign Language (ArSL) images. Publicly available and free for researchers, it contains 54,049 images of 32 ArSL signs and alphabets, collected from 40 participants of different age groups. It is essential for developing automated systems for deaf individuals using machine learning algorithms and computer vision [73].

- **ASL-LEX 2.0:** A large-scale lexical database for American Sign Language (ASL), including detailed phonological descriptions, measures of phonological density and complexity, among others. This database is publicly accessible and can be explored interactively online [95].

- **ISLTranslate:** A translation dataset for Indian Sign Language (ISL), including 31,000 ISL-English phrase/sentence pairs. It is one of the largest datasets for continuous ISL translation, useful for validating the performance of sign language to spoken language translation systems [66].

3.1.2 Libras Datasets

In Brazil, besides Libras, there is the Brazilian Kaapor Sign Language, demonstrating the plurality within the country. In Portugal, Portuguese Sign Language (LGP) is the official sign language, while in Angola and Mozambique, we have Angolan Sign Language (LAS) and Mozambican Sign Language (LMS), respectively. Each of these languages has its own grammatical rules, vocabulary, and cultural expressions, highlighting the importance of developing specific datasets for each context.

In the specific case of Libras, we find several datasets that serve as references to promote research and development of ASLR technologies in Libras:

- **LIBRAS-34:** The LIBRAS-34 [10] is a set of Libras signs aimed at benchmarking for ASLR and gestures. The dataset contains 34 distinct signs for words/sentences such as person, spread, copy, grab, gather, disappear, look, fair, truth, weight, justice, who, nothing, believe, forget, love, distress, celebrate, resentment, assembly meeting, compare, shout, speak, absorb, fatten, fight, shrewd, shine, maid, replace, prison, television, yesterday, and future. Each sign was recorded 5 times by a single signer, totaling a database of 170 samples. The signals were captured using an RGB-D sensor (Microsoft Kinect) and processed by the nuiCaptureAnalyze software. This dataset is publicly available. These data are valuable for the development and evaluation of Libras and gesture recognition algorithms.

- **Elias Dataset:** The Elias Dataset [107] is a Libras database created for continuous sign language recognition in a healthcare environment to help deaf and hard-of-hearing people access essential information and services. It contains 26 Libras signs recorded five times by two signers, totaling 166 samples. The signs were captured using the Microsoft Kinect V1 sensor, including RGB-D image and skeleton articulation information. The RGB and depth videos have a resolution of 640x480, while the body coordinates of the signer are in a text file format. The background is solid color and provides a great contrast with the signers' clothing. The Elias Dataset is a database with great potential for use in ASLR tasks in Libras.

- **LIBRAS-10:** The LIBRAS-10 [9] consists of a set of Libras signs aimed at benchmarking for ASLR and gestures. The dataset contains 10 action signs: calm, accuse, annihilate, in love, fatten, happiness, thin, lucky, surprise, and angry. Each sign was recorded 10 times by a single signer, resulting in a total of 100 samples. The signals were captured using an RGB-D sensor (Microsoft Kinect) and processed by the nuiCaptureAnalyze software. The dataset includes RGB-D images, skeleton images, RGB-D face images, and .mat files containing all the information obtained by the Kinect software. These data provide a valuable source for the development and evaluation of Libras and gesture recognition algorithms.

- **CEFET-Libras:** The CEFET-Libras [53] is a collection of Libras data composed of 24 classes, including signs and sentences. Although the CEFET-Libras has a more limited number of words, it has a broader number of samples per sign.

- **MINDS-Libras:** The MINDS-Libras dataset [92] was developed to fill a gap in the academic community interested in Libras, providing a standardized and challenging resource for artificial intelligence (AI) research. Composed of 20 Libras signs recorded by 12 different signers, including deaf and hearing men and women, the dataset offers a wide variety of samples. Using an RGB camera and an RGB-D sensor to capture videos and body coordinates of the signer, MINDS-Libras provides detailed information about body points and facial data, facilitating the development of Libras sign recognition algorithms. With an organized structure and accessible file formats, this dataset promotes collaboration and innovation in Libras research, representing a valuable resource for the deaf community and AI researchers.

- **LIBRAS-UFOP-ISO:** The LIBRAS-UFOP dataset [23] is a valuable contribution to the field of Libras recognition, publicly available and challenging for researchers. Recorded with the Microsoft Kinect V1 sensor, it offers complete RGB-D and skeleton articulation information. Composed of 56 signs divided into four distinct categories, the dataset presents significant intraclass variations, including different movements, articulation points, and hand configurations. Each category was carefully selected and validated by a Libras expert. The recordings were made in front of the Kinect sensor, following the guidance of a LIBRAS specialist, and the dataset includes variations in lighting, sign execution speeds, and the use of one or both hands by the signers. In summary, the LIBRAS-UFOP dataset provides a comprehensive Kinect data set for the research and development of LIBRAS sign recognition algorithms.

- **SILFA:** This dataset [98] was obtained through stimuli from well-validated sentences to elicit a variety of grammatical and affective facial expressions, with manual annotation of facial actions using Facial Action Coding System (FACS). This work also promotes the exploration of discriminative features in subtle facial expressions in SL, providing a deeper understanding of the relationship between the dynamics of grammatical facial expression classes and facial action units, as well as providing protocols and benchmarks for the automated recognition of facial action units for sign language research.

These datasets are fundamental for continuous research and the development of more robust Libras recognition systems, contributing to the inclusion and accessibility of the deaf community. However, our focus in this work is on lesser-known datasets with characteristics that impair model performance.

3.1.3 Variety of Data Types

Sign language datasets vary significantly in the types of data they include. This variety is essential to address different aspects of sign language recognition, such as movement, facial expressions, hand configurations, and context. Here are some of the main categories of data types found in these datasets:

Video Data

Videos are the most common form of data used in sign language datasets. They capture the complexity of movement and facial expressions, which are essential for communication in sign language. Examples include:

- **ASLLVD:** Contains high-quality videos of American Sign Language signs with multiple camera angles and detailed annotations.
- **RWTH-PHOENIX:** Includes real videos from a weather channel with a corpus of 1,081 words and 7,000 sentences in German Sign Language.
- **UTA-ASL:** Presents a dataset of video and depth data based on Microsoft Kinect for research in body part detection and tracking.
- **ASL-LEX 2.0:** A large-scale lexical database for ASL, with detailed phonological descriptions and videos.
- **How2Sign:** A dataset that includes videos of ASL signs with detailed annotations and audio synchronization.

Static Images

Some datasets use static images that capture specific moments of the signs. These datasets can be useful for analyzing hand configurations and specific positions. Examples include:

- **ArASL:** Contains 54,049 images of 32 signs and alphabets of Arabic Sign Language.

3D and Depth Data

3D and depth data provide a more detailed view of hand and body positions and movements. They are particularly useful for developing models that need to understand spatial orientation. Examples include:

- **MINDSLibras:** Uses an RGB camera and an RGB-D sensor to capture videos and body coordinates of the signer, offering detailed information about body points and facial data [92].
- **LIBRAS-UFOP:** Recorded with the Microsoft Kinect V1 sensor, it provides complete RGB-D and skeleton articulation information, offering a comprehensive dataset for the research and development of Libras sign recognition algorithms [23].

Skeleton Data

Skeleton data capture the articulations and positions of the limbs, providing an abstract model of body movements. These data are often used in conjunction with machine learning algorithms for gesture recognition. Examples include:

- **Elias:** Includes RGB-D image and skeleton articulation information captured using the Microsoft Kinect V1 sensor [107].

Facial Expression Data

Facial expressions play a crucial role in sign language, providing grammatical and emotional information. Some datasets focus specifically on capturing these facial nuances. Examples include:

- **SILFA:** A video database of grammatical facial expressions in Libras, manually annotated for facial actions using FACS [98].

3.2 Sign Language Classification

The area of ASLR has received significant attention in recent years due to its structural complexity and its application in testing computer vision algorithms [7, 26, 98, 113]. The gestures and movements in sign languages involve different parts of the body, with the hands, body, and FE being key elements [112]. However, dealing with the diversity of signs and their variability is a significant challenge [7, 26, 116]. Various techniques have been explored in this field, such as CNN, Deep Learning (DL), Hidden Markov Model (HMM), and Linear Discriminant Analysis (LDA) [7, 26].

Hand Configuration Recognition (HCR) techniques have been widely implemented in the field of vision-based sign recognition, particularly for static gesture classification. Recent studies explore the use of CNNs for this task. For example, Zhan [121] proposed a CNN-based algorithm that achieved an average accuracy of 98.76% in recognizing nine hand gestures. Additionally, Poon et al. [85] utilized a multi-camera approach to reduce ambiguity caused by self-occlusion of the hand in bimanual gestures.

In Static Gesture Recognition (SGR) for Libras, different approaches have been proposed. Bastos et al. [17] achieved high recognition rates using shape descriptors, such as Histogram of Oriented Gradients (HOG) and Zernike Invariant Moment (ZIM), along with a two-stage neural network classifier. On the other hand, Costa et al. [28] used novelty and KNN classifiers to recognize 61 hand configurations in Libras, achieving high accuracies. Additionally, Caiafa et al. [21] explored the use of CNNs with different architectures, such as AlexNet, VGG16, VGG19, InceptionV3, and ResNet50, to classify static signs in Libras, obtaining promising results. de Carvalho et al. [33] also explored the use of CNNs on a dataset with 91 classes of manual configurations in Libras, called Libras91.

Dynamic gesture classification involves recognizing temporal sequences of signs, which adds an extra layer of complexity due to the need to capture the evolution of movements over time. Techniques such as HMMs and Recurrent Neural Networks (RNN), including LSTMs and GRUs, have been successfully applied in this area [7, 26].

3.2.1 Hybrid Techniques and Recent Advances

During our review, various studies were analyzed (Table 3.2), addressing the use of different techniques for automatic sign recognition, such as hand gestures and sign language. Vision Transformer (ViT) have shown significant advantages in terms of capturing linguistic context and sequential representation, surpassing previous methods. Additionally, techniques such as LAT and the use of body pose as input for ViT models were also explored. However, few studies have applied these techniques specifically to Libras, which opens research opportunities to explore the potential of these approaches in this context.

In summary, the area of sign language classification is rapidly evolving, with various techniques being continuously improved to address the challenges posed by the complexity of gestures and the diversity of signs. Our study proposes a reference approach in the field, combining motion analysis techniques and deep learning models for the recognition of Libras.

Table 3.2: Works on sign language recognition. Values in “Results” column correspond to the metrics indicated in “Metrics” column.

Paper	**Type**		**Mode**		**Data provided**			**SL**	**Body Part**	**Models**	**Metrics**	**Augmentation**	**Results**
	Dynamic	**Static**	**Real Time**	**F2F**	**RGB**	**Depth**	**Skeleton**						
Bastos et al. [17]		✓			✓			Libras	Hand	HOG, ZIM	MSE		96.77%
Costa et al. [28]		✓				✓		Libras	Hand	KNN	ACC		96.31%
Caiafa et al. [21]		✓			✓	✓		Libras	Hand	CNN	ACC		97.98%
de Carvalho et al. [33]		✓	✓	✓	✓			Libras	Hand	CNN, SVM	ACC		99.0%
Castro et. al. [120]	✓			✓	✓	✓		Libras	Hand, torso	3DCNN, CNN	ACC	✓	72.6%
Escovedo et. al. [47]	✓		✓	✓	✓	✓	✓	Libras	Full body	CNN, SVM, HCM	Recognition rate		67.37%
Zhang et al. [122]	✓		✓		✓				Hand	DL	MSE, AP	✓	25.7% 95.7%
Zhan [121]		✓	✓	✓	✓			ASL	Hand	CNN	ACC	✓	98.2%
Poon et al. [85]		✓			✓				Hand	SVM, LDA	ACC		99.0%
Kagirov et al. [68]	✓	✓		✓	✓	✓		RSL	Full body	3DCNN, CNN, LSTM	ACC		73.25%
Devineau et al. [40]	✓		✓	✓	✓	✓	✓		Hand	CNN, LSTM, Droput	ACC, F1		91.28% 84.35%
Montazerin et al. [79]									Hand	ViT	ACC		84.62%
Du et al. [44]	✓		✓	✓	✓			ASL	Hand	CNN, KNN, TF	ACC	✓	90.57%
Guo et al. [56]	✓		✓	✓	✓			GSL, CSL	Full body	TF	BLEU, ROUGE	✓	49.16% 49.74%
Guo et al. [56]	✓		✓	✓	✓			GSL, CSL	Full body	TF	BLEU, ROUGE	✓	98.27% 98.13%
Hinrichs et al. [60]	✓		✓	✓	✓			GSL	Multi-Hand	TF	WER	✓	<10%
Bohacek and Hruz [19]	✓			✓	✓	✓	✓	LSA, ASL	Full body	ViT	ACC	✓	LSA64: 100% WLASL100: 63.18%
Al-Hammadi et al. [8]	✓		✓	✓	✓			ARSL, ASL	Full body	SGD, LSTM, 3DCNN	ACC		ArSL: 100% KSU-SSL: 96.69% RVL-SLLL: 76.67%
Passos et al. [83]	✓		✓	✓	✓	✓	✓	Libras	Full body	KNN	ACC	✓	CEFET: 85.40% MINDS: 84.66% UFOP: 64.91%

3.3 Data Augmentation for Sign Language

It is well-known in the literature that having insufficient data can lead to overfitting, where a model fails to generalize and only adapts to the training dataset. Data augmentation techniques emerge as an effective solution to this problem, helping address issues such as class imbalance, accuracy, robustness, regularization, and data scarcity. Zhan [121] demonstrated that data augmentation plays a crucial role in achieving superior performance in ASLR for ASL.

3.3.1 Classical Data Augmentation Techniques

Data augmentation involves using techniques that generate new data from a limited dataset. These techniques can be applied to any type of data. In the case of images, these techniques include classical transformations that modify geometry, such as rotation, translation, and scaling, as well as photometric transformations, such as changes in lighting, contrast, and saturation [69, 96]. These techniques are widely used for their simplicity and efficiency in creating variations of the original data without the need to generate new data from scratch.

3.3.2 Advanced Data Augmentation Techniques

Beyond classical transformations, there are deep learning-based approaches, such as Generative Adversarial Network (GAM), Neural Style Transfer [69, 96], and more recently, diffusion models Denoising Diffusion Probabilistic Modelss (DDPMs) [62], as demonstrated by Chen et al. [25], Xiao et al. [114], Azizi et al. [14]. Techniques like MixUp [123] and CutMix [118] also fall within this scope, offering variations such as TokenMix[1], used in image preprocessing for ViTs. Other approaches combine multiple techniques to provide broader results, such as Neural Augmentation [84], Auto Augmentation [29], and Smart Augmentation [74]. Figure 3.1 illustrates a basic taxonomy of these techniques, and Table 3.3 presents the data augmentation techniques reviewed for this dissertation.

[1] A variation of CutMix at the token level.

Table 3.3: Works on data augmentation.

Paper	Date	Technique	Type	Tested model	Dataset
Chawla et al. [24]	2002	SMOTE	Basic	DesitionTree, SVM, KNN	∼
Perez and Wang [84]	2017	Neural Augmentation	Basic, DL	CNNs, DenseNet	imagenet-200, MNIST
Lemley et al. [74]	2017	Smart Augmentation	Basic, DL	CNNs, VGG16	∼
DeVries and Taylor [41]	2017	Feature Space	Features	Rede neural	Arabic Digits
Zhang et al. [123]	2018	MixUp	Basic	ResNet, Wide ResNet, DenseNet	CIFAR-10, CIFAR-100, SVHN e ImageNet
Yun et al. [118]	2019	CutMix	Basic	ResNet, DenseNet e EfficientNet.	CIFAR-10, CIFAR-100, SVHN e ImageNet
Cubuk et al. [29]	2019	Auto Augment	DL	ResNet, Wide ResNet, DenseNet	CIFAR-10, CIFAR-100, SVHN, ImageNet
Yun et al. [119]	2020	VideoMix	Basic	SlowOnly-50, SlowFast-50, I3D, T-CAM, SlowFast-50	Kinetics-400, Mini-Kinetics, AVA.
Bai et al. [15]	2020	TaCo	DL	3DCNNs, 3D-ResNet18	UCF-101, HMDB-51
Sinha et al. [100]	2021	NDA	Basic	GANs	CIFAR-10, CIFAR-100, SVHN, ImageNet, COCO, UCF101.
Wang et al. [108]	2021	ISDA	DL	CNNs, RNNs	CIFAR-10, CIFAR-100, SVHN e ImageNet.
Dablain et al. [31]	2021	DeepSMOTE	DL	CNNs, DNNs, RNNs	MNIST, CIFAR-10, CIFAR-100, SVHN, CelebA.
Ye et al. [117]	2023	XmDA	DL	XmDA	PHOENIX-2014T, CSL-Daily
Liu et al. [77]	2023	TokenMix	DL	DeiT-S, DeiT-B, Swin-T, ViT-L16.	ImageNet-1K, ADE20K

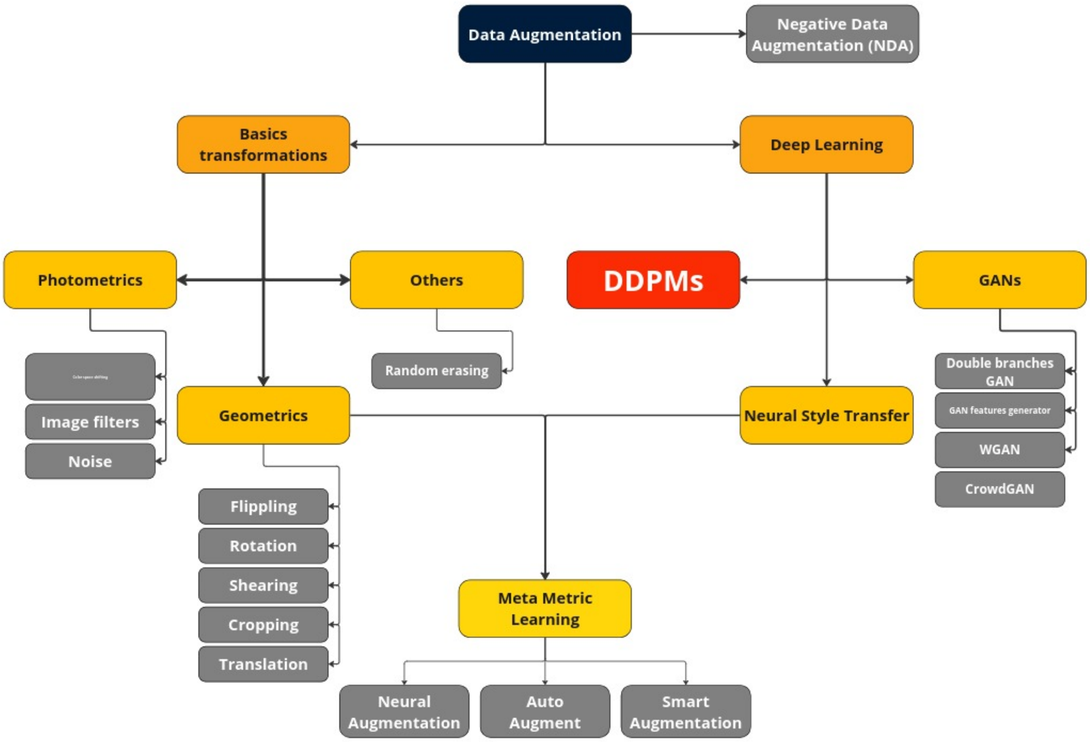


Figure 3.1: Basic taxonomy of techniques involved in the image data augmentation. In red represents the data augmentation used in this work.

3.3.3 Data Augmentation for Sign Language

In ASLR tasks, the complexity is higher because it is necessary to preserve both temporal characteristics and the grammatical and syntactical structure of a sign. Various authors choose techniques with spatio-temporal transformations [120, 121], heuristics[2] [80], or approaches like Cross Modality Data Augmentation (XmDA) [117] to preserve these characteristics that give meaning to the sign.

3.3.4 Avatars and Generative Models

Another approach creates an entire area called Sign Language Production (SLP) [88, 89], which proposes generating language typically simulated through avatars as a way to augment data for ASLR tasks. One of the recent works proposing this is by Nguyen et al. [81]. This work presents a technique to automatically generate a 3D sign language avatar from skeleton data for real-time interpretation. It also discusses the architecture of the CNN model used and the metric for evaluating the accuracy of hand sign classification.

However, avatars do not preserve the non-manual gestures of a sign, which means that deaf individuals may not fully understand the message [102]. With the advancement

[2] Lemmatization of spoken words: for example, the word “running” would be lemmatized to “run”. Exclusion of random and “POS”-dependent words: for example, articles and prepositions can be excluded more frequently than nouns and verbs. Random permutation of words: helps increase the syntactic diversity of the generated synthetic data.

of GANs, new alternatives have emerged to achieve more natural results. One of the first known works for sign language production is by Stoll et al. [102]. This work allows for the production of sign language videos from spoken language sentences, requiring minimal gloss annotations at the skeletal level. Later, Stoll et al. [103] improved on the previous work, proposing a combination of CNNs and GANs to map glosses to sequences of sign poses and sign videos, enabling the model to capture the complexities of sign movements and generate high-resolution videos. Xiao et al. [115] also use GANs in a bidirectional communication process between deaf and hearing individuals. This work focuses on generating skeleton sequences of Chinese Sign Language (CSL) using a Bi-LSTM-based discriminator, while the generator has a two-level probability model based on encoding and decoding random samples from the distribution of a sign.

3.3.5 Generative Data Augmentation

Similarly, video generation works based on DDPMs have gained high representativity. These models allow for generating high-quality videos from a text input, as shown by Singer et al. [99] and Ho et al. [63]. Or copying the style of an image to an input sequence, as demonstrated by Esser et al. [48]. Although these diffusion models are prominent, there are few reference works focused on sign language. However, it is important to mention the work of Zhang et al. [124], which focuses on generating pose sequences from text, being relevant to our work.

We reviewed works that apply different forms of data augmentation in the context of sign language (Table 3.4). We found works from different areas, techniques, models, and training data, mostly utilizing DL approaches. However, in the case of Libras, the scarcity of works oriented towards this type of task has allowed our work to make significant contributions to the field.

Table 3.4: Works on data augmentation in this context.

Paper	Date	Application domain	Technique	Data type	Type	Tested model	Dataset	Metric
Stoll et al. [102]	2018	Sign language production	∼	GSL	DL	GANs	PHOENIX-2014	BLEU-4
Zhan [121]	2019	Hand gesture recognition	Spatio-temporal	Black and white hand gesture image	Basic	CNN	∼	Acc
Zanon de Castro et al. [120]	2019	Libras sign recognition	Spatio-temporal	Libras sign videos	Basic	3DCNN	∼	Mean Acc
Wu et al. [111]	2019	Human activity recognition	GANs	Human activity	DL	CNN, 3DCNN	UCF101, KTH	∼
Zhang et al. [125]	2019	Video classification	GANs	Ações humanas	DL	3DCNN	HMDB51, UCF101	Acc
Li et al. [75]	2019	Dynamic signal recognition	Temporal Cropping, Rotation, Translation, Spacial Cut	Hand action	Basic	3DCNN, md CNN	VIVA Dataset	∼
Zhang et al. [126]	2020	Video classification	GANs	Human action	DL	3DCNN, Inception-v3	HMDB51, UCF101	Acc
Pu et al. [86]	2020	Sign Language Recognition	XmDA	GSL	DL	GoogLeNet, BLSTM[3], CTC[4]	PHOENIX-2014	WER
Stoll et al. [103]	2020	Sign language video generation	Spatio-temporal	GSL	DL	gloss2pose, pose2video, GANs	PHOENIX-2014	MSE, Own metric
Xiao et al. [115]	2020	Sign language generation and recognition	Jittering	CSL, skeleton	DL	Bi-LSTM, GANs	CSL Continuous Sign Language Dataset	Acc
Tan et al. [104]	2021	Hand gesture recognition	9 Techniques	Hand positions	Basic	EDenseNet	UB-HDB, UST-HK, UW-Madison	Acc
Moryossef et al. [80]	2021	Language Translation	Heuristic	ASL, GSL, Glosses	Other	NMT[5]	NCSLGR dataset, DGS, PHOENIX-2014	BLEU, METEOR
Nguyen et al. [81]	2021	Sign language avatar generation	Mirroring	GSL	DL	CNN, ResNet-50	DGS Corpus	Acc
Ho et al. [63]	2022	Text-conditioned video generation	∼	Text-video	DL	DDPM	LAION-400M	FID, FVD, CLIP
Zhang et al. [124]	2022	Text-conditioned human moviment generation	∼	Human actions	DL	DDPM	KIT dataset, HumanML3D	FID, R Precision, Recall, Acc, F1 Score.
Chen et al. [25]	2022	Image generation	DDPM	Weed images	DL	VGG16, InceptionV3, DenseNet161, ResNet50	CottonWeedID15	Top-1%, Precision, Recall
Hu et al. [64]	2022	Image Classification	Attention-guided image cropping	Images	DL	ViT	CUB-200-2011, Stanford Dogs	Top-1%
Singer et al. [99]	2022	Video creation	∼	Video clips with captions	DL	Make-A-Video	WebVid-10M, HD-VILA-10M, MSR-VTT	FID[6], CLIPSIM, AMT[7]
Kothadiya et al. [71]	2023	Sign language recognition	Spatio-temporal	Hand position	DL	TF	∼	Recall, Precision, Accuracy
Azizi et al. [14]	2023	Data augmentation	DDPM	High-resolution images	DL	ResNet-50, ResNet-152, ViT	ImageNet	FID, Inception Score, Acc
Xiao et al. [114]	2023	Image caption generation	DDPM	Imagens and captions	DL	Fully Convolutional (FC) model	COCO	BLEU-4, METEOR, ROUGE, CIDEr, SPICE
Trabucco et al. [105]	2023	Data augmentation	DDPM	Object images	DL	ResNet50	Caltech101, Flowers102	∼
Esser et al. [48]	2023	Video creation	DDPM		DL			

Chapter 4

Methodology

In this chapter, we present the methodology proposed in this dissertation, which is summarized in the diagram depicted in Figure 4.1. In the following sections, we present each step of the methodology, including the used datasets, the preprocessing employed on those data, the data augmentation procedure, the model training procedure and the experimental protocols.

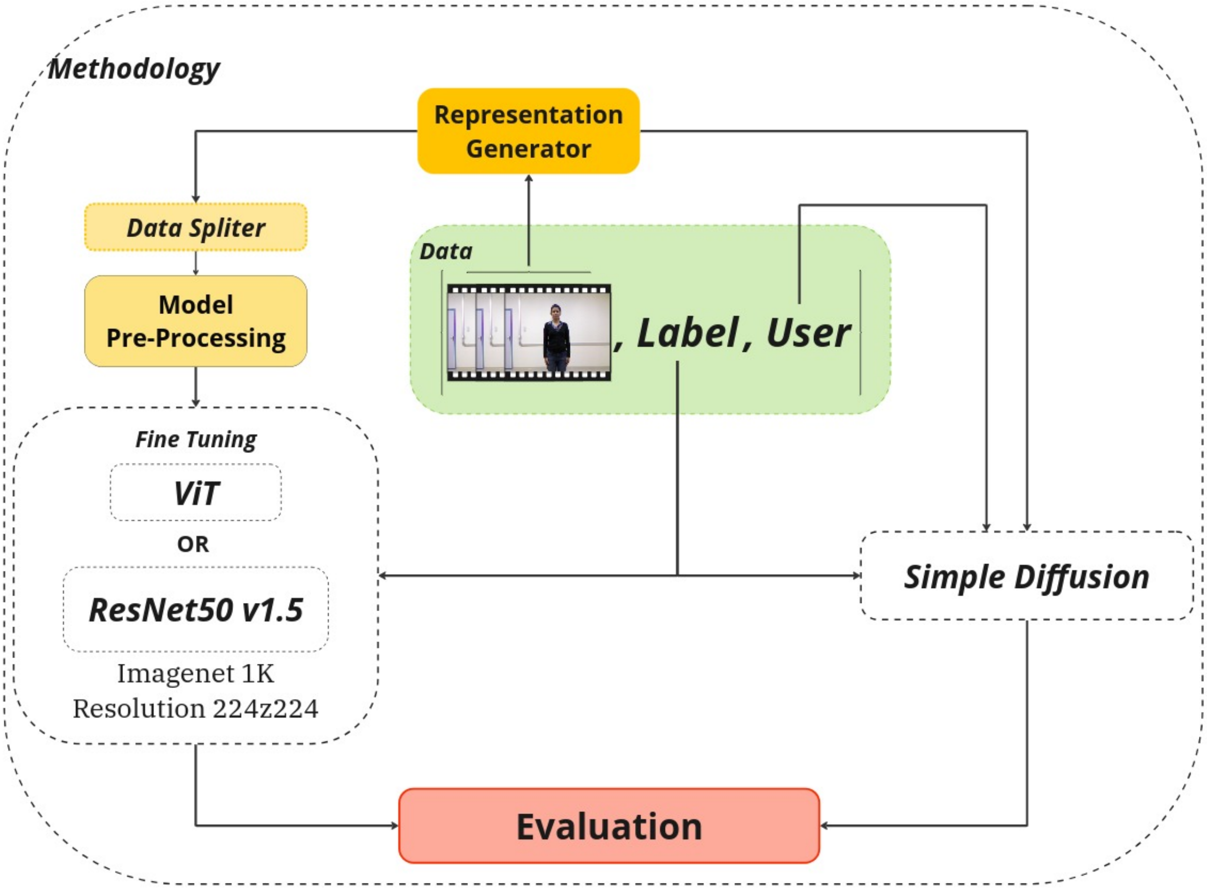


Figure 4.1: Pipeline of the methodology proposed in this dissertation. The image shows the data flow for each stage of our methodology, on the left is the classification, and the right is the data augmentation.

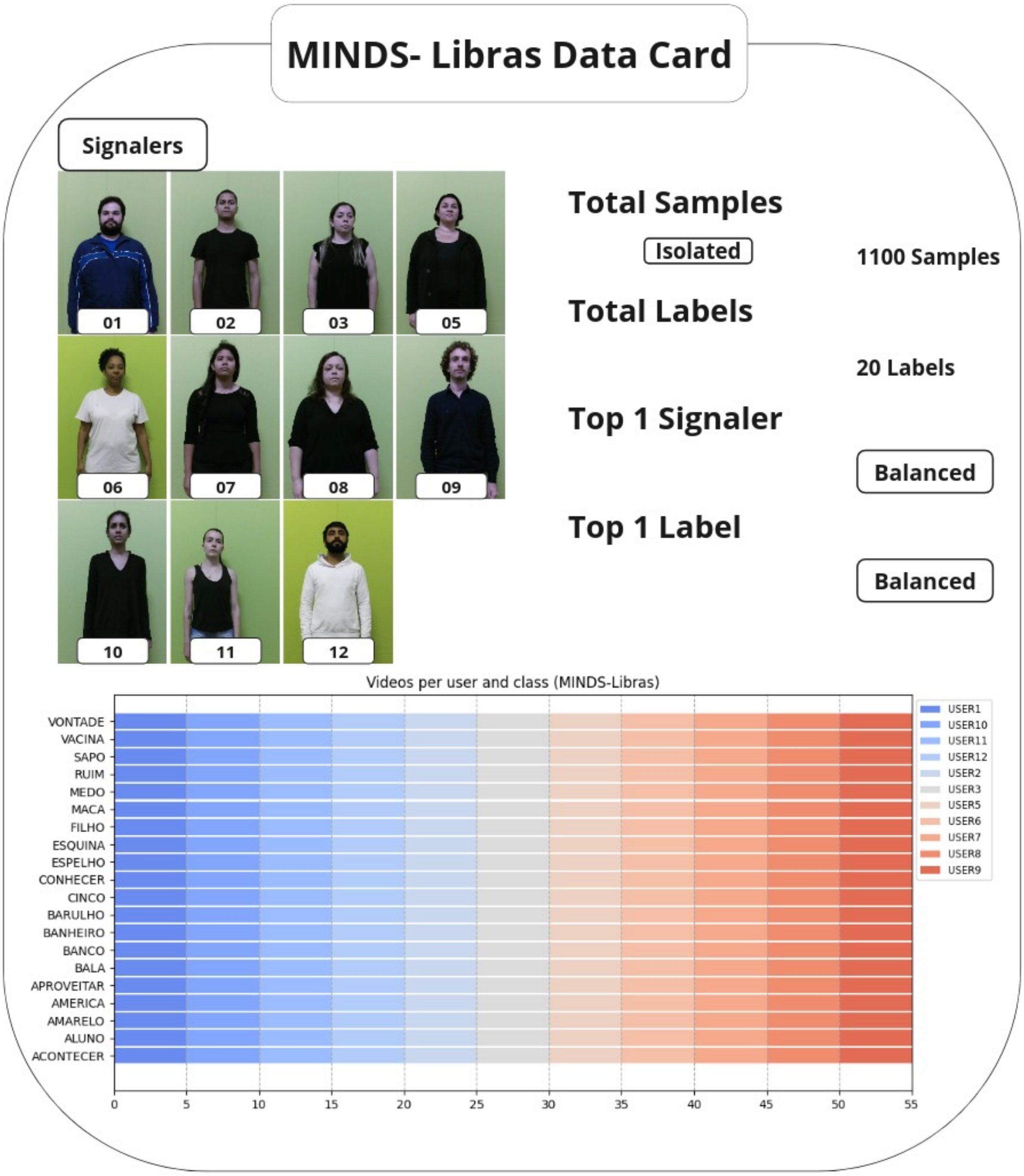


Figure 4.2: MINDS-libras data card. The image shows the characteristics of the dataset and how the classes are distributed for each Signer.

4.1 Dataset Choice

We chose the dataset based on various criteria, including availability, video quality, sample size, and number of signers. Our goal is to assess the model's performance in a variety of settings, ranging from low-quality datasets (with few samples, few signers, and unbalanced) to high-quality datasets.

Evaluating the model's performance under adverse situations is critical for understanding its limitations and identifying areas for improvement. Low-quality datasets can

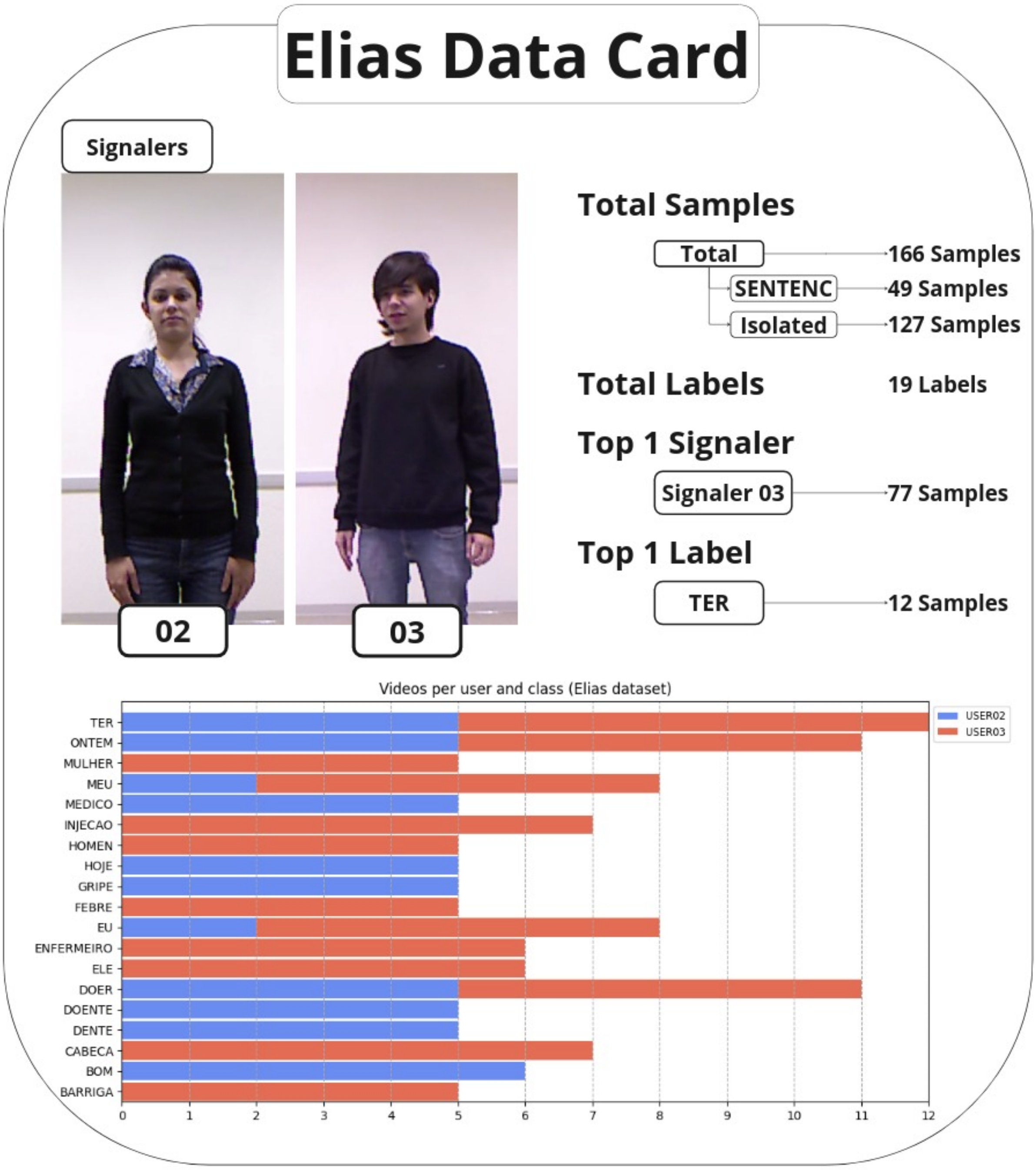


Figure 4.3: Elias data card. The image shows the characteristics of the dataset and how the classes are distributed for each Signer.

provide substantial issues, such as lighting fluctuations, inconsistent movements, and a lack of signer diversity, all of which can have a severe impact on the model's accuracy. In contrast, more robust and balanced datasets provide a more controlled and varied environment, allowing for a more comprehensive assessment of the model's effectiveness in accurately recognizing indicators.

Furthermore, the diversity of the signers, including variances in gender, age, and level of expertise with sign language, is critical to ensuring that a learning model can generalize well in real-world scenarios. We also assessed the representativeness of the signs in terms of phonological features, ensuring that the dataset encompasses a wide range of movements and hand positions.

We selected Elias Dataset and MINDS-Libras among the datasets discussed in Section 3.1.2.

- **MINDS-Libras:** This dataset, also discussed in Section 3.1.2, is superior to Elias Dataset in several aspects, although both cover a similar vocabulary. It is a balanced dataset made up of 12 signers, with the same amount of samples per signer and a consistent number of labels. All of its classifications correspond to standalone signs. To maintain the balanced qualities of this dataset, all samples from signer four will be deleted, leaving 1100 of a total of 1155. Figure 4.2 illustrates the dataset's details.

- **Elias Dataset:** As previously described in Section 3.1.2, this dataset was built for healthcare situations, with a vocabulary focused on generating statements specific to the scenario. Figure 4.3 shows only two signers of different genders and age differences. It includes 49 and 127 videos for sentences and isolated signs, respectively. The collection includes RGB, depth videos, and skeleton sequences. The first two are in *.MP4* format, while the skeletal sequences are in *.mat* format. Both have a resolution of 640x480. We will only use the 19 classes in this dataset that correspond to isolated signs, eliminating phrases, for a total of 26 classes. Also, we can see in Figure 4.3 depicts an imbalanced class distribution with insufficient examples for some classes per signer. We found that one signer had almost 50% more samples than the other. We're interested in this dataset because of its unique properties.

4.2 Pre-processing

Our approach intends to leverage pre-trained vision models that accept an image as input. However, our data is in video format. Based on Passos et al.'s work [83], we employ the GEI representation for motion analysis to maintain gesture dynamics from a sequence of photos. Thus, our pre-processing comprises the following steps: i) body segmentation, ii) body part selection, iii) per-frame masking, and iv) GEI creation. The method is shown in Figure 4.4. As a first step, we execute label extraction. This phase is not part of our pre-processing pipeline but is detailed in this section.

Label Extraction

The datasets do not include a structured label table. However, the video name correlates to the label that reflects each sign's class and the signer's sample number. At this point, the goal is to develop a data table that correlates the labels with the appropriate video URLs and then include it in our procedure. This will allow for proper label access during model training and evaluation.

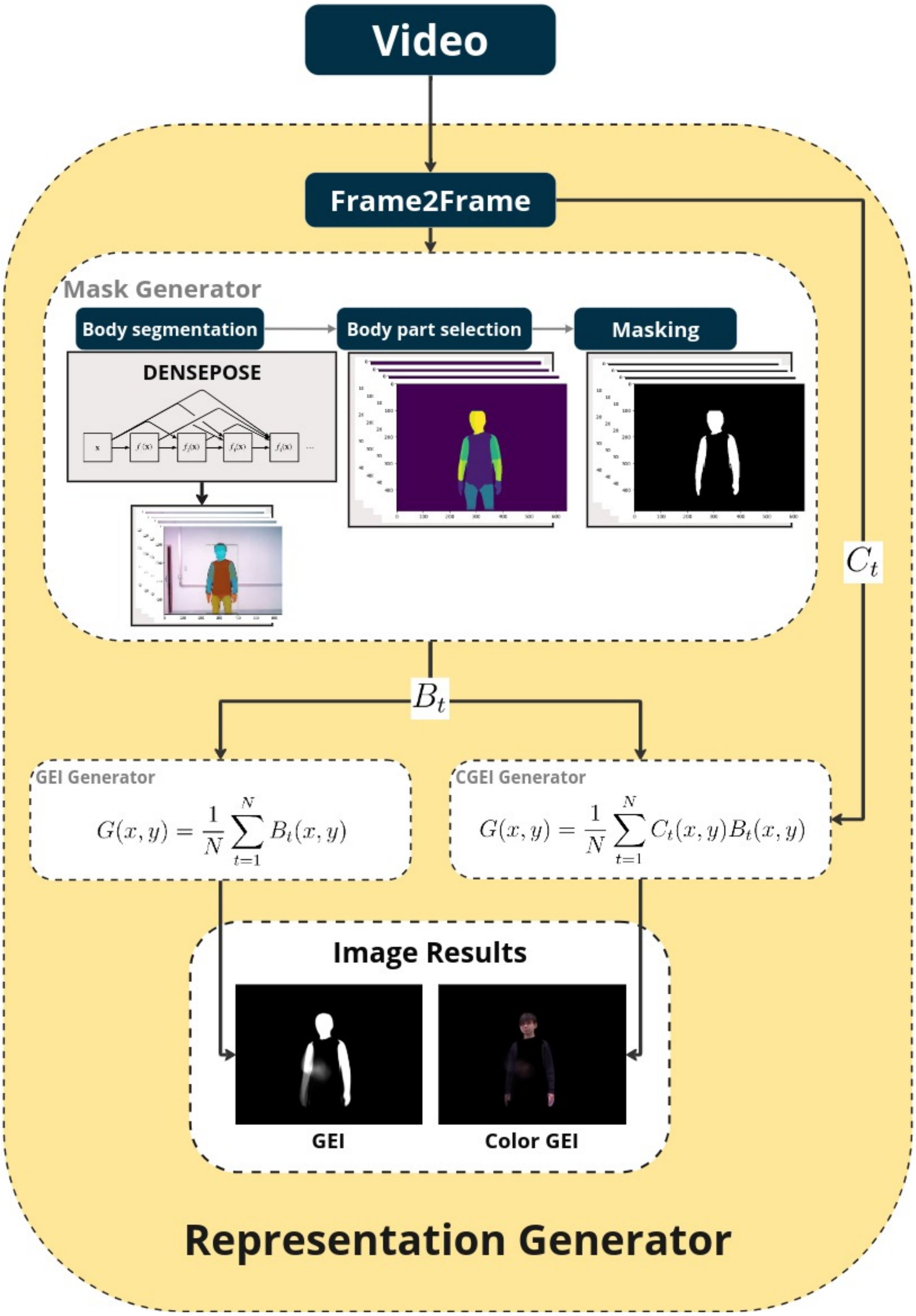


Figure 4.4: Pre-processing pipeline. The image shows each stage of our pre-processing.

Body Segmentation

Body segmentation is the initial stage in preparing video data for motion analysis. In this step, we employ image segmentation techniques to separate the person's body from the video background. We refer to the work of Passos et al. [83], who propose the usage of *DensePose* [57] for this purpose. *DensePose* recognizes bodily components including head, torso, upper and lower arms, upper and lower legs, hands, and feet. This process is depicted in Figure 4.4.

Body Part Selection

After separating the body from the background, the following step is to choose the body portions of interest. This procedure entails identifying and isolating key body areas, such as the head, arms, and hands, that are important for motion analysis. Accurate selection of body parts enables more precise and concentrated analysis of movements, ensuring that the dynamics of individual gestures are preserved and accurately portrayed in later steps.

Masking per Frame

We use the specified body components to make masks for each video sequence frame. Each mask highlights the regions of interest, enabling accurate observation of motion variations without focusing on the time. This stage involves applying masks on original frames, resulting in a series of frames that only show the specified body parts. Masking each frame aids in the detection of specific motion patterns, resulting in a clearer and more informative portrayal of gesture dynamics.

GEI Creation

The GEI is created from the sequence of masked frames. To generate the GEI, we use the formula 2.2.1, where the average pixel intensity values at each position (x, y) are applied over all frames in the sequence.

4.2.1 Color Gait Energy Image (CGEI)

The CGEI representation in Figure 2.1e, is an innovative extension of the traditional Gait Energy Image (GEI) that incorporates color information to enrich the visualization and analysis of gait patterns. While conventional GEI relies on grayscale to highlight intensity variations and capture the average silhouette of movement, CGEI introduces an additional dimension by assigning specific colors to different body regions. The application of this technique is given by:

$$G(x, y) = \frac{1}{N} \sum_{t=1}^{N} C_t(x, y) B_t(x, y), \tag{4.2.1}$$

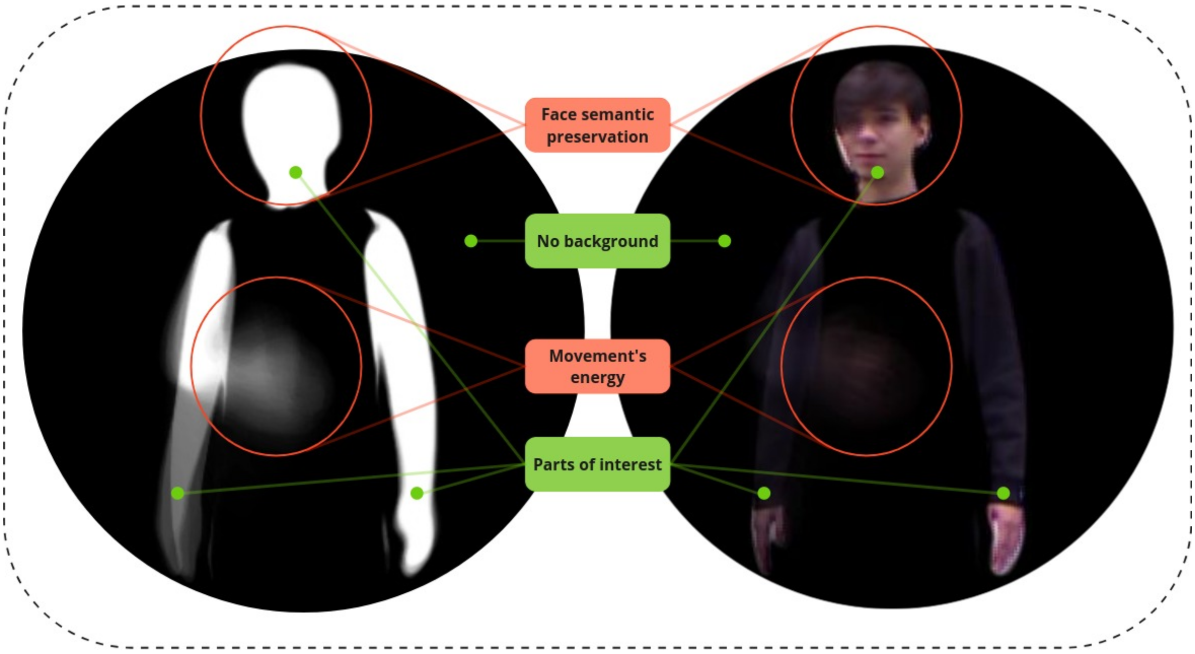
where B_t and C_t represent, respectively, the sequence of masks and the sequence of original frames. As in Equation 4.2.1, x, y, and N are the coordinates of each pixel in each image of the sequence and the total number of frames in the video.

This inclusion of chromatic information allows for a more detailed and precise representation, capturing specific configurations and orientations of the hands and other body parts and subtle variations in movement that might go unnoticed in a black-and-white image.

We propose CGEI as an alternative for motion analysis in the context of Libras. By combining the advantages of traditional GEI with the richness of information provided by color, CGEI becomes a powerful tool for analyzing human motion. In our work, we apply CGEI to observe its effects on our models and whether it can improve performance. Additionally, in the context of data augmentation, CGEI offers data types that enrich our qualitative analysis. This proposal aims to advance the development of more accurate and accessible technologies for the deaf community, contributing to the recognition and preservation of Libras.

CGEI Creation

The CGEI is also created from the sequence of masked frames. To add color information to our GEI representation, we use the formula 4.2.1, where before applying the average pixel intensity values at each position (x, y) over all frames in the sequence, we perform an extractive operation by multiplying the binary mask sequence by the image sequence that makes up the videos.



(a) GEI representation (b) CGEI representation

Figure 4.5: Pre-processing results for the label "FOME" in the Elias dataset. The image shows the differences between a) GEI Representation and b) CGEI Representation.

Finally, Figure 4.5, shows the characteristics and differences of the GEI (Figure 4.5a) and ColorGEI (Figure 4.5b) representations. A preliminary qualitative assessment shows that the ColorGEI representation has more visual complexity than the traditional GEI representation, keeping some facial features and hand shapes. Additionally, these features can be positive for providing gender variations to our model. Another positive aspect of both representations is the elimination of visual semantic context that is not relevant to our case, focusing only on the body parts where the movements are executed.

4.3 Data Split Protocol

To train the models, we defined two data split protocols, each designed to evaluate the model's performance in contrasting scenarios: *in-distribution* and *out-of-distribution*. The objective is to challenge the model both in controlled environments and under conditions that require greater generalization capability.

Protocol 1: Controlled Stratified Split (CSS)

The *Controlled Stratified Split (CSS)* protocol uses a stratified split of the data with proportions of 60%, 20%, and 20% for the training, validation, and test sets, respectively. We assigned this name because we made sure that all labels were present in the training set. This protocol seeks to simulate a controlled scenario where the samples are uniformly distributed without considering the variability between the users performing the gestures. The complexity of this protocol is low, as it does not present significant challenges in terms of model generalization, but it allows us to analyze its performance under standard and well-distributed conditions.

Protocol 2: Controlled User Split (CUST)

We propose The *Controlled User Split (CUST)* protocol. This follows a variant of the *Leave-One-Person-Out (LOPO)* method, presenting a scenario of greater complexity in terms of generalization. In practical applications, it is crucial that the model learns from a limited set of users and adapts to significant variations in the data distributions across different users. To address this need, in the Elias Dataset dataset, we selected one user who holds 40% of the samples for validation, leaving the model learning from a limited diversity. In the MINDS-Libras dataset, we chose seven users to compose the training set and four users for the validation set, maintaining a 60%-40% ratio for these sets. The test groups for both datasets were formed by splitting the validation set into two equal parts. This protocol introduces significant complexity, as it exposes the model to out-of-distribution cases, forcing it to generalize correctly to users it has not seen before.

4.4 Classification Models

In this section, we present the machine learning models used in our experiments, including the Vision Transformer (ViT) and the ResNet50. Both models were employed to classify the GEI representations containing signs. Below, we describe the architecture of each model and its justification in the context of our study.

4.4.1 Deep Residual Network (ResNet50)

The ResNet50 is a widely established Convolutional Neural Network (CNN) architecture, consisting of 50 layers. It was designed to address the vanishing gradient problem by incorporating residual connections. These connections allow the network to learn identity mappings, facilitating the training of deeper networks.

Pretrained Model

For our experiments, we used the pretrained model *microsoft/resnet-50* available on the Hugging Face platform. This model has been trained on the ImageNet-1k dataset, which contains a wide range of images and labels, and has demonstrated solid performance on image classification tasks.

Model Architecture

The ResNet50 architecture includes:

- **Initial Convolutional Layer:** This layer performs an initial convolution to extract basic features from the image;
- **Residual Blocks:** The network is composed of 16 residual blocks, each with multiple convolutional layers and residual connections that facilitate learning identity mappings. Each residual block allows information to pass through shortcut connections, improving the network's ability to learn complex features;
- **Pooling Layer:** After the residual blocks, global pooling is performed to reduce dimensionality and obtain a compact representation of the image;
- **Final Fully Connected Layer:** This layer performs the final classification based on the features learned by the network.

The hyperparameters chosen for this model are detailed in Table 4.2, which were optimized through an extensive search, as described in the Experiments section.

4.4.2 Vision Transformer (ViT)

The Vision Transformer (ViT) is a state-of-the-art model that applies the transformer architecture, originally developed for natural language processing, to image classification tasks (see details in Section 2.3). For our experiments, we used the pre-trained model *google/vit-base-patch16-224-in21k* from Hugging Face, which has been trained on the ImageNet-21k dataset.

Pretrained Model

The *google/vit-base-patch16-224-in21k* model is a version of ViT pre-trained on the ImageNet-21k dataset, which contains a wide variety of images and 1,000 labels. This model has demonstrated robust performance in image classification tasks and provides a solid foundation for fine-tuning on Libras datasets.

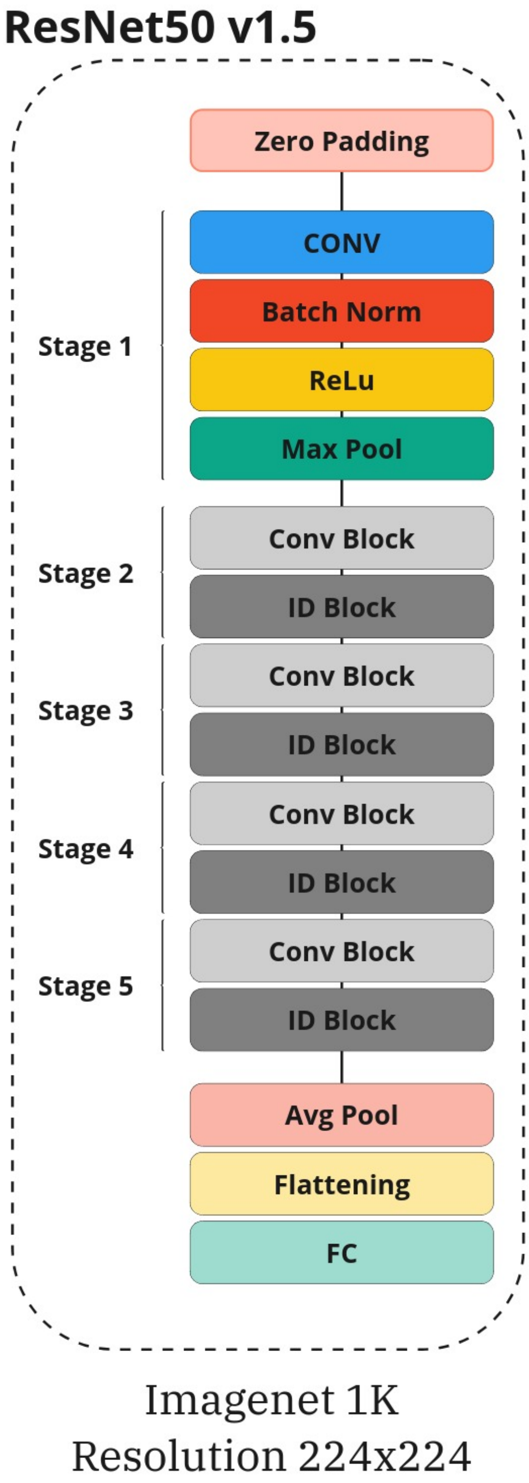


Figure 4.6: ResNet50 Architecture.

ViT Architecture

The ViT architecture is based on dividing the image into fixed-size patches (16x16 pixels in this case) and treating these patches as sequences, similar to words in a sentence. The

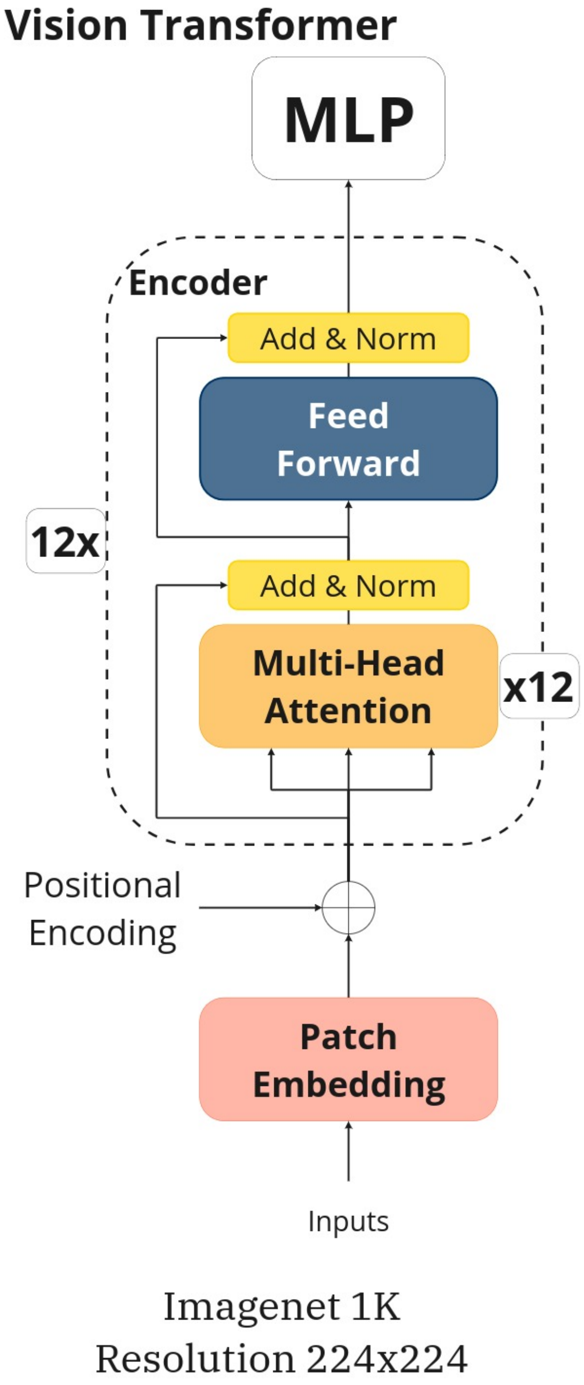


Figure 4.7: ViT model architecture.

architecture includes an embedding layer that converts the image patches into vectors, followed by multiple transformer encoder layers. Each encoder layer consists of multi-head self-attention mechanisms and feed-forward neural networks, allowing the model to capture global dependencies within the image (see details in Section 2.3).

The ViT architecture (Fig. 4.7) includes:

- **Image Patch:** 16x16 pixels;

- **Embedding Dimension:** 768;
- **Number of Encoder Layers:** 12;
- **Number of Self-Attention Heads:** 12;
- **Feed-Forward Dimension:** 3072.

The hyperparameters selected for this model are detailed in Table 4.2, following an extensive hyperparameter search as described in the Experiments section.

4.4.3 Evaluation Metrics

To evaluate the performance of the classification models, we use several standard metrics in the field of machine learning. In this study, we chose to start our evaluation with recall due to its relevance in the context of sign language classification, where the ability to correctly identify gestures is crucial. Subsequently, we use the F1-score to ensure that the model not only detects a large number of gestures but also maintains acceptable precision. Finally, we present accuracy as an overall view of performance but contextualize it after evaluating the more specific metrics of recall and F1-score to prevent metrics like accuracy from masking underlying issues in minority classes or imbalanced scenarios.

Below, we describe each of these metrics and their relevance for evaluating our models.

Recall

Recall, also known as sensitivity or completeness, is defined as the number of true positives divided by the sum of true positives and false negatives. This metric measures the model's ability to correctly identify all positive instances:

$$Recall = \frac{TP}{TP + FN}, \tag{4.4.1}$$

where TP is the number of true positives and FN is the number of false negatives. Recall is crucial in our context because we are interested in ensuring that the model correctly identifies as many correct gestures as possible, especially when working with imbalanced data or in scenarios where failing to recognize a gesture could have significant consequences.

Precision

Precision is the ratio of true positives to the total number of instances predicted as positive (the sum of true positives and false positives). This metric tells us how many of the positive predictions made by the model were correct:

$$Precision = \frac{TP}{TP + FP}, \tag{4.4.2}$$

where FP is the number of false positives. In the context of sign language classification, precision helps us understand if the model is being cautious when classifying gestures, minimizing incorrect predictions.

F1-Score

The F1-score is the harmonic mean between precision and recall. This metric provides a balance between precision and completeness, and is useful when a single metric is needed to evaluate performance in class imbalance contexts:

$$\begin{aligned} F1 &= 2 \times \frac{\mathit{Precision} \times \mathit{Recall}}{\mathit{Precision} + \mathit{Recall}} \\ &= 2 \times \frac{TP}{2 \cdot TP + FP + FN}, \end{aligned} \tag{4.4.3}$$

where FP is the number of false positives. The F1-score focuses on balancing recall and precision. This is important because, while we want the model to recognize many gestures (high recall), we also need to ensure that those gestures are classified correctly. The F1-score helps us evaluate whether the model is managing this balance well, showing whether high recall is accompanied by high precision.

Accuracy

Overall accuracy, or *accuracy*, is defined as the total number of correct predictions divided by the total number of instances. This metric provides a general overview of the model's performance:

$$\mathit{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{4.4.4}$$

where TN is the number of true negatives. Although accuracy is a metric that is often presented first due to its simplicity, it is more useful after evaluating recall and F1-score in this case. This is because accuracy can be misleading if we do not consider how the model performs on minority classes. Accuracy gives a general overview of performance but should be considered after better understanding the model's behavior in terms of detection and precision through recall and F1-score.

4.4.4 Qualitative Evaluation with Vision Transformer Attention Maps

To complement the quantitative analysis of the models in terms of standard metrics such as recall, F1-score, and accuracy, we propose conducting a qualitative evaluation using the attention maps generated by the ViT. Attention maps allow us to visualize the regions of the image that the model deems most relevant during the classification task, providing a more detailed view of its decision-making process.

The qualitative analysis will include the following aspects:

- **Exploration of Attention Points in Complex Gestures:** We will use attention maps to examine whether the ViT can focus on critical areas of the image during more complex gestures, which may involve multiple simultaneous or sequential movements. This analysis will help us verify if the model is adequately capturing the most important spatial features;

- **Comparison of GEI and CGEI Representations:** Attention maps will be used to compare the behavior of the ViT when using different types of representations (GEI and CGEI). We will evaluate whether the model shows different attention patterns depending on the type of input, which might suggest differences in how representations affect the classification process;

- **Identification of Distinctive Signals:** We will focus on analyzing how the ViT identifies distinctive signals within images. We will evaluate if the model concentrates on specific visual features that are crucial for distinguishing visually similar gestures that have different meanings.

To generate the attention maps, we will extract the outputs from the attention heads of the ViT. These maps will be overlaid on the input images, allowing for direct visual inspection of the areas influencing the model's predictions. This methodology will allow us to verify the effectiveness of the model and interpret and diagnose potential failures in the ViT's decisions.

The results of this qualitative evaluation will provide a deeper understanding of the ViT's performance, allowing us to identify the strengths and limitations of the model in different classification contexts.

4.5 Generative Models

4.5.1 Simple Diffusion Model

In this work, we propose the *Simple Diffusion* model, designed to generate new images through a forward and reverse diffusion process. Our approach follows the paradigm of stochastic diffusion models, in which progressively increasing levels of noise are added to an image until it becomes pure noise, and then neural networks are trained to reverse this process and generate new images from the noise.

Model Architecture

The *Simple Diffusion* model (In Figure 4.8) uses a modified *U-Net*-based architecture, whose input includes:

- The noisy image $\mathbf{x}_t$.

- The *timestep (t)*, indicating the current stage of the diffusion process.

- The label associated with the image class.

- The user identifier, allowing the model to learn variations related to different users in the dataset.

The *U-Net* consists of convolutional layers with *skip connections* between the encoding and decoding layers, facilitating the preservation of important details during the image reconstruction process.

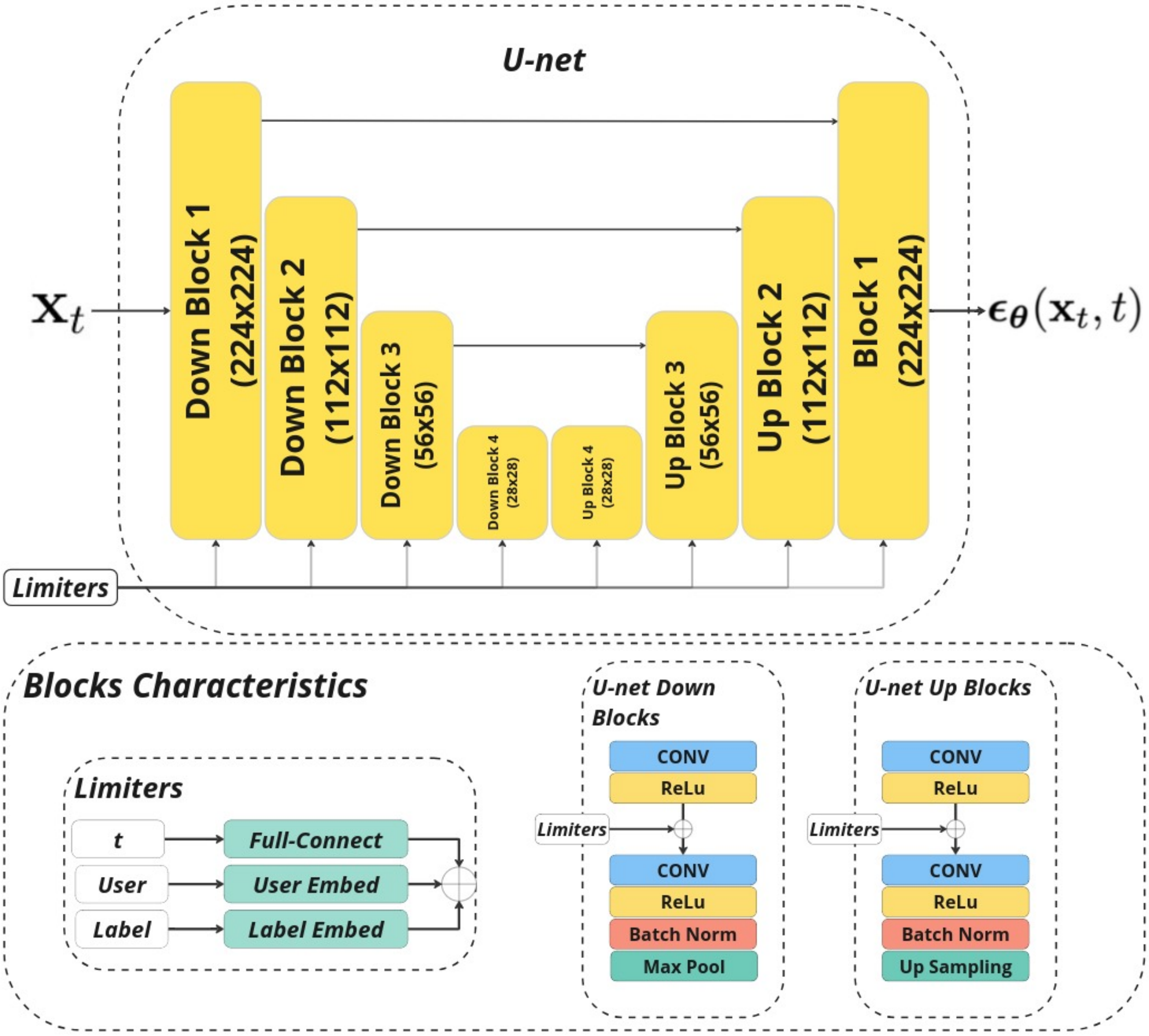


Figure 4.8: Simple Diffusion. The image shows a U-Net with 224×224 input and output blocks, the composition of the downlink and uplink blocks, and the composition of the limiter block used for this work.

4.5.1.1 Schedulers Used

We explored two types of *schedulers* in the reverse diffusion process:

- **Cosine Scheduler**: This scheduler adjusts the diffusion step according to a cosine function, allowing a smooth transition between different stages of the process. We recreated the step schedule as described by Nichol and Dhariwal [82], as follows:

$$\bar{\alpha} = \frac{f(t)}{f(0)}, \qquad f(t) = \cos\left(\frac{t/T + s}{1 + s} \cdot \frac{\pi}{2}\right)^2. \tag{4.5.1}$$

  A β_t value (noise coefficient of time step t) can be expressed as:

$$\beta_t = 1 - \frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}. \tag{4.5.2}$$

In practice, the values are constrained to $\beta_t < 0.999$ to avoid singularities around $t = T$.

- **Linear Scheduler**: This scheduler adjusts the diffusion step linearly, implying a constant change in the amount of noise removed at each step.

Both *schedulers* were tested to evaluate their impact on the quality of the generated images.

4.5.2 Evaluation of the Simple Diffusion Model

Instead of using traditional quantitative metrics such as Frechet Inception Distance (FID) or Inception Score (IS), we opted for a qualitative evaluation of the images generated by the *Simple Diffusion* model. This decision is based on the fact that these metrics do not specifically assess the composition of the generated gesture, focusing more on the overall structure of the image. Preliminary tests with these metrics on the dataset itself showed no significant differences in the images, leading us to conclude that they were not suitable for capturing the nuances of the gesture itself.

Therefore, the qualitative evaluation focused on visually inspecting the generated images to verify:

- The visual coherence of the generated images compared to the original data;
- Consistency in the representation of gestures;
- The model's ability to preserve important visual details throughout the diffusion process;
- Preservation of key features of the gesture in relation to the associated label and user;
- Diversity of the generated images, especially regarding inter- and intra-class variations;
- The overall quality of the visual composition.

This evaluation provided us with a deeper understanding of the model's capabilities to generate visually coherent gestures under different noise conditions and data variations.

4.6 Experiments

In this section, we present the experiments conducted to address our research questions. The aim is to compare the performance of a ViT (RQ1) and to evaluate the feasibility of augmenting a Libras dataset with data from other Libras datasets (RQ2).

4.6.1 Generative Data Augmentation

To address RQ1 of this dissertation, we designed three experiments using our *Simple Diffusion* model for generating new data from the Elias Dataset dataset, and in a third experiment, with the inclusion of a new external dataset. As previously mentioned in Section 4.1, the Elias Dataset dataset includes two users (User 1 and User 2) and 19 signs. Six signs are performed by User 1, eight by User 2, and only five signs are common to both users. Coincidentally, User 2 represents 60% of the dataset. Initially, our goal is to evaluate the quality of the generation in terms of user features and labels and to determine if it is possible to fill in the missing labels for one user using those from the other (*Cross-User*) and effectively balance the dataset.

Experiment 1: Generation with Unknown User

In the first experiment, we trained the *Simple Diffusion* model using only data from User 2. Subsequently, we performed inferences with both known and unknown users and labels. This setup allows us to assess the model's ability to generate new signs in a context where certain users and labels are not present in the training set.

Experiment 2: Inclusion of Known and Unknown Users

For the second experiment, we incorporated two samples of each class from User 1 into the training dataset and trained the model with data from both users (User 1 and User 2) to increase the variety of signs available during training. We then performed inferences not only with these two users but also with a third unknown user. This increases the complexity of the experiment, allowing us to observe the model's ability to generalize gesture generation to users who were not present in the training set.

In both experiments, we conducted a qualitative evaluation of the generated results, analyzing the coherence of the gestures in relation to the expected signs. We chose this qualitative evaluation because traditional metrics, such as FID and IS, do not allow for the assessment of the specific composition of gestures. In preliminary tests, these metrics did not indicate a significant change in the visual structure of the images, which justifies their omission in favor of a more detailed visual assessment.

4.6.2 Vision Transformers Performance

To address RQ2, we compared the performance of the ViT with other deep learning models, such as ResNet50, using two Libras datasets. Additionally, our results were compared with those obtained in other works in the literature, such as Passos et al. [83].

These comparisons were made from different perspectives: model comparison (ViT vs ResNet), dataset comparison (Elias vs Minds), representation comparison (GEI vs CGEI), and data splitting protocol comparison (Protocol 1 vs Protocol 2).

We also subjected the models to training processes with degraded data to observe their behavior with varying amounts of data, thus evaluating their robustness and generalization capability under limited conditions.

Experimental Design

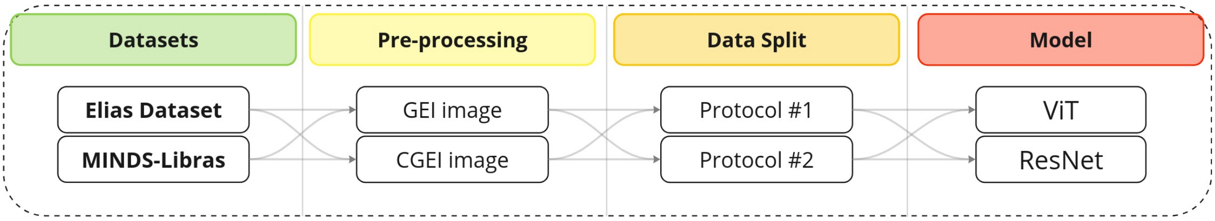


Figure 4.9: Scheduler for classification performance experiments.

The experimental design, shown in Figure 4.9, structured our list of experiments, detailed in Table 4.1. This scheme guided us in subjecting each model to different data scenarios. The selected datasets were used independently and, subsequently, the obtained results were analyzed and compared.

Table 4.1: Experiments list: Each experiment's name is composed of each element in line.

	Model		**Dataset**		**Pre-pro.**		**Split**		
ID	***ViT***	***ResNet***	***Elias***	***Minds***	***GEI***	***CGEI***	***P_1***	***P_2***	**Experiment Name**
1	✓		✓		✓		✓		*ViT-Elias-GEI-P_1*
2	✓		✓		✓			✓	*ViT-Elias-GEI-P_2*
3	✓		✓			✓	✓		*ViT-Elias-CGEI-P_1*
4	✓		✓			✓		✓	*ViT-Elias-CGEI-P_2*
5	✓			✓	✓		✓		*ViT-Minds-GEI-P_1*
6	✓			✓	✓			✓	*ViT-Minds-GEI-P_2*
7	✓			✓		✓	✓		*ViT-Minds-CGEI-P_1*
8	✓			✓		✓		✓	*ViT-Minds-CGEI-P_2*
9		✓	✓		✓		✓		*ResNet-Elias-GEI-P_1*
10		✓	✓		✓			✓	*ResNet-Elias-GEI-P_2*
11		✓	✓			✓	✓		*ResNet-Elias-CGEI-P_1*
12		✓	✓			✓		✓	*ResNet-Elias-CGEI-P_2*
13		✓		✓	✓		✓		*ResNet-Minds-GEI-P_1*
14		✓		✓	✓			✓	*ResNet-Minds-GEI-P_2*
15		✓		✓		✓	✓		*ResNet-Minds-CGEI-P_1*
16		✓		✓		✓		✓	*ResNet-Minds-CGEI-P_2*

The hyperparameter search was initially conducted for the eight experiments involving the Elias Dataset dataset. The best hyperparameters obtained from this search were then used for analogous experiments with the MINDS-Libras dataset. The tuned hyperparameters included: batch size, learning rate, optimizer, scheduler, warm-up ratio, and weight decay. The ranges and values explored during this search are presented in Table 4.2.

We opted for a Bayesian method for hyperparameter search, performing 50 iterations per experiment, with a duration of approximately 16 hours per search. The number of epochs was fixed at 32, as preliminary experiments showed that at least the ViT converged within this time. After the search, fine-tuning was performed using the best hyperparameters obtained (see Table 4.2).

To ensure the robustness of the results, each model was trained 10 times in each of the experiments, and the average results of the 10 runs were calculated. These experiments were conducted on an Nvidia A6000 GPU, with an approximate duration of 2 to 3 hours per experiment.

Table 4.2: Results of hyperparameter tuning for different models and datasets. Each row shows specific settings, including batch size, learning rate, optimizer, scheduler, warmup ratio, and weight decay.

Experiment Name	Batch size	Learning Rate	Optimizer	Scheduler	Warmup Ratio	Weight Decay
ViT-Elias-GEI-P_ 1	32	0.00096735	AdamW	CosineWarmupLr	0.22327	0.0042706
ViT-Elias-CGEI-P_ 1	16	0.00074534	AdamW	CosineWarmupLr	0.23634	0.015257
ViT-minds-GEI-P_ 1	64	0.00096735	AdamW	CosineWarmupLr	0.22327	0.0042706
ViT-minds-CGEI-P_ 1	64	0.00074534	AdamW	CosineWarmupLr	0.23634	0.015257
ResNet-Elias-GEI-P_ 1	16	0.00085242	AdamW	CosineWarmupLr	0.16932	0.06999
ResNet-Elias-CGEI-P_ 1	16	0.00092494	AdamW	CosineWarmupHr	0.22383	0.14474
ResNet-minds-GEI-P_ 1	64	0.00085242	AdamW	CosineWarmupLr	0.16932	0.06999
ResNet-minds-CGEI-P_ 1	64	0.00092494	AdamW	CosineWarmupHr	0.22383	0.14474
ViT-Elias-GEI-P_ 2	16	0.00007256	SGD	CosineWarmupLr	0.21701	0.25085
ViT-Elias-CGEI-P_ 2	16	0.000050607	SGD	CosineWarmupLr	0.21258	0.14791
ViT-minds-GEI-P_ 2	64	0.00007256	SGD	CosineWarmupLr	0.21701	0.25085
ViT-minds-CGEI-P_ 2	64	0.000050607	SGD	CosineWarmupLr	0.21258	0.14791
ResNet-Elias-GEI-P_ 2	32	0.00015006	SGD	CosineWarmupHr	0.25197	0.094596
ResNet-Elias-CGEI-P_ 2	8	0.00013392	SGD	CosineWarmupLr	0.11083	0.0029143
ResNet-minds-GEI-P_ 2	32	0.00015006	SGD	CosineWarmupHr	0.25197	0.094596
ResNet-minds-CGEI-P_ 2	8	0.00013392	SGD	CosineWarmupLr	0.11083	0.0029143

Finally, the obtained results were compared across different models, datasets, representations, and data splitting protocols. This detailed analysis allowed us to identify behavior patterns in each experimental context, providing a comprehensive view of the models' performance in different scenarios and offering a solid foundation to discuss the effectiveness of ViT compared to other deep learning architectures in sign language recognition.

Chapter 5

Results and Discussion

In this chapter, we present and discuss the main findings in our computational experiments. First, we show and discuss the results obtained in the classification results; in the sequence, we also present and discuss results from the data augmentation experiments.

5.1 Classification Baseline Results

We have established this baseline as a fundamental part of understanding the initial behavior of the proposed models on the dataset shown in Section 4.1. In this study, we have trained and evaluated the models using two distinct protocols, as described in Chapter 4. The analysis of these results addresses the second research question (RQ2) formulated in Chapter 1. We conducted several experiments using different models, datasets, representations, and data split protocols to understand how they impact classification performance. We evaluate CSS and CUST protocols and present the detailed results and comparative analysis below.

5.1.1 Analysis for the CSS Protocol

The CSS protocol, described in Section 4.3, ensures a stratified and controlled data split. This protocol guarantees that both classes and users are evenly represented in the training and testing partitions, allowing for a balanced and generalizable model evaluation. The results obtained under this protocol were thoroughly analyzed to identify performance patterns and areas for improvement.

The analysis was conducted at various levels, starting with a general comparison of the models used, followed by a more specific analysis considering the different datasets, and finally, a more detailed analysis considering preprocessing techniques. This approach allows us to isolate and understand how each element contributes to the model's performance and, in turn, provides key insights into the factors that most affect system behavior in sign language classification tasks.

Table 5.1: General Model Comparison Results for the CSS Protocol.

Model	Recall (%)	F1-Score (%)	Acc (%)
ViT	$\mathbf{97.2 \pm 2.0}$	$\mathbf{96.9 \pm 2.0}$	$\mathbf{97.2 \pm 2.0}$
ResNet	96.2 ± 1.6	95.9 ± 1.6	96.2 ± 1.6

Model Comparison: ViT vs ResNet

Since these models are designed with different architectures and learning mechanisms, the analysis focuses on how each model handles the classification task under the conditions of the CSS protocol.

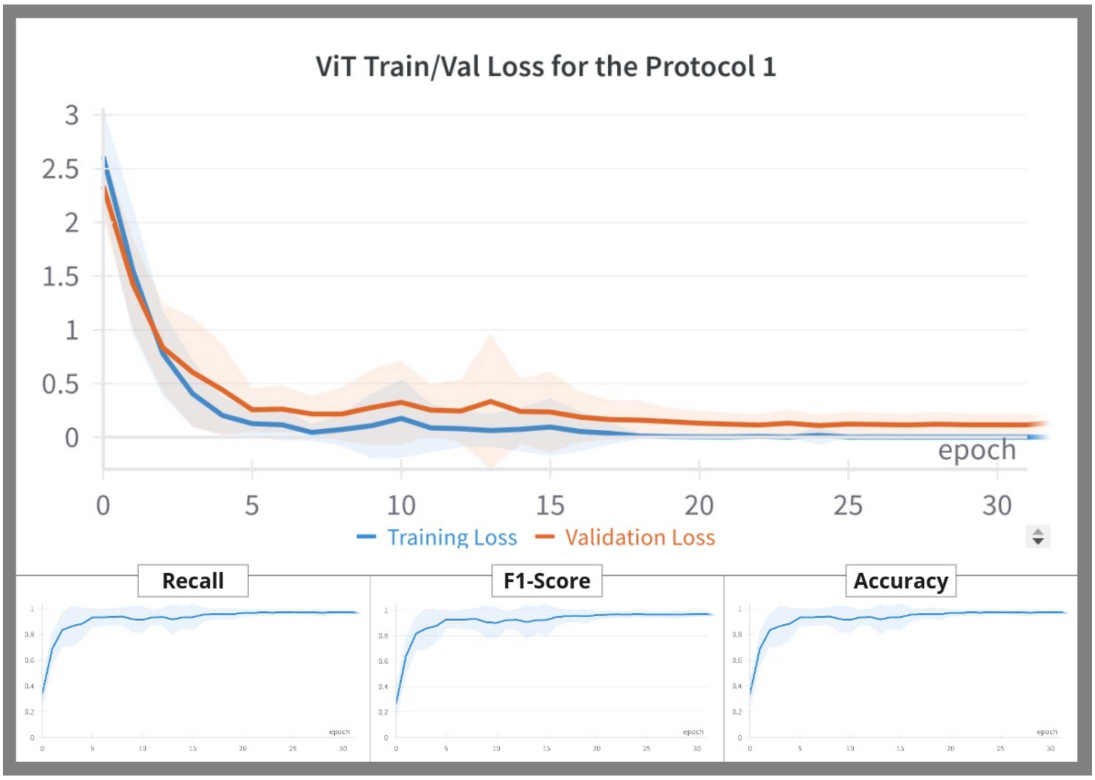


Figure 5.1: Train and Validation Loss for the CSS Protocol in ViT model.

In general, we observe that the ViT model slightly outperformed the ResNet50 model across all evaluated metrics. As shown in Table 5.1, ViT achieved an average *Recall* of 97.18%, an *F1-Score* of 96.96%, and an *Accuracy* of 97.18%, with respective standard deviations indicating the consistency of these results. Conversely, ResNet50 attained slightly lower values, with an average *Recall* of 96.18%, an *F1-Score* of 95.95%, and an *Accuracy* of 96.18%. These metrics reflect averages calculated across trials, focusing solely on model performance without additional factors.

Initially, the results suggest that both models are capable of classifying the signals with great accuracy. In fact, in Figure 5.1 and 5.2, we can observe a slight advantage of ViT over ResNet50, as ViT converges much faster and, overall, exhibits better validation loss stabilization compared to ResNet50. This advantage can also be perceived in the standard deviation of the curves (shaded areas), where the training and validation losses of ViT tend to converge with less variation in the final stages of the process, unlike ResNet50,

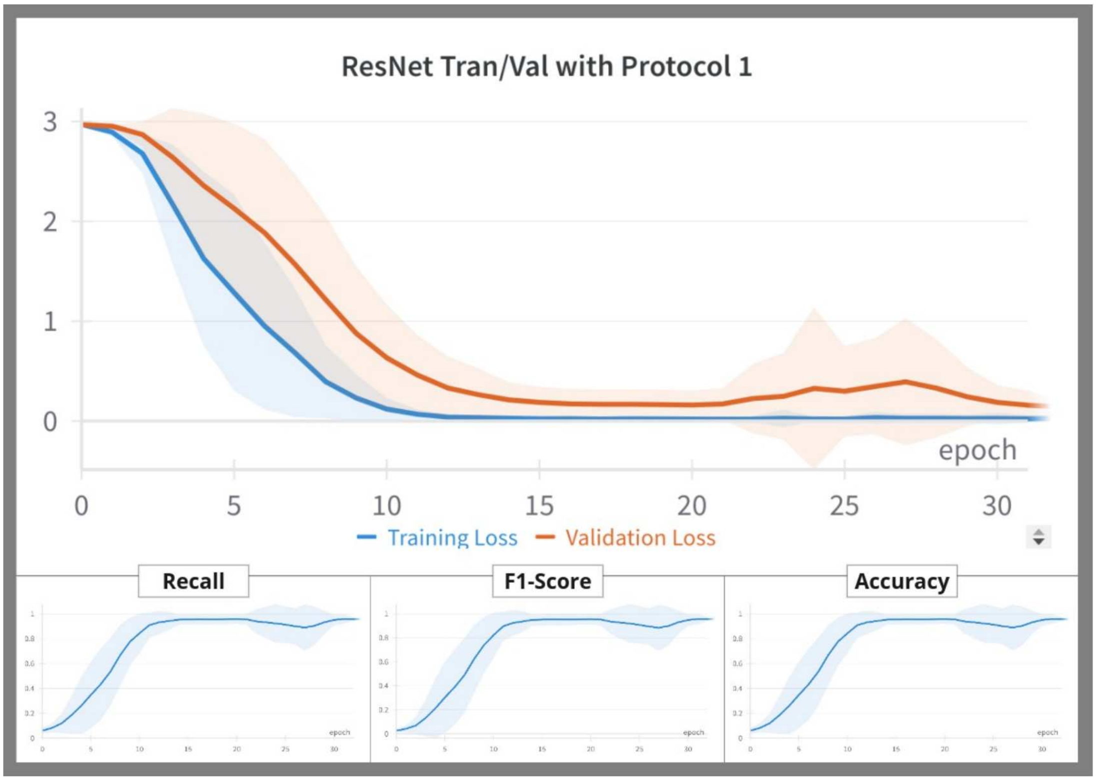


Figure 5.2: Train and Validation Loss for the CSS Protocol in ResNet50 model.

which still requires adjustments.

Another point of comparison between the models is their metric variations. In Figure 5.3, we can see how the *recall* of ViT varies in comparison to ResNet50. It is evident that the range of variation for ViT is smaller than that of ResNet50.

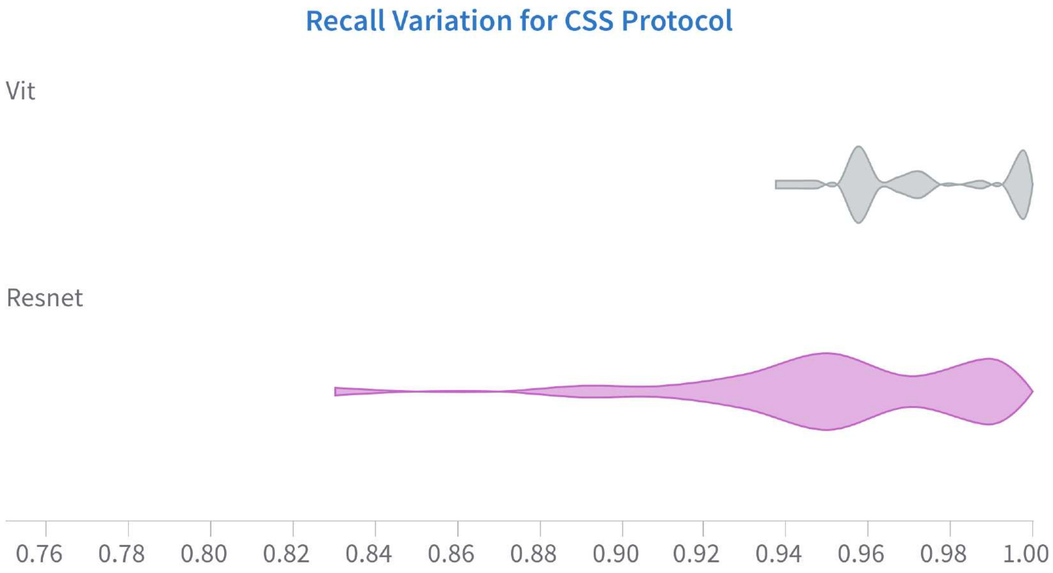


Figure 5.3: Recall Variation for the CSS Protocol. a) ViT and b) ResNet50.

Although we can initially assume good performance from both models, as we delve deeper into the analysis, we find this is not entirely the case.

Comparison Concerning the Datasets

In Section 4.1, the characteristics of each dataset used in this study were described in detail. It is evident that there are significant differences between Elias Dataset and MINDS-Libras, which go beyond the number of samples, signers, and classes. These differences include the diversity of the signals and the internal structure of the datasets, posing an additional challenge when training models that generalize effectively.

One feature that deserves special attention is the balance and homogeneity of the datasets. Elias Dataset, while having fewer classes, shows significant imbalance. This imbalance manifests in the unequal distribution of samples per class, the presence of classes with an insufficient number of examples, and the lack of intersection between some classes and different signers. Such imbalance may introduce biases in the model, hindering its ability to generalize to new signals. The limited number of signers and samples further exacerbates this issue, suggesting that the model could overfit to the specific characteristics of the dataset rather than learning generalizable patterns.

In contrast, MINDS-Libras presents more robust characteristics. Although it also has limitations, such as a relatively low number of samples per user, its structure is more homogeneous and balanced in terms of class and user distribution. This homogeneity is crucial for enabling the model to learn more effectively and generalize better to new or unseen signals.

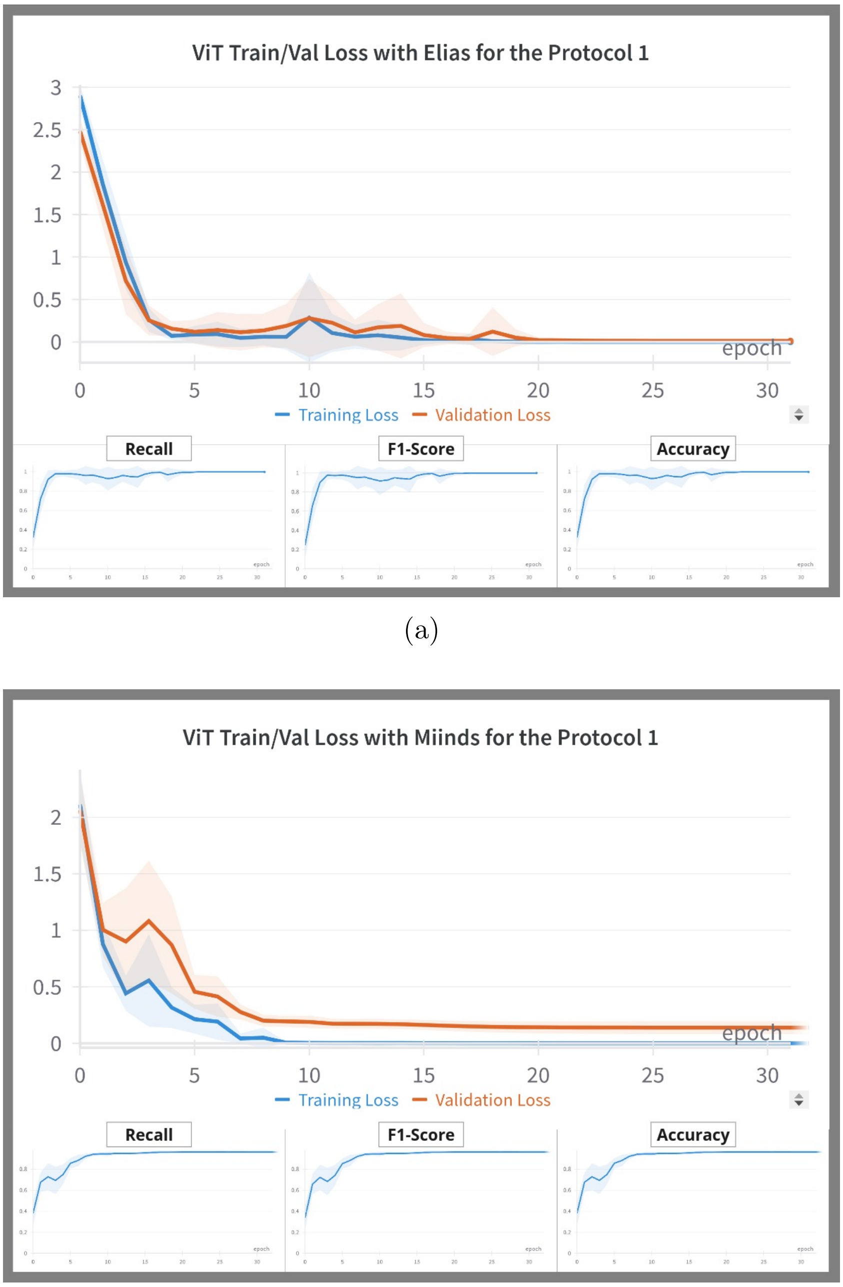
Table 5.2: Model Comparison Results Considering the Datasets for the CSS Protocol.

Model	**Dataset**	**Recall (%)**	**F1-Score (%)**	**Acc (%)**
ViT	*Elias*	98.0 ± 2.0	97.6 ± 2.0	98.0 ± 2.0
ViT	*Minds*	96.3 ± 1.6	96.3 ± 1.6	96.3 ± 1.6
ResNet	*Elias*	96.4 ± 5.0	95.9 ± 5.7	96.4 ± 5.0
ResNet	*Minds*	95.9 ± 1.6	95.9 ± 1.7	95.9 ± 1.6

This analysis is structured into two parts. In the first, we examine how each model performs on each dataset separately; in the second, we compare the relative performance between the datasets. Table 5.2 presents the general results of the metrics for each model across the different datasets.

For the Elias Dataset, ViT achieved high values, with a *Recall* of 97.99%, an *F1-Score* of 97.59%, and an *Accuracy* of 97.99%. While these results suggest that ViT is capable of capturing relevant patterns even in an imbalanced dataset with a limited number of samples, it is important to note that approximately 50% of the trained models showed signs of overfitting. Similarly, in Figure 5.4a, the convergence of the loss functions is nearly complete, indicating that the model is vulnerable to the low complexity of the data, suggesting that the task is relatively easy to solve due to the dataset structure in controlled scenarios.

On the other hand, the results obtained with MINDS-Libras were slightly lower but still high, achieving a *Recall* of 96.36%, an *F1-Score* of 96.34%, and an *Accuracy* of 96.36%. These values indicate that despite the mentioned limitations, MINDS-Libras provides a



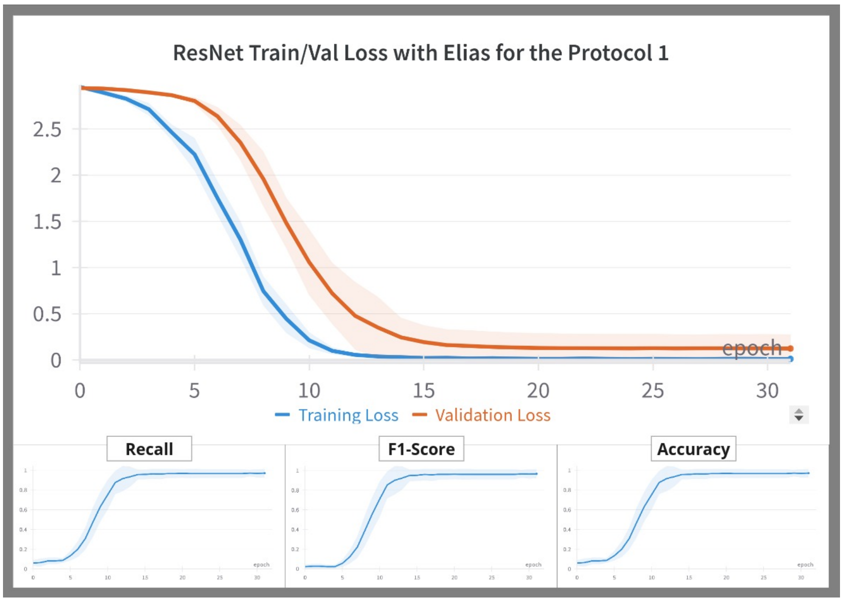
(a)

(b)

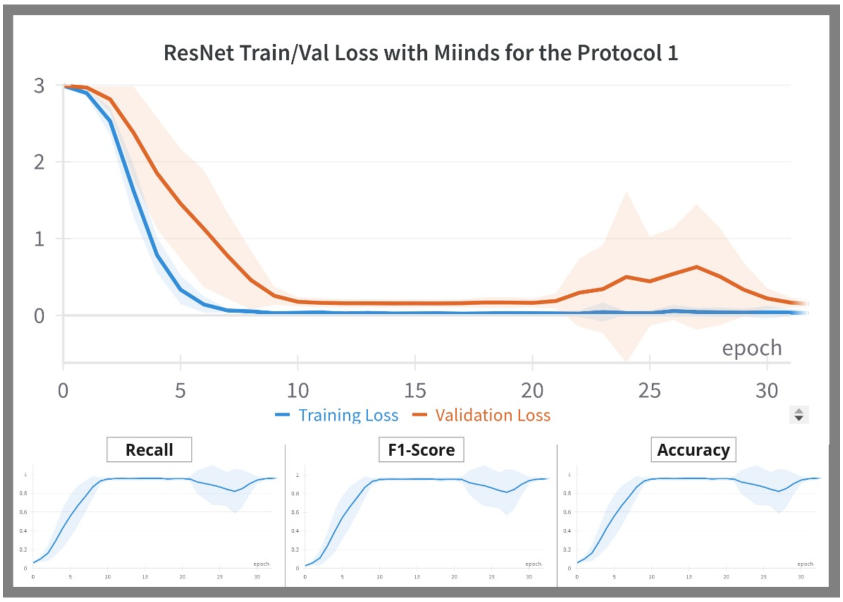
Figure 5.4: Train and Validation Loss Considering Models and Datasets for the CSS Protocol Part 1. a) ViT with Elias Dataset, b) ViT with MINDS-Libras.

more favorable environment for model generalization in controlled scenarios, possibly due to its better balance and homogeneity compared to Elias Dataset. In Figure 5.4b, the loss functions' convergence is not complete, suggesting that the complexity of MINDS-Libras is slightly higher, providing a more challenging structure for the model. Unlike Elias Dataset, only 10% of the models trained showed signs of overfitting with this dataset.

Regarding the performance of the ResNet50 model on the Elias Dataset, the results show slightly lower metrics compared to ViT, but still respectable, with a *Recall* of 96.39%, an *F1-Score* of 95.96%, and an *Accuracy* of 96.39%. Although ResNet50 also captures important patterns in an imbalanced dataset with a limited number of samples, its lower



(a)



(b)

Figure 5.5: Train and Validation Loss Considering Models and Datasets for the CSS Protocol Part 2. a) ResNet50 with Elias Dataset and b) ResNet50 with MINDS-Libras

generalization capacity suggests that this model is even more prone to overfitting. In this case, about 70% of the trained models showed signs of overfitting, reinforcing the hypothesis that the low complexity of the dataset facilitates excessive model fitting to the training data. The convergence of the loss functions follows a similar trend to that observed with ViT, although it is slightly slower, as shown in Figure 5.5a. This confirms that Elias Dataset presents a structure implying a lower complexity of the task.

The performance of ResNet50 on the MINDS-Libras dataset was also high, though slightly lower than that of ViT, achieving a *Recall* of 95.97%, an *F1-Score* of 95.95%, and an *Accuracy* of 95.97%. These metrics indicate good generalization in controlled settings.

In Figure 5.5b, a temporary fluctuation in performance is observed between epochs 21 and 27, followed by stabilization. This pattern may suggest the occurrence of a double descent phenomenon, where the model experiences a brief period of overfitting before reaching a more stable generalization. Further analysis would be needed to confirm if double descent or other model-specific factors are contributing to this behavior.

The variability between datasets directly influences the performance of the models. ViT showed slightly better performance than ResNet50 on both datasets, reaffirming its effectiveness in classification tasks regardless of data complexity.

Comparison of Representations: GEI vs. CGEI

Another aspect we examined is the impact of the type of signal representation on the model's performance. The representations GEI and CGEI, as discussed in Chapter 4, are compared in terms of how they facilitate or hinder the classification task for the ViT and ResNet50 models.

Table 5.3: Model Comparison Results Considering Datasets and Representations for the CSS Protocol.

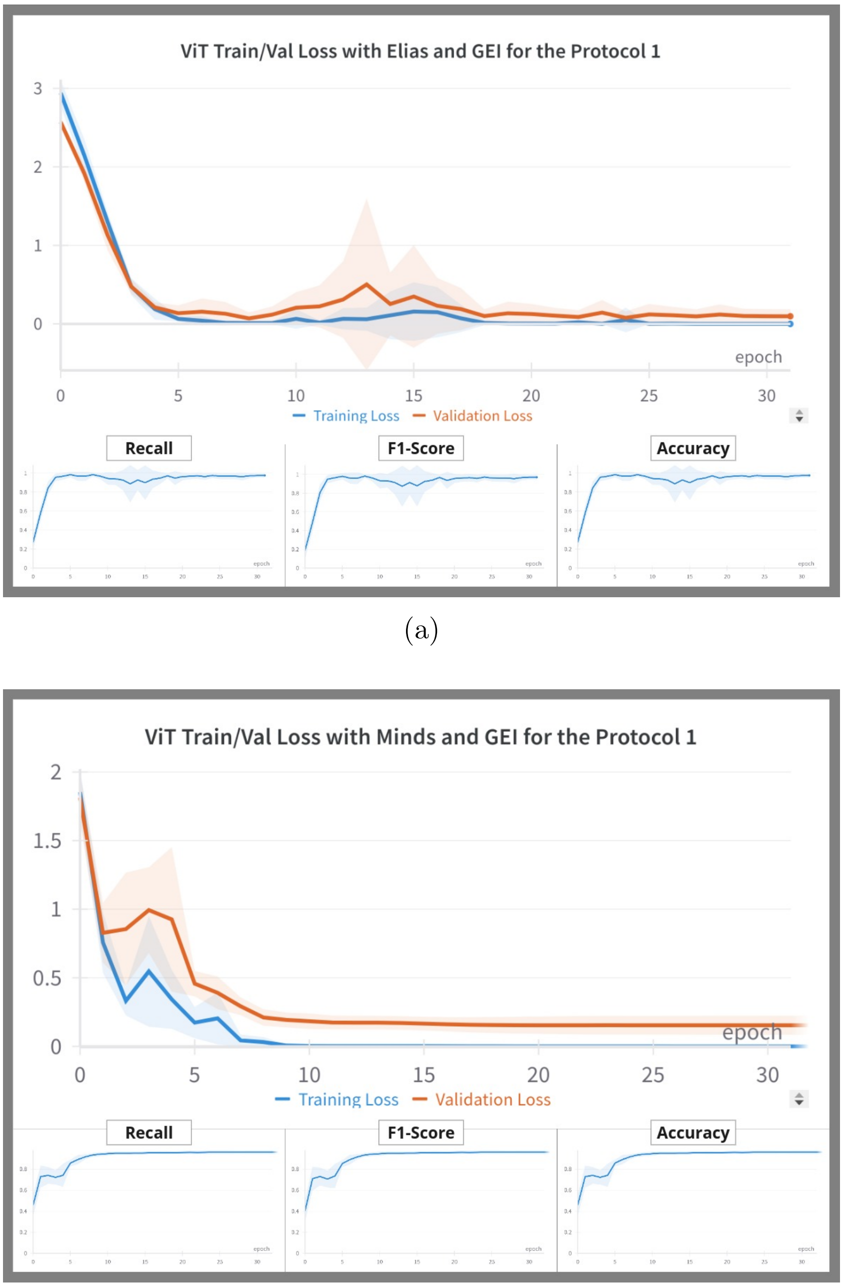
Model	Dataset	Pre-pro	Recall (%)	F1-Score (%)	Acc (%)
ViT	*Elias*	*GEI*	97.6 ± 2.0	97.1 ± 2.5	97.6 ± 2.0
ViT	*Minds*	*GEI*	95.9 ± 1.8	95.9 ± 1.8	95.9 ± 1.8
ViT	*Elias*	*CGEI*	98.4 ± 2.0	98.1 ± 2.5	98.4 ± 2.0
ViT	*Minds*	*CGEI*	**96.8 ± 2.0**	96.8 ± 2.0	96.8 ± 2.0
ResNet	*Elias*	*GEI*	98.4 ± 2.7	98.2 ± 3.2	98.4 ± 2.7
ResNet	*Minds*	*GEI*	96.7 ± 1.2	96.6 ± 1.3	96.7 ± 1.2
ResNet	*Elias*	*CGEI*	94.4 ± 5.0	93.7 ± 5.8	94.4 ± 5.0
ResNet	*Minds*	*CGEI*	95.3 ± 1.9	95.2 ± 1.9	95.3 ± 1.9

Table 5.3 presents the results obtained in each case, taking into account the model-dataset-representation relationships.

In the case of ViT, the model's performance varies depending on the representation used. When using GEI, ViT shows solid results across both datasets, although its performance is slightly better on Elias Dataset, with a *Recall* and *Accuracy* of 97.60% and an *F1-Score* of 97.09%. Notably, as mentioned earlier, 50% of the models trained on the Elias Dataset exhibited signs of overfitting, which may reflect the dataset's lower complexity. In contrast, on MINDS-Libras, the model experiences a slight decrease in performance, with values around 95.91%, suggesting that this dataset is more complex and varied.

When using CGEI, a representation incorporating chromatic information, ViT improves its performance on both datasets. On Elias Dataset, the *Recall* and *Accuracy* reach 98.40%, and the *F1-Score* rises to 98.08%, indicating that the CGEI representation facilitates better classification. On MINDS-Libras, although the improvement is more modest, the performance with CGEI remains superior to that of GEI, with a *Recall*, *Accuracy*, and *F1-Score* of 96.81%, reflecting that ViT can leverage the greater complexity of CGEI to capture the patterns present in the data better.

On the other hand, ResNet50 shows high performance with the GEI representation, although there are also differences depending on the dataset. In Elias Dataset, the model



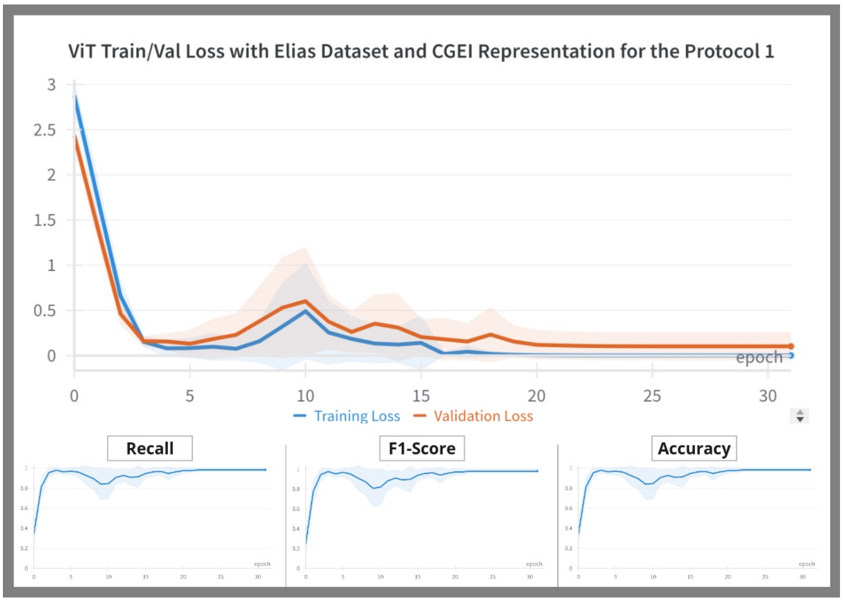
(a)

(b)

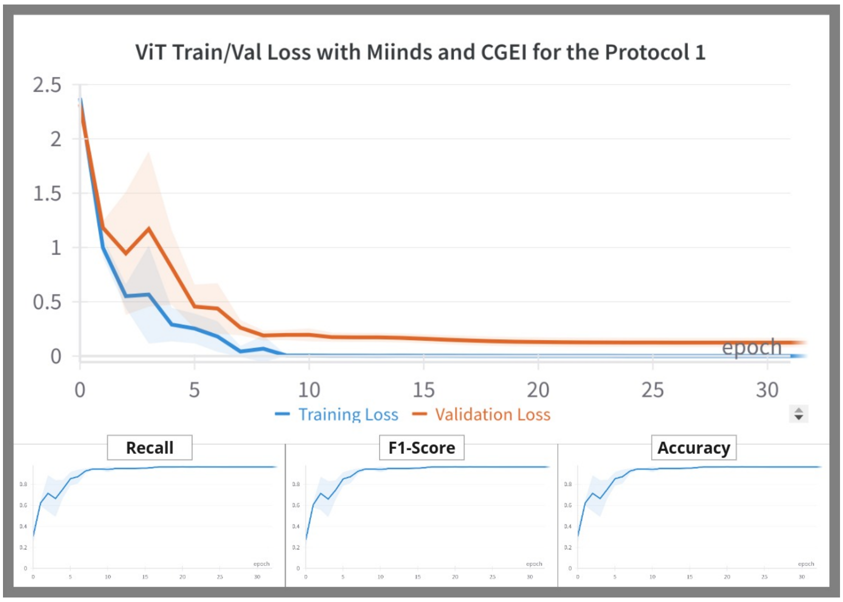
Figure 5.6: Train and Validation Loss Considering Models, Datasets and Representations for the CSS Protocol. a) ViT with Elias Dataset and GEI, and b) ViT with MINDS-Libras and GEI.

achieves a *Recall* and *Accuracy* of 98.40% and an *F1-Score* of 98.18%, while in MINDS-Libras, the results slightly decrease, with a *Recall* of 96.68% and an *F1-Score* of 96.63%. This suggests that while ResNet50 handles signals well with GEI, it faces greater challenges with MINDS-Libras due to the diversity of the signals.

However, when using CGEI, ResNet50 does not experience the same improvements as ViT. In Elias Dataset, performance drops, with a *Recall* of 94.40% and an *F1-Score* of 93.73%, indicating that the addition of chromatic information is not as beneficial for this model. In MINDS-Libras, although the decline is less pronounced, with a *Recall* of 95.27% and an *F1-Score* of 95.26%, it still does not reach the performance observed with



(a)



(b)

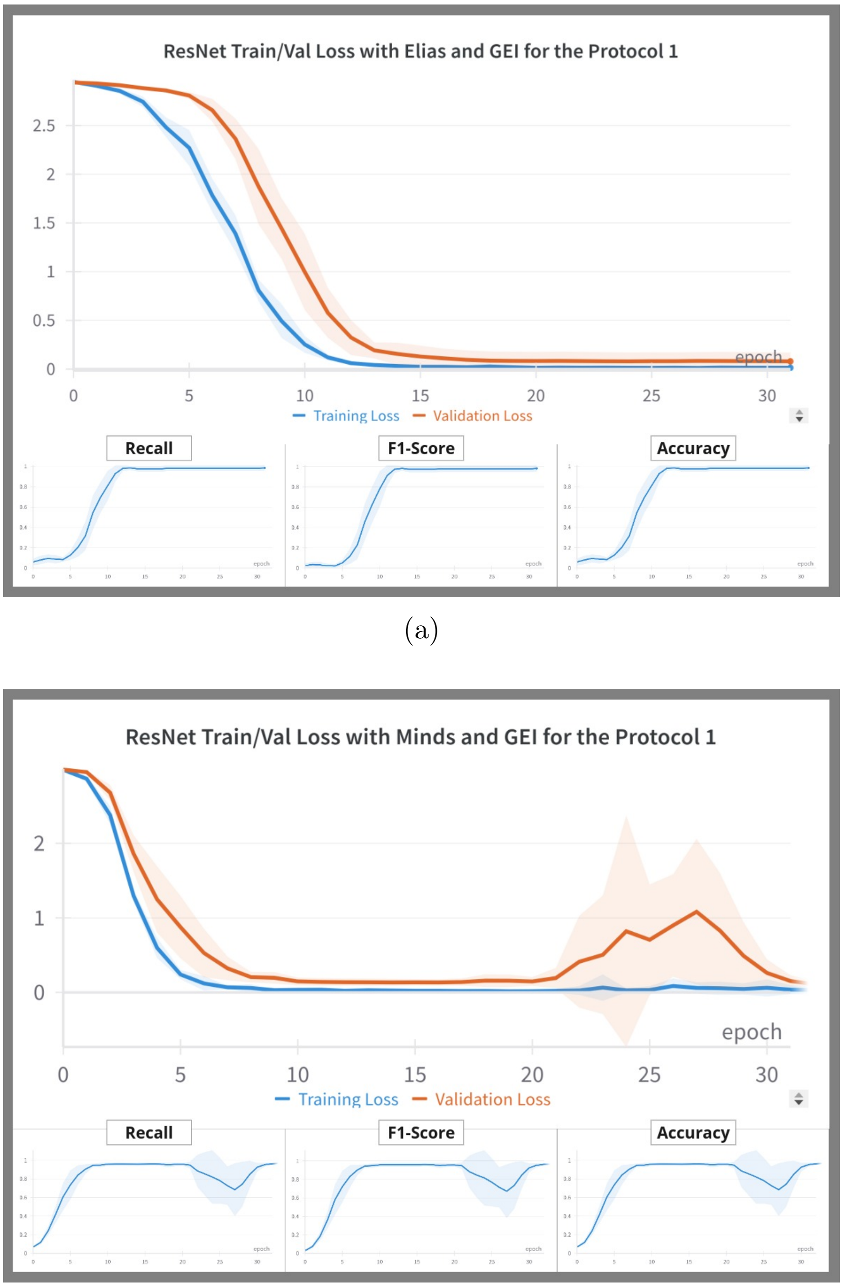
Figure 5.7: Train and Validation Loss Considering Models, Datasets and Representations for the CSS Protocol. a) ViT with Elias Dataset and CGEI, and b) ViT with MINDS-Libras and CGEI.

GEI.

In Figure 5.6, 5.7, 5.8, and 5.9, the loss curves for all experiments are displayed. It is easy to observe that in cases involving the Elias Dataset, the validation losses converge almost exactly with the training loss. However, in situations where this convergence is less evident, as in Figure 5.9a, the standard deviation (represented by the shaded area) shows significant variability, confirming the strong presence of overfitting in both models.

In contrast, with the MINDS-Libras dataset, it can be seen that the ViT model exhibits more stable behavior compared to its counterpart, ResNet50.

Overall, this analysis reveals that in controlled scenarios, ViT significantly benefits

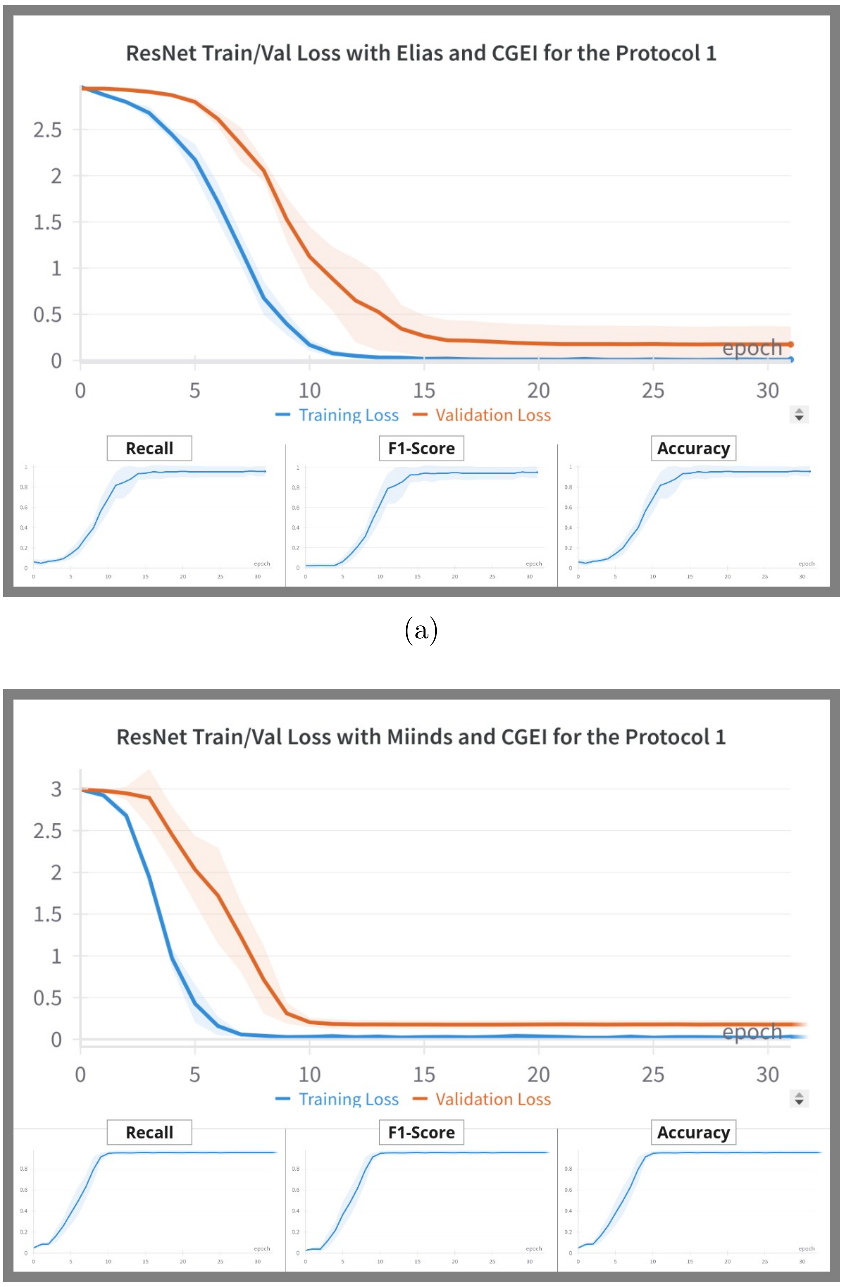


(a)

(b)

Figure 5.8: Train and Validation Loss Considering Models, Datasets and Representations for the CSS Protocol. a) ResNet50 with Elias Dataset and GEI, and b) ResNet50 with MINDS-Libras and GEI.

from the CGEI representation. This could be attributed to the pre-training of the model. On the other hand, ResNet50 does not seem to take advantage of this additional complexity, performing better with GEI. Although both models face greater challenges with the MINDS-Libras dataset, CGEI helps ViT mitigate some of these difficulties. This suggests that CGEI is more advantageous for models capable of processing additional information, while ResNet50 appears to adapt better to simpler representations like GEI. Nevertheless, it is important to note that the results are influenced to some extent by the overfitting observed in both models.



(a)

(b)

Figure 5.9: Train and Validation Loss Considering Models, Datasets and Representations for the CSS Protocol. a) ResNet50 with Elias Dataset and CGEI, and b) ResNet50 with MINDS-Libras and CGEI.

5.1.2 Analysis for the CUST Protocol

The CUST protocol, described in Section 4.3, subjects the models to a scenario where they must generalize to users not seen during training, posing a significant challenge in terms of performance. The primary goal of this protocol is to force the models to handle out-of-distribution data, which simulates real-world situations where inference systems must adapt to unknown users. The analysis was conducted at multiple levels, similar to Section 5.1.1.

Model Comparison: ViT vs ResNet

Table 5.4: General Model Comparison Results for the CUST Protocol.

Model	Recall (%)	F1-Score (%)	Acc (%)
ViT	8.1	4.9	8.1
ResNet	5.1	2.5	5.1

The results in Table 5.4 reflect a drastic drop in performance for both models under the CUST protocol. For the ViT, the metrics of *Recall*, *F1-Score*, and *Accuracy* barely reach 8.10%, 4.90%, and 8.10%, respectively. On the other hand, the ResNet50 model yields even worse results, with a *Recall* of 5.12%, an *F1-Score* of 2.52%, and an *Accuracy* of 5.12%.

This behavior indicates that both models struggle significantly to generalize to unseen users, a problem that becomes evident when working with datasets that exhibit substantial variation in patterns among users. However, it is important to note that, despite the poor performance of both models, ViT still slightly outperforms ResNet50 across all evaluated metrics. In Figure 5.10 and 5.11, we see the general behavior of the loss functions for training and validation.

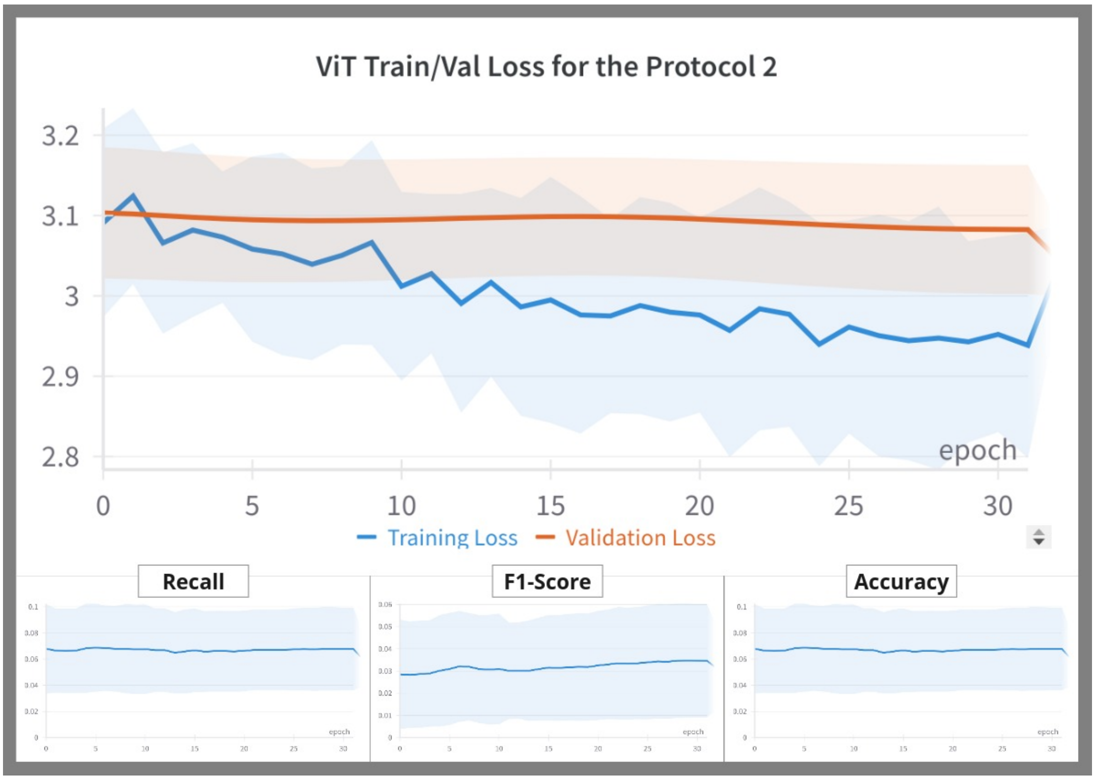


Figure 5.10: Train and Validation Loss Considering Models for the CUST Protocol in ViT.

The obtained *Accuracy* values for both models (8.10% for ViT and 5.12% for ResNet50) are notably low, suggesting that the models are operating only slightly above what could be considered a *random classifier*, which would have an accuracy close to the chance level (5% for each dataset in our case). This result highlights that, under the CUST protocol,

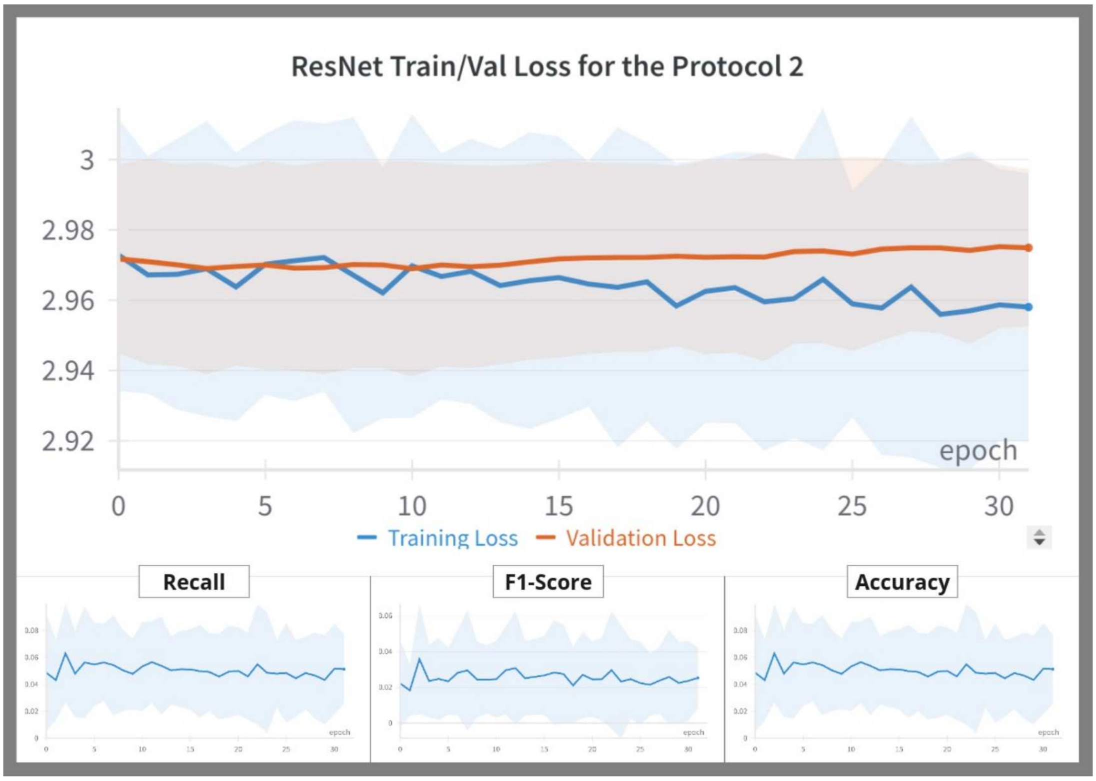


Figure 5.11: Train and Validation Loss Considering Models for the CUST Protocol in ResNet50.

the models perform almost randomly when faced with unseen users, revealing the models' inability to adapt to new data distributions in this scenario.

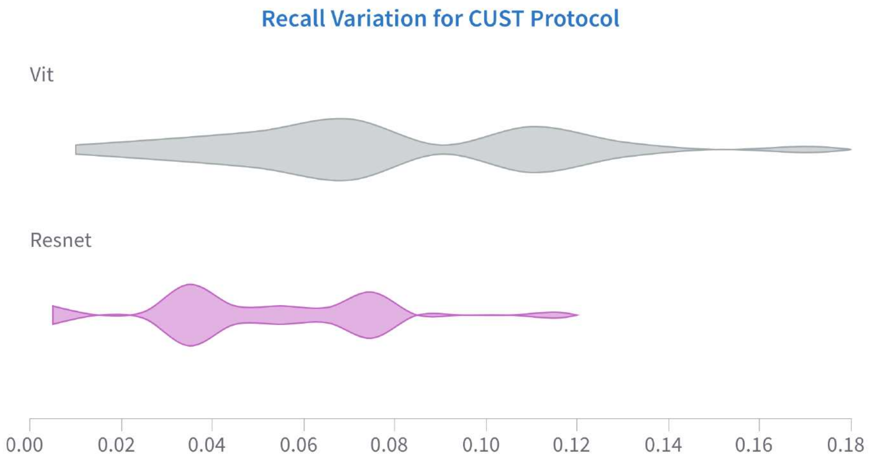


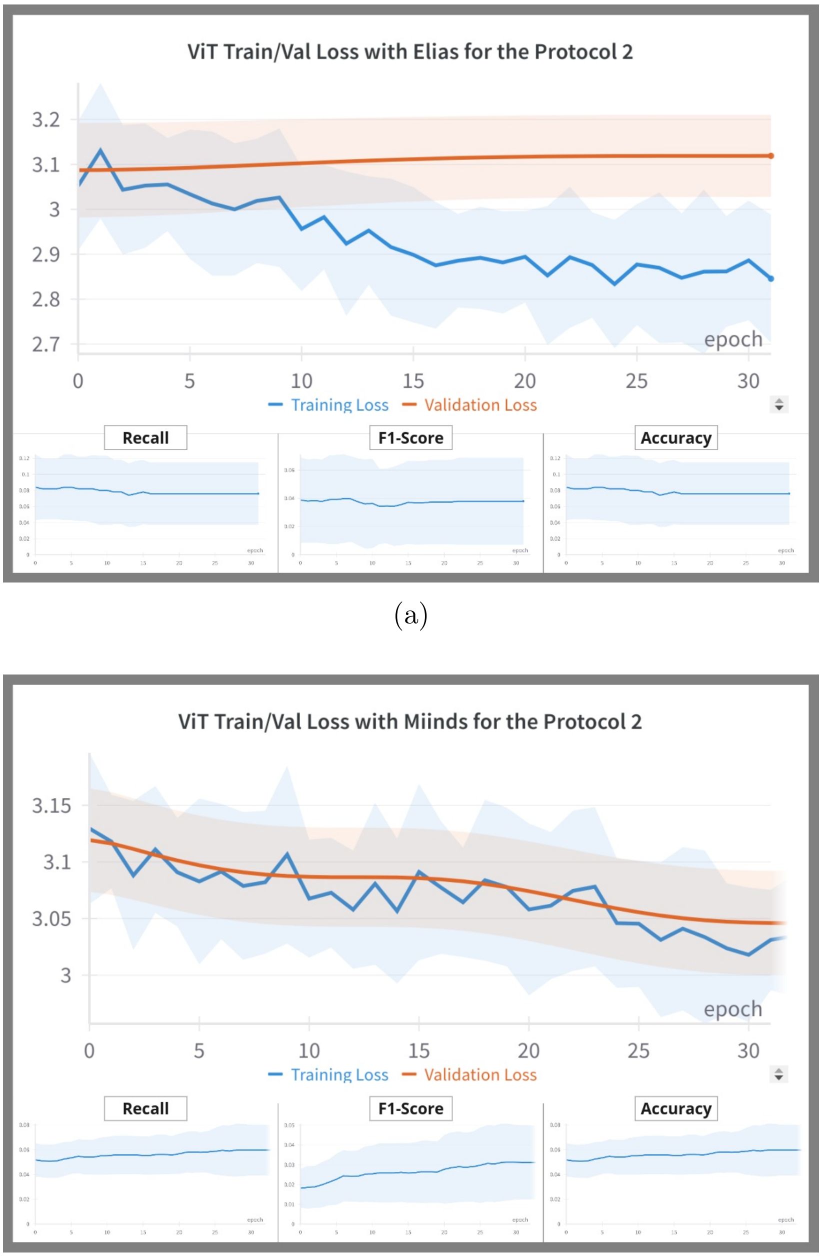
Figure 5.12: Train and Validation Loss Considering Models for the CUST Protocol. a) ViT and b) ResNet50.

In Figure 5.12, we can observe how the *recall* of ViT varies in comparison to ResNet50. We can see that the variation range of ResNet50 tends more towards values below the expected values for a random classifier (5% for both datasets in our case). In contrast, ViT tends to surpass this value.

Comparison concerning the Datasets

Table 5.5: Model Comparison Results Considering the Datasets for the CUST Protocol.

Model	Dataset	Acc (%)	F1-Score (%)	Recall (%)
ViT	*Elias*	7.6	3.8	7.6
ViT	*Minds*	8.6	5.9	8.6
ResNet	*Elias*	4.8	2.5	4.8
ResNet	*Minds*	5.4	2.5	5.4

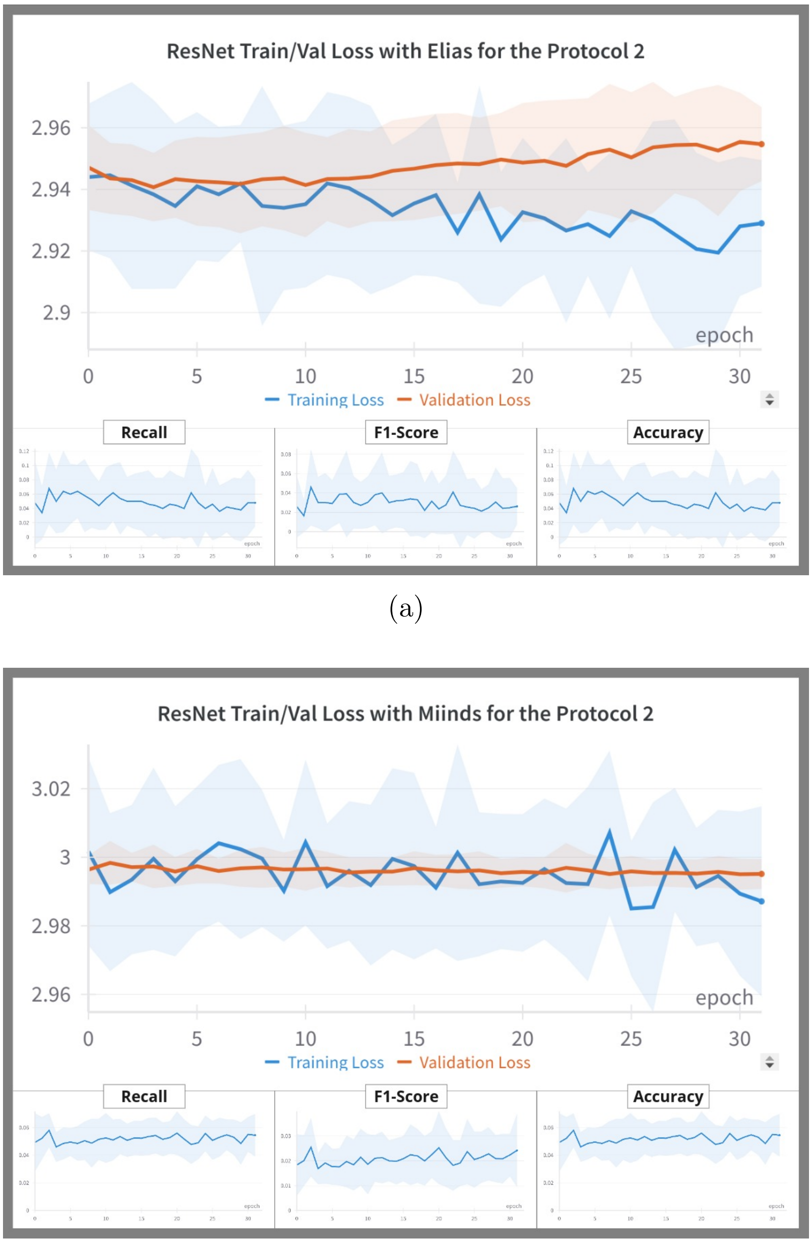


(a)

(b)

Figure 5.13: Train and Validation Loss Considering Models and Datasets for the CUST Protocol. a) ViT with Elias Dataset, and b) ViT with MINDS-Libras.

When examining the results of ViT on the Elias Dataset and MINDS-Libras datasets in Table 5.5, we observe that although the performance is low in both cases, the model achieves slightly better metrics on the MINDS-Libras dataset, with an *Accuracy* of 8.60% and an *F1-Score* of 5.89%. In contrast, on Elias Dataset, the metrics drop even further, with an *Accuracy* of 7.60% and an *F1-Score* of 3.80%, indicating that ViT faces greater difficulties in generalizing on this dataset.

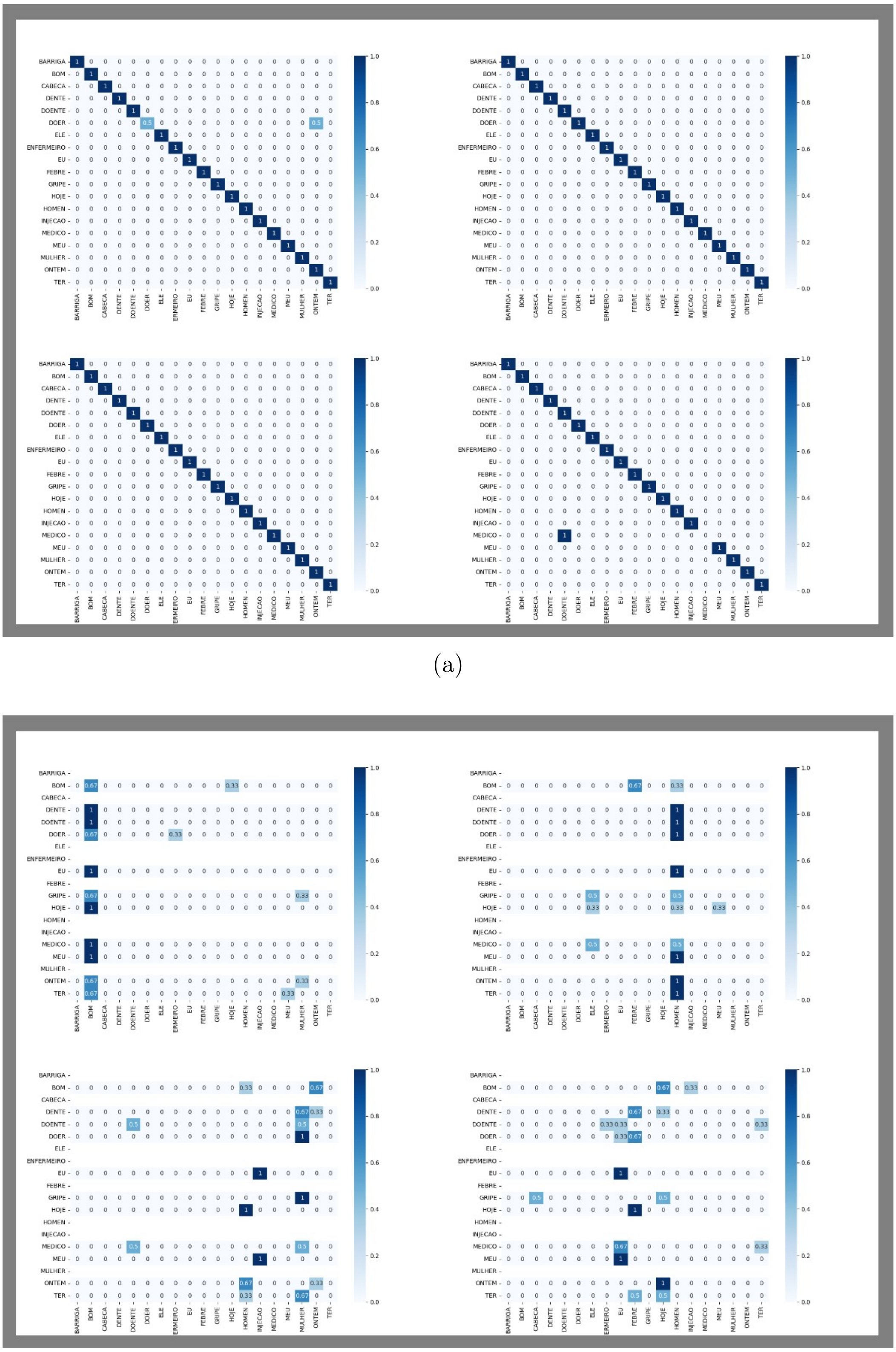


(a)

(b)

Figure 5.14: Train and Validation Loss Considering Models and Datasets for the CUST Protocol. a) ResNet50 with Elias Dataset, and b) ResNet50 with MINDS-Libras.

On the other hand, ResNet50 shows a similar trend, although with lower performance than ViT. On the Elias Dataset dataset, ResNet50 achieves an *Accuracy* of 4.80% and an *F1-Score* of 2.56%, while on MINDS-Libras, it improves slightly, with an *Accuracy* of 5.45% and an *F1-Score* of 2.51%. Despite the slight differences, the results reflect a low generalization capacity in both datasets.



(a)

(b)

Figure 5.15: Confusion Matrix examples for visualizing classes prediction. a) ViT with Elias Dataset and CGEI for CSS protocol, and b) ViT with Elias Dataset and CGEI for CUST protocol.

These results reveal that, although there are minimal differences between the datasets, both models exhibit poor generalization performance under the CUST protocol. It is

Table 5.6: Comparison of the best results achieved in this work against other studies using the MINDS-Libras dataset. Note: All results are obtained under controlled conditions with CSS protocol, ensuring a consistent and fair comparison.

Study	**Acc. (%)**	**F1-Score (%)**
Alves et al. [12]	93.00	93.00
de Castro et al. [34]	91.00	90.00
Passos et al. [83]	84.60	–
Our ViT+GEI	**95.9**	**95.9**
Our ViT+CGEI	**96.8**	**96.8**
Our ResNet50+GEI	**96.7**	**96.6**
Our ResNet50+CGEI	**95.3**	**95.2**

important to note that ViT tends to adapt better to the MINDS-Libras dataset, while ResNet50 faces similar difficulties across both datasets, with consistently low performance, as shown in Figure 5.13 and 5.14.

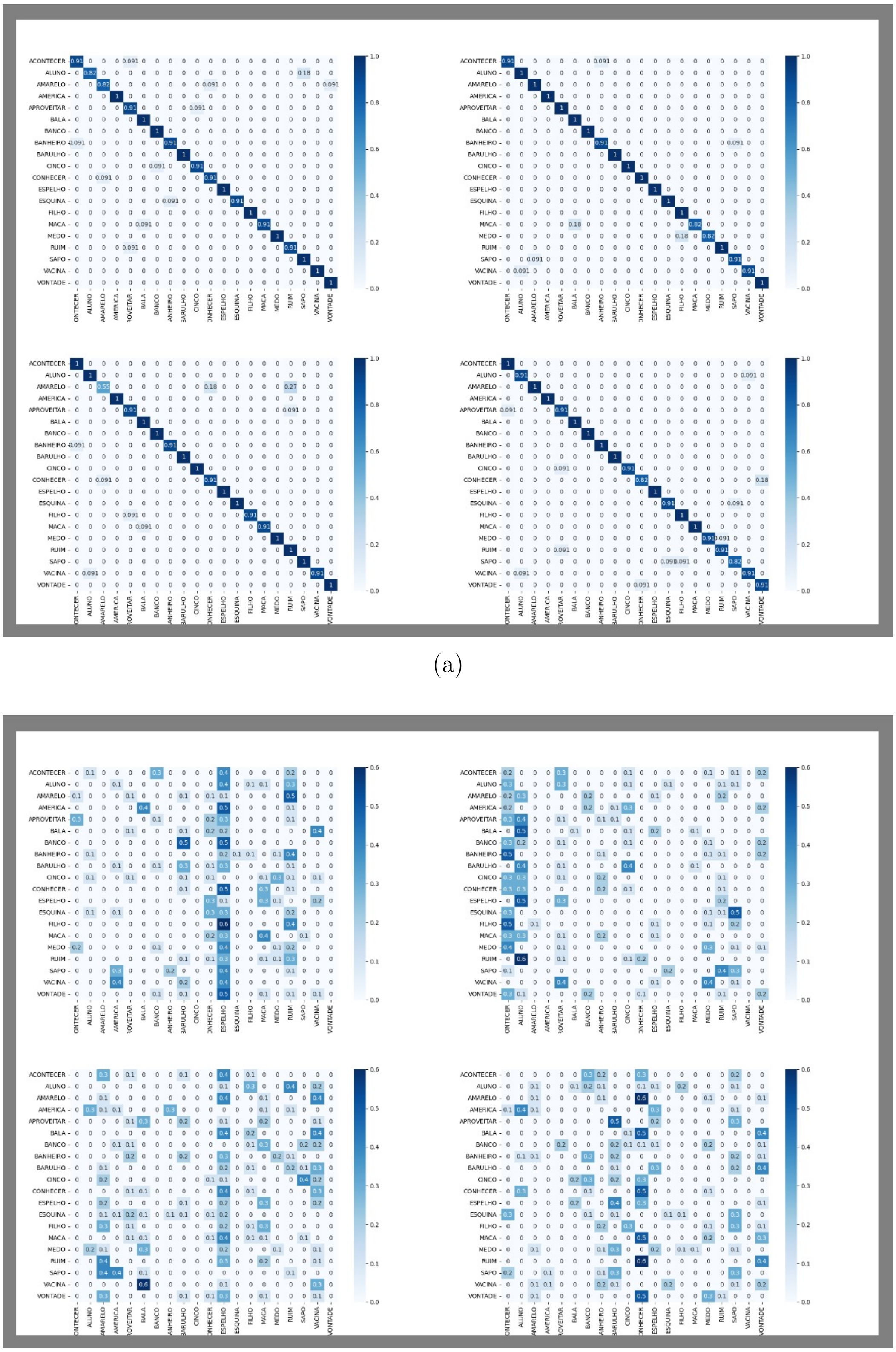
In Fig. 5.15 and 5.16, we show examples of how the confusion matrices behaved for some of the models in each protocol. This analysis allows us to conclude that the overall behavior of the models in this protocol is dominated by their low ability to adapt to unseen users. Additionally, no significant improvements are observed when analyzing the results in more detail, as the trend of poor performance persists when broken down by representations (GEI or CGEI) and models. This indicates that the generalization issues do not depend on the representation used, but rather on the inherent limitations of the models under this evaluation protocol.

In Table 5.6, we present a comparison of the best results obtained in this study with those reported by other works on the MINDS-Libras dataset. This comparison highlights the advantages of our approach, particularly the integration of ViT and ResNet50 models with different data representations, such as GEI and CGEI. Our results demonstrate consistent improvements in both accuracy and F1-score compared to prior works, with our best-performing configuration (ViT+CGEI) achieving 96.81% accuracy and 96.79% F1-score. Importantly, these results were achieved under controlled conditions to ensure a fair and consistent comparison. This outcome suggests that the enhancements made in data representation and model selection significantly impact performance, especially when applied to the nuanced task of sign language gesture recognition in MINDS-Libras.

5.1.3 Qualitative Evaluation: Attention Maps of the Vision Transformer

In the previous analyses, we observed that ViT consistently achieves better results compared to ResNet50 across all evaluated scenarios. This motivates us to delve deeper into the analysis of the attention maps within ViT, to better understand the patterns it detects and how it distributes its attention across the different signal representations.

For this qualitative analysis, we grouped the signals based on certain characteristics that allow us to observe how the model handles gestures that are similar or different in



(a)

(b)

Figure 5.16: Confusion Matrix examples for visualizing classes prediction. a) ViT with MINDS-Libras and CGEI for CSS protocol, and b) ViT with MINDS-Libras and CGEI for CUST protocol.

terms of execution and meaning (See Figure 5.17):

- Group 1: Signals with similar executions and linguistic meanings: “EU” and “MEU”.

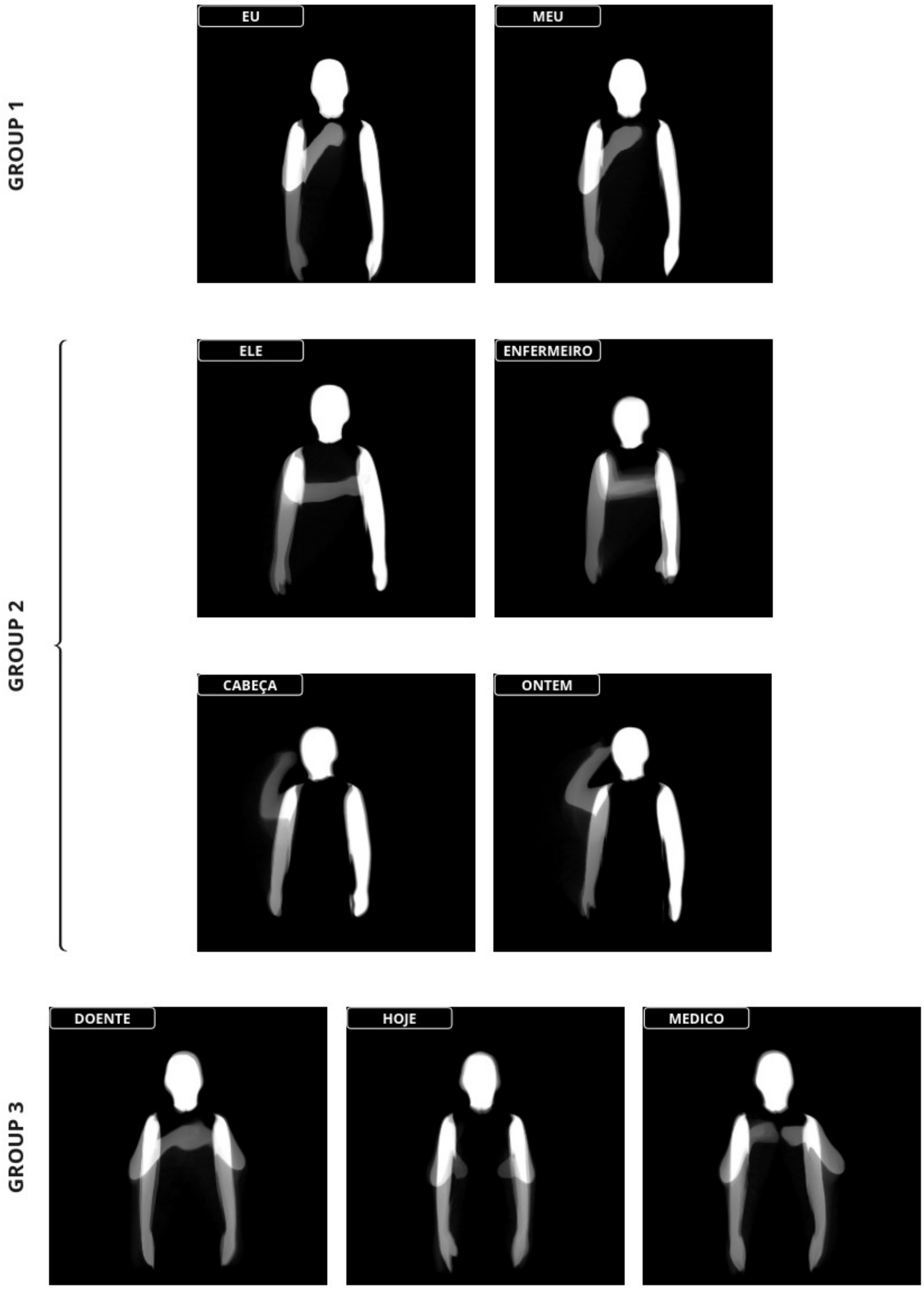
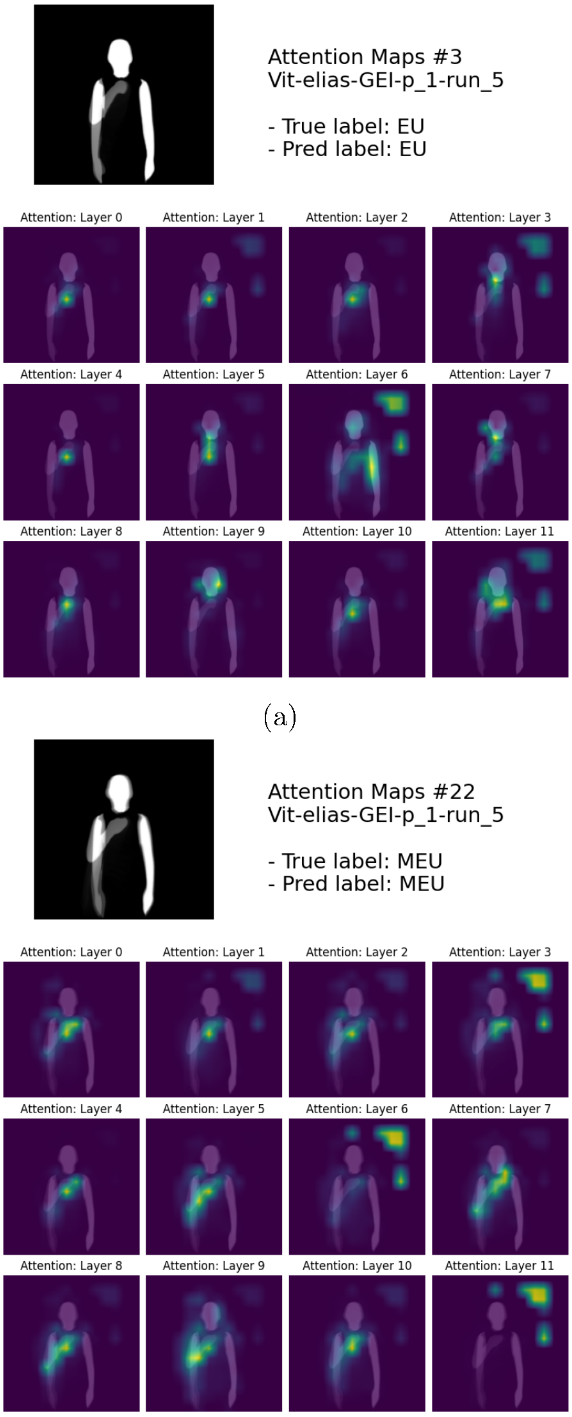


Figure 5.17: Samples in groups. In the image, examples in GEI representation for each group formed for the analysis.

- Group 2: Signals with similar executions but different linguistic meanings: “ENFERMEIRO” and “ELE", “CABEÇA” and “ONTEM”.
- Group 3: Signals involving the use of both hands: “DOENTE”, “HOJE”, and “MÉDICO”.

These groupings allow us to explore the relationship between gesture location and the model’s attention.

Group 1

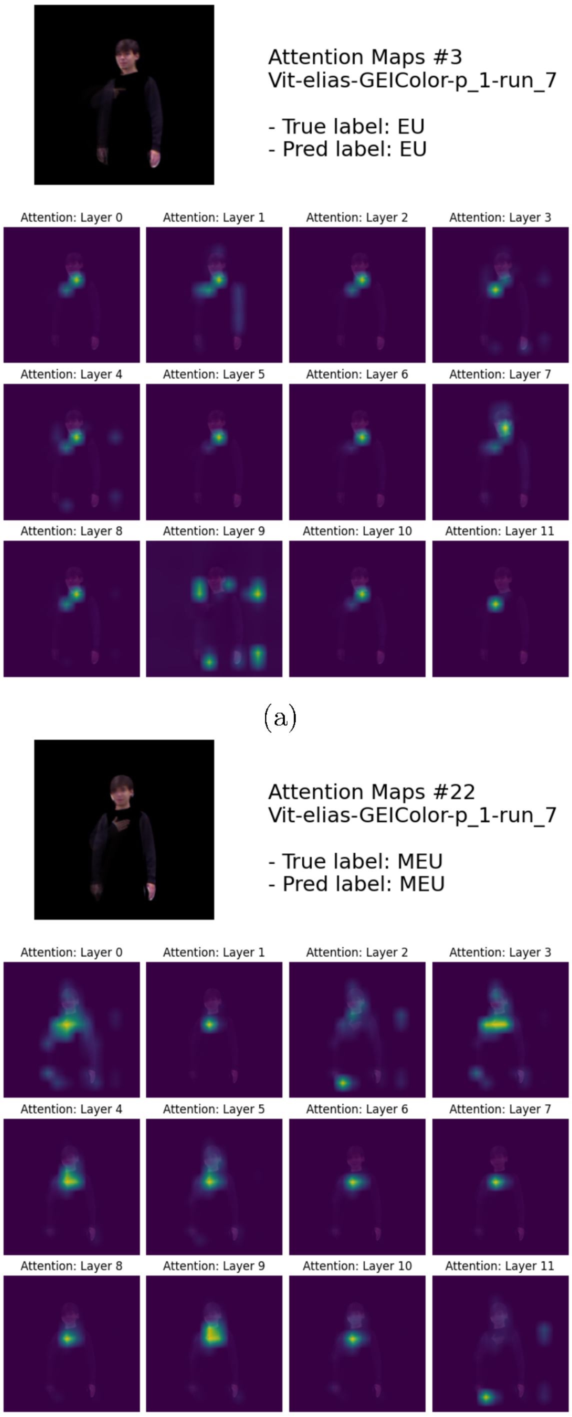


(a)

(b)

Figure 5.18: Attention maps of the GEI from Elias Dataset after finetuning. a) “EU” and b) “MEU”.

In Figure 5.18, we observe the attention map of a ViT for the classes “EU” and “MEU” in Elias Dataset using the GEI representation. It can be seen that the model manages to focus attention on the areas where the movement is being executed: some layers concentrate on the hands, others on the arms, some on the head, and others capture elements of the background. We note that, although the meanings of these words share similar components related to the first person and the images are alike, the model generates completely different attention maps for each class.



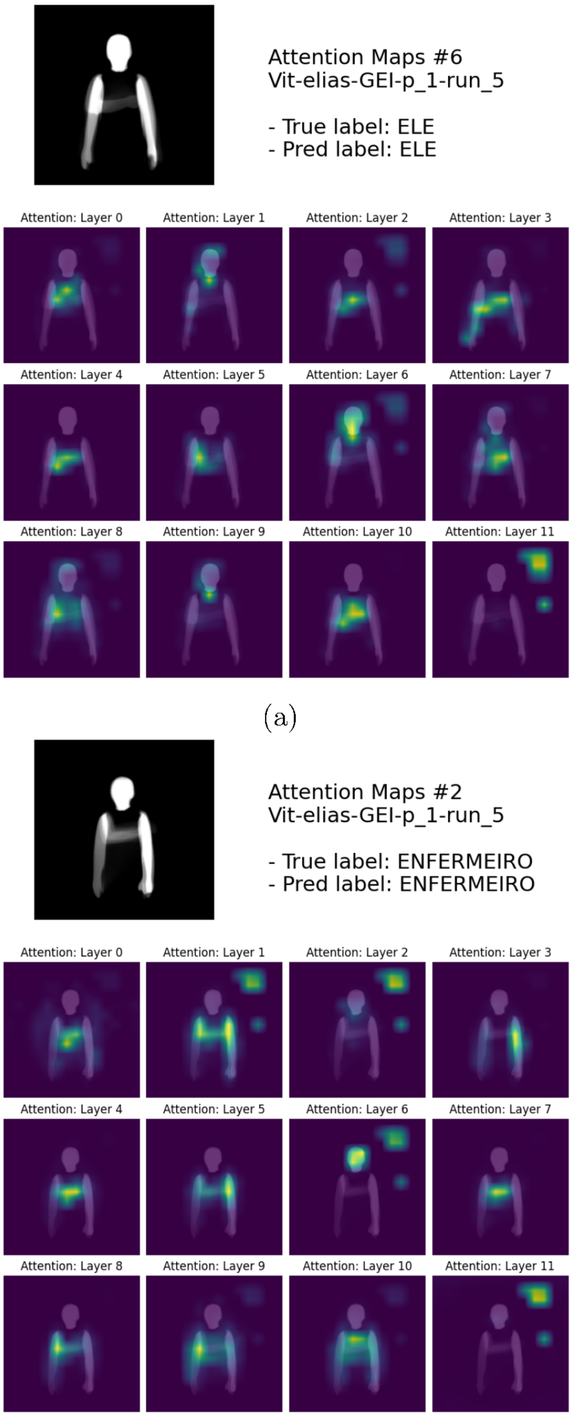
(a)

(b)

Figure 5.19: Attention maps of the CGEI from Elias Dataset after finetuning. a) “EU” and b) “MEU”.

In Figure 5.19, we observe the attention map of a ViT for the classes “EU” and “MEU” in Elias Dataset, this time using the CGEI representation. One of the characteristics we observe is that with the CGEI representation, the model shows more focused attention on specific features, such as the hands. This can be explained by the fact that the inclusion of chromatic information provides an additional layer of discrimination, enhancing the model's ability to more accurately identify the key areas of the gesture.

Group 2

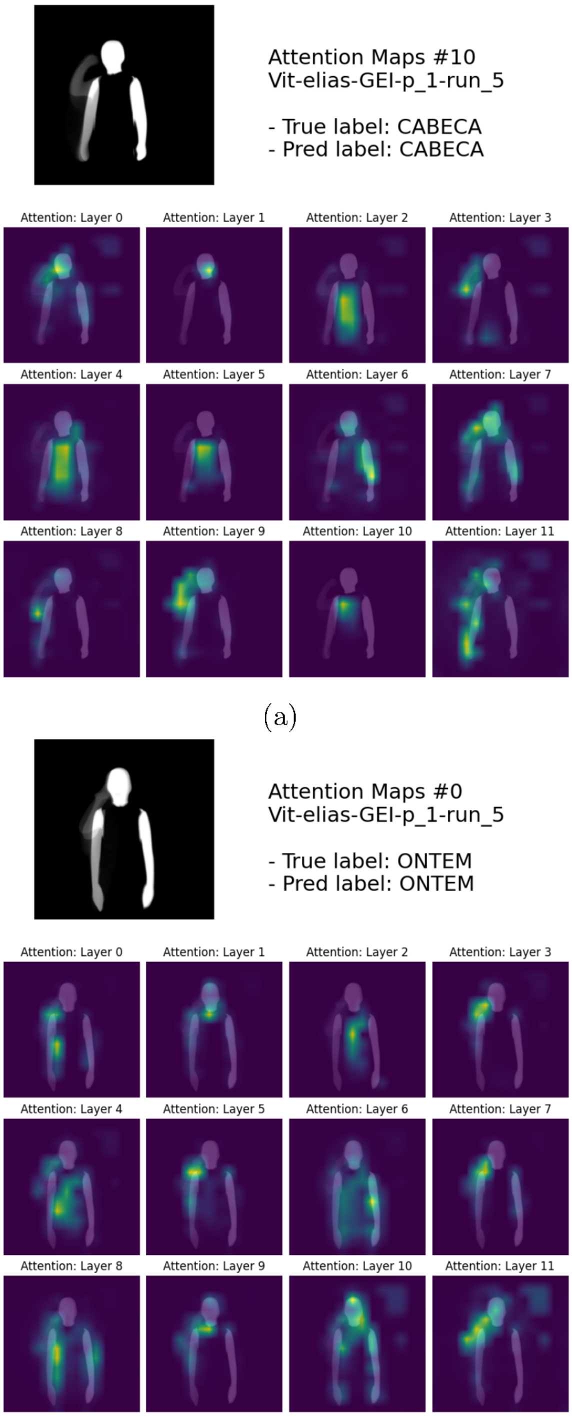


(a)

(b)

Figure 5.20: Attention maps of the GEI from Elias Dataset. a) "ELE" and b) "ENFERMEIRO".

In Figure 5.20 and Figure 5.21, we observe the attention map of a ViT for the classes "ELE", "ENFERMEIRO", "CABEÇA", and "ONTEM" in Elias Dataset using the GEI representation. These classes have no linguistic similarity; however, the first two have similar representations (the gesture is performed on the torso), while the latter two gestures are performed on the head. In both cases, we can see how the ViT managed to capture the differences quite effectively.

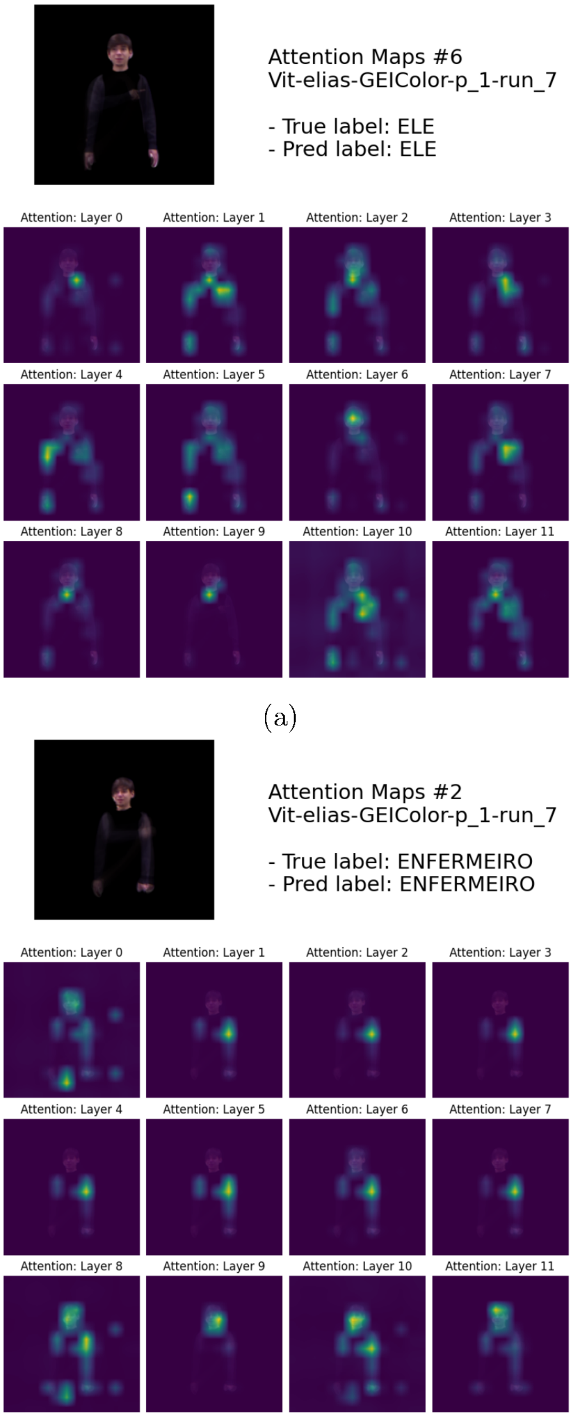


(a)

(b)

Figure 5.21: Attention maps of the GEI from Elias Dataset. a) "CABEÇA" and b) "ONTEM".

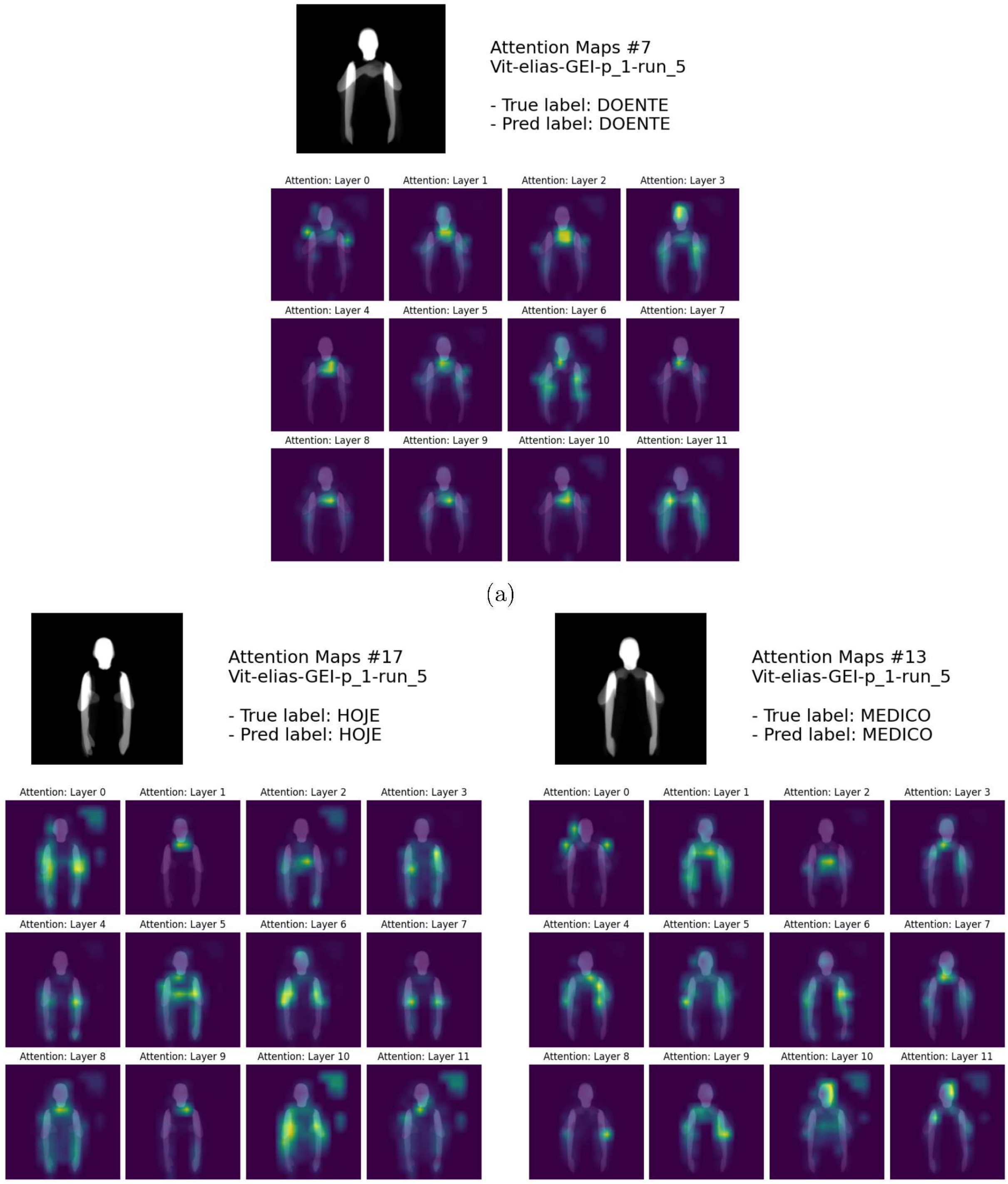
We found that for their counterparts with the CGEI representation, once again the attention maps are more focused on the areas of interest of the gesture, as seen in Figure 5.22. This same behavior is observed across all examples, where, when contrasting the representations, CGEI is consistently better focused than GEI in capturing the different characteristics of the image.



(a)

(b)

Figure 5.22: Attention maps of the CGEI from Elias Dataset after finetuning. a) “ELE” and b) “ENFERMEIRO”.



(a)

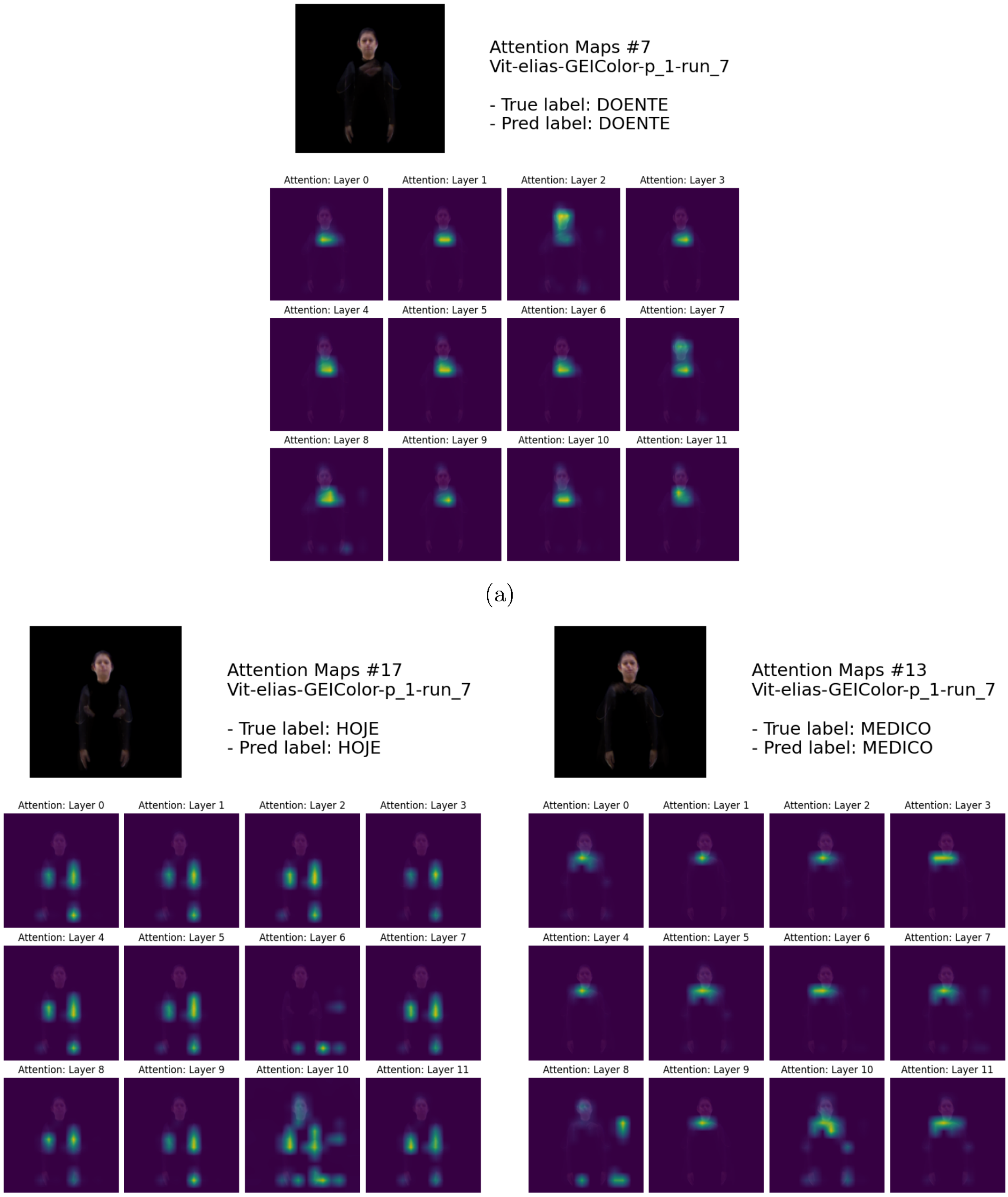
(b)

(c)

Figure 5.23: Attention maps of the GEI from Elias Dataset. a) “DOENTE”, b) “HOJE” and c) “MÉDICO”.

Group 3

In Figure 5.23, we find the third group. These signals involve gestures performed with both hands in different parts of the torso. The ViT demonstrates an attention distribution that reflects this characteristic, focusing on both hands simultaneously. This observation indicates that the model detects the symmetry in hand movements, suggesting a capability to identify more complex gesture patterns.



(a)

(b)

(c)

Figure 5.24: Attention maps of the CGEI from Elias Dataset. a) “DOENTE”, b) “HOJE” and c) “MÉDICO”.

In Figure 5.24, we see that with the CGEI representation, the results are significantly more precise and focused, which justifies the previously obtained results.

Overall, this qualitative analysis of the ViT attention maps reveals that the model tends to focus more on the spatial region where the gesture is executed, regardless of the semantic meaning of the signal. These findings suggest that, while ViT is effective in gesture-based pattern classification, its ability to discern the semantic meaning of gestures may depend more on spatial representation than on specific linguistic content. This opens the door for future research on how to integrate deeper semantic information into the attention process.

Additionally, the use of color, with the CGEI representation, highlights subtle differences between classes that may not be as apparent in the GEI representation, allowing the model to focus its attention more efficiently on the most relevant parts of the signal. Color helps ViT better differentiate similar visual signals, increasing the specificity of its predictions.

5.1.4 Discussion

The detailed analysis of model performance under the CSS protocol reveals several significant findings:

- Model Comparison: The results indicate that ViT slightly outperforms ResNet50 in terms of recall, F1-score, and accuracy across the evaluated datasets. This advantage is evident in ViT's faster and more stable convergence, indicating a greater ability to generalize and consistently learn patterns. However, the behavior observed in Elias Dataset suggests that both models may overfit due to the dataset's simplicity and imbalance, which could explain the high but potentially non-generalizable results.

- Dataset Impact: The differences between the Elias Dataset and MINDS-Libras datasets highlight the influence of dataset quality and structure on model performance. While Elias Dataset presents challenges due to its imbalance and limited number of samples, MINDS-Libras offers a more balanced and complex environment that supports better generalization. Models tend to overfit more in Elias Dataset, whereas, in MINDS-Libras, their performance is more consistent, reflecting a greater ability to handle the dataset's complexity.

- Representation Effect: The comparison between GEI and CGEI representations reveals that the addition of chromatic information in CGEI significantly enhances ViT's performance, particularly on the Elias Dataset dataset. This improvement suggests that ViT's ability to capture subtle gesture details benefits from the additional information provided by CGEI. In contrast, ResNet50 does not exhibit notable improvement with CGEI, and its performance even declines, indicating that chromatic information does not provide a meaningful advantage for this model and may introduce noise.

The analysis under the CUST protocol, which evaluates the models' ability to generalize to unseen users, highlights significant challenges in model adaptability:

- Model Comparison: The results reveal a dramatic drop in performance for both models under this protocol, in stark contrast to their results with the CSS protocol. Although ViT achieves slightly higher metrics (recall, F1-score, and accuracy) than ResNet50, both models operate close to chance levels, highlighting substantial limitations in generalizing to new users. This pattern suggests that, regardless of architectural or representational differences (GEI vs. CGEI), both models encounter similar challenges when adapting to unseen data distributions. While ViT initially demonstrated strong performance, it suffers considerably under out-of-distribution conditions, underscoring the need for methods better suited to generalize beyond training data.

- Dataset Impact in the CUST Protocol: The results show that ViT performs slightly better on the MINDS-Libras dataset compared to Elias Dataset, although in both cases, the metrics remain very low. This indicates that MINDS-Libras, despite being more complex and balanced under the CSS protocol, does not provide a significant advantage when facing unseen users. The low generalization ability in both datasets suggests that the issue does not lie in the specific characteristics of the datasets but rather in the models' overall inability to handle variability in the input data when confronted with new users.
- Independence of Representation: Unlike the results obtained under the CSS protocol, where the CGEI representation improved ViT's performance, in the CUST protocol, there is no significant improvement in performance using CGEI over GEI. This suggests that the enhancements provided by the CGEI representation are not sufficient to address the fundamental limitations of the models in terms of generalization to unseen users. The adaptation challenges seem to be more fundamental rather than specific to the representation used.

The results from the qualitative evaluation of the attention maps of the ViT reveal important insights into how the model processes and focuses on gesture signals. Overall, the ViT demonstrates a remarkable ability to concentrate its attention on the relevant areas for interpreting gestures, regardless of linguistic similarities or differences between the signals.

- Attention on Similar Signals: In the case of signals with similar executions and linguistic meanings ("EU" and "MEU") in *Group 1*, the model shows the ability to differentiate between the signals despite their similarity, generating significantly different attention maps. This suggests that ViT is sensitive to subtle movement nuances and not just the general characteristics of the signals.
- Similar Executions but Different Meanings: When analyzing signals with similar executions but different meanings, such as "ENFERMEIRO" and "ELE" in *Group 2*, the model continues to focus its attention on the relevant areas of the gesture, despite differences in meaning. This indicates that ViT can distinguish between gestures based on specific spatial features beyond the semantic context.
- Signals Involving Both Hands: For gestures involving both hands, the model shows more distributed attention, focusing on both hands simultaneously. This reflects ViT's ability to identify complex patterns in gesture execution, such as the symmetry of hand movements.

- Impact of the CGEI Representation: The comparison between GEI and CGEI representations highlights that the inclusion of chromatic features in CGEI provides an additional layer of discrimination. The attention maps generated with CGEI are more focused, suggesting that color enhances the model's ability to identify key gesture areas with greater precision. This increased specificity may help separate visually similar signals and improve model accuracy.

Our findings suggest that the use of chromatic information in CGEI significantly influences the model's focus on key gesture areas, as evidenced by both quantitative and qualitative analyses. Across more than 80 experimental trials using the MINDS-Libras dataset, we observed that in 99% of cases, the CGEI representation directed model attention toward specific gesture points of interest, enhancing discrimination between gestures. By contrast, the GEI representation led to more dispersed attention, often focusing on broader, less relevant regions of the image. These patterns indicate that CGEI not only improves quantitative performance metrics, such as recall and F1-score, but also refines the spatial precision of attention maps, which is critical for accurately interpreting complex gestures.

5.2 Data Augmentation Experimental Results

The computational experiments carried out in this section aimed to pursue the answer to Research Question 1 (RQ1), proposed in Chapter 1. Here, we report results from two types of experiments: in the first one, we assess a simple diffusion model in generating gestures from an unknown user; in the second one, we evaluate model performance for an increasing variety of training data. Finally, we conclude this section (and this chapter) with some discussion on the obtained results.

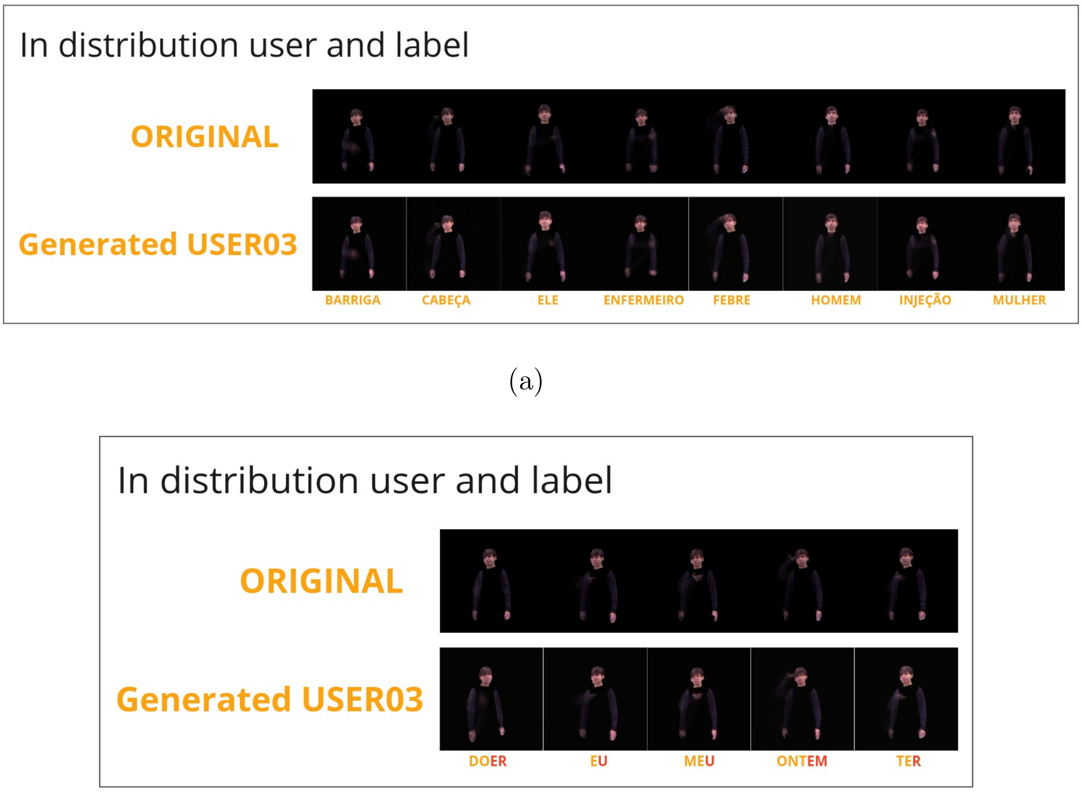
5.2.1 Results of Experiment 1: Generation with Unknown User

In this experiment, we analyzed the ability of the *Simple Diffusion* model to generate gestures with an unknown user and both known and unknown signs. We evaluated four cases: gestures generated for known signs and known users, gestures generated for known signs and unknown users, gestures generated for unknown signs and known users, and gestures generated for unknown signs and unknown users.

Known Users and Known Labels

In Figure 5.25, we show the cases where inference was performed using data known to the model. In Figure 5.25a, the labels are specific to User 3, and in Figure 5.25b, the inference was performed on labels that both users share in common, but the model did not see any example of that sign from the other user (User 2).

We observe that the model was able to generate coherent signs for both known users and labels. The characteristics of the user are preserved, and it seems to introduce variations, as seen with the label "DOER" in Figure 5.25b, where the subject appears mirrored. In terms of gesture capture, the representation of the sign remains consistent



(a)

(b)

Figure 5.25: a) Only Classes of the User 3 and b) Common Classes but do not exist another user in the training dataset.

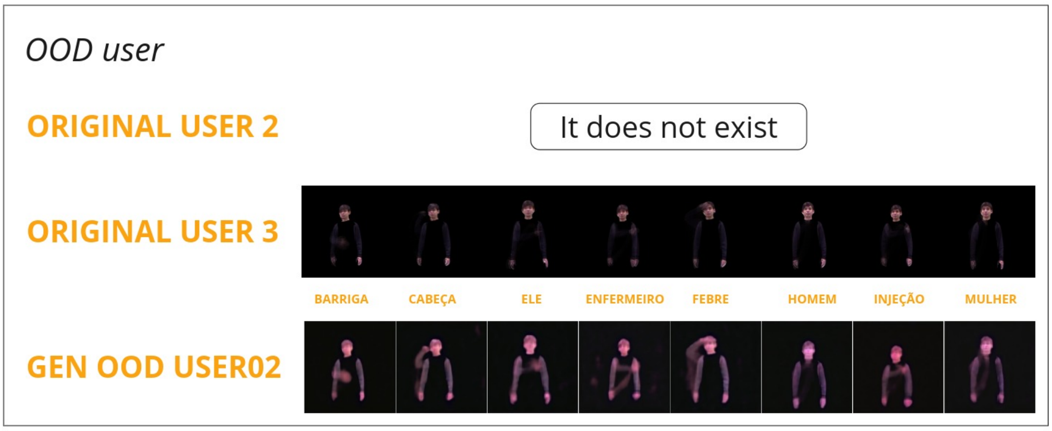
across all cases. However, the model behaved as expected, since this case falls within the model's training distribution.

Unknown users and known labels

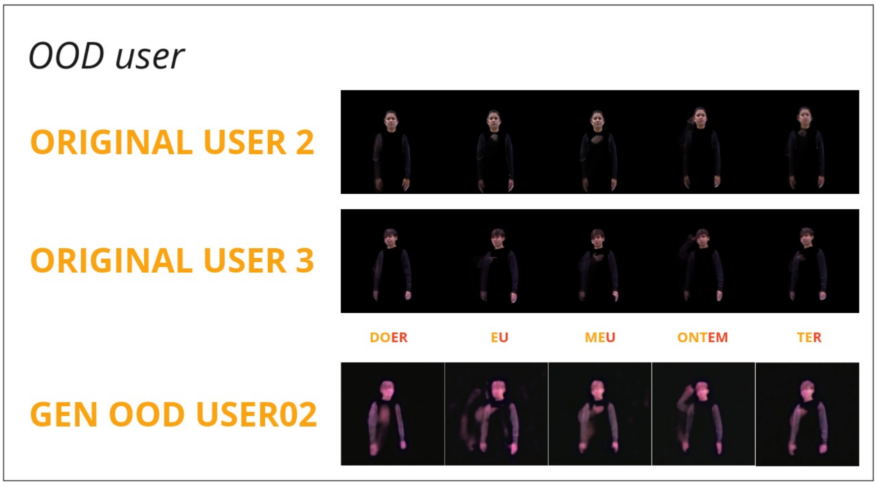
Increasing the complexity slightly, we asked the model to generate known labels with users not seen during training. We observed that, although the model understands the difference, it fails to converge to a user different from those seen during training. In Figure 5.26, we again show the cases with known labels, both common and individual. We note that the model converges to the characteristics of the known user but introduces variations in color. Regarding the sign, the model maintains gesture coherence and seems to understand the differences between signs with similar gestures and completely different meanings, as seen in Figure 5.26 with pairs like "Injeção-Enfermeiro", "Eu-Meu", "Febre-Cabeça". We consider this result as partially out-of-distribution.

Known users and unknown labels

Here, we evaluate how well the model can generate gestures for labels not seen during training. We asked the model to generate gestures for User 3, who was part of the training, but with missing labels. This scenario was set up to observe potential variations



(a)



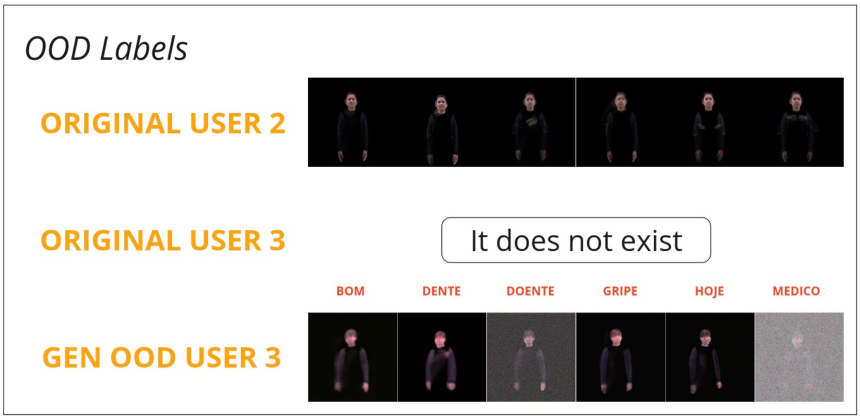
(b)

Figure 5.26: a) Out Of the Distribution User when label not exist in the other user and b) Out Of the Distribution User in common labels.

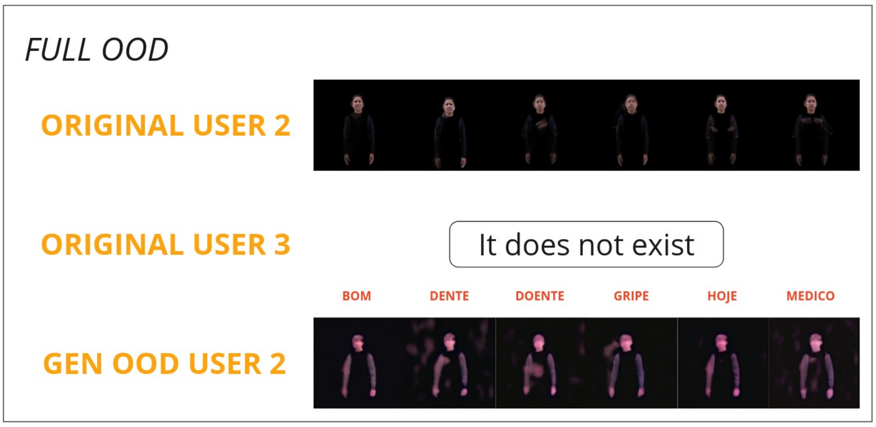
in generation. In Figure 5.27a, we include the original image of User 2 as a visual reference for the gesture, since there was no reference for these labels for the training user. In *GEN OOD USER 3* from Figure 5.27a, we can see the result generated by the model. Although it did not generate an accurate representation of the requested gesture, it attempted to recreate the characteristics of the known user. This result is expected, given that the model was not trained with a sufficient variety of users. Here, the label is out of distribution.

Unknown users and unknown labels

Finally, we present the results where the model has seen neither the user nor the labels. In Figure 5.27b, we show the results for the fully out-of-distribution case. As expected, the model does not generate any significant or relevant variations in these cases.



(a)



(b)

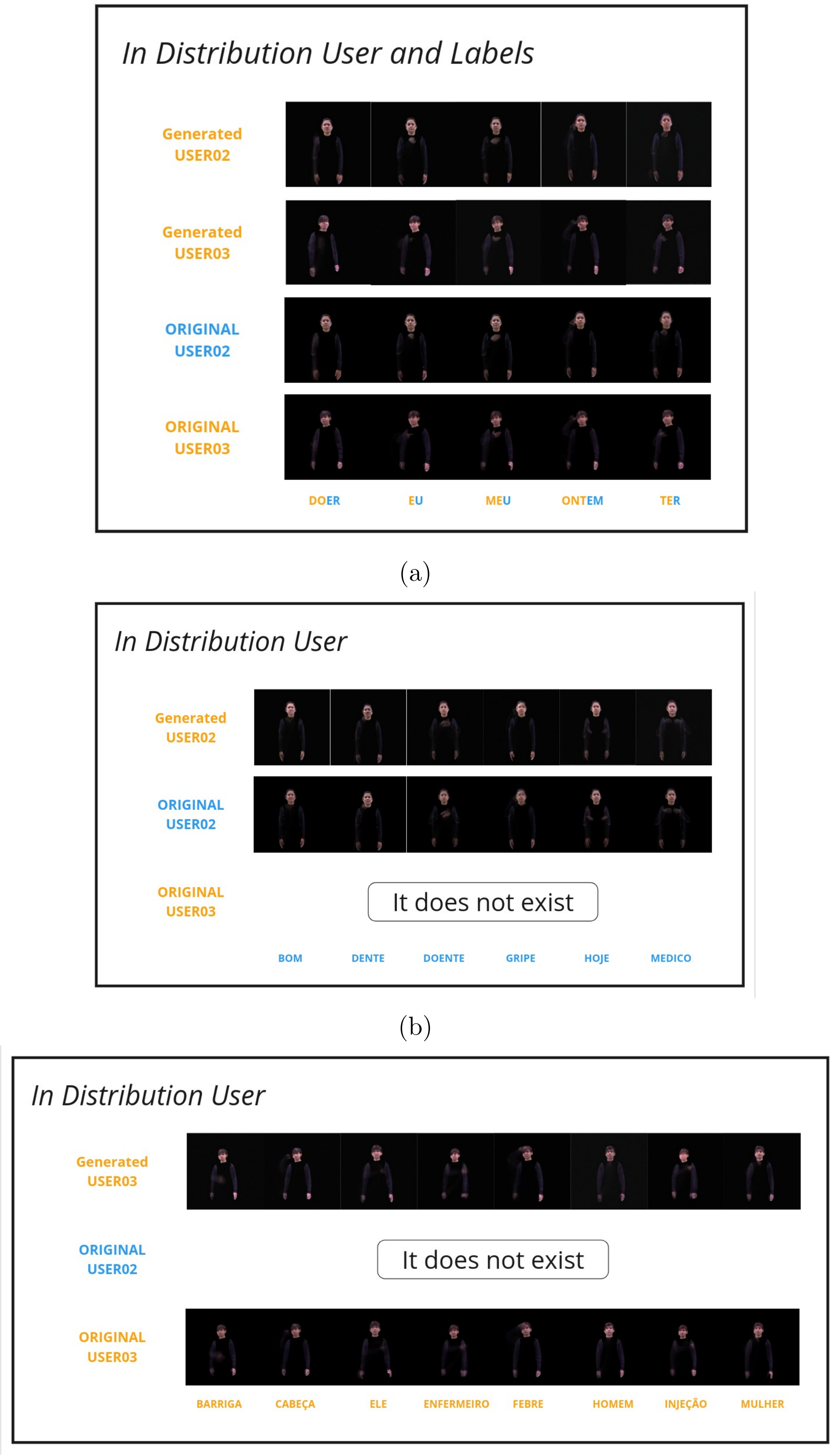
Figure 5.27: a) Out of the Distribution Label when user exist and b) Full Out of the Distribution User and Labels.

5.2.2 Results of Experiment 2: Inclusion of Users

In this experiment, we evaluate how the model behaves when increasing the variety of training data. We ensured that all classes were present and performed inferences to evaluate the following cases: known user and label, cross-validation between users, and fully out-of-distribution cases.

Known User and Label

In this scenario, we have three cases. In Figure 5.28a, the inference results for both users with intersecting labels are shown. The images generated by the model are consistent with the desired gesture. For the sign "Doer" for example, the model generated one of the variants included in the training. Upon verifying these results, we observe that the model can generate the variations seen during training, which is an expected outcome. Throughout this analysis, it has become clear that the model is very good at reproducing the data distribution it was trained on, with slight variations that do not significantly deviate from the distribution. These results are replicated in the cases shown in Figure 5.28b and Figure 5.28c.



(a)

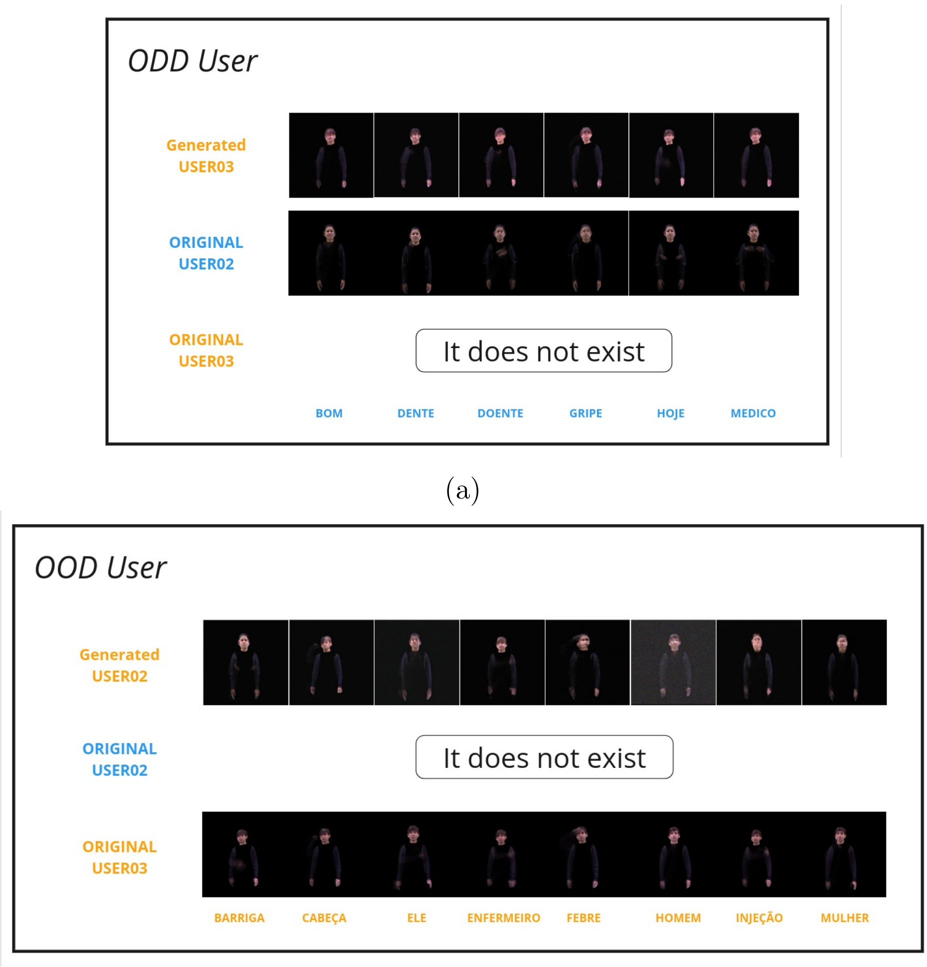
(b)

(c)

Figure 5.28: In distribution generated samples.

Cross-Validation Between Users

We wanted to evaluate the model's ability to generate different users from those seen during training, using the labels that belonged to one of the trained users. In the infer-



(a)

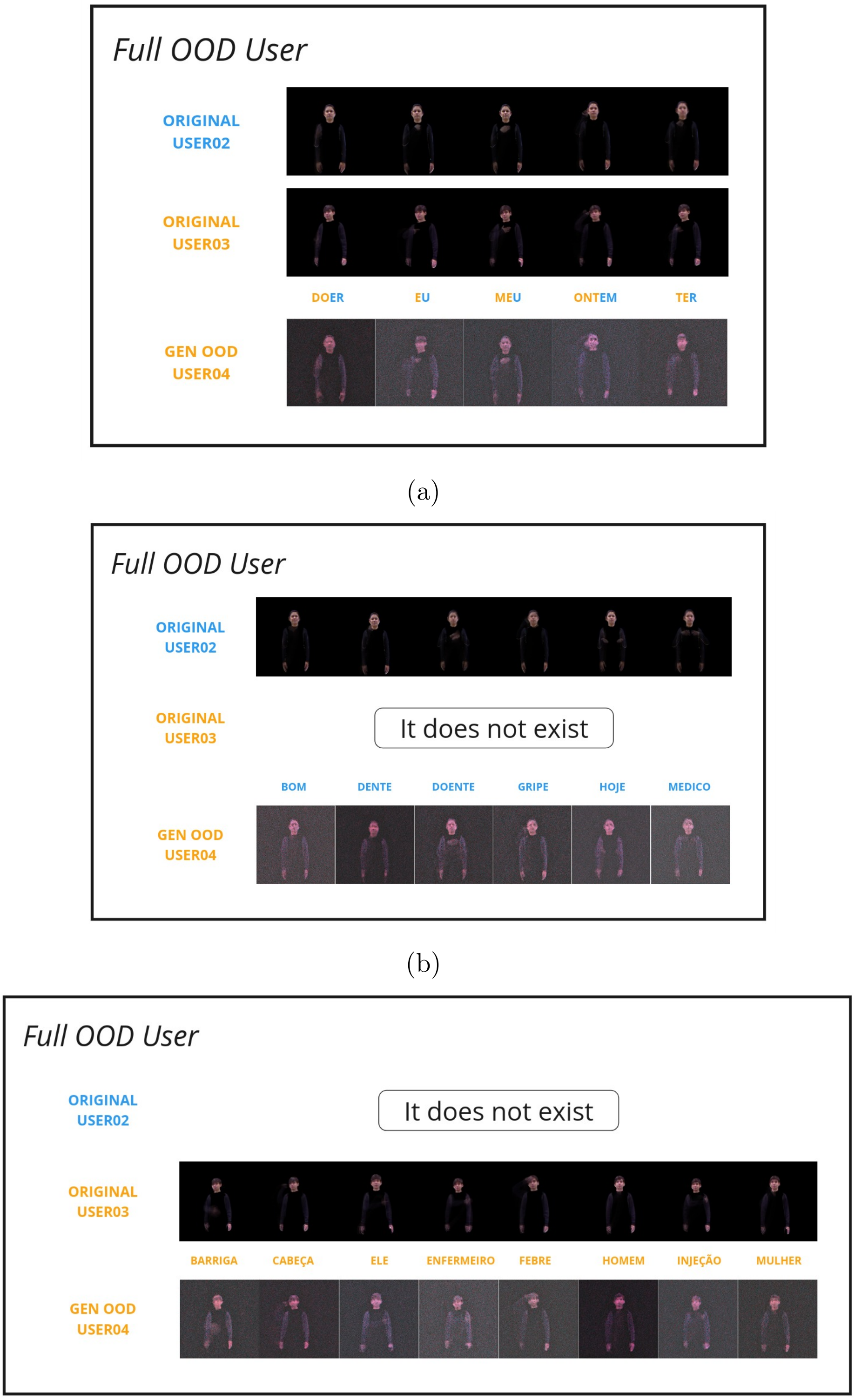
(b)

Figure 5.29: Out of the distribution user when do not exist samples for someone users.

ences, we asked the model to generate gestures with the labels of one user seen during training, but with the characteristics of another user who did not have those labels. In Figure 5.29a, we generated User 3 with User 2's labels. Although the result was not entirely satisfactory, we highlight the model's ability to capture the differences between users. In Figure 5.29b, we did the reverse case, asking the model to generate User 2 with User 3's labels. Here, the result was slightly different. In some cases, the model attempted to generate the sign with the correct user's characteristics, such as in the signs "Febre" "Injeção" and "Mulher" where it is evident that the model tries to represent the requested user with the appropriate gesture. This is a relevant finding, as it demonstrates the model's potential to diversify data and adapt to new users.

Fully Out-of-Distribution Cases

Finally, we evaluated the model using a completely unknown user, aiming to analyze its ability to generalize and diversify in out-of-distribution scenarios. In Figure 5.30, the results of these cases are presented, where the model had not previously seen the user. Despite the difficulty, we again highlight the model's ability to approximate the requested gesture correctly. However, the model tends to generate images with a certain noise level.



(a)

(b)

(c)

Figure 5.30: Full Out of distribution user.

In general, the results indicate that the model attempts to combine semantic elements from the users it learned during training, suggesting its capacity to incorporate diversity into the generated data.

An interesting aspect is that, although the gestures generated for the third user show coherence with the expected signs, some visual details indicate that the model still relies on information from previously learned users during the generation process. This highlights both the model's potential to extrapolate to new cases and the limitations it faces due to insufficient diversity of users during training.

5.2.3 Discussion

Throughout the two experiments presented, we have identified key patterns and behaviors of the *Simple Diffusion* model in different generation scenarios, both with known and unknown users and signs. These results provide valuable insights into the model's strengths and limitations in terms of generalization and diversification.

In **Experiment 1**, where we evaluated the model's ability to generate gestures in scenarios with unknown users, we observed that the model performs reasonably well when generating coherent gestures for known signs and users. However, when unknown users were introduced, the model tended to generate coherent gestures, but replicated characteristics of the known users, showing a dependency on the training data. The gestures generated for unknown signs did not show precision in their execution, highlighting the model's limitations in situations completely outside the training distribution.

On the other hand, in **Experiment 2**, by increasing the diversity in the training set, the model showed improvements in its ability to generalize to new users and signs. In cross-validation scenarios, where the model was forced to generate signs for one user with the characteristics of another, the results were mixed. Although it did not always capture all the characteristics of the target user, the model showed clear signs of attempting to differentiate between users and adapt its predictions based on the labels. This behavior suggests potential improvement if more user variety is included in the training.

In the more challenging *Full Out of Distribution* cases, where neither the users nor the signs had been seen before, the model produced interesting results. Although it did not fully capture the expected diversity, there were instances where gesture coherence was maintained, indicating that the model is beginning to combine semantic elements of the known users. This highlights both the model's potential to extrapolate to new scenarios and its inherent dependence on the data seen during training.

The qualitative findings from both experiments reveal that *Simple Diffusion* is competent at generating signs for known cases but still faces significant challenges in situations completely out of the training distribution. While there are signs that the model can generate diversity from new users and gestures, it still relies heavily on previously learned information. To improve its performance in more challenging scenarios, it would be essential to increase the diversity of the dataset and explore regularization techniques that help reduce the model's dependence on known data.

Chapter 6

Conclusions

In this chapter, we provide some conclusion notes on the work presented in this Master's dissertation. First, we present the answers for the research questions raised in Chapter 1. In the sequence, we describe the participation of the author in academic events. Finally, we provide some final remarks about the research.

6.1 Answers to the Research Questions

After accomplishing the work described in the previous chapters, we can revisit the research questions raised in Chapter 1 to address them:

RQ1
For a given Libras dataset (augmented or not), how does the performance of a Vision Transformer (ViT) compare to the machine learning models previously used for this problem?

- **Answer:**

  To address the question of how the performance of a Vision Transformer (ViT) compares to machine learning models previously used for the classification of Libras signs, we can rely on the experiments conducted with the CSS and CUST protocols, which assess performance in controlled and generalization scenarios. It is important to note that only natural data from the datasets were used in these experiments, without augmentation.

  Under the CSS Protocol, in non-augmented Libras datasets, the ViT demonstrated a **clear advantage** over traditional models like ResNet50. This is reflected in its **ability to learn complex patterns** and a **faster and more stable convergence** of the loss function. This behavior suggests that in controlled scenarios, where both users and signs are known, the ViT offers superior performance in terms of accuracy and efficiency compared to previous models used for this problem.

  Additionally, the ViT **benefits significantly from richer representations**, such as the inclusion of chromatic information in CGEI, improving gesture classification accuracy. This contrasts with models like ResNet50, which do not show notable improvement or may even suffer a performance drop when using such representations. This difference highlights the ViT's ability to exploit additional spatial and visual features better, placing it in an advantageous position over previous approaches, even under conditions of insufficient and imbalanced data.

In the CUST protocol, where the ability to generalize to unseen users is evaluated, the performance of the ViT, although competitive, **does not show a considerable advantage** over previous models like ResNet50. Both models exhibit significant challenges when generalizing in scenarios where the users and signs are unknown. This result indicates that while it may be more effective in controlled scenarios, the limitations of the ViT in terms of generalization are similar to those of earlier models.

Notably, in the CUST protocol, chromatic representation does not provide a significant advantage, suggesting that the improvements observed in controlled classification scenarios do not carry over to environments with greater data variability. The fundamental limitations in generalization ability persist regardless of the representations employed.

In summary, for a non-augmented Libras dataset, the ViT **outperforms traditional deep learning models** like ResNet50 in controlled scenarios, exhibiting better performance in learning complex patterns and efficiency in convergence. However, in situations that require broader generalization, such as in the CUST protocol, the ViT **does not offer a significant improvement** over previous models. This finding highlights the **need to enhance the generalization capacity of models**, potentially through data augmentation techniques or approaches that better address user and sign variability.

Finally, the experiments revealed an interesting aspect of the ViT: **its potential for segmentation and data annotation tasks**. This potential arises from the **attention maps generated by the model**, which show promising behavior in identifying key regions within gestures. Although this was not the primary objective of the study, the ViT's ability to precisely focus its attention suggests that it may be a useful model for more complex tasks involving not only classification but also the segmentation of gestural signals.

RQ2

Is it possible to augment data from a Libras database using data from other Libras datasets?

- **Answer:**

  Based on the partial results obtained in the experiments presented, it is possible to formulate a well-founded hypothesis regarding the question of the feasibility of augmenting a Libras dataset using information from other Libras datasets.

  Firstly, the results indicate that the *Simple Diffusion* model shows limited generalization capability when exposed to signs and users not part of the original training distribution. This observation suggests that the model may struggle when faced with data from other Libras datasets, especially if these datasets exhibit significantly different characteristics from the original data. However, the model's ability to combine semantic features from seen users suggests potential for the use of external data, provided that it offers a sufficiently diverse representation of signs and users.

On the other hand, the cross-user validation experiment has shown that the model can partially adapt to new users, further supporting the idea that, under certain conditions, data from other Libras datasets could improve the model's performance by contributing greater diversity and aiding in its generalization. Nevertheless, the current results highlight the model's dependence on familiar data, demonstrating that the success of this data augmentation strategy will largely depend on the variability the new datasets can offer, in terms of both signs and users.

Lastly, the preliminary results also suggest that, while the model faces limitations when extrapolating gestures in completely new scenarios, there is a degree of extrapolation capability that can be harnessed and improved through more diversified training. This indicates that the use of other Libras datasets could have a positive effect if implemented correctly.

In summary, although the current experiments do not explicitly address the question of augmenting data by using other Libras datasets, the results obtained thus far suggest that this could be a viable strategy. However, the final "cross dataset" experiment must be conducted to conclusively evaluate the impact of this strategy, especially in scenarios where there are no users or signs in common between the datasets used.

6.2 Participation in Events

Initial results of this work were presented on a poster at the *International Meeting on Artificial Intelligence and its Applications - RIIAA* Carrillo et al. [22] held in Quito, Ecuador from February 19 to 25, 2024.

6.3 Final Remarks

Our work presents results that provide a foundation for future research. Although we have made significant progress with the baseline and initial data augmentation experiments, this study should be regarded as a first step. We are motivated to continue advancing this line of research and to refine the results obtained thus far. The current conclusions establish an initial framework and guide the path for future analysis and improvements.

Bibliography

[1] Lei nº 10.436, 2002. URL https://www.planalto.gov.br/ccivil_03/leis/2002/l10436.htm.

[2] Decreto nº 7387, 2010. URL https://www.planalto.gov.br/ccivil_03/_ato2007-2010/2010/decreto/d7387.htm.

[3] American signal language (asl) corpus database, 2018. URL https://www.csdeagles.com/apps/pages/aslcorpus.

[4] Corpus libras, 2019. URL https://corpuslibras.ufsc.br/.

[5] As Dificuldades da Pessoa Surda na Sociedade Brasileira, 2020. URL https://rodrigomagnorm.jusbrasil.com.br/artigos/1176514129/as-dificuldades-da-pessoa-surda-na-sociedade-brasileira.

[6] Samah Abbas, Hassanin Al-Barhamtoshy, and Fahad Alotaibi. Towards an Arabic Sign Language (ArSL) *corpus* for deaf drivers. *PeerJ Computer Science*, 7:e741, November 2021. ISSN 2376-5992. doi: 10.7717/peerj-cs.741. URL https://peerj.com/articles/cs-741.

[7] Fahmid Al Farid, Noramiza Hashim, Junaidi Abdullah, Md Roman Bhuiyan, Wan Noor Shahida Mohd Isa, Jia Uddin, Mohammad Ahsanul Haque, and Mohd Nizam Husen. A Structured and Methodological Review on Vision-Based Hand Gesture Recognition System. *Journal of Imaging*, 8(6):153, June 2022. ISSN 2313-433X. doi: 10.3390/jimaging8060153. URL https://www.mdpi.com/2313-433X/8/6/153. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

[8] Muneer Al-Hammadi, Ghulam Muhammad, Wadood Abdul, Mansour Alsulaiman, Mohamed A. Bencherif, and Mohamed Amine Mekhtiche. Hand gesture recognition for sign language using 3dcnn. *IEEE Access*, 8:79491–79509, 2020. ISSN 2169-3536. doi: 10.1109/ACCESS.2020.2990434. Conference Name: IEEE Access.

[9] Silvia Grasiella Moreira Almeida, Tamires Martins Rezende, Andreia Chagas Rocha Toffolo, and Cristiano Leite De Castro. Libras-10 Dataset, May 2019. URL https://zenodo.org/record/3229958. Type: dataset.

[10] Sílvia Grasiella Moreira Almeida. Libras-34 Dataset (Kinect v1). "https://zenodo.org/record/4451526", September 2014. URL https://zenodo.org/record/4451526. Version Number: 1 Type: dataset.

[11] Abdulaziz Almohimeed, Mike Wald, and Robert Damper. An Arabic Sign Language Corpus for Instructional Language in School. *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, May 2010.

[12] Carlos Eduardo G. R. Alves, Francisco de Assis Boldt, and Thiago M. Paixão. Enhancing Brazilian Sign Language Recognition through Skeleton Image Representation, April 2024. URL http://arxiv.org/abs/2404.19148. arXiv:2404.19148 [cs].

[13] Vassilis Athitsos, Carol Neidle, Stan Sclaroff, Joan Nash, Alexandra Stefan, Quan Yuan, and Ashwin Thangali. The American Sign Language Lexicon Video Dataset. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, June 2008. doi: 10.1109/CVPRW.2008.4563181. ISSN: 2160-7508.

[14] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic Data from Diffusion Models Improves ImageNet Classification. April 2023. doi: 10.48550/arXiv.2304.08466. URL http://arxiv.org/abs/2304.08466. arXiv:2304.08466 [cs].

[15] Yutong Bai, Haoqi Fan, Ishan Misra, Ganesh Venkatesh, Yongyi Lu, Yuyin Zhou, Qihang Yu, Vikas Chandra, and Alan Yuille. Can Temporal Information Help with Contrastive Self-Supervised Learning?, November 2020. URL http://arxiv.org/abs/2011.13046. arXiv:2011.13046 [cs].

[16] Sagor Chandro Bakchy, Md. Rabiul Islam, M. Rasel Mahmud, and Faisal Imran. Human gait analysis using gait energy image, 2022. URL https://arxiv.org/abs/2203.09549.

[17] Igor L.O. Bastos, Michele F. Angelo, and Angelo C. Loula. Recognition of Static Gestures Applied to Brazilian Sign Language (Libras). In *2015 28th SIBGRAPI Conference on Graphics, Patterns and Images*, pages 305–312, August 2015. doi: 10.1109/SIBGRAPI.2015.26. ISSN: 2377-5416.

[18] Ursula Bellugi and Susan Fischer. A comparison of sign language and spoken language. *Cognition*, 1:173–200, 1 1972. ISSN 0010-0277. doi: 10.1016/0010-0277(72)90018-2.

[19] Matyas Bohacek and Marek Hruz. Sign Pose-based Transformer for Word-level Sign Language Recognition. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 182–191, Waikoloa, HI, USA, January 2022. IEEE. ISBN 978-1-66545-824-5. doi: 10.1109/WACVW54805.2022.00024. URL https://ieeexplore.ieee.org/document/9707552/.

[20] BSLCorpus. The british sign language (bsl) corpus, 2022. URL http://bslcorpusproject.org.temp.link/.

[21] Eros Caiafa, Fabiana Fonseca, Amaro Lima, Gabriel Araujo, and Eduardo Silva. Aprendizado profundo no reconhecimento de sinais estáticos de Libras. In *Anais de XXXVIII Simpósio Brasileiro de Telecomunicações e Processamento de Sinais*. Sociedade Brasileira de Telecomunicações, 2020. doi: 10.14209/SBRT.2020.1570660513. URL http://biblioteca.sbrt.org.br/articles/2318.

[22] Wladimir Arturo Garces Carrillo, Emely Pujoli da Silva, and Marcelo da Silva Reis. Performance of vision transformers in sign language identification. Poster presented at the Reunión Internacional de Inteligencia Artificial y sus Aplicaciones (RIIAA), 2024. Conference poster.

[23] Lourdes Ramirez Cerna, Edwin Escobedo Cardenas, Dayse Garcia Miranda, David Menotti, and Guillermo Camara-Chavez. A multimodal LIBRAS-UFOP Brazilian sign language dataset of minimal pairs using a microsoft Kinect sensor. *Expert Systems with Applications*, 167:114179, April 2021. ISSN 0957-4174. doi: 10.1016/j.eswa.2020.114179. URL https://www.sciencedirect.com/science/article/pii/S0957417420309143.

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002. ISSN 1076-9757. doi: 10.1613/jair.953. URL http://arxiv.org/abs/1106.1813. arXiv:1106.1813 [cs].

[25] Dong Chen, Xinda Qi, Yu Zheng, Yuzhen Lu, and Zhaojian Li. Deep Data Augmentation for Weed Recognition Enhancement: A Diffusion Probabilistic Model and Transfer Learning Based Approach, October 2022. URL http://arxiv.org/abs/2210.09509. arXiv:2210.09509 [cs].

[26] Ming Jin Cheok, Zaid Omar, and Mohamed Hisham Jaward. A review of hand gesture and sign language recognition techniques. *International Journal of Machine Learning and Cybernetics*, 10 (1):131–153, January 2019. ISSN 1868808X. doi: 10.1007/S13042-017-0705-5/TABLES/5. URL https://link.springer.com/article/10.1007/s13042-017-0705-5. Publisher: Springer Verlag.

[27] Christopher Conly, Paul Doliotis, Pat Jangyodsuk, Rommel Alonzo, and Vassilis Athitsos. Toward a 3D body part detection video dataset and hand tracking benchmark. In *Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments*, pages 1–6, Rhodes Greece, May 2013. ACM. ISBN 978-1-4503-1973-7. doi: 10.1145/2504335.2504337. URL https://dl.acm.org/doi/10.1145/2504335.2504337.

[28] Cicero Ferreira Fernandes Costa, Robson Silva de Souza, Jonilson Roque dos Santos, Bárbara Lobato dos Santos, and Marly Guimarães Fernandes Costa. A fully automatic method for recognizing hand configurations of brazilian sign language. *Research on Biomedical Engineering*, 33:78–89, March 2017. ISSN 2446-4732, 2446-4740. doi: 10.1590/2446-4740.03816. URL http://www.scielo.br/j/reng/a/LwYDtw8r9MbYZ89mmyXfpTr/abstract/?lang=en. Publisher: Sociedade Brasileira de Engenharia Biomédica.

[29] Ekin D. Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. AutoAugment: Learning Augmentation Strategies From Data, 2019. URL https://ieeexplore.ieee.org/document/8953317/.

[30] Emely Pujólli Da Silva, Paula D Paro Costa, Kate M Oliveira Kumada, and Mario De Martino. Classification of Facial Action Units in Brazilian Sign Language. Technical report, University of Campinas, 2019.

[31] Damien Dablain, Bartosz Krawczyk, and Nitesh V. Chawla. DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data. May 2021. doi: 10.48550/arXiv.2105.02340. URL http://arxiv.org/abs/2105.02340. arXiv:2105.02340 [cs].

[32] Murtaza Dalal, Alexander C. Li, and Rohan Taori. Autoregressive models: What are they good for?, 2019. URL https://arxiv.org/abs/1910.07737.

[33] Gabriel Peixoto de Carvalho, André Luiz Brandão, and Fernando Teubl Ferreira. Handarch: A deep learning architecture for libras hand configuration recognition. *Anais do XVII Workshop de Visão Computacional (WVC 2021)*, pages 19–24, 11 2021. doi: 10.5753/wvc.2021.18883.

[34] Giulia Zanon de Castro, Rúbia Reis Guerra, and Frederico Gadelha Guimarães. Automatic translation of sign language with multi-stream 3D CNN and generation of artificial depth maps. *Expert Systems with Applications*, 215:119394, April 2023. ISSN 0957-4174. doi: 10.1016/j.eswa.2022.119394. URL https://www.sciencedirect.com/science/article/pii/S0957417422024125.

[35] Patrícia Teixeira De Matos, Giovana Maria Belém Falcão, Pedro Claesen Dutra Silva, Tania Maria De Sousa França, M. Marinho, and Gabrielle Silva Marinho. Comparative analysis of the impact of libras on development of the deaf: A case study in a city in northeastern brazil. *International Journal of Research*, 8:72–83, 2020. doi: 10.29121/granthaalayah.v8.i2.2020.186.

[36] Ronice Muller De Quadros, Aline Lemos Pizzio, and Patrícia Luiza Ferreira Rezende. Universidade federal de santa catarina licenciatura e bacharelado em letras-libras na modalidade a distância. 2009.

[37] Ronice Müller De Quadros and Alexandre Melo De Sousa. Brazilian sign language corpus: Acre libras inventory/corpus da língua brasileira de sinais: inventário de libras do acre. *REVISTA DE ESTUDOS DA LINGUAGEM*, 29(2):805, March 2021. ISSN 2237-2083, 0104-0588. doi: 10.17851/2237-2083.29.2.805-828. URL http://www.periodicos.letras.ufmg.br/index.php/relin/article/view/17344.

[38] Ronice Müller De Quadros, Bruna Crescêncio Neves, Deonísio Schmitt, Juliana Tasca Lohn, and Marcos Luchi. *Língua Brasileira de Sinais: Patrimônio Linguístico Brasileiro.* Garapuvu, 2018.

[39] Ronice Müller De Quadros, Christian Rathmann, Johanna Mesch, and Jair Barbosa Da Silva. Documentação de línguas de sinais. *Fórum Linguístico*, 17(4):5444–5456, December 2020. ISSN 1984-8412, 1415-8698. doi: 10.5007/1984-8412.2020.e77336. URL https://periodicos.ufsc.br/index.php/forum/article/view/77336.

[40] Guillaume Devineau, Fabien Moutarde, Wang Xi, and Jie Yang. Deep learning for hand gesture recognition on skeletal data. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 106–113, May 2018. doi: 10.1109/FG.2018.00025.

[41] Terrance DeVries and Graham W. Taylor. Dataset Augmentation in Feature Space, February 2017. URL http://arxiv.org/abs/1702.05538. arXiv:1702.05538 [cs, stat].

[42] Francisco Aulísio dos Santos Paiva, Plínio Almeida Barbosa, José Mario De Martino, Ackley Dias Will, Márcia Regina Nepomuceno dos Santos Oliveira, Ivani Rodrigues Silva, and André Nogueira Xavier. Analysis of the role of non manual expressions in intensification processes in brazilian sign language. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, 34:1135–1158, 12 2018. ISSN 0102-4450. doi: 10.1590/0102-445069907579551549. URL https://www.scielo.br/j/delta/a/ZkwFT3Nh3ncD4z8TpzjDK4d/abstract/?format=html&lang=en.

[43] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. June 2020. doi: 10.48550/arXiv.2010.11929. URL http://arxiv.org/abs/2010.11929. arXiv:2010.11929 [cs] version: 1.

[44] Yao Du, Pan Xie, Mingye Wang, Xiaohui Hu, Zheng Zhao, and Jiaqi Liu. Full transformer network with masking future for word-level sign language recognition. *Neurocomputing*, 500:115–123, August 2022. ISSN 0925-2312. doi: 10.1016/j.neucom.2022.05.051. URL https://www.sciencedirect.com/science/article/pii/S0925231222006178.

[45] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[46] Sergio Escalera, Vassilis Athitsos, and Isabelle Guyon. Challenges in multimodal gesture recognition. *Journal of Machine Learning Research*, 17:1–60, 4 2017. ISSN 15337928. doi: 10.1007/978-3-319-57021-1_1/FIGURES/14. URL https://link.springer.com/chapter/10.1007/978-3-319-57021-1_1.

[47] Edwin Jonathan Escobedo Cardenas and Guillermo Camara Chavez. Multimodal hand gesture recognition combining temporal and pose information based on CNN descriptors and histogram of cumulative magnitudes. *Journal of Visual Communication and Image Representation*, 71:102772, August 2020. ISSN 1047-3203. doi: 10.1016/j.jvcir.2020.102772. URL https://www.sciencedirect.com/science/article/pii/S1047320320300225.

[48] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and Content-Guided Video Synthesis with Diffusion Models, February 2023. URL http://arxiv.org/abs/2302.03011. arXiv:2302.03011 [cs].

[49] Tanya A Felipe. Curso bÁsico livro do estudante. 2007. URL www.feneis.org.br.

[50] Cristiane Lima Terra Fernandes. Libras in law and school practice. *Momento - Diálogos em Educação*, 2022. doi: 10.14295/momento.v31i02.14507.

[51] Lucinda Ferreira-Brito. Uma abordagem fonológica dos sinais da lscb. *Espaço Informativo técnico-científico do INES*, 1990. URL https://seer.ines.gov.br/index.php/revista-espaco/article/view/35/29.

[52] Lucinda Ferreira-Brito. Por uma gramática de língua de sinais. *Tempo Brasileiro*, 2010.

[53] Priscila V. Gameiro, Wesley L. Passos, Gabriel M. Araujo, Amaro A. De Lima, Jonathan N. Gois, and Anna R. Corbo. A brazilian sign language video database for automatic recognition. *2020 Latin American Robotics Symposium, 2020 Brazilian Symposium on Robotics and 2020 Workshop on Robotics in Education, LARS-SBR-WRE 2020*, November 2020. doi: 10.1109/LARS/SBR/WRE51543.2020.9307017. Publisher: Institute of Electrical and Electronics Engineers Inc. ISBN: 9780738111537.

[54] governodigital. Vlibras — governo digital, 2020. URL https://www.gov.br/governodigital/pt-br/vlibras.

[55] C. Guimaraes and Rita Cassia Maestri. Non-manual expression – sign language as l2. *International Journal for Innovation Education and Research*, 2018. doi: 10.31686/ijier.vol6.iss10.1169.

[56] Zihui Guo, Yonghong Hou, Chunping Hou, and Wenjie Yin. Locality-Aware Transformer for Video-Based Sign Language Translation. *IEEE Signal Processing Letters*, 30:364–368, 2023. ISSN 1558-2361. doi: 10.1109/LSP.2023.3263808. Conference Name: IEEE Signal Processing Letters.

[57] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild, 2018.

[58] J. Han and Bir Bhanu. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):316–322, 2006. doi: 10.1109/TPAMI.2006.38.

[59] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[60] Reemt Hinrichs, Angelo Sitcheu, and Jörn Ostermann. Continuous Sign-Language Recognition using Transformers and Augmented Pose Estimation:. pages 672–678, 2023. doi: 10.5220/0011709100003411. URL https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0011709100003411.

[61] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

[62] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. December 2020. doi: 10.48550/arXiv.2006.11239. URL http://arxiv.org/abs/2006.11239. arXiv:2006.11239 [cs, stat].

[63] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, and Tim Salimans. IMAGEN VIDEO: HIGH DEFINITION VIDEO GENERATION WITH DIFFUSION MODELS. 2022. URL https://imagen.research.google/video.

[64] Chao Hu, Liqiang Zhu, Weibin Qiu, and Weijie Wu. Data Augmentation Vision Transformer for Fine-grained Image Classification. November 2022. doi: 10.48550/arXiv.2211.12879. URL http://arxiv.org/abs/2211.12879. arXiv:2211.12879 [cs].

[65] Abhinav Joshi, Ashwani Bhat, Pradeep S, Priya Gole, Shashwat Gupta, Shreyansh Agarwal, and Ashutosh Modi. CISLR: Corpus for Indian Sign Language Recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10357–10366, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.707.

[66] Abhinav Joshi, Susmit Agrawal, and Ashutosh Modi. Isltranslate: Dataset for translating indian sign language, 2023.

[67] José Arnor De Lima Júnior, Indira Simionatto Stedile, Assis Moura, Sédina Dos Santos, Jales Ferreira, Ana Elilia Trigueiro, and Barros Cavalcanti. De mÃos amarradas para as mÃos que falam: A histÓria da educaÇÃo dos surdos no mundo e no brasil, 2021. URL https://editorarealize.com.br/artigo/visualizar/81751.

[68] Ildar Kagirov, Dmitry Ryumin, and Alexandr Axyonov. Method for Multimodal Recognition of One-Handed Sign Language Gestures Through 3D Convolution and LSTM Neural Networks. In Albert Ali Salah, Alexey Karpov, and Rodmonga Potapova, editors, *Speech and Computer*, Lecture Notes in Computer Science, pages 191–200, Cham, 2019. Springer International Publishing. ISBN 978-3-030-26061-3. doi: 10.1007/978-3-030-26061-3_20.

[69] Nour Eldeen Khalifa, Mohamed Loey, and Seyedali Mirjalili. A comprehensive survey of recent trends in deep learning for digital images augmentation. *Artificial Intelligence Review*, 55(3):2351–2377, March 2022. ISSN 1573-7462. doi: 10.1007/s10462-021-10066-4. URL https://doi.org/10.1007/s10462-021-10066-4.

[70] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, December 2015. ISSN 1077-3142. doi: 10.1016/j.cviu.2015.09.013. URL https://www.sciencedirect.com/science/article/pii/S1077314215002088.

[71] Deep R. Kothadiya, Chintan M. Bhatt, Tanzila Saba, Amjad Rehman, and Saeed Ali Bahaj. SIGNFORMER: DeepVision Transformer for Sign Language Recognition. *IEEE Access*, 11:4730–4739, 2023. ISSN 2169-3536. doi: 10.1109/ACCESS.2022.3231130. Conference Name: IEEE Access.

[72] Kate Mamhy Oliveira Kumada. Libras: Língua brasileira de sinais. *Londrina: Editora e Distribuidora Educacional SA*, 2016.

[73] Ghazanfar Latif, Nazeeruddin Mohammad, Jaafar Alghazo, Roaa AlKhalaf, and Rawan AlKhalaf. Arasl: Arabic alphabets sign language dataset. *Data in Brief*, 23:103777, 4 2019. ISSN 23523409. doi: 10.1016/j.dib.2019.103777.

[74] Joseph Lemley, Shabab Bazrafkan, and Peter Corcoran. Smart augmentation learning an optimal data augmentation strategy. *IEEE Access*, 5:5858–5869, 2017. ISSN 2169-3536. doi: 10.1109/ACCESS.2017.2696121.

[75] Jie Li, Mingqiang Yang, Yupeng Liu, Yanyan Wang, Qinghe Zheng, and Deqiang Wang. Dynamic Hand Gesture Recognition Using Multi-direction 3D Convolutional Neural Networks. 2019.

[76] H. J. D. Lima. Processos produtivos de sinais nocionalmente nomes na libras. 2:6–23, 2017. doi: 10.5216/RS.V1I1.47164.

[77] Jihao Liu, Boxiao Liu, Hang Zhou, Hongsheng Li, and Yu Liu. TokenMix: Rethinking Image Mixing for Data Augmentation in Vision Transformers. April 2023. doi: 10.48550/arXiv.2207.08409. URL http://arxiv.org/abs/2207.08409. arXiv:2207.08409 [cs].

[78] Silke Matthes, Thomas Hanke, Anja Regen, Jakob Storz, Satu Worseck, Eleni Efthimiou, Athanasia-Lida Dimou, Annelies Braffort, John Glauert, and Eva Safar. Dicta sign building a multilingual sign language corpus. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon. Satellite Workshop to the eighth International Conference on Language Resources and Evaluation (LREC-2012)*, 2012. URL https://hal.science/hal-03404012.

[79] Mansooreh Montazerin, Soheil Zabihi, Elahe Rahimian, Arash Mohammadi, and Farnoosh Naderkhani. ViT-HGR: Vision Transformer-based Hand Gesture Recognition from High Density Surface EMG Signals. January 2022. doi: 10.48550/arXiv.2201.10060. URL http://arxiv.org/abs/2201.10060. arXiv:2201.10060 [cs, eess].

[80] Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. Data Augmentation for Sign Language Gloss Translation. *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages, AT4SSL 2021*, pages 1–11, May 2021. doi: 10.48550/arxiv.2105.07476. URL https://arxiv.org/abs/2105.07476v1. arXiv: 2105.07476 Publisher: Association for Machine Translation in the Americas.

[81] Lan Thao Nguyen, Florian Schicktanz, Aeneas Stankowski, and Eleftherios Avramidis. Automatic generation of a 3D sign language avatar on AR glasses given 2D videos of human signers. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 71–81, Virtual, August 2021. Association for Machine Translation in the Americas. URL https://aclanthology.org/2021.mtsummit-at4ssl.8.

[82] Alex Nichol and Prafulla Dhariwal. Improved Denoising Diffusion Probabilistic Models, February 2021. URL http://arxiv.org/abs/2102.09672. arXiv:2102.09672 [cs, stat].

[83] Wesley L. Passos, Gabriel M. Araujo, Jonathan N. Gois, and Amaro A. de Lima. A gait energy image-based system for brazilian sign language recognition. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 68(11):4761–4771, 2021. doi: 10.1109/TCSI.2021.3091001.

[84] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning, 2017.

[85] Geoffrey Poon, Kin Chung Kwan, and Wai-Man Pang. Occlusion-robust bimanual gesture recognition by fusing multi-views. *Multimedia Tools and Applications*, 78(16):23469–23488, August 2019. ISSN 1573-7721. doi: 10.1007/s11042-019-7660-y. URL https://doi.org/10.1007/s11042-019-7660-y.

[86] Junfu Pu, Wengang Zhou, Hezhen Hu, and Houqiang Li. Boosting Continuous Sign Language Recognition via Cross Modality Augmentation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1497–1505, October 2020. doi: 10.1145/3394171.3413931. URL http://arxiv.org/abs/2010.05264. arXiv:2010.05264 [cs].

[87] Elakkiya R and Natarajan B. ISL-CSLTR: Indian Sign Language Dataset for Continuous Sign Language Translation and Recognition. 1, January 2021. doi: 10.17632/kcmpdxky7p.1. URL https://data.mendeley.com/datasets/kcmpdxky7p/1. Publisher: Mendeley Data.

[88] Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, and Mohammad Sabokrou. Sign language production: A review. *CoRR*, abs/2103.15910, 2021. URL https://arxiv.org/abs/2103.15910.

[89] Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, Vassilis Athitsos, and Mohammad Sabokrou. All you need in sign language production. *CoRR*, abs/2201.01609, 2022. URL https://arxiv.org/abs/2201.01609.

[90] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows, 2016. URL https://arxiv.org/abs/1505.05770.

[91] Tamires Rezende, Cristiano Castro, and Silvia Almeida. An approach for brazilian sign language recognition based on facial expression and k-nn classifier. 10 2016.

[92] Tamires Martins Rezende, Sílvia Grasiella Moreira Almeida, and Frederico Gadelha Guimarães. Development and validation of a brazilian sign language database for human gesture recognition. *Neural Computing and Applications*, 33:10449–10467, 8 2021. ISSN 14333058. doi: 10.1007/S00521-021-05802-4/FIGURES/22. URL https://link.springer.com/article/10.1007/s00521-021-05802-4.

[93] Ailton José Rodrigues. V-LIBRASIL : uma base de dados com sinais na língua brasileira de sinais (Libras), August 2021. URL https://repositorio.ufpe.br/handle/123456789/43491. Accepted: 2022-03-23T19:51:38Z Publisher: Universidade Federal de Pernambuco.

[94] ANGÉLICA NIERO MENDES DOS SANTOS and CÁSSIA GECIAUSKAS SOFIATO. A educaÇÃo de surdos no sÉculo xix e a circulaÇÃo da lÍngua de sinais no imperial instituto de surdos-mudos. *Educação em Revista*, 37:e288663, 2021. ISSN 0102-4698. doi: 10.1590/0102-4698288663. URL https://doi.org/10.1590/0102-4698288663.

[95] Zed Sevcikova Sehyr, Naomi Caselli, Ariel M Cohen-Goldberg, and Karen Emmorey. The asl-lex 2.0 project: A database of lexical and phonological properties for 2,723 signs in american sign language. *The Journal of Deaf Studies and Deaf Education*, 26:263–277, 3 2021. ISSN 1081-4159. doi: 10.1093/deafed/enaa038. URL https://doi.org/10.1093/deafed/enaa038.

[96] Connor Shorten and Taghi M. Khoshgoftaar. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1):60, July 2019. ISSN 2196-1115. doi: 10.1186/s40537-019-0197-0. URL https://doi.org/10.1186/s40537-019-0197-0.

[97] Ala Addin I. Sidig, Hamzah Luqman, Sabri Mahmoud, and Mohamed Mohandes. KArSL: Arabic Sign Language Database. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 20(1):14:1–14:19, March 2021. ISSN 2375-4699. doi: 10.1145/3423420. URL https://doi.org/10.1145/3423420.

[98] Emely Pujólli Da Silva, Paula Dornhofer, Paro Costa, Kate Mamhy, Oliveira Kumada, and José Mario De Martino. Silfa: Sign language facial action database for the development of assistive technologies for the deaf. Technical report, Unicamp, 2021.

[99] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-A-Video: Text-to-Video Generation without Text-Video Data. September 2022. doi: 10.48550/arXiv.2209.14792. URL http://arxiv.org/abs/2209.14792. arXiv:2209.14792 [cs].

[100] Abhishek Sinha, Kumar Ayush, Jiaming Song, Burak Uzkent, Hongxia Jin, and Stefano Ermon. Negative Data Augmentation. February 2021. doi: 10.48550/arXiv.2102.05113. URL http://arxiv.org/abs/2102.05113. arXiv:2102.05113 [cs].

[101] William C. Stokoe. Sign Language Structure. *Annual Review of Anthropology*, 9(1):365–390, 1980. doi: 10.1146/annurev.an.09.100180.002053. URL https://doi.org/10.1146/annurev.an.09.100180.002053. _eprint: https://doi.org/10.1146/annurev.an.09.100180.002053.

[102] Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. Sign Language Production using Neural Machine Translation and Generative Adversarial Networks. *Proceedings of the 29th British Machine Vision Conference (BMVC 2018)*, 2018. URL https://openresearch.surrey.ac.uk/esploro/outputs/conferencePresentation/Sign-Language-Production-using-Neural-Machine/99511566102346#file-0.

[103] Stephanie Stoll, Simon Hadfield, and Richard Bowden. SignSynth: Data-Driven Sign Language Video Generation. In Adrien Bartoli and Andrea Fusiello, editors, *Computer Vision – ECCV 2020 Workshops*, Lecture Notes in Computer Science, pages 353–370, Cham, 2020. Springer International Publishing. ISBN 978-3-030-66823-5. doi: 10.1007/978-3-030-66823-5_21.

[104] Yong Soon Tan, Kian Ming Lim, and Chin Poo Lee. Hand gesture recognition via enhanced densely connected convolutional neural network. *Expert Systems with Applications*, 175:114797, August 2021. ISSN 0957-4174. doi: 10.1016/j.eswa.2021.114797. URL https://www.sciencedirect.com/science/article/pii/S0957417421002384.

[105] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective Data Augmentation With Diffusion Models, May 2023. URL http://arxiv.org/abs/2302.07944. arXiv:2302.07944 [cs].

[106] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[107] José Elías Yauri Vidalón and José Mario De Martino. Continuous Sign Recognition of Brazilian Sign Language in a Healthcare Setting. *Journal of Communication and Information Systems*, 30 (1), October 2015. ISSN 1980-6604. doi: 10.14209/jcis.2015.10. URL https://jcis.sbrt.org.br/jcis/article/view/99. Number: 1.

[108] Yulin Wang, Gao Huang, Shiji Song, Xuran Pan, Yitong Xia, and Cheng Wu. Regularizing Deep Networks with Semantic Data Augmentation, June 2021. URL http://arxiv.org/abs/2007.10538. arXiv:2007.10538 [cs].

[109] Lilian Weng. From autoencoder to beta-vae. *lilianweng.github.io*, 2018. URL https://lilianweng.github.io/posts/2018-08-12-vae/.

[110] Lilian Weng. What are diffusion models? *lilianweng.github.io*, Jul 2021. URL https://lilianweng.github.io/posts/2021-07-11-diffusion-models/.

[111] Di Wu, Junjun Chen, Nabin Sharma, Shirui Pan, Guodong Long, and Michael Blumenstein. Adversarial Action Data Augmentation for Similar Gesture Action Recognition. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, July 2019. doi: 10.1109/IJCNN.2019.8851993. ISSN: 2161-4407.

[112] Ying Wu and Thomas S. Huang. Human hand modeling, analysis and animation in the context of hci. *IEEE International Conference on Image Processing*, 3:6–10, 1999. doi: 10.1109/ICIP.1999.817058.

[113] Ying Wu and Thomas S. Huang. Vision-based gesture recognition: A review. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1739:103–115, 1999. ISSN 16113349. doi: 10.1007/3-540-46616-9_10/COVER. URL https://link.springer.com/chapter/10.1007/3-540-46616-9_10.

[114] Changrong Xiao, Sean Xin Xu, and Kunpeng Zhang. Multimodal Data Augmentation for Image Captioning using Diffusion Models. May 2023. URL http://arxiv.org/abs/2305.01855. arXiv:2305.01855 [cs].

[115] Qinkun Xiao, Minying Qin, and Yuting Yin. Skeleton-based Chinese sign language recognition and generation for bidirectional communication between deaf and hearing people. *Neural Networks*, 125:41–55, May 2020. ISSN 0893-6080. doi: 10.1016/j.neunet.2020.01.030. URL https://www.sciencedirect.com/science/article/pii/S089360802030040X.

[116] Ruiduo Yang, Sudeep Sarkar, and Barbara Loeding. Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming. *IEEE transactions on pattern analysis and machine intelligence*, 32:462–477, 2010. ISSN 1939-3539. doi: 10.1109/TPAMI.2009.26. URL https://pubmed.ncbi.nlm.nih.gov/20075472/.

[117] Jinhui Ye, Wenxiang Jiao, Xing Wang, Zhaopeng Tu, and Hui Xiong. Cross-modality Data Augmentation for End-to-End Sign Language Translation. May 2023. URL http://arxiv.org/abs/2305.11096. arXiv:2305.11096 [cs].

[118] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features, 2019.

[119] Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, and Jinhyung Kim. Videomix: Rethinking data augmentation for video classification, 2020.

[120] Giulia Zanon de Castro, Rúbia Reis Guerra, Moises Mendes de Assis, Tamires Martins Rezende, Gabriela Tolentino Boaventura de Almeida, Sílvia Grasiella Moreira Almeida, Cristiano Leite de Castro, and Frederico G. Guimarães. Desenvolvimento de uma base de dados de sinais de libras para aprendizado de máquina: Estudo de caso com cnn 3d. *Anais do 14º Simpósio Brasileiro de Automação Inteligente*, December 2019. doi: 10.17648/SBAI-2019-111451. Publisher: Galoa Events Proceedings.

[121] Felix Zhan. Hand gesture recognition with convolution neural networks. In *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 295–298, July 2019. doi: 10.1109/IRI.2019.00054.

[122] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. MediaPipe Hands: On-device Real-time Hand Tracking, June 2020. URL http://arxiv.org/abs/2006.10214. arXiv:2006.10214 [cs].

[123] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond Empirical Risk Minimization, April 2018. URL http://arxiv.org/abs/1710.09412. arXiv:1710.09412 [cs, stat] version: 2.

[124] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model, August 2022. URL http://arxiv.org/abs/2208.15001. arXiv:2208.15001 [cs].

[125] Yumeng Zhang, Gaoguo Jia, Li Chen, Mingrui Zhang, and Junhai Yong. Self-Paced Video Data Augmentation with Dynamic Images Generated by Generative Adversarial Networks, September 2019. URL http://arxiv.org/abs/1909.12929. arXiv:1909.12929 [cs].

[126] Yumeng Zhang, Gaoguo Jia, Li Chen, Mingrui Zhang, and Junhai Yong. Self-Paced Video Data Augmentation by Generative Adversarial Networks with Insufficient Samples. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1652–1660, Seattle WA USA, October 2020. ACM. ISBN 978-1-4503-7988-5. doi: 10.1145/3394171.3414003. URL https://dl.acm.org/doi/10.1145/3394171.3414003.