A Lightweight I3D-Based Approach for Real-Time Brazilian Sign Language Recognition

Victor Costa
vcrcosta@inf.ufpel.edu.br
PPGC, ViTech - UFPel
Pelotas/RS, Brazil

Leandro Tavares
lwtavares@inf.ufpel.edu.br
PPGC, ViTech - UFPel
Pelotas/RS, Brazil

Ruhan Conceição
radconceicao@inf.ufpel.edu.br
PPGC, ViTech - UFPel
Pelotas/RS, Brazil

Luciano Agostini
agostini@inf.ufpel.edu.br
PPGC, ViTech - UFPel
Pelotas/RS, Brazil

Brenda Salenave Santana
bssalenave@inf.ufpel.edu.br
PPGC - UFPel
Pelotas/RS, Brazil

Tatiana Bolivar Lebedeff
tblebedeff@gmail.com
PPGL - UFPel
Pelotas/RS, Brazil

Guilherme Corrêa
gcorrea@inf.ufpel.edu.br
PPGC, ViTech - UFPel
Pelotas/RS, Brazil

ABSTRACT

The demand for accessible technologies to support the deaf and hard-of-hearing community in Brazil is significant. However, many state-of-the-art deep learning models are too computationally intensive for practical, real-time applications. This study addresses this gap by proposing an efficient, lightweight pipeline approach for isolated Brazilian Sign Language (LIBRAS) recognition. We fine-tune a pre-trained Inflated 3D ConvNet (I3D) model on the MINDS-Libras dataset using an end-to-end methodology that operates directly on raw RGB videos, circumventing the need for heavy pre-processing steps like skeleton extraction. To ensure a realistic evaluation of the model's generalization capabilities, we adopt a strict signer-independent protocol, where test subjects are completely unseen during training. Our proposed model achieves a competitive accuracy of 92.5% and is able to perform sign recognition in real-time, demonstrating strong performance comparable to more complex architectures. This work establishes a new, robust benchmark for signer-agnostic LIBRAS recognition, highlighting that an end-to-end approach can effectively balance high accuracy with the pipeline efficiency required for deployable, real-world accessibility tools.

KEYWORDS

Brazilian Sign Language, LIBRAS, Real-Time Recognition, Deep Learning, I3D, Signer-Independent

1 INTRODUCTION

Modern technology has the potential to bridge long-standing social gaps, yet individuals with disabilities are still often excluded from many of its benefits. Among these groups, the deaf and hard-of-hearing community continues to face significant barriers in communication, education, and access to services. In Brazil, accessibility is a legal right, guaranteed by the legislation [11, 12], which recognizes LIBRAS (*LIngua BRAsileira de Sinais* – Brazilian Sign Language) as a legal means of communication and expression. Recently, this legal framework was expanded with the introduction of Bilingual Education for the Deaf as a new modality of Brazilian education, establishing LIBRAS as the first language and written Portuguese as the second for deaf students [2]. This significant advancement in educational policy underscores the urgent need for technological solutions that can support quality education tailored to the linguistic and cultural specificities of the deaf community. Approximately 5% of the Brazilian population lives with some degree of hearing impairment [9], highlighting the urgency of inclusive solutions. Therefore, scientific and technological research must actively address such inequalities, particularly by developing tools that support natural, inclusive communication.

Recent advances in deep learning, particularly with the emergence of large transformer-based models [16], have led to promising results in video understanding and language translation. However, such models are computationally intensive and often impractical for real-time use, especially in resource-constrained environments such as smartphones, embedded systems, or public service terminals [18]. To develop truly inclusive technologies, we must also consider their deployability. This calls for solutions that are accurate and efficient in terms of complexity.

In response to this challenge, our work introduces an efficient, end-to-end deep learning model for LIBRAS recognition. We fine-tune the Inflated 3D ConvNet (I3D) architecture [4], pre-trained on the Kinetics-400 dataset [8], directly on raw RGB videos from the MINDS-Libras dataset [13], a design that strategically circumvents the computational bottlenecks of methods reliant on pre-processing steps like skeleton extraction. Another important contribution of this work is demonstrating the model's performance under a strict signer-independent evaluation protocol. By testing the system exclusively on signers unseen during training, we provide an accurate

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2025). Rio de Janeiro, Brazil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2966-2753

measure of its generalization capability. The trained model can recognize LIBRAS signs in real-time with 92.5% accuracy. Therefore, this study establishes a new benchmark for the dataset that balances sign recognition accuracy and complexity, allowing for practical deployment of the processing pipeline.

This paper is organized as follows. Section 2 reviews related work in the field. Section 3 presents the proposed solution and details the dataset-splitting strategy, data pre-processing, training, and the final model architecture. Section 4 reports experimental results in accuracy and complexity, comparing this work to previous solutions. Finally, Section 5 concludes the paper.

2 RELATED WORK

Silva et al. [5] explored data synthesis for sentence-level Sign Language Translation (SLT), using pre-computed I3D features [4] for feature extraction. However, their approach does not address end-to-end fine-tuning for isolated sign classification or generalization across different signers.

Fanucchi et al. [6] applied advanced Vision Transformer architectures to LIBRAS recognition in an Augmented Reality (AR) interpretation system. They fine-tuned a Video Masked Autoencoder (VideoMAE) [15], achieving 84% accuracy when trained with 10 signers. Optimal performance occurred at epoch 4, after which validation loss increased, indicating overfitting. The study demonstrates the feasibility of Vision Transformers for LIBRAS and highlights opportunities for improvement through stronger training strategies and exploration of alternative video architectures.

Alves et al. (2024) [1] propose an Isolated Sign Language Recognition (ISLR) approach where body, hands, and facial landmarks are extracted from RGB video frames using OpenPose [3]. These spatio-temporal landmarks are then encoded into a single 2D image, which serves as input to a ResNet-18 [7] pre-trained on the ImageNet Dataset. The work presents a key limitation: the significant time imposed by the OpenPose-based landmark extraction (approximately 36 seconds per video sequence), hindering real-time performance.

Rezende et al. [14] detail the development and validation of the MINDS-Libras [13] dataset used in our study. The paper introduces a multi-modal database, as described in Section 3.1. To validate the dataset's utility for gesture recognition tasks, the authors present a baseline study employing a 3D Convolutional Neural Network (3D CNN) trained on the RGB videos. They report a high classification accuracy of 93.3% in their experiments. However, it is noteworthy that this result was obtained using a "per-sign" data split, meaning that video samples from the same signer could be present in both the training and testing sets. This work provides not only a valuable public resource for the research community but also an important performance benchmark for the MINDS-Libras dataset.

Despite significant recent advances in LIBRAS recognition, most studies do not address computational complexity or the feasibility of real-time processing. Among the reviewed works, only Alves et al. (2024) [1] report execution time, with approximately 36 seconds per video sequence – clearly unsuitable for real-time applications. Moreover, other approaches rely on computationally intensive strategies, such as the use of Vision Transformers, as in [6]. This highlights a gap in the literature and the need for more efficient methods that balance performance with practical applicability in interactive and real-time systems.

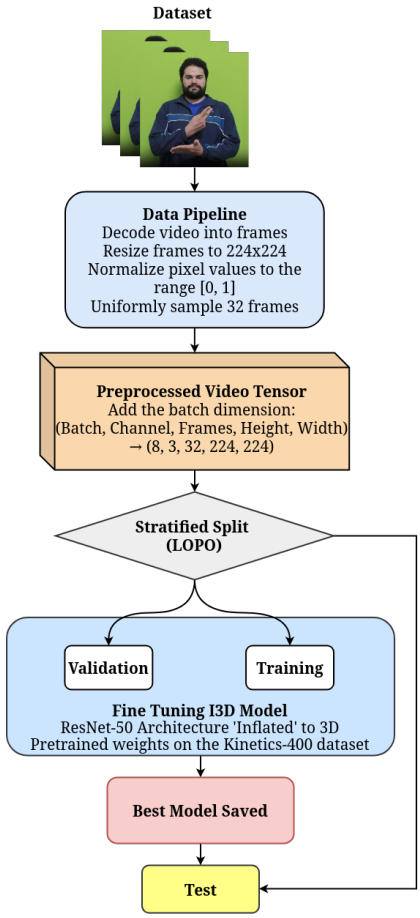

Figure 1: Methodology employed for training and testing the I3D-based LIBRAS recognition model.

3 LIGHTWEIGHT LIBRAS RECOGNITION MODEL

This section describes the approach presented in Fig. 1, employed to develop the proposed lightweight recognition system for LIBRAS. The solution was developed using Python 3.9 and the PyTorch deep learning framework version *2.7.0+cu128* [10].

3.1 Dataset and Splitting Strategy

We employed the MINDS-Libras dataset [13], which comprises 20 isolated signs, each performed up to five times by 12 signers with varying LIBRAS proficiency, resulting in 1,155 RGB videos. Besides conventional 2D video, the dataset also includes depth, body-joint, and facial landmark data. MINDS-Libras signs were chosen to reflect phonological diversity and recorded in controlled conditions with both RGB and RGB-D cameras. This dataset supports evaluation under realistic signer-independent protocols.

Depth and pose data were not used in our approach, which relies solely on 2D video information to maintain low model complexity – one of the central goals of this work. To assess model generalization across unseen individuals, we adopted a *Leave One Person Out* (LOPO) strategy, in which the system must operate without prior knowledge of the current signer. Specifically, signers #11 and #12 were reserved for the test set, comprising 200 samples (2 signers × 20 words × 5 repetitions).

Training and validation were performed using samples from the remaining 10 signers. We applied a stratified split by class, allocating 85% of the data for training and 15% for validation. This resulted in 814 training samples and 144 validation samples, ensuring balanced representation of all sign classes across both subsets.

3.2 Data Pre-processing and Augmentation

Video frames were spatially resized from 1920 × 1080 pixels to 224 × 224 pixels and normalized to the $[0, 1]$ range. Also, videos were temporally subsampled to a fixed length of 32 frames. Both resizing and temporal subsampling are necessary to match the input dimensions expected by the I3D architecture and to reduce computational complexity. To improve model generalization, data augmentation was applied only during the training phase. With a 50% probability, a horizontal flip was applied to the entire video tensor.

3.3 I3D Architecture Adaptation and Fine Tuning

The proposed sign language classification solution employs the *Inflated 3D ConvNet* (I3D) architecture with a ResNet-50 backbone [4]. The I3D architecture [4] extends 2D convolutional networks into the spatio-temporal domain by inflating kernels to capture both spatial and temporal features, enabling end-to-end video analysis. Pre-training on large-scale datasets such as Kinetics-400 [4] provides strong feature representations that can be fine-tuned for domain-specific tasks like LIBRAS recognition. In this work, we adopt an I3D with a ResNet-50 backbone and adapted it for our 20-class classification problem by replacing the original classification head, specifically the final fully-connected (FC) layer, with a new FC layer with 20 outputs, followed by a *Softmax* function.

The model was initialized with pre-trained weights from the Kinetics-400 dataset [4], except for the final FC layer, which followed a standard random initialization strategy. This fine-tuning approach explores the spatio-temporal feature learning capabilities acquired from pre-training on a large-scale action recognition dataset, but now adapts the model to the specific context of Brazilian Sign Language.

3.4 Training Configuration

The training process was performed using an NVIDIA GeForce RTX 4090D GPU. The operating system was Ubuntu 24.04.2 LTS, running on a machine with an Intel® Core™ i7-11700 @ 2.50 GHz processor and 48 GB of RAM.

Fine-tuning was conducted using the Adam optimizer with a learning rate of 1×10^{-4}, selected to provide a balance between convergence speed and stability. The model was trained for up to 100 epochs, using *cross-entropy loss*, which is well-suited for multi-class classification tasks involving one-hot encoded labels. Early stopping was applied with a patience of 10 epochs, monitoring validation loss to prevent overfitting and ensure better generalization.

The *batch size* was set to 8, considering both memory constraints and convergence behavior on our hardware configuration. Gradients were backpropagated through the entire network, with all layers fine-tuned rather than freezing earlier convolutional blocks.

4 EXPERIMENTAL RESULTS

This section reports both the training/validation progress and the final evaluation of the model on the unseen test set. The same hardware configuration described in the training phase (Section 3.4) was used in all reported experiments.

4.1 Training and Validation

Training and validation of the fine-tuning process was monitored on the MINDS-Libras dataset. We monitored the model performance over epochs using both loss and accuracy metrics on the training and validation sets. The goal was to ensure that the model not only learned to fit the training data but also generalized well to unseen validation samples, minimizing overfitting.

Figure 2 illustrates the accuracy of the fine-tuned I3D model on the training and validation sets over 61 epochs. The model demonstrates rapid learning, with both training and validation accuracies converging to near-optimal values within the first 15 epochs. The training accuracy (blue line) quickly approaches 100% and remains around this value, indicating the model had sufficient capacity to fit the training data. The validation accuracy (red line) largely mirrors this high performance, but exhibits several sharp, transient drops. These momentary decreases are likely attributable to the model's performance on specific, challenging batches within the validation set, with accuracy recovering swiftly in subsequent epochs, which underscores the model's overall strong ability to generalize. The training process was automatically concluded at epoch 61 following the early stopping criterion with a patience of 10, as the validation loss did not improve upon the minimum value achieved at epoch 51.

4.2 Evaluation on Unseen Signers

As previously mentioned, the test was performed using the LOPO strategy with signers #11 and #12 (unseen in the training/validation process), comprising 200 samples. Table 1 presents the detailed classification performance of the model on the test set, including precision, recall, and F1-score for each individual class. These results enable a deeper analysis of the model's behavior, particularly for identifying misclassifications and understanding class-specific performance patterns.

The results demonstrate that the model achieved excellent performance on most signs, with 12 out of 20 classes reaching perfect F1-scores (1.0000). These signs were always correctly predicted and never misclassified, nor were other signs incorrectly labeled as them. Signs W3–W7, W9–W12, W15, W17 and W20 achieved this perfect performance.

In contrast, sign W16 exhibited the lowest precision, primarily due to a high false positive rate, as the model frequently misclassified instances of sign W14 as W16. This confusion also negatively

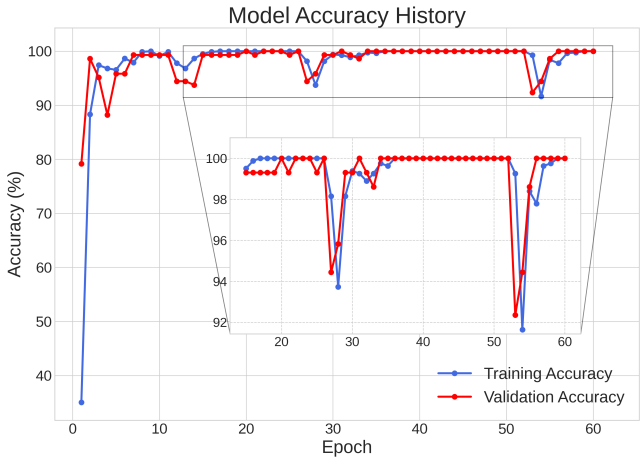

Figure 2: Training (blue) and validation (red) accuracy across 61 epochs. The inset highlights the convergence behavior during the final training phase.

Table 1: Per-class classification results on the test set

Class	Precision	Recall	F1-Score
W1 (*Acontecer*/To Happen)	1.0000	0.6000	0.7500
W2 (*Aluno*/Student)	1.0000	0.5000	0.6667
W3 (*Amarelo*/Yellow)	1.0000	1.0000	1.0000
W4 (*America*/America)	1.0000	1.0000	1.0000
W5 (*Aproveitar*/To Enjoy)	1.0000	1.0000	1.0000
W6 (*Bala*/Candy)	1.0000	1.0000	1.0000
W7 (*Banco*/Bank)	1.0000	1.0000	1.0000
W8 (*Banheiro*/Bathroom)	0.9000	0.9000	0.9000
W9 (*Barulho*/Noise)	1.0000	1.0000	1.0000
W10 (*Cinco*/Five)	1.0000	1.0000	1.0000
W11 (*Conhecer*/To Know)	1.0000	1.0000	1.0000
W12 (*Espelho*/Mirror)	1.0000	1.0000	1.0000
W13 (*Esquina*/Corner)	0.7143	1.0000	0.8333
W14 (*Filho*/Son)	1.0000	0.5000	0.6667
W15 (*Maçã*/Apple)	1.0000	1.0000	1.0000
W16 (*Medo*/Fear)	0.6667	1.0000	0.8000
W17 (*Ruim*/Bad)	1.0000	1.0000	1.0000
W18 (*Sapo*/Frog)	0.9091	1.0000	0.9524
W19 (*Vacina*/Vaccine)	0.7143	1.0000	0.8333
W20 (*Vontade*/Will)	1.0000	1.0000	1.0000
Average	**0.9452**	**0.9250**	**0.9201**

impacted the recall of W14, which was not correctly predicted in all of its occurrences. This misclassification can be attributed to the visual similarity between the signs W14 (*Filho*/Son) and W16 (*Medo*/Fear), as illustrated in Fig. 3. This similarity can be further analyzed through the lens of the phonological parameters of sign languages, which include hand configuration, location, movement, palm orientation, and non-manual markers (e.g., facial expressions and movements of the head and torso) [17]. In the case of W14 and W16, the visual ambiguity is exacerbated because both signs

Figure 3: Five frames of Signer #1 performing the signs (a) W14 (*Filho*/Son) and (b) W16 (*Medo*/Fear).

share the same location on the body and the same palm orientation (facing the signer). This inward-facing orientation obscures the precise hand configuration and finger movements from a direct frontal view, depriving the model of key distinctive features. The primary contrast between the two signs, therefore, arises from non-manual markers; the expression of "fear" (Medo) relies heavily on facial cues, which are not essential for the sign "son" (Filho) [17]. However, it is uncertain whether a dataset of this scale provides sufficient examples for a model to learn to distinguish signs based solely on such subtle facial expressions with high accuracy.

In addition to the confusion between W14 and W16, other signs also exhibited lower recall values, which reduced their F1-scores despite perfect precision. For instance, the lower recall for W1 (0.6000) was due to the model frequently predicting its instances as W13. Similarly, in the case of W2 (0.5000), it was often misclassified as W7 and W18.

Overall, the model achieved an average precision of 0.9452, an average recall of 0.9250 and an average F1-score of 0.9201. The overall accuracy of the model was 92.5%, demonstrating strong generalization to unseen signers while also highlighting opportunities for targeted improvement in specific classes.

4.3 Comparison With Related Work

Table 2 presents a comparison of our proposed approach against related works using the MINDS-Libras [13] dataset. While [14] and [1] reported slightly higher accuracies of 93.3% and 93% respectively, our model's accuracy of 92.5% is highly competitive, especially when considering our methodological advantages and the required inference complexity. Specifically, our strict signer-independent evaluation provides a more realistic measure of generalization than the "per-sign" split used in [14]. Furthermore, our I3D-based approach operates directly on raw pixels, offering a more efficient pipeline by circumventing the computational overhead from pre-processing steps like the skeleton extraction used in [1].

Although [6] and [14] do not provide complexity metrics, [1] reports a pre-processing time of about 36,000 ms per sign on an NVIDIA GeForce RTX 4070 GPU. In comparison, our solution is significantly more efficient, requiring only 824 ms per sign on an NVIDIA GeForce RTX 4090D GPU. For non-accelerated hardware, we evaluated our pipeline on an Intel® Core™ i7-11700 CPU processor and 48 GB of RAM, where total processing time was 2,642 ms.

Table 2: Comparison with works using the MINDS-Libras dataset

Work	Accuracy (%)	Total End-to-End Time
Fanucchi et al. [6]	84.00	Not Available
Rezende et al. [14]	93.30	Not Available
Alves et al. [1]	93.00	~ 36,000 ms*
Ours (GPU)	**92.50**	**~ 824 ms**
Ours (CPU)	**92.50**	**~ 2,642 ms**

**Reported pre-processing time (landmark extraction).*

Given that each sign accounts for approximately 4,000 seconds of video in the MINDS-Libras dataset, our solution is capable of real-time processing for each 4-second video sample, even when running on a CPU. The performance gap – our CPU-only execution being over 13 times faster than the competing method – demonstrates the practicality of our end-to-end approach for LIBRAS recognition systems.

5 CONCLUDING REMARKS

This study addressed the critical challenge of developing a practical and generalizable system for Brazilian Sign Language (LIBRAS) recognition. By fine-tuning a pre-trained I3D architecture directly on RGB videos from the MINDS-Libras dataset, we achieved a competitive accuracy of 92.5% on a completely unseen test set. The trained model is able to recognize LIBRAS signs in 824 ms on average, which enables real-time operation in practical scenarios.

The first contribution of this work lies in its rigorous, signer-independent evaluation protocol, ensuring no overlap between test set the training/validation data. This approach eliminates any chances of data leakage and provides a true measure of the model's ability to generalize. Secondly, our end-to-end methodology introduces a lightweight pipeline that contrasts with other state-of-the-art methods relying on computationally expensive models or pre-processing steps like vision transformers and skeleton extraction. Our results show that high accuracy can be achieved without these heavy pre-processing frameworks.

As future work, we intend to extend this approach to continuous sign language recognition and test its scalability on larger, more diverse datasets with greater environmental variability. Additionally, we also intend to explore model optimization techniques, such as quantization and pruning, further reducing the computational footprint of the I3D model, enhancing its deployability on low-cost edge devices, such as smartphones and tablets.

ACKNOWLEDGMENTS

The authors of this work would like to thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) Finance Code 001, CNPq, and FAPERGS for funding this research.

REFERENCES

[1] Carlos Eduardo G. R. Alves, Francisco De A. Boldt, and Thiago M. Paixão. 2024. Enhancing Brazilian Sign Language Recognition Through Skeleton Image Representation. In *2024 37th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. 1–6. https://doi.org/10.1109/SIBGRAPI62404.2024.10716301

[2] Brasil. 2021. Lei n.º 14.191, de 3 de agosto de 2021. Diário Oficial da União. Altera a Lei de Diretrizes e Bases da Educação Nacional para dispor sobre educação bilíngue de surdos. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/L14191.htm.

[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2021. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields . *IEEE Transactions on Pattern Analysis & Machine Intelligence* 43, 01 (Jan. 2021), 172–186. https://doi.org/10.1109/TPAMI.2019.2929257

[4] João Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4724–4733. https://doi.org/10.1109/CVPR.2017.502

[5] David Vinicius da Silva, Valter Estevam, and David Menotti. 2024. Less is More: Concatenating Videos for Sign Language Translation from a Small Set of Signs. In *2024 37th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. 1–6. https://doi.org/10.1109/SIBGRAPI62404.2024.10716311

[6] Rodrigo Zempulski Fanucchi, Arlindo Rodrigues Galvão Junior, Gabriel da Mata Marques, Lucas Brandão Rodrigues, Anderson da Silva Soares, and Telma Woerle Lima Soares. 2024. Fine-Tuning a Video Masked Autoencoder to Develop an Augmented Reality Application for Brazilian Sign Language Interpretation. In *Proceedings of the 26th Symposium on Virtual and Augmented Reality*. 275–278.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[8] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. arXiv:1705.06950 [cs.CV] https://arxiv.org/abs/1705.06950

[9] Ministério da Educação do Brasil. 2023. Dia Nacional do Surdo destaca importância da inclusão e da acessibilidade. https://portal.mec.gov.br/component/tags/tag/33784. Estimates that about 5% of the Brazilian population has some form of hearing impairment..

[10] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems* 32 (2019). https://papers.nips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

[11] Presidência da República do Brasil. 2002. Lei nº 10.436, de 24 de abril de 2002. http://www.planalto.gov.br/ccivil_03/leis/2002/l10436.htm. Recognizes the Brazilian Sign Language (Libras) as a legal means of communication and expression..

[12] Presidência da República do Brasil. 2005. Decreto nº 5.626, de 22 de dezembro de 2005. http://www.planalto.gov.br/ccivil_03/_Ato2004-2006/2005/Decreto/D5626.htm. Regulates Lei nº 10.436/2002 and mandates the provision of accessibility for deaf people in education, media, and public services..

[13] Tamires Martins Rezende. 2021. *Reconhecimento automático de sinais da Libras: desenvolvimento da base de dados MINDS-Libras e modelos de redes convolucionais*. Ph. D. Dissertation. Universidade Federal de Minas Gerais.

[14] Tamires Martins Rezende, Sílvia Grasiella Moreira Almeida, and Frederico Gadelha Guimarães. 2021. Development and validation of a Brazilian sign language database for human gesture recognition. *Neural Computing and Applications* 33, 16 (2021), 10449–10467.

[15] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems* 35 (2022), 10078–10093.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)*. 5998–6008.

[17] Andre Xavier. 2014. *UMA OU DUAS? EIS A QUESTÃO!: UM ESTUDO DO PARÂMETRO NÚMERO DE MÃOS NA PRODUÇÃO DE SINAIS DA LÍNGUA BRASILEIRA DE SINAIS (LIBRAS)*. Ph. D. Dissertation. https://doi.org/10.13140/2.1.3136.1922

[18] Mingjie Xu, Zhenyu Huang, Hailong Shen, Fei Wu, Zhaoyang Chen, Xudong Zhao, and Xiaofei Wang. 2023. A Survey on Efficient Transformer Architectures and Applications on Edge Devices. *ACM Computing Surveys (CSUR)* 56, 3 (2023), 1–36. https://doi.org/10.1145/3582043