Sign Language Translation

Final Project Report

Mahdi Ranjbar
mahdi.ranjbar@ens-paris-saclay.fr

January 30, 2024

Abstract

This research project focuses on the translation of continuous sign language into spoken language through video interpretation. It addresses the multifaceted challenges associated with this task, which include the multi-modal nature of sign language, its complexity and diverse expressions, and the need for advanced transformer encoder designs that necessitate resource-intensive training, compounded by the usual scarcity of adequate training data. The study encompasses several key activities: (1) an assessment of the existing checkpoint, (2) the development and training of a new model from scratch, (3) an exploration of the model's effectiveness in adapting from I3D to SWIN extracted video features, and (4) the introduction and detailed examination of novel evaluation metrics, specifically ROUGE and BERT-score, to assess the model's performance.

1. Introduction

Sign Languages are the native languages of the Deaf and their main medium of communication. As visual languages, they utilize multiple complementary channels to convey information. This includes manual features, such as hand shape, movement and pose as well as non-manuals features, such as facial expression, mouth and movement of the head, shoulders and torso. Sign language translation (SLT) is the task of translating continuous sign language videos into spoken language sentences.
Generating spoken language sentences given sign language videos is therefore a spatio-temporal machine translation task. Such a translation system requires us to accomplish several sub-tasks, which are currently unsolved:

Sign Segmentation: Firstly, the system needs to detect sign sentences, which are commonly formed using topic-comment structures, from continuous sign language videos. This is trivial to achieve for text based machine translation tasks, where the models can use punctuation marks to separate sentences. Speech-based recognition and translation systems, on the other hand, look for pauses, e.g. silent regions, between phonemes to segment spoken language utterances. However so far, there is no study which uses sign segmentation for realizing continuous sign language translation.

Sign Language Recognition and Understanding: Following successful segmentation, the system needs to understand what information is being conveyed within a sign sentence. From a computer vision perspective, this is the most challenging task. Considering the input of the system is high dimensional spatio-temporal data, i.e. sign videos, models are required that understand what a signer looks like and how they interact and move within their 3D signing space. Moreover, the model needs to comprehend what these aspects mean in combination.

Sign Language Translation: Once the information embedded in the sign sentences is understood by the system, the final step is to generate spoken language sentences. SLT is a challenging task as the grammar of sign and spoken languages are very different. These differences include: different word ordering, multiple channels used to convey concurrent information and the use of direction and space to convey the relationships between objects. Put simply, the mapping between speech and sign is complex and there is no simple word-to-sign mapping. As such, this problem truly represents a machine translation task.

In this work, we focus on Sign Language Translation. A recent study by Larre's et al. [14] introduced a pioneering baseline model[1], using a standard transformer encoder-decoder architecture. This model was trained on

[1] Project code-base: https://github.com/imatge-upc/slt_how2sign_wicv2023

the comprehensive How2Sign dataset [5], encompassing 80 hours of video covering 10 distinct topics. Their model achieved a BLEU score of 8.03, indicating significant potential for further enhancement. Crucially, their approach leveraged video features extracted from an inflated 3D convolutional neural network (I3D), initially pre-trained on ImageNet and subsequently fine-tuned on the Kinetics-300 [7] and How2Sign datasets. However, the current trend in the vision community is shifting from CNNs to Transformer models. These pure Transformer architectures have demonstrated superior accuracy across major video recognition benchmarks, including the Kinetics-300 dataset. Therefore, substituting the I3D video features with Transformer video features appears to be a promising avenue for boosting performance. Additionally, the study highlighted the criticality of selecting appropriate metrics for evaluation. Some metrics might present misleading interpretations of model performance. Thus, integrating a variety of measures is recommended for a more accurate assessment.

2. Related Work

Sign language video understanding has been addressed from a variety of tasks: sign language recognition (SLR) over isolated or continuous signs, sign language translation (SLT), sign language production (SLP) [13] or retrieval. This work focuses on sign language translation.

Gloss-based SLT [2] uses an intermediate textual representation between the input video sequence and the output text. These tokens are named glosses. Glosses are a type of transcription of sign languages that must be produced by trained sign language linguists and that are available in some SLT datasets. Glosses provide supervision that helps models in their training, but their acquisition is also very time-consuming and expensive because of the scarcity of annotators.

On the other hand, gloss-free SLT addresses the raw task of converting the video into text, without any intermediate gloss. In this project, the second approach—gloss-free SLT is considered.

SLT was initially approached with rule-based systems [16] and statistical methods. Since 2018, virtually all related work has basically applied the advances in deep learning to sign language translation datasets. Given that SLT can be formulated as an input sequence of video frames that is transformed into a sequence of words, it fits perfectly in the popular sequence-to-sequence (seq2seq) formulation widely adopted by the Machine Translation field which employs an encoder-decoder architecture to transform the input sequence into the output one. This project studies specifically the work done in [14].

3. Methodology

3.1. Neural architecture

The architecture proposed by [14] is depicted in Figure 1. The input video stream is tokenized with a pre-trained I3D feature extractor. These tokens are fed into the encoding layers of the Transformer. The decoder of the Transformer operates with lowercase and tokenized textual representations. In this work video tokenization is not done and the input data is directly the provided extracted video features by previous work that will be discussed in the following sections.

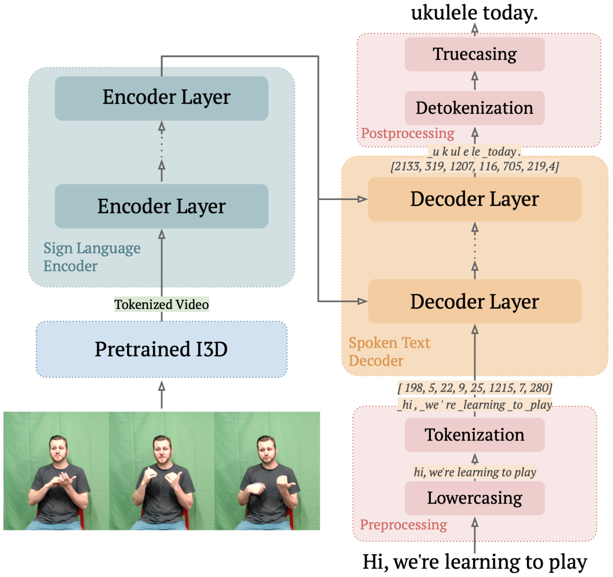

Figure 1. The input video sequence is fed into a Transformer to generate the output text sequence.

3.2. Sign Video Embedding

3.2.1 Two-Stream Inflated 3D ConvNets (I3D)

The I3D model [3] represents a significant advancement in video understanding. It adapts the architecture of 2D Convolutional Networks (ConvNets), widely used for deep image classification, into the 3D domain. This transformation is achieved by expanding the filters and pooling kernels of these ConvNets from 2D to 3D. Such an expansion allows for the effective learning of spatio-temporal features directly from videos.

As shown in Figure 2, there are two streams configuration with one I3D network trained on RGB inputs, and another on flow inputs that carries smooth flow information.

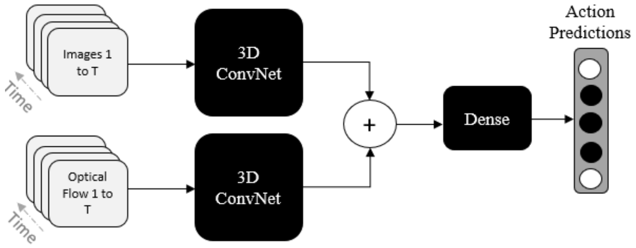

Figure 2. Two-Stream Inflated 3D ConvNets (I3D) architecture.

3.2.2 I3D Features

In [6], the authors implemented a sign recognition model to obtain sign video embeddings. Their model is a 3D convolutional neural network instantiated with an I3D architecture pretrained on BOBSL dataset [1] and fine-tuned on the How2Sign dataset [5] using their automatic sign spotting annotations. Specifically, they used the outputs corresponding to the spatio-temporally pooled vector before the last (classification) layer. This produces a 1024-dimensional real-valued vector for each 16 consecutive RGB frames. These features were extracted densely, using a stride of 1, from the How2Sign sign language sentences to obtain the sequence of sign video embeddings. In this project we use these provided features.

3.3. Metrics

3.3.1 Bleu Score

The BLEU score [11] is a key metric for assessing the quality of machine translation. It measures how much the candidate translation aligns with reference translations by comparing overlapping phrases, or n-grams. Essentially, it evaluates the translation based on two aspects: the precision (matching n-grams with the reference) and the adequacy of translation length to avoid overly short translations. A significant aspect of BLEU is the ’brevity penalty’. This component of the score penalizes translations that are too short, ensuring that the translation is not just accurate, but also complete. In summary, the BLEU score provides a balance between the accuracy of the translation (in terms of n-gram precision) and its fluency (appropriate length and completeness).

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

In this formula, BP represents the brevity penalty, and p_n is the precision of n-grams.

3.3.2 Reduced Bleu Score

The Reduced Bleu Score, or rBleu, introduces an innovative twist to the conventional BLEU score assessment. This metric involves the creation of a blacklist containing words that are common in the training data but relatively insignificant in conveying the core meaning of sentences. These typically include function words such as articles, prepositions, and pronouns – for example, ’the’, ’it’, ’at’, ’more’, among others. The rBleu metric operates by excluding these words from both the reference and predicted translations prior to calculating the BLEU score. This adjustment aims to focus the evaluation more on the meaningful content of the translations, potentially offering a more nuanced insight into the translation’s quality, particularly in terms of its substantive accuracy and relevance.

3.3.3 ROUGE Score

The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score [8] is another important metric, particularly for evaluating the quality of text summarization and machine translation. Unlike BLEU, which focuses primarily on precision, ROUGE emphasizes recall — the ability of the translation to capture as much of the content from the reference text as possible. ROUGE evaluates translations by comparing the overlap of n-grams, word sequences, and word pairs between the candidate and reference texts. The most commonly used variant, ROUGE-N, calculates the overlap of n-grams between the translation and the reference, thus measuring the extent to which the translation captures important content from the reference. ROUGE-L, another variant, focuses on the longest common subsequence, offering insights into the fluency and sentence-level structure of the translation. ROUGE scores are particularly relevant for assessing the comprehensiveness of translations, ensuring that the translated content reflects as much of the reference material as possible.

3.3.4 BERT Score

The BERT score [15] represents a more recent advancement in translation evaluation, leveraging the capabilities of the Bidirectional Encoder Representations from Transformers (BERT) language model. This metric computes the semantic similarity between the candidate and reference texts by using contextual embeddings from BERT. Unlike traditional metrics that rely on surface-level text overlap, the BERT score evaluates the degree of semantic and contextual alignment between texts. It computes the cosine similarity between the embeddings of words in the candidate translation and the reference text, capturing the deep semantic relationships in the language. This approach allows for a more nuanced and context-sensitive assessment of translation quality, especially in cases where literal n-gram overlap might be low, but the translated text is semantically appropriate and contextually accurate. The BERT score is increasingly seen as a valuable tool for evaluating translations, as it addresses some of the limitations of earlier metrics by focusing on the underlying meaning rather than just the literal text overlap.

4. Experiments

4.1. Code Set-up

The foundational framework for this project is derived from the work presented in [14]. The initial phase of the project involved a comprehensive set-up of the code. I encountered several challenges in this phase, primarily due to the disorganized state of the existing code base. There were significant mismatches between the provided documentation and the actual source code. Furthermore, issues with package compatibility added to the initial hurdles. Subsequent to the data download phase, it was observed that the dataset's address was hard-coded, necessitating manual intervention for rectification.

My approach involved a detailed examination and systematic restructuring of the code base. This process included reorganizing certain elements of the code and downgrading several packages to ensure compatibility. After these adjustments, I successfully set up a stable code foundation, ready for further experimental work.

4.2. Evaluation Using the Provided Checkpoint

A model checkpoint from a previously trained model was made available for this study. The initial phase involved replicating the results presented in the original research. This was accomplished by loading the model from the provided checkpoint and conducting evaluations on both the validation and test datasets. The outcomes of this process are presented in Table 1. A notable observation is that the performance metrics derived from the provided checkpoint are marginally lower than those reported in the original study. This discrepancy leads to the hypothesis that the checkpoint provided may not correspond precisely to the one utilized in the research publication.

4.3. Training from Scratch

Following the analysis of the initial model's checkpoint results, I started implementing a novel transformer model, building it from scratch. The architecture of this model was inspired by the original study, with its detailed architecture outlined in Table 2. The training process spanned 35 epochs, amounting in a total duration of one and a half days, roughly translating to about an hour per epoch. This duration emphasizes a computational constraint of the process.

Performance metrics, specifically the BLEU and ReducedBLEU scores, are presented in Table 3. It is noteworthy to mention that while the BLEU score aligns closely with that of the checkpoint model trained for 108 epochs, the ReducedBLEU score is considerably lower. Additionally, the progression of the BLEU and ReducedBLEU score performances along the 35 epochs is represented in Figures 3 and 4, respectively. These figures provide a graphical illustration of the model's performance metrics across the training epochs.

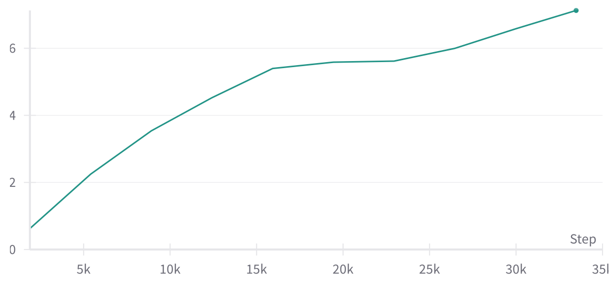

Figure 3. Bleu score on validation set along 35 epochs.

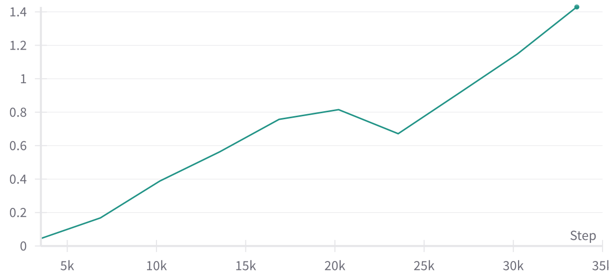

Figure 4. ReducedBleu score on validation set along 35 epochs.

4.4. Video Features Impact Analysis

4.4.1 Video Swin Transformer

Recent advancements in the field of computer vision have seen a paradigm shift from Convolutional Neural Networks (CNNs) to Transformer-based architectures. This transition was initiated by the introduction of the Vision Transformer (ViT) [4], which demonstrated that purely Transformer-based architectures could achieve state-of-the-art accuracy on major video recognition benchmarks. These video models primarily use Transformer layers to create global connections among patches in both spatial and temporal dimensions. Building on this progress, the Video Swin Transformer [10] introduces a significant enhancement by incorporating the concept of locality in video Transformers. This

	val					test				
	rBLEU	BLEU-1	BLEU-2	BLEU-3	BLEU	rBLEU	BLEU-1	BLEU-2	BLEU-3	BLEU
Reported.	2.79	35.20	20.62	13.25	8.89	2.21	34.01	19.30	12.18	8.03
Ckpt Inf.	2.73	34.80	20.45	13.18	8.84	2.06	32.98	18.79	11.85	7.77

Table 1. Performance Evaluation on the How2Sign Dataset for Sign Language Translation: The table delineates three sets of results: 'Reported' refers to the results as detailed in the original study and 'Ckpt Inf' relates to the outcomes using the model checkpoint provided in the paper.

Parameter	Value
Encoder Embed Dim	256
Encoder FFN Dim	1024
Encoder Attention Heads	4
Encoder Layers	6
Decoder Embed Dim	256
Decoder FFN Dim	1024
Decoder Attention Heads	4
Decoder Layers	3
Layer Norm Embed	True
Activation Func	Relu
Dropout	0.3

Table 2. Trained model's architecture.

Model	Bleu	rBleu
Trained Model.	7.1	1.42

Table 3. Bleu and ReducedBleu score on test set of 35 epoch trained model.

approach results in an improved balance between speed and accuracy. The Video Swin Transformer adapts the Swin Transformer [9], originally designed for image processing, for video recognition tasks. This adaptation allows for the continued utilization of pre-trained image models, thereby leveraging their power and efficiency. The detailed architecture of the Video Swin Transformer is illustrated in Figure 5.

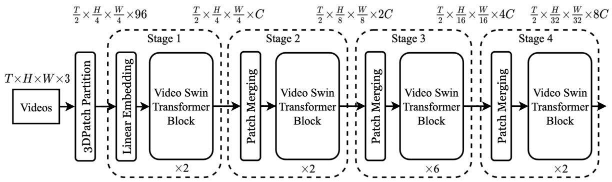
$\frac{T}{2}\times\frac{H}{4}\times\frac{W}{4}\times 96$
$\frac{T}{2}\times\frac{H}{4}\times\frac{W}{4}\times C$
$\frac{T}{2}\times\frac{H}{8}\times\frac{W}{8}\times 2C$
$\frac{T}{2}\times\frac{H}{16}\times\frac{W}{16}\times 4C$
$\frac{T}{2}\times\frac{H}{32}\times\frac{W}{32}\times 8C$
Stage 1
Stage 2
Stage 3
Stage 4
$T\times H\times W\times 3$
Videos
3DPatch Partition
Linear Embedding
Video Swin Transformer Block
Patch Merging
Video Swin Transformer Block
Patch Merging
Video Swin Transformer Block
Patch Merging
Video Swin Transformer Block
×2
×2
×6
×2

Figure 5. Overall architecture of Video Swin Transformer.

4.4.2 Substituting I3D features with Swin features

An intriguing approach involves substituting the I3D features, as discussed in Section 3.2.2, with Swin Transformer features that have been pre-trained on the BOBSL dataset, following the methodology of Prajwal et al. [2022] [12]. This substitution is aimed at evaluating the influence of different input video features on the model's performance. The acquired Swin Transformer features were initially converted to match the I3D format, which is the format expected by the base code. This conversion was essential to facilitate their integration into the existing training pipeline, enabling the training process to use these newly formatted features. More specifically, adjustments were made to the Transformer model, previously detailed in Table 2. These modifications included altering the dimensions of the linear layer that receives the input features from 1024 to 768. Subsequently, the network underwent training for a duration of 35 epochs, which approximately equated to one and a half days. The outcomes of this experiment are presented in Table 4. It is noteworthy that, when using Swin Transformer features, there is a significant decline in performance metrics, the Bleu and ReducedBleu scores, compared to the results achieved with I3D features. For instance, the Bleu score experienced a drastic reduction, falling from 7.1 to a mere 0.96. This marked decrease in performance can be attributed to the lack of specific fine-tuning of the Video Swin Transformer for the How2Sign dataset, unlike the I3D backbone which had undergone such optimization. Additionally, the progression of the BLEU and ReducedBLEU score performances along the 35 epochs is represented in Figures 6 and 7, respectively.

Model	Bleu	rBleu
I3D Model.	7.1	1.42
Swing Model.	0.96	0.003

Table 4. Bleu and ReducedBleu score on test set of 35 epoch trained models with I3D features versus Swing features.

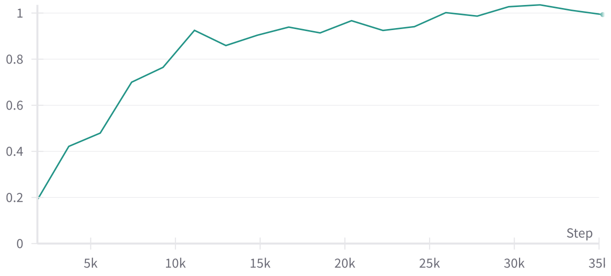

Figure 6. Bleu score on validation set along 35 epochs using Swin features.

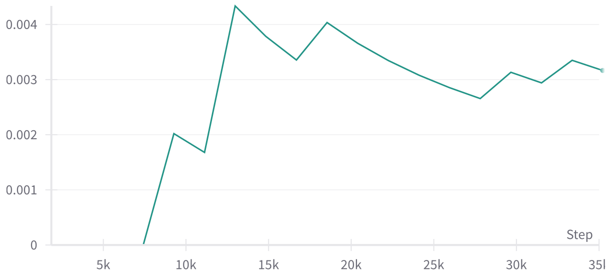

Figure 7. ReducedBleu score on validation set along 35 epochs using Swin features.

4.5. Metrics Analysis

In the original study, the authors assessed two metrics for the task of sign language translation: the BLEU (Bilingual Evaluation Understudy) score and the ReducedBleu score. In the subsequent analysis, I have also evaluated additional metrics, specifically ROUGE and BERT-score.

4.5.1 Building a Benchmark

With the integration of BERT and ROUGE scores into the evaluation framework, the model, as described in Section 4.3, was used for inference. The outcomes of this process are documented in Table 5, presenting a comprehensive benchmark for assessing the efficacy of the translation model.

Model	Bleu	rBleu	Rouge	Bert
Trained Model.	7.1	1.42	0.29	0.87

Table 5. Evaluation of trained model on I3D features. Rouge refers to RougeL and Bert to average precision.

4.5.2 Qualitative Results

In scenarios where a prediction achieves perfection, that is, when it aligns identically with the reference sentence, the metrics—namely Bleu, rBleu, Rouge, and Bert attain their maximal values, being 100, 100, 1, and 1 respectively. To visually demonstrate the qualitative impact exerted by these various metrics, a set of four sentences was randomly selected and is presented in Table 6. It's important to highlight a specific limitation in the computation of the Bleu score. When the sentence in question is reduced to less than four words, the ReducedBleu score automatically drops to zero, regardless of whether the words match perfectly. Additionally, the analysis shows that while shorter sentences tend to translate well, translating longer sentences poses a greater challenge.

5. Discussion

The experiments carried out have yielded enlightening results. As mentioned in Section 4.4.2, exploring the impact of fine-tuning the Video Swin Transformer using the How2Sign dataset to extract video features would have been intriguing. There is a necessity to replicate the outcomes with fine-tuned features to comprehend the significance of this fine-tuning; however, this entails a computationally intensive training process. This scenario presents an avenue for future research endeavors.

Additionally, the qualitative analysis demonstrated that a model trained for merely 35 epochs is capable of generating significant translations. It was also observed that the ReducedBleu score, as suggested by the originating study, may not be the most appropriate metric, particularly due to its limitations in evaluating short sentences. In future investigations, the Bert-score, which considers the underlying semantics rather than just literal textual overlap, could be a more suitable alternative.

6. Conclusion

This research project represents a step forward in the field of sign language translation using video interpretation. The study successfully navigates through various challenges, from setting up a stable code base, evaluating the given checkpoint, training a new model from scratch, the adaptation of advanced video feature extraction methods, to the incorporation of novel evaluation metrics. Key conclusions drawn from this study are:

Adaptation of Transformer Models: While the adoption of Transformer-based models like the Video Swin Transformer presents a promising direction, the study demonstrates the necessity for specific fine-tuning and adaptation to specialized datasets. The contrast in performance be-

Sentences	BLEU	rBLEU	ROUGE	BERT
Reference: you can also **step** back. —— **Prediction:** we're also going to **step** back.	27.48	0.00	0.42	0.95
Reference: let's relax them on your **legs**. —— **Prediction: let's** be **breathing** on the **legs**.	15.61	0.00	0.57	0.92
Reference: you can use **red peppers** if you like to get a little bit **color** in your **omelet**. —— **Prediction:** you can also have a **red pepper**, you want to **add** a **color**.	6.07	18.99	0.61	0.90
Reference: so, what we're going to **start** is our **feet facing straight ahead**, **knees bent**, **belly button pulled** in, **chest lifted**, **shoulders** back and down, so as you **step forward**, you want to **shift** all your **weight forward**. —— **Prediction:** what we're going to do is **lift** your **feet forward**, **knees forward**, **knees** are going to **lift** your **knees** up, **chest**, **chest**, **chest**, **chest**, **knees** are going to want all the **weight**.	9.37	3.36	0.32	0.88

Table 6. Qualitative examples of the trained model on test set. In bold the words remaining to compute rBLEU.

tween I3D and Swin Transformer features highlights the importance of dataset-specific optimizations.

Importance of Evaluation Metrics: The integration of ROUGE and BERT-score alongside metrics like BLEU and ReducedBleu offers a more holistic approach to evaluating translation models. This approach acknowledges the complexity of language translation and the need for diverse metrics to fully capture model performance.

References

[1] Samuel Albanie et al. "Bbc-oxford british sign language dataset". In: *arXiv preprint arXiv:2111.03635* (2021).

[2] Necati Cihan Camgoz et al. "Neural Sign Language Translation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018.

[3] Joao Carreira and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset". In: *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6299–6308.

[4] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[5] Amanda Duarte et al. "How2sign: a large-scale multimodal dataset for continuous american sign language". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 2735–2744.

[6] Amanda Duarte et al. "Sign language video retrieval with free-form textual queries". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 14094–14104.

[7] Will Kay et al. "The kinetics human action video dataset". In: *arXiv preprint arXiv:1705.06950* (2017).

[8] Chin-Yew Lin. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81. URL: `https://aclanthology.org/W04-1013`.

[9] Ze Liu et al. "Swin transformer: Hierarchical vision transformer using shifted windows". In: *Proceedings*

of the IEEE/CVF international conference on computer vision. 2021, pp. 10012–10022.

[10] Ze Liu et al. "Video swin transformer". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 3202–3211.

[11] Kishore Papineni et al. "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Ed. by Pierre Isabelle, Eugene Charniak, and Dekang Lin. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://aclanthology.org/P02-1040.

[12] KR Prajwal et al. "Weakly-supervised Fingerspelling Recognition in British Sign Language Videos". In: *arXiv preprint arXiv:2211.08954* (2022).

[13] Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. "Everybody Sign Now: Translating Spoken Language to Photo Realistic Sign Language Video". In: *CoRR* abs/2011.09846 (2020). arXiv: 2011.09846. URL: https://arxiv.org/abs/2011.09846.

[14] Laia Tarrés et al. "Sign Language Translation from Instructional Videos". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 5624–5634.

[15] Tianyi Zhang et al. "Bertscore: Evaluating text generation with bert". In: *arXiv preprint arXiv:1904.09675* (2019).

[16] Liwei Zhao et al. "A machine translation system from English to American Sign Language". In: *Proceedings of the Fourth Conference of the Association for Machine Translation in the Americas: Technical Papers*. Ed. by John S. White. Cuernavaca, Mexico: Springer, Oct. 2000, pp. 54–67. URL: https://link.springer.com/chapter/10.1007/3-540-39965-8_6.