# Research Proposal - SL Translation

Daniela Winter
*Computer Vision Master Project - Dr. Ehsan Yaghoubi*
*Computer Science Department*
University of Hamburg, Germany
daniela.winter@studium.uni-hamburg.de

*Abstract*—**Sign Language Translation (SLT) using machine translation techniques is an important help in communication for deaf and hard-of-hearing people. Sign Language (SL) is often their most important means of communication. But hearing people seldom have the required language skills to understand and communicate in SL. By enabling fluent communication, SLT applications break through a great obstacle in everyday life for both sides of the interaction. One of the main challenges in current SLT research includes achieving meaningful encodings of sign videos without training the models with gloss supervision. This is highly useful for two different reasons. The first is that common textual gloss representations come with a great lack of information since they fail to capture all details of the visual sign. The second reason is that datasets with gloss transcriptions are very expensive compared to datasets that only contain SL videos and spoken language translations.**

**The current works done in SLT with Transformers use a CNN as a backbone for feature extraction. Previous research however has shown that using a Pure Transformer architecture, that does not rely on the embeddings from a CNN, results in better performances. Therefore, in this work, we propose an SLT architecture that consists of a Pure Transformer that directly uses video frames without the need for a CNN. For this, we are going to adapt the current standard Transformer architecture for SLT and combine it with the approach of the Video Swin Transformer. The decoder part of the SLT Transformer will be kept unchanged while the encoder will be partly replaced by the Video Swin Transformer model. This Pure Video Transformer captures the biases of visual data especially well and is therefore a good fit for the SLT task. The model will be evaluated using the smaller but commonly used dataset PHOENIX-Weather 2014T and the bigger, more recent DGS Corpus dataset. The results will be compared to the end-to-end SLT Transformer introduced in 2020 that has been used as a baseline since then.**

## I. Introduction

### A. Motivation and Challenges

Sign Languages (SL) are an important means of communication for deaf and hard-of-hearing people around the world. Same as with spoken languages, communication between languages is a great challenge, may it be different SLs or SL and spoken language.

Machine learning techniques help create applications to simplify communication between languages. Current Neural Machine Translation (NMT) systems using Transformer networks make striking progress. NMT for SLs on the other hand still lags. Video–to–text Sign Language Translation (SLT) achieves way lower BLEU scores than text-to-text translations of spoken languages [Camgöz et al., 2020b] [Zhu et al., 2020]. While the idea of translating between languages is the same, the task of SLT differs a lot from NMT for spoken languages. As SLs are visual languages, their translation poses challenges from both the NLP and CV domains. NMT of SLs has to deal with visual and textual data and with visuospatial linguistic principles that are unique in SL [Crasborn, 2015]. In video-to-text translation, the encoder part of the SLT Transformer deals with the visual aspects of the language. The effective encoding of visuospatial principles is crucial for the preservation of the passed information - and therefore plays a big role in trying to reach current levels of spoken language NMT applications. A very popular implementation of a video-to-text SLT Transformer was introduced by Camgöz et al. in 2020 [Camgöz et al., 2020b]. This model has since been used by other researchers trying to improve the performance with changes that mostly keep the same architecture [Camgöz et al., 2020a] [Voskou et al., 2021]. That architecture is a standard NMT Transformer with the difference that it uses CNN embeddings of the video frames as input instead of word embeddings. Recent research, however, has shown that Pure Video Transformer architectures, that do not rely on features extracted by CNNs, can reach a better performance when enough data is available [Arnab et al., 2021] [Liu et al., 2022].

This brings up another weakness of the current research. There is almost no research done with much bigger datasets than PHOENIX-Weather 2014T [Camgöz et al., 2018] which is considered tiny in comparison to other NLP datasets [De Coster et al., 2021]. However, bigger datasets are available now in some SLs [Camgöz et al., 2021] [Schulder et al., 2020]. One of the reasons why NMT of spoken languages is so successful is because the networks can be trained on a huge amount of textual data. It is not far-fetched to assume that more data would greatly improve the performance of SLTs as well. The creation of SL datasets, however, is very expensive. This is mostly due to the necessity of creating glosses, a form of textual representation of SLs. Datasets without glosses can be created fast and inexpensively since signers can translate spoken language in real-time. Gloss data is currently needed for training most better-performing SLT models since it provides guidance for how the Transformer needs to encode the visual data [Camgöz et al., 2020b]. Therefore, it can be assumed that if the encoding is more efficient, gloss data would not be as crucial.

### B. Objectives and Contributions

In this research, we are aiming to improve the video-to-text SLT performance, to help produce better SL to spoken language translators. In order to achieve this, we are planning to improve the encoding of the visual data to make up for the backlog in the translation of those visual languages. For our new model, we propose a change of the encoder part of the SLT Transformer [Camgöz et al., 2020b]. Instead of using CNNs to produce embeddings that can then be input in a standard NMT Transformer, we are going to realize a model that uses a Pure Video Transformer as encoder. The so-called Video Swin Transformer has proven to be a successful Pure Video Transformer [Liu et al., 2022]. It combines the mechanism of self-attention with the visuospatial properties of a CNN. Instead of applying self-attention globally, it applies it only locally in spatiotemporal neighborhoods of the video. This induces a beneficial local bias, that captures a property of visual data. Pixels closer in space and time correlate more with each other than farther ones. This different way of encoding the SL videos could increase the quality of the encoding of special SL features and therefore the overall performance of SLT.

Moreover, we want to examine the performance of our new model as well as the current standard model [Camgöz et al., 2020b] on a bigger dataset [Schulder et al., 2020] to see how much of the current performance weaknesses are due to the usage of too small datasets. The bigger dataset and the different way of encoding the sign videos might allow it to forgo the need of using glosses for training the encoder and achieve better performances than previous rather weakly-performing no-gloss SLT models [Camgöz et al., 2020a] [Voskou et al., 2021].

### C. Evaluation

We will evaluate our new model on the widely used PHOENIX-Weather 2014T dataset [Camgöz et al., 2018] to allow comparability with previous models as well as on the bigger DGS Corpus dataset [Schulder et al., 2020]. Evaluation will be done using the BLEU, ROUGE, and METEOR scores which are standard metrics in NLP tasks. We will compare the performance of our model with the baseline model of Camgöz et al. [Camgöz et al., 2020b]. For this, we will test both models on differently sized datasets as well as by training them with and without gloss supervision.

## II. Related Work

### A. SLT Transformers

The first implementation of a Transformer for end-to-end SLT has been published in 2020 [Camgöz et al., 2020b]. The main idea of this work was to use a Transformer to perform end-to-end SLT from video data to text data without the intermediate step over gloss representations, since enforcing that step might be a bottleneck for the overall performance. However, glosses are still used as supervision for the training of the Transformer – without them, the performance dropped. This implementation performed way better than earlier models without attention. But the BLEU score still lies below spoken language NMT levels and the dependency on gloss datasets limited the training to smaller datasets.

To address this issue, in their next work, the authors changed the previous architecture to using multiple input channels for the encoder according to the different SL articulators [Camgöz et al., 2020a]. With this idea, they were able to achieve better performance for training without gloss supervision than in their previous work. However, that performance is still worse than when glosses are used and the qualitative results show that often wrong information is translated. Another work that tried to adapt the original end-to-end SLT Transformer to get better performances without the need for glosses was presented by Voskou et al. [Voskou et al., 2021]. They changed the standard ReLU units of the network with units that use a stochastical local winner-takes-all principle. While they perform better than previous approaches without gloss supervision, the qualitative results are still only acceptable. In some cases don't even follow a good grammatical structure, and sometimes wrong information is translated. While the grammar could be improved by pre-training the models with spoken language data, the wrong translation of words, like for example mix-ups of weekdays, could be due to deficient quality of the encodings. A common property of all the mentioned models is that they use CNNs for feature extraction. Those features are then fed into the encoder part of the network. Improving the encoding of the signs in the videos could help ensure correct information transfer.

### B. Pure Video Transformers

In trying to follow the success of Transformers in NLP, different approaches have been introduced to apply Transformers directly to visual data. First for image data [Dosovitskiy et al., 2021] [Liu et al., 2021] and later adapted for video input as well. The first successful of those implementations for videos is the ViViT (Video Vision Transformer) which is based on the ViT (Vision Transformer) [Arnab et al., 2021]. Visual data has specific properties that are not necessarily found in text data, such as spatial locality and translation invariance. The encoding of the data needs to preserve those properties to allow meaningful interpretation of the data. To induce the visual bias of spatial locality, ViViT, similar to 3D CNNs, splits the data into 3D blocks. Since pixels of the video that are close to each other in space or time correlate more with each other than farther pixels, combining them preserves the spatial locality property. Each of those 3D blocks performs as a token for the input of the Transformer encoder. However, these blocks are of fixed size, which ignores that visual elements can be of very differing sizes. Moreover, translational invariance is not included in this model.

To deal with this shortcoming, the Video Swin Transformer uses hierarchical feature maps by merging 3D blocks in later layers [Liu et al., 2022]. In addition to that, the Swin Transformer includes connections across blocks by alternating between two overlapping blocks in consecutive layers. Self-attention is then applied within the blocks. With this strategy, the Video Swin Transformer captures the biases of visual data better and results in a better performance on image classification tasks.

Generally, both Pure Video Transformers show that their performance exceeds the levels of CNN-based models if enough data is available. Pure Video Transformers have in common that the attention mechanism of Transformers is directly applied to the data instead of on embeddings of the frames that were obtained by CNNs. This different way of encoding the data seems to be beneficial for standard Computer Vision tasks such as image classification, semantic segmentation [Liu et al., 2022], and video captioning. As a video-to-text task, video captioning is similar to SLT. For this task, Pure Video Transformers were successfully applied using a video Swin Transformer as encoder for the input data and a BERT model as a decoder to produce an output sentence [Lin et al., 2022].

## III. Proposed Method

### A. Overview

Following the success of Pure Video Transformers, we propose combining the above-mentioned approaches, the SLT Transformer with the Video Swin Transformer (see Figure 1). We will use this architecture to perform end-to-end video-to-text SLT. The proposed architecture is made up of three main parts: the Video Swin Transformer and the gloss prediction part on the encoder side of the network and the decoder part from the standard SLT Transformer.

The Video Swin Transformer which takes the input and uses several Transformer blocks to produce a spatiotemporal representation of the data in which the visual properties are contained. These encodings are then passed to the gloss prediction part of the network as well as to the decoder.

The gloss prediction is an optional part of the network that can be included depending on the experiment setup. Based on the video encodings it predicts a gloss transcription of this video. This can be used to compute a gloss loss which serves as optional intermediate supervision of the encoder.

The decoder uses the previously predicted words as well as the encodings it receives from the encoder to predict the spoken language words that translate the sign language of the input video. In the following, the three parts of the architecture are explained in more detail.

### B. Encoder

Since we follow the architecture of the Video Swin Transformer for our encoder network, the raw input video frames will be used as direct input for the encoder. The Video Swin Transformer [Liu et al., 2022] takes the input video of the size $T \times H \times W \times 3$ (T: number of frames) and splits it into 3D blocks of size 2x4x4x3 that are flattened. Each of them counts as an input token. The number of tokens hence depends on the size of the video. Each 3D token is a 96-dimensional vector. But a linear embedding layer allows transferring the input to a variable dimension size C. The Video Swin Transformer consists of several stages. The stages represent the hierarchy of the processing scale. In each stage, in the patch merging layer, the patch dimension is changed by 2x spatial downsampling. Spatially neighboring patches are grouped by 2x2 and the concatenated features are then downsampled to half their dimension size. This results in a final dimension size of 8C after the last block of the encoder network (see Figure 1). In each stage Video Swin Transformer blocks are applied multiple times (see Figure 1) [Liu et al., 2022]. This module includes multi-head self-attention that incorporates the 3D

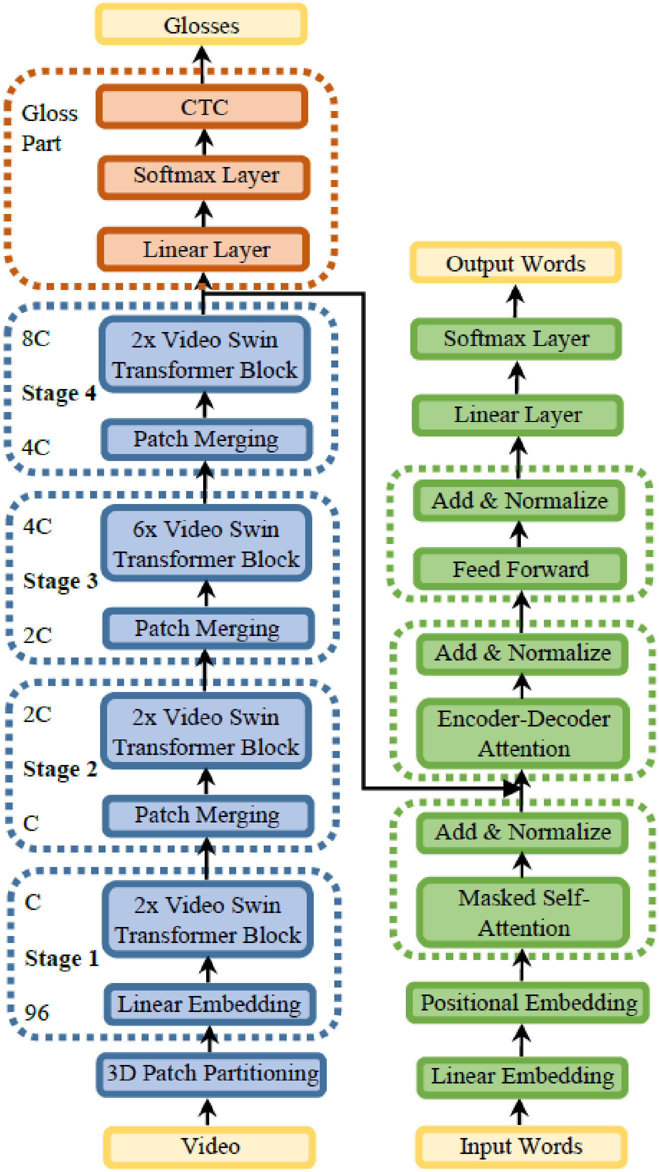


Fig. 1: This is our proposed model. We follow the architecture of the Video Swin Transformer [Liu et al., 2022] and the SLT Transformer [Camgöz et al., 2020b]. The blue parts depict the Video Swin Transformer encoder, the red parts the gloss computation, the green parts the Transformer decoder and the yellow parts the inputs/outputs. In the encoder, the different stages show the input and output dimensions of the features.

shifted window approach of the Swin Transformer to enable cross-patch connections [Liu et al., 2021]. The outputs of the encoder network are spatiotemporal encodings that are forwarded to the decoder network.

### C. *Gloss Prediction*

The above-mentioned encodings can optionally also be used to predict glosses. A linear layer followed by a softmax layer predict possible words for the gloss transcription. Using those predictions, the Connectional Temporal Classification (CTC) layer predicts alignments of the gloss transcription sequence to the parts of the video. The CTC layer tackles the alignment difficulty of having a long input sequence (video frames) and a relatively short output sequence (gloss transcriptions). It produces different possible gloss alignments with a probability assigned to them. The loss of the CTC layer is calculated as the sum of the possible gloss alignment probabilities [PyTorch Contributors, 2022]. It can be used to additionally guide the training of the encoder, not only with the loss from the output of the decoder. With the gloss supervision, the training of the encoder is focused on learning representations that are closer to glosses. This can help steer the training in the right direction. However, it needs to be mentioned that glosses lack important information that is contained in the sign videos [Camgöz et al., 2020b]. Therefore, we will analyze how the performance of the network will be without the usage of gloss supervision for training.

### D. *Decoder*

For our decoder, we follow the implementation used for the SLT Transformer by Camgöz et al. [Camgöz et al., 2020b]. Before inputting them into the decoder network, the one-hot encoded input words are converted into a denser representation by the word embedding layer. Positional encoding is then added to keep track of the order of the words in the sentence. After that, we apply a self-attention layer that masks the inputs during training so that only previous inputs and not later ones are known by the network when trying to predict the next word. The resulting representation of the spoken language context and the sign video representations yielded from the encoder are then combined for the encoder-decoder attention layer. Then, we use the outputs from that attention layer to pass through a non-linear feed forward layer. Each block of the decoder has residual connections and normalization applied after the layers. Finally, a linear layer and a softmax layer are used to predict the probability distributions of the output words over the vocabulary. This is done word by word until the end-of-sentence token is reached.

For the training, we compute a cross-entropy loss from the predicted words. The complete network can be trained with a combination of gloss and word prediction loss or only with the word prediction loss.

## IV. Evaluation and Comparison

### A. *Comparison to Baseline*

We are going to compare our new model's performance with the performance of the state-of-the-art SLT Transformer by Camgöz et al. [Camgöz et al., 2020b] as a baseline. For both models, performances for the state-of-the-art dataset PHOENIX-Weather 2014T [Camgöz et al., 2018] will be evaluated first. Afterwards, as an optional extension, we would like to compare the performances for the bigger DGS Corpus dataset [Schulder et al., 2020] as well to see how well both those models scale to a larger amount of data. In addition to that, we will check the performance of both models without the usage of gloss supervision. Having models that perform well without needing training on gloss datasets would be a big breakthrough in SLT. For that reason, it is important to see how the different approaches perform without gloss data as well. With this comprehensive experiment setup, we will perform two different trainings per dataset, on each, the baseline model as well as our new model.

### B. *Datasets*

*1) PHOENIX-Weather 2014T:* The data in the PHOENIX-Weather 2014T dataset [Camgöz et al., 2018] consists of videos from the weather forecast of the Phoenix TV channel in German SL (originally published by Forster et al. [Forster et al., 2012]) and their spoken language and gloss annotations. Since the videos are all from weather forecasts, the topic and the SL vocabulary are quite restricted. The resolution of the videos is 210x260px with a frame rate of 25fps. Nine different people are used as signers and in total 1078 different words were used [Ananthanarayana et al., 2021]. The big advantage of this dataset is, that it has a controlled environment which makes it possible to easily detect the signer in front of a constant, homogeneous, undetailed background (see Figure 2). The dataset has been the state-of-the-art continuous SLT dataset with both gloss and spoken language annotations.

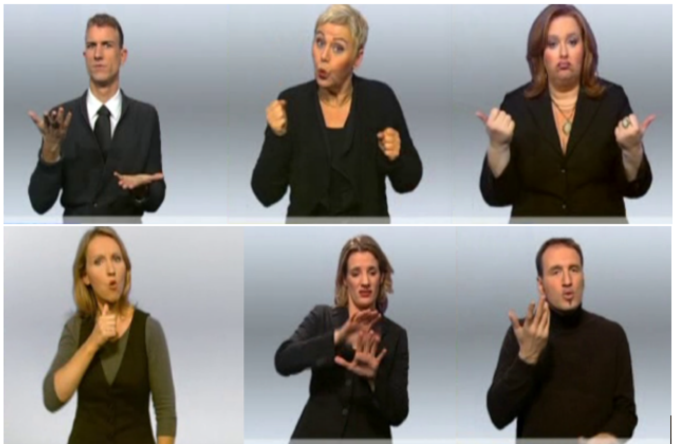


Fig. 2: These are samples from the PHOENIX-Weather 2014T dataset [Ananthanarayana et al., 2021]. It can be seen that the dataset shows the signers in front of an un-detailed background. The background is grey and the signers wear dark clothing to ensure contrast.

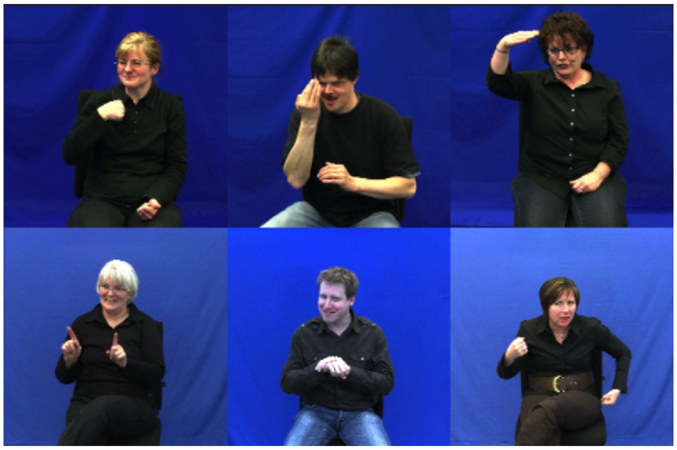


Fig. 3: This figure shows frames from different videos in the DGS Corpus [Schulder et al., 2020]. The background is blue and the signers wear black/dark clothing to ensure sufficient contrast to the background.

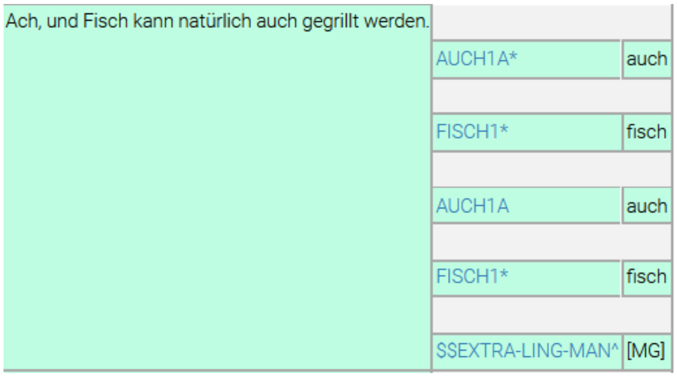
*2) DGS:* The publicly available, annotated part of the DGS Corpus [1] consists of a total of 50h of videos about different topics collected by researchers of the Institute for German SL and Communication of the Deaf at Hamburg University. The videos include 330 different signers from different areas of Germany and 374800 different words were used in the publicly available dataset [Isard and Konrad, 2022]. This makes it considerably larger than the previously used datasets as the PHOENIX-Weather 2014T dataset which only uses about 2800 different words in about 0.95M frames total. The published videos were downscaled to a resolution of 640x360px at 50fps [Schulder et al., 2021], which results in circa 9M frames in total. The background of the videos is consistently in blue while the signers wear dark clothing to ensure contrast to the background (see Figure 3 [2]). An exemplary annotation including gloss and spoken language transcription can be seen in Figure 4 [3].

### C. Metrics

The metrics that are going to be used as evaluation are the BLEU, ROUGE and METEOR Scores.

*1) BLEU (Bilingual Evaluation Understudy):* The BLEU Score is a standard in NLP tasks and also in SLT evaluations. It describes the n-gram precision of the predicted translation when compared to the ground truth [Google, 2022].

*2) ROUGE (Recall-Oriented Understudy for Gisting Evaluation):* The ROUGE evaluation outputs multiple values which include the n-gram overlap precision, recall, and F1-Score separately. This makes it more interpretable than other scores [Lin, 2004].



| | | |
|---|---|---|
| Ach, und Fisch kann natürlich auch gegrillt werden. | | |
| | AUCH1A* | auch |
| | | |
| | FISCH1* | fisch |
| | | |
| | AUCH1A | auch |
| | | |
| | FISCH1* | fisch |
| | | |
| | $$EXTRA-LING-MAN^ | [MG] |

Fig. 4: This is an example of the spoken language (left) and gloss (right) annotations of the DGS Corpus [Schulder et al., 2020]. The right side shows the gloss as it is written in the dictionary in the middle column in addition to the mouth gestures (MG) in the rightmost column. In the online tool, the gloss annotations are linked to the respective SL dictionary entry of this sign to provide more information.

*3) METEOR (Metric for Evaluation of Translation with Explicit Ordering):* In addition to those metrics we are going to evaluate the models on the METEOR Score. The METEOR-Score is an evaluation metric that tries to tackle the problem of the BLEU-Score which does only reward exact overlaps. In the METEOR evaluation also words with the same stem, synonyms, and paraphrases are counted. For this reason, METEOR-Scores are closer to human evaluations [Denkowski and Lavie, 2014].

## V. Work Plan

Figure 5 shows the planned project schedule. In the beginning of the project, we start by downloading and testing the code of the two models that we base our new model on. Afterwards follows the collection and preprocessing of the first dataset. Then, we will start with evaluating the baseline model [Camgöz et al., 2020b] in both experiment setups, with and without gloss supervision. The following step will be an iterative process of implementing and testing the proposed model by combining the Video Swin Transformer with the SLT Transformer. Once this is finished successfully, we can test the new model in the different experiment setups as well. Depending on the progress of the project, we keep it optional for the project to evaluate the performances for the bigger dataset as well. The progress will be continuously updated on our GitHub page [4]. Project deadline steps are marked in red.

[1] https://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/welcome.html

[2] retrieved from https://www.sign-lang.uni-hamburg.de/meinedgs/ling/start_de.html, 13.12.2022

[3] retrieved from https://www.sign-lang.uni-hamburg.de/meinedgs/ling/start-name_de.html, 13.12.2022

[4] https://github.com/daniewin/cv-project

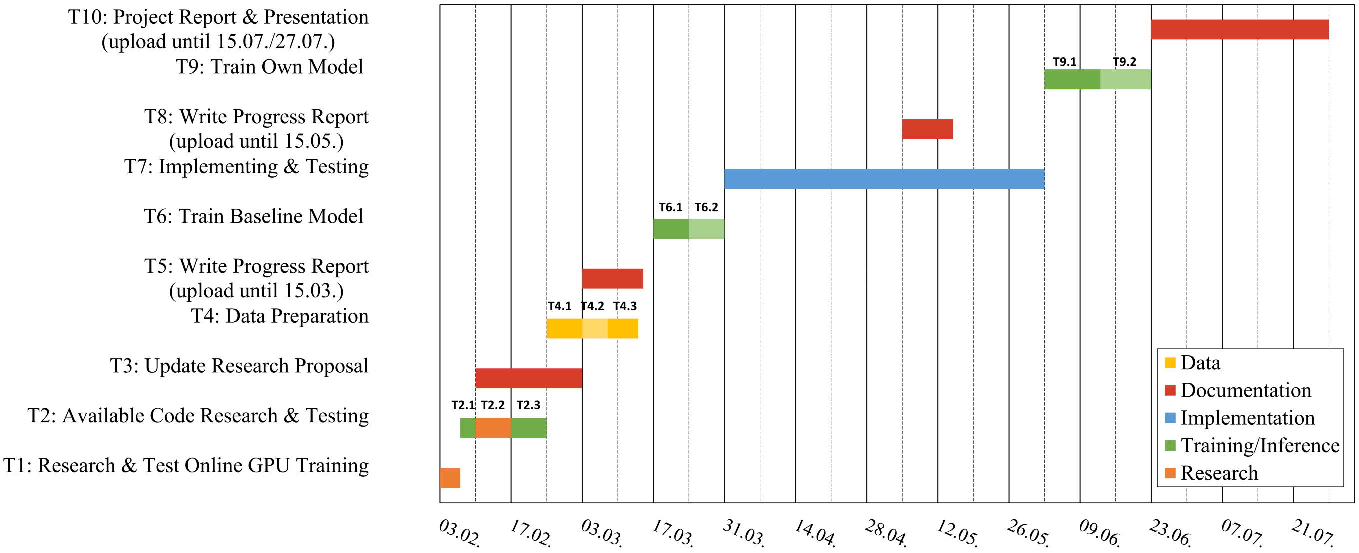


Fig. 5: Gantt-diagram of the preliminary project schedule. The optional extension of the project of evaluating the models on the DGS Corpus is not included in this preliminary project schedule. Tasks include regular documentation on our GitHub page. For explanation of abbreviations see Table I.

| Task | Explanation |
|---|---|
| T2.1: | Run Inference for SLT Model with original parameters (add to Preliminary Results in Research Proposal) |
| T2.2: | Download and Understand Video Swin Transformer Code |
| T2.3: | Run Inference for Video Swin Transformer Code if Weights are Available |
| T4.1: | Download and Process PHOENIX-Weather 2014T Data |
| T4.2: | Clean Data & Choose Subsets for Testing |
| T4.3: | Write Dataloader |
| T6.1: | Train Baseline Model (PHOENIX, gloss) |
| T6.2: | Train Baseline Model (PHOENIX, no gloss) |
| T9.1: | Train Own Model (PHOENIX, gloss) |
| T9.2: | Train Own Model (PHOENIX, no gloss) |

TABLE I: Explanations for the abbreviations used in the Gantt-diagram (see Figure 5).

## VI. Conclusion

In this paper, we present a research project to further improve the performance of machine translations for SLT with Transformers. SLT does not yet reach the level of spoken language machine translation performances. It is likely, that part of the reason for this lies in the main differences between those two areas. For one, only very small datasets compared to the ones for spoken language have been used to train the SLT models. That is why we aim to use a way bigger, more recently available dataset to see how well the models scale. The second big difference between sign and spoken language is their modality. SLs are visual languages and therefore are fundamentally different from spoken languages. Spoken languages can be represented in written form without much loss of information. Gloss transcriptions are an attempt to achieve the same for SLs, however, they are not able to capture the ambiguity, context dependency, and visuospatial properties of signs well. Therefore, it is desirable to encode SL data without using glosses as an information bottleneck [Camgöz et al., 2020b]. Instead, other ways of encoding the visual data need to be explored to find better representations for the encoder part of the Transformer network. To address this issue, we present a Pure Video Transformer network which does not rely on CNNs features but uses the sign videos as direct input to the encoder. Pure Video Transformers have proven to encode visual data in a way that leads to better performance on various CV tasks, including video-to-text processing. Hence, the proposed architecture is a promising approach to improving SLT performances.

# REFERENCES

[Ananthanarayana et al., 2021] Ananthanarayana, T., Srivastava, P., Chintha, A., Santha, A., Landy, B., Panaro, J., Webster, A., Kotecha, N., Sah, S., Sarchet, T., et al. (2021). Deep learning methods for sign language translation. *ACM Transactions on Accessible Computing*, 14(4):1–30.

[Arnab et al., 2021] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., and Schmid, C. (2021). Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846.

[Camgöz et al., 2018] Camgöz, N. C., Hadfield, S., Koller, O., Ney, H., and Bowden, R. (2018). Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7784–7793.

[Camgöz et al., 2020a] Camgöz, N. C., Koller, O., Hadfield, S., and Bowden, R. (2020a). Multi-channel transformers for multi-articulatory sign language translation. In *European Conference on Computer Vision*, pages 301–319. Springer.

[Camgöz et al., 2020b] Camgöz, N. C., Koller, O., Hadfield, S., and Bowden, R. (2020b). Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[Camgöz et al., 2021] Camgöz, N. C., Saunders, B., Rochette, G., Giovanelli, M., Inches, G., Nachtrab-Ribback, R., and Bowden, R. (2021). Content4all open research sign language translation datasets. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–5. IEEE.

[Crasborn, 2015] Crasborn, O. A. (2015). Transcription and notation methods. *Research Methods in Sign Language Studies: A Practical Guide*, pages 74–88.

[De Coster et al., 2021] De Coster, M., D'Oosterlinck, K., Pizurica, M., Rabaey, P., Verlinden, S., Van Herreweghe, M., and Dambre, J. (2021). Frozen pretrained transformers for neural sign language translation. In *18th Biennial Machine Translation Summit*, pages 88–97. Association for Machine Translation in the Americas.

[Denkowski and Lavie, 2014] Denkowski, M. and Lavie, A. (2014). METEOR universal: Language specific translation evaluation for any target language. In *Proceedings of the 9th Workshop on Statistical Machine Translation*, pages 376–380.

[Dosovitskiy et al., 2021] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations*.

[Forster et al., 2012] Forster, J., Schmidt, C., Hoyoux, T., Koller, O., Zelle, U., Piater, J., and Ney, H. (2012). RWTH-PHOENIX-weather: A large vocabulary sign language recognition and translation corpus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 3785–3789.

[Google, 2022] Google (2022). AutoML translation documentation: Evaluating models. https://cloud.google.com/translate/automl/docs/evaluate#interpretation. Accessed: 2022-12-06.

[Isard and Konrad, 2022] Isard, A. and Konrad, R. (2022). MY DGS – ANNIS: ANNIS and the public DGS corpus. In Efthimiou, E., Fotinea, S.-E., Hanke, T., Hochgesang, J. A., Kristoffersen, J., Mesch, J., and Schulder, M., editors, *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 73–79, Marseille, France. European Language Resources Association (ELRA).

[Lin, 2004] Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

[Lin et al., 2022] Lin, K., Li, L., Lin, C.-C., Ahmed, F., Gan, Z., Liu, Z., Lu, Y., and Wang, L. (2022). SwinBERT: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17949–17958.

[Liu et al., 2021] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.

[Liu et al., 2022] Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., and Hu, H. (2022). Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211.

[PyTorch Contributors, 2022] PyTorch Contributors (2022). CTCLoss - PyTorch documentation. https://pytorch.org/docs/stable/generated/torch.nn.CTCLoss.html. Accessed: 2023-10-03.

[Schulder et al., 2020] Schulder, M., Blanck, D., Hanke, T., Hofmann, I., Hong, S.-E., Jeziorski, O., König, L., König, S., Konrad, R., Langer, G., et al. (2020). Data statement for the public DGS corpus. Technical report, Project Note AP06-2020-01, DGS-Corpus project, IDGS, Hamburg University.

[Schulder et al., 2021] Schulder, M., Blanck, D., Hanke, T., Hofmann, I., Hong, S.-E., Jeziorski, O., König, L., König, S., Konrad, R., Langer, G., Nishio, R., and Rathmann, C. (2021). Data statement for the public DGS corpus.

[Voskou et al., 2021] Voskou, A., Panousis, K. P., Kosmopoulos, D., Metaxas, D. N., and Chatzis, S. (2021). Stochastic transformer networks with linear competing units: Application to end-to-end SL translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11946–11955.

[Zhu et al., 2020] Zhu, J., Xia, Y., Wu, L., He, D., Qin, T., Zhou, W., Li, H., and Liu, T.-Y. (2020). Incorporating BERT into neural machine translation. In *Proceedings of the 8th International Conference on Learning Representations*.