mBART for Supervised Gloss-Free Sign Language Translation: Integrating RGB and Facial Keypoint Images

Jiannan Mao[1]* Chenchen Ding[2] Hour Kaing[2] Hideki Tanaka[2]
Masao Utiyama[2] Tadahiro Matsumoto[1]
[1]Gifu University, Gifu, Japan [2]ASTREC, UCRI, NICT, Japan
{mao, tad}@mat.info.gifu-u.ac.jp
{chenchen.ding, hour_kaing, hideki.tanaka, mutiyama}@nict.go.jp

Abstract

Sign language translation (SLT) has traditionally depended on gloss annotations, which are costly and time-consuming to produce. This work presents a gloss-free SLT framework that integrates raw RGB video input with facial keypoint features, enabling richer visual representations. We leverage a two-stage approach: first aligning visual and textual features with a frozen multilingual mBART encoder, then refining translation through the mBART decoder. Evaluations on the PHOENIX-2014T dataset show performance gains over baselines, yielding a +0.64 BLEU improvement. These results confirm that incorporating facial keypoints strategy can significantly improve gloss-free sign language translation.

1 Introduction

Sign languages are visual signals used for communication among the Deaf or Hard of Hearing (DHH). These languages are primarily expressed through manual articulations, but they are also greatly enriched by the movement of the body, mouth, eyes, and eyebrows. This visual complexity not only enhances the expressiveness of sign languages but also helps convey thoughts more clearly [1, 2].

For those of us with intact hearing and speaking abilities, there is often a misconception that "DHH individuals prefer reading spoken language; therefore, when communicating with DHH, it is sufficient to rely solely on spoken language, using written text, either on paper or via smart devices [3]." This perspective fails to account for a critical fact: for many DHH individuals, sign language is not merely a communication tool but their primary and most natural language, deeply intertwined with their identity and culture [4, 5]. Unlike written spoken languages, which may feel secondary or foreign to many DHH individuals, sign language provides a more direct and expressive connection to their thoughts and emotions [5].

In this context, Sign Language Translation (SLT) systems are essential to bridge the communication gap between DHH individuals and the hearing population, enabling more meaningful interactions and fostering a society that truly values diversity, equity, and inclusion [6].

To address this challenge, researchers have explored various approaches inspired by the Neural Machine Translation (NMT) framework, adapted to handle visual inputs. Typically, a sequence of video frames is processed by a visual network to either predict glosses, or extract image features, which are then mapped to spoken language using NMT [7, 8, 9, 10], as shown in Figure 1. While glosses enhance translation accuracy, their production demands costly, time-intensive manual annotation. As a result, gloss-free SLT has emerged as a trend, aiming to directly translate raw video into text [9, 11, 12].

This work focuses on gloss-free sign language translation, employing the mBART model as a teacher model to supervise the outputs of a visual network through the encoder outputs of mBART. For the visual network, we utilize a combination of RGB images and facial keypoints, enable the model to capture more detailed facial features during the learning process. Our experimental results show a significant improvement, with an increase of 0.64 BLEU points over the baseline. This approach not only confirms the effectiveness of combining RGB images with facial keypoints for visual feature extraction but also underscores

* This work was done during the first author's internship at National Institute of Information and Communications Technology, Kyoto, Japan.

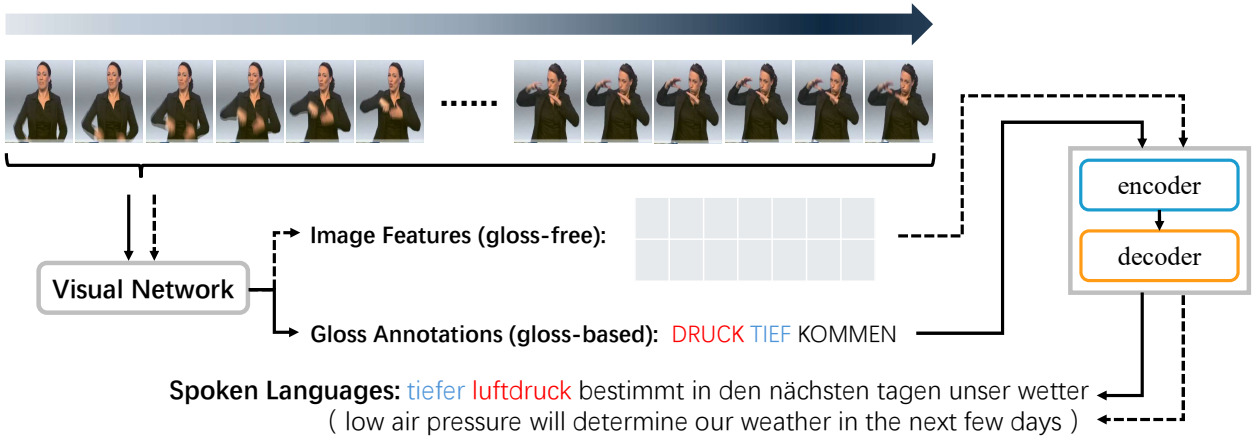

Figure 1 The difference between gloss-free and gloss-based approaches in sign language translation. In the example, we demonstrated the difference in word order between sign and spoken languages : DRUCK corresponds to luftdruck, and TIEF corresponds to tiefer.

the efficacy of employing mBART as a supervisory model.

2 Related Works

Gloss-Free Glosses are written labels used to represent gestures in sign language, providing a stable representation units by segmenting continuous gestures into discrete lexical elements. For instance, the gesture for 'Put the book on the table' might be glossed as 'PUT BOOK TABLE'. While glosses act as a bridge between sign and spoken language, they are not equivalent to spoken words. They follow the syntax of sign language, which can differ from spoken language order, as shown in Figure 1.

Currently, the majority of gloss-free sign language translation studies rely on datasets without gloss annotations [13, 14, 15]. Our work follows this gloss-free research direction. For comparability with the baseline [9], we use a dataset containing gloss annotations [16] but completely disregard the gloss information.

Facial Keypoints Visual feature extraction is crucial for sign language translation. Combining keypoints with RGB images has been shown to improve recognition accuracy by offering richer visual representations [8, 17]. In particular, incorporating facial keypoints enhances these representations, as they provide fine-grained semantic cues, such as expressions, which help distinguish ambiguous gestures.

In this work, we extract facial keypoints to enhance the model's ability to learn detailed facial features, which are often overlooked in existing studies.

mBART in SLT mBART is a pre-trained multilingual model that using denoising autoencoding to capture universal linguistic features [18]. It has demonstrated strong performance in low-resource tasks, including sign language translation [7, 8, 9, 10].

We adopt mBART as a teacher model to supervise the visual networks, aligning its outputs with the encoder representations of mBART. This strategy bridges the gap between visual features and linguistic understanding, resulting in improved translation performance.

3 Method

Our method employs a two-stage framework for gloss-free sign language translation, as shown in Figure 2. For the visual feature extraction network, we adopt the approach proposed by Zhou et al. [9], which combines ResNet [19], temporal blocks [20], and parts of the mBART encoder [18] to form the visual network. Two separate visual network are utilized: one processes the raw RGB image sequences, while the other processes facial keypoints. These visual features are aligned with the frozen mBART encoder in the first stage, and further integrated and trained with the mBART decoder in the second stage, enabling the translation from sign language to spoken language.

Data Preparation The original dataset D contain RGB sequences d_{rgb} and corresponding text d_{text}: $D = (d_{rgb}, d_{text})$. We extract the facial keypoints sequence d_{key} for each d_{rgb} using HRNet [21], forming $D' = (d_{rgb}, d_{key}, d_{text})$. Here, d_{rgb}, d_{key} are sequences in $\mathbb{R}^{N \times L^{rgb}}$, and $d_{text} \in \mathbb{R}^{N \times L^{text}}$, where N is the dataset size. These augmented data provide richer visual cues that support more nuanced interpretation of signs image sequences.

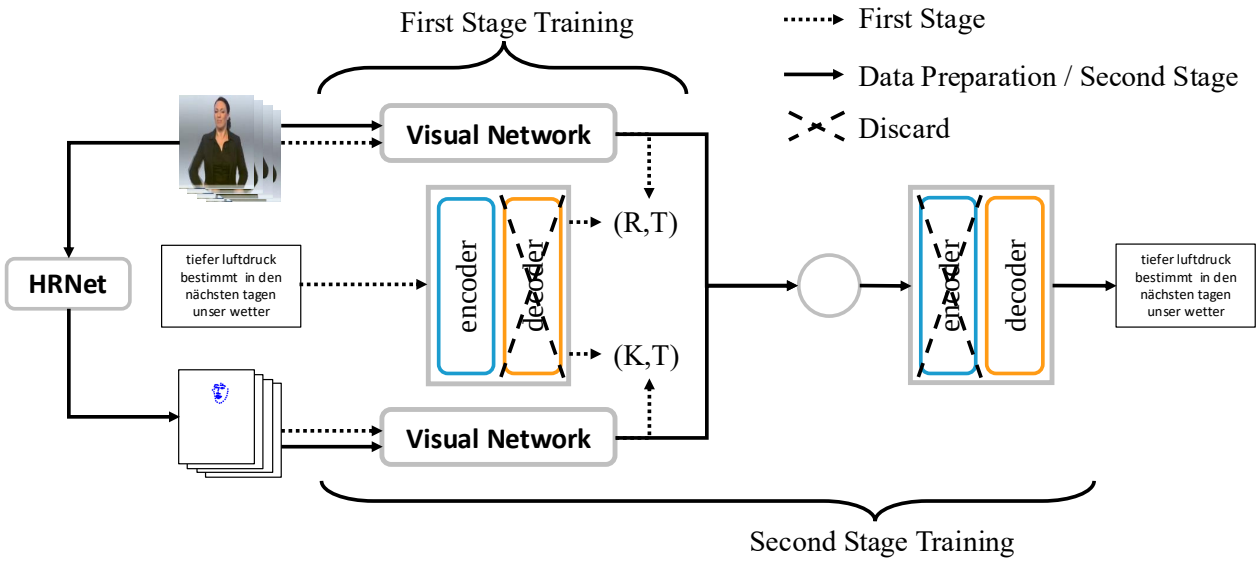

Figure 2 Overview of our two-stage framework. In the first stage, we align visual (RGB + facial keypoints) and textual features using a frozen mBART encoder. In the second stage, we fuse the learned visual representations and train the mBART decoder to produce the final translations. Colored elements in background correspond to the main focus at each stage.

First Stage: Feature Alignment Sign and spoken language sequences differ in length ($L^{rgb} \neq L^{text}$), making direct alignment challenging. To address this, we introduce special markers $\langle rgb \rangle$, $\langle key \rangle$, and $\langle EOS \rangle$ at the end of each respective sequence. These markers produce a global summary vector for each modality.

We define two visual networks: f_{rgb} for RGB inputs and f_{key} for facial keypoints. Both employ ResNet [19], temporal blocks [20], and parts of the mBART encoder architecture [18] as per [9]. Extracted features are:

$$R = f_{rgb}(d_{rgb}, \langle rgb \rangle), K = f_{key}(d_{key}, \langle key \rangle)$$

Here, R and K represent the feature vectors extracted from the respective positions of the markers $\langle rgb \rangle$ and $\langle key \rangle$, which aggregate the information from the entire sequence.

For the text sequence, we employ the mBART encoder, $mBART_{encoder}$, as a feature extractor. During this phase, $mBART_{encoder}$ remains frozen to serve as a teacher model, to ensure robust and stable supervision through text embeddings. The text sequence is processed as follows:

$$T = mBART_{encoder}(d_{text}, \langle EOS \rangle)$$

Here, T is the feature vector extracted at the position of $\langle EOS \rangle$, capturing the overall semantics of the text.

To ensure effective alignment between visual and textual modalities, the loss function maximizes their dot-product similarity in the shared latent space, as follows:

$$\mathcal{L} = -\frac{1}{2}\left(\sum \log \text{sim}(R,T) + \sum \log \text{sim}(K,T)\right)$$

Here, $sim(R,T)$ measures the similarity between feature vectors. By maximizing these alignments, the model learns to capture the semantic correspondences between the visual and text modalities in the first stage of training.

Second Stage: Translation At this stage, we utilize the RGB image visual network f_{rgb} and the facial keypoint visual network f_{key}, obtained from the first stage of training, to extract features from the input data d_{rgb} and d_{key}. Subsequently, we fuse these two features using the Visual-Language Mapper [7], a fully-connected MLP with two hidden layers, as follows:

$$F_{rgb} = f_{rgb}(d_{rgb}), F_{key} = f_{key}(d_{key})$$
$$F_{fusion} = VLM([F_{rgb}, F_{key}])$$

The fused features F_{fusion} are fed into the mBART decoder for training, and the loss for this process is defined as follows:

$$\mathcal{L} = -\sum logP(d^{i}_{text} | F_{fusion}, d^{<i}_{text})$$

The second stage bridges visual and textual modalities by directly optimizing the translation task. Leveraging the fused features F_{fusion} as input to the mBART decoder, it ensures fluent, semantically accurate spoken language translations. While the first stage focuses on feature alignment through similarity maximization, this stage refines end-to-end translation quality by fine-tuning the decoder and visual network within a shared representation space. Together, these two stages enable progressive learning of

semantic alignment, resulting in robust, accurate gloss-free sign language translation.

4 Experiments

4.1 Settings

The experiments were conducted on the PHOENIX-2014T dataset [16], which contains 8,257 German Sign Language (DGS) videos paired with corresponding German translations drawn from weather forecast broadcasts. The dataset is split into training (7,096), development (519), and test (642) sets. We strictly follow a gloss-free scenario by not using any provided gloss annotations.

In the first training stage, we utilize only the mBART-cc25 [18] encoder (frozen as a teacher model) to provide textual supervision.[1)] In the second stage, we employ the first three layers of the mBART-cc25 decoder to generate final translations. All other training hyperparameters closely follow Zhou et al. [9] to ensure consistency.

We evaluate our model using standard automatic metrics: BLEU-4 [22] and ROUGE [23], allowing direct comparison to previous work.

4.2 Results and Discussions

Table 1 presents a comparison of our gloss-free sign language translation method against both gloss-based and gloss-free baselines on the PHOENIX-2014T dataset. For gloss-based approaches, methods such as MMTLB [7], TS-SLT [8], and CV-SLT [10] achieve relatively high performance, with CV-SLT scoring the highest BLEU-4 and ROUGE values (29.27 and 54.33, respectively). These methods benefit from annotated gloss intermediates, which provide an explicit linguistic bridge between sign and spoken language, thus improving translation quality.

In contrast, our work, along with GFSLT [9], focuses on a gloss-free scenario, which is more challenging due to the absence of explicitly annotated sign glosses. Within this setting, GFSLT-rgb relies solely on raw RGB video frames, while GFSLT-key substitutes RGB input with whole human body keypoints extracted via HRNet. Interestingly, the GFSLT-key variant, which encodes the entire body skeletal motion, achieves lower scores (16.08 BLEU-4, 35.21 ROUGE) than GFSLT-rgb (21.44 BLEU-4, 42.49 ROUGE). This suggests that while body keypoints provide skeletal motion cues, they may lose important visual details (e.g., subtle body movements, hand shape nuances) that contribute to accurate sign interpretation.

Table 1 Results on the PHOENIX-2014T dataset. Improve represents the gains of our method compared to GFSLT-rgb[9].

Gloss	Method	BLEU-4	ROUGE
based	MMTLB [7]	28.39	52.65
	TS-SLT [8]	28.95	53.48
	CV-SLT [10]	**29.27**	**54.33**
free	GFSLT-rgb [9]	21.44	42.49
	GFSLT-key	16.08	35.21
	rgb+key_facial(our)	**22.08**	**44.12**
Improve	-	+0.64	+1.63

Our proposed method, denoted as rgb+key_facial(our), combines the strengths of raw RGB input with facial keypoints. This fusion outperforms both GFSLT-rgb and GFSLT-key, improving BLEU-4 by +0.64 and ROUGE by +1.63 over the RGB-only baseline. The improvement indicates that integrating detailed facial cues with the global scene information from RGB frames leads to more semantically aligned and contextually richer representations, ultimately enhancing translation performance.

Despite these gains, gloss-free methods, including ours, still exhibit a performance gap compared to gloss-based approaches. Nevertheless, our results demonstrate that carefully selecting and fusing multiple visual cues can mitigate the challenges posed by the lack of gloss annotations.

5 Conclusion and Future Work

This study extends existing efforts in gloss-free sign language translation, an area in sign language processing where no intermediate gloss annotations are used. We propose an approach that integrates RGB frames with facial keypoint data. By exploiting complementary information from these modalities, our method better captures complex spatial and temporal patterns of sign language. This multimodal strategy improves translation accuracy compared to relying solely on RGB inputs and may facilitate more accurate, context-aware translations.

In future work, we plan to explore additional visual cues, refine fusion strategies, and incorporate more powerful language models to further enhance gloss-free sign language translation.

1) huggingface.co/facebook/mbart-large-cc25

Acknowledgement

This work was financially supported by JST SPRING, Grant Number JPMJSP2125. The author (Initial) would like to take this opportunity to thank the "THERS Make New Standards Program for the Next Generation Researchers."

References

[1] Annika Herrmann and Markus Steinbach. **Nonmanuals in sign language**, Vol. 53. John Benjamins Publishing, 2013.

[2] Kathleen, Paul, and Dot Sign Language. British sign language. https://bsl.surrey.ac.uk/principles/f-non-manual-features.

[3] Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, Vassilis Athitsos, and Mohammad Sabokrou. A survey on recent advances in sign language production. **Expert Systems with Applications**, Vol. 243, p. 122846, 2024.

[4] H.D.L. Bauman. **Open Your Eyes: Deaf Studies Talking**. University of Minnesota Press, 2008.

[5] A. Mindess. **Reading Between the Signs: Intercultural Communication for Sign Language Interpreters**. Intercultural Press, 2011.

[6] United Nations. Convention on the Rights of Persons with Disabilities (CRPD). https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-rights-persons-disabilities.

[7] Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. A simple multi-modality transfer learning baseline for sign language translation. In **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**, pp. 5120–5130, 2022.

[8] Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. Two-stream network for sign language recognition and translation. **NeurIPS**, 2022.

[9] Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. Gloss-free sign language translation: Improving from visual-language pretraining. In **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)**, pp. 20871–20881, October 2023.

[10] Biao Fu Cong Hu Jinsong Su Yidong Chen Rui Zhao, Liang Zhang. Conditional variational autoencoder for sign language translation with cross-modal alignment. In **Proceedings of the AAAI Conference on Artificial Intelligence**, 2024.

[11] Jinhui Ye, Xing Wang, Wenxiang Jiao, Junwei Liang, and Hui Xiong. Improving gloss-free sign language translation by reducing representation density. 2024.

[12] Zhigang Chen, Benjia Zhou, Yiqing Huang, Jun Wan, Yibo Hu, Hailin Shi, Yanyan Liang, Zhen Lei, and Du Zhang. C^2rl: Content and context representation learning for gloss-free sign language translation and retrieval, 2024.

[13] Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, et al. Bbc-oxford british sign language dataset. **arXiv preprint arXiv:2111.03635**, 2021.

[14] 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.

[15] Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. Open-domain sign language translation learned from online video. In **EMNLP**, 2022.

[16] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In **Proceedings of the IEEE conference on computer vision and pattern recognition**, pp. 7784–7793, 2018.

[17] Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. Spatial-temporal multi-cue network for continuous sign language recognition. In **Proceedings of the AAAI conference on artificial intelligence**, Vol. 34, pp. 13009–13016, 2020.

[18] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 726–742, 2020.

[19] 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**, pp. 770–778, 2016.

[20] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arxiv. **arXiv preprint arXiv:1803.01271**, Vol. 10, , 2018.

[21] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Vol. 43, No. 10, pp. 3349–3364, 2021.

[22] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th annual meeting of the Association for Computational Linguistics**, pp. 311–318, 2002.

[23] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In **Text summarization branches out**, pp. 74–81, 2004.