

Visual Alignment Pre-training for Sign Language Translation

Peiqi Jiao^{1,2}, Yuecong Min^{1,2}, and Xilin Chen^{1,2}

¹ Key Laboratory of AI Safety of CAS, Institute of Computing Technology,
Chinese Academy of Sciences (CAS), Beijing, China

² University of Chinese Academy of Sciences, Beijing, China
peiqi.jiao@vipl.ict.ac.cn, {minyuecong, xlchen}@ict.ac.cn

Abstract. Sign Language Translation (SLT) aims to translate sign videos into text sentences. While gloss sequences, the written approximation of sign videos, provide informative alignment supervision for visual representation learning in SLT, the associated high cost of gloss annotations hampers the scalability. Recent works have yet to achieve satisfactory results without gloss annotations. In this study, we attribute the challenge to the flexible correspondence between visual and textual tokens, and aim to address it by constructing a gloss-like constraint from text sentences. Specifically, we propose a Visual Alignment Pre-training (VAP) scheme to exploit visual information by aligning visual and textual tokens in a greedy manner. The VAP scheme enhances visual encoder in capturing semantic-aware visual information and facilitates better adaptation with translation modules pre-trained on large-scale corpora. Experimental results across four SLT benchmarks demonstrate the effectiveness of VAP, which can generate reasonable alignments and significantly narrow the performance gap with gloss-based methods.

Keywords: Sign Language Translation · Vision-Language Pre-training · Visual Embedding

1 Introduction

Sign languages, serving as the predominant way of communication for the Deaf communities, convey information through both manual signals (*e.g.*, hand shape, orientation, place of articulation and movement) and non-manual signals (*e.g.*, lip pattern, facial expression, head and upper body orientation) [12, 24]. Its distinctive grammar and vocabulary pose challenges for non-sign language users. To address the communication barrier between the Deaf and hearing people, Sign Language Translation (SLT) aims to convert sign language videos into text sentences, garnering considerable attention in recent years [6–8, 54].

As a visual language, the quality of visual representation plays a critical role in SLT. Recent advancements in SLT network architectures [7, 8, 50, 52, 59] typically comprise a visual encoder and a translation module. These approaches can be broadly classified into two categories based on the utilization of gloss annotations. **Gloss-based methods** [9, 50, 54] often involve pre-training the visual

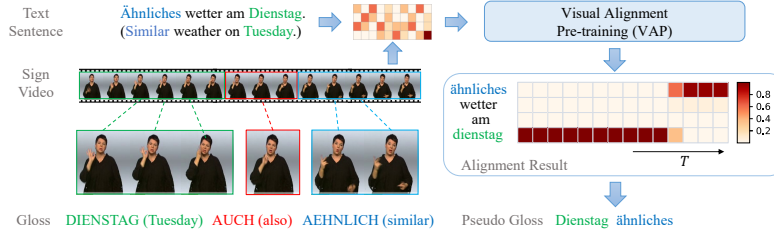


Fig. 1: Illustration of the alignment generated by VAP for a given sign video and its text sentence. The corresponding parts among different sequences are indicated by the identical color.

encoder on Sign Language Recognition (SLR) datasets or integrating an auxiliary SLR branch to improve the quality of visual representation. They primarily focus on exploring the translatability between sign and spoken languages and significantly improve translation performance. However, the scalability of these methods is hindered by the considerable expense associated with gloss labeling.

To leverage broad online sign language sources, **gloss-free methods** have received increasing attention these years [25, 27, 34, 52, 59]. However, developing a robust visual encoder without the aid of gloss annotations presents a significant challenge, and gloss-free methods [52, 59] often exhibit performance inferior to that of gloss-based methods. Inspired by the success of recent Vision-Language Pre-training (VLP) works [20, 51, 58], Zhou *et al.* [59] propose a two-stage gloss-free pipeline: initially aligning visual and semantic representations through a pre-training process, and subsequently leveraging the pre-trained weights for fine-tuning the entire SLT model. However, the proposed GFSLT-VLP exclusively focuses on sentence-wise alignment, which falls short of providing supervision as strong as gloss annotations. Besides, how to approach the role of gloss annotations under gloss-free setting remains unclear.

In this paper, we first discuss the role of gloss annotation in SLT and attribute the worse performance of recent gloss-free works to the flexible correspondence between visual and textual tokens, which fails to provide valid feedback for the visual encoder. To address this challenge, we hypothesize that the approximate meaning of each sign in the video can be found in the text sentence, and propose a Visual Alignment Pre-training (VAP) scheme to facilitate the utilization of visual information by constructing a gloss-like constraint from text sentences. Specifically, VAP encourages each feature to find its most relevant paired feature and pulls them closer in a contrastive manner, which approximates the optimal alignment in a greedy manner. As shown in Fig. 1, VAP can establish reasonable alignments between visual and textual tokens and play a role similar to gloss annotations, thereby aiding the visual encoder in learning semantic-aware visual representations and enhancing compatibility with translation modules pre-trained on large-scale corpora to further improve translation quality. In addition, we propose two approaches to assess the quality of the generated alignments and the alignment method itself, respectively.

In summary, this paper focuses on the exploitation of visual information in the absence of gloss annotations. We propose VAP to enhance the visual encoder through aligning visual and textual tokens during the pre-training phase, and experimental results on public datasets demonstrate the effectiveness of the proposed VAP. Notably, the VAP approaches the performance of gloss-based methods while providing satisfactory alignment results.

The main contributions are summarized as follows:

- Approaching the role of gloss annotation in SLT and showing the importance of exploiting visual information under gloss-free setting.
- Proposing Visual Alignment Pre-training (VAP) to facilitate visual representation learning by aligning visual and textual tokens in a greedy manner.
- Conducting thoughtful experiments about the training of gloss-free SLT models. Experimental results show the effectiveness of VAP, which can generate reliable alignment results and approach gloss-based methods.

2 Related Works

Recent sign language understanding tasks can be categorized into two main classes: SLR and SLT. SLR [16, 19, 46, 55, 57] focuses on transcribing sign language videos into ordered gloss sequences while SLT endeavors to translate sign videos into corresponding text sentences. The inherent distinctions in word order and vocabulary between sign and spoken languages pose challenges in using text sentences as the supervision for the visual encoder. Based on the use of gloss annotations, recent SLT research can be broadly classified into two categories: gloss-based SLT [6–8, 33, 53, 61] and gloss-free SLT [25, 52, 59].

Gloss-based SLT methods utilize the monotonic alignment between sign videos and gloss annotations, often enhancing the visual encoder with an auxiliary SLR branch [7, 8, 61] or directly employing gloss sequences as intermediate representations of sign videos [6, 33, 53]. The primary objective of gloss-based SLT is to investigate the translatability between sign and spoken languages, a pursuit often limited by the availability of parallel corpora. To solve this problem, SignBT [60] employs back-translation as a data augmentation strategy to enrich training data. MMTLB [8] introduces pre-trained translation modules on large-scale multilingual corpora into SLT with transfer learning. SLTUNet [54] narrows the gap between sign video and text through jointly modeling various SLT-related tasks. While these methods significantly improve translation performance, scalability is hindered by the expansive annotation cost.

Gloss-free SLT methods investigate more flexible correspondence between sign videos and text sentences, exploring the potential of scaling with online sign language resources [1, 40, 44]. However, it is hard to provide sufficient supervision for the visual encoder without gloss annotations. Recent works try to alleviate this problem through various constraint designs [52] and leveraging powerful pre-trained models [25, 59]. TSPNet [25] pre-trains the visual encoder on sign gesture videos and employs a hierarchical feature learning method to explore the temporal structures of sign videos. GASLT [52] proposes gloss attention to

assist the visual encoder in focusing on local semantic units within sign videos. GFSLT-VLP [59] incorporates supervision from text sentences by performing a CLIP-like vision-language pre-training. There are also some works [10, 22, 48, 57] leverage contrastive learning to capture fine-grained information in sign language understanding tasks. Different from them, we attempt to identify the factors hindering the learning of visual encoder from text sentences and show the effectiveness of aligning visual and textual tokens under gloss-free setting. Different to the concurrent work Sign2GPT [49], which mainly focuses on the effective utilization of pre-trained vision and language models, VAP provides a more general pre-training mechanism with a more effective aggregation method, making it not limited by the architecture of the pre-trained models.

Vision-Language Pre-training has been extensively investigated in recent image-based or video-based works [20, 38, 41, 51, 58] to leverage large-scale multi-modal data. Pre-trained models exhibit remarkable generalization capacities on downstream tasks, such as cross-modal matching/retrieval [10, 20, 38] and zero-shot transfer [38]. Pioneering studies like CLIP [38] and ALIGN [20] demonstrate the feasibility of transferring the success of pre-training in Natural Language Processing (NLP) to the visual domain by aligning visual and textual features into a unified semantic space. Different from the instance-level alignment, FILIP [51] and RegionCLIP [58] enforce fine-grained alignment between image patches and textual words and surpass CLIP on multiple downstream tasks with less training data. Our work draws inspiration from FILIP [51], but we directly employ the word embedding to guide the learning of the visual encoder, which makes it learn a semantic-aware visual embedding space.

Visual Embedding aims to utilize low-dimensional continuous vectors for representing visual inputs, with the objective that similar visual inputs reside closely within the embedding space [4, 43]. Inspired by the success of word embedding in NLP [23, 28, 30, 37], many efforts [11, 17, 32, 41, 45] have been done to explore the unified visual embedding. Compared to general images and videos, sign videos have natural semantic structures, and some studies [18, 34] seek to leverage the pre-trained visual encoder to enhance visual embeddings. Different from these approaches, we utilize word representations pre-trained on large-scale corpora to guide the learning of visual embedding, achieving competitive results with gloss-based methods efficiently.

3 Method

In this section, we first discuss the role of gloss annotation in SLT in Sec. 3.1. Then we propose the Visual Alignment Pre-training (VAP) scheme, which provides valid feedback for the visual encoder by aligning visual and textual tokens in Sec. 3.2. After that, we present the entire fine-tuning procedure of the SLT model with the proposed VAP in Sec. 3.3. Finally, two approaches are proposed to assess the quality of generated alignments and the alignment method itself in Sec. 3.4, and Sec. 3.5 provides the training and inference details. An overview of the proposed VAP is presented in Fig. 2.

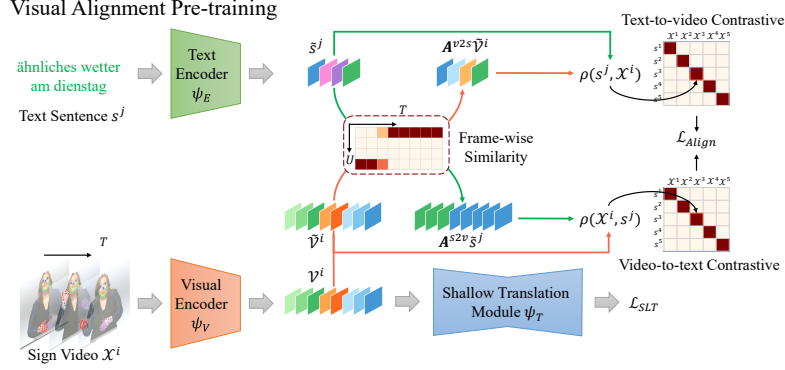


Fig. 2: Overview of the proposed visual alignment pre-training. The $\mathcal{L}_{Align}$ encourages each feature to find its most relevant paired feature and pulls them closer in a contrastive manner. And the $\mathcal{L}_{SLT}$ ensures the temporal consistency of the alignment with a shallow translation module employed.

3.1 The Role of Gloss Annotation in SLT

Sign language translation aims to translate sign videos to corresponding text sentences. As shown in Fig. 2, a general SLT network can be divided into a visual encoder ψ_V and a translation module ψ_T . Given a sign video or other type of inputs (*e.g.*, skeleton data) $\mathcal{X}$ and the corresponding text sentence $\mathbf{s} = \{s_1, \dots, s_U\}$ with U words, ψ_V undertakes to extract visual features $\mathcal{V} = \{v_1, \dots, v_T\} = \psi_V(\mathcal{X})$ from the video and ψ_T is responsible for predicting $\mathbf{s}$ based on $\mathcal{V}$. The network is optimized by minimizing the negative log-likelihood:

$$\mathcal{L}_{SLT} = -\log p(\mathbf{s}|\mathcal{V}; \theta_{\psi_V}, \theta_{\psi_T}). \quad (1)$$

Word alignment [2, 5] in NLP typically refers to the process of indicating the corresponding words in a parallel text $\mathbf{a} = \{a_i\}_{i=1}^m$ and $\mathbf{b} = \{b_j\}_{j=1}^n$, which can be formulated as a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, where $\mathbf{A}_{ij} \in \{0, 1\}$ represents whether a_i corresponds to b_j . Considering the alignment $\mathbf{A}$ between $\mathcal{V}$ and $\mathbf{s}$, the loss function used in recent gloss-based SLT works [8, 9, 50] can be reformulated as:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{SLT} + \mathcal{L}_{SLR} \\ &= -\log \sum_{\mathbf{A}} p(\mathbf{A}|\mathcal{V}; \theta_{\psi_V}) p(\mathbf{s}|\mathcal{V}, \mathbf{A}; \theta_{\psi_V}, \theta_{\psi_T}) - \log p(g|\mathcal{V}; \theta_{\psi_V}), \end{aligned} \quad (2)$$

where $g = \{g_1, \dots, g_M\}$ represents gloss annotation with M glosses, which is monotonically aligned with $\mathcal{V}$. As presented in Eq. (2), the supervision of $\mathcal{L}_{SLT}$ on ψ_V comes from two aspects: the alignment itself and the translation coupled with the alignment. The gloss-based methods improve the utilization of visual information by incorporating additional supervision for ψ_V . Therefore, we attribute the inferior performance of previous gloss-free works [25, 34, 52, 59] to the flexible correspondence between $\mathcal{V}$ and $\mathbf{s}$, which hinders the finding of the optimal alignment and fails to provide enough supervision for the visual encoder.

From the above analysis, we believe the key to solve this problem lies in constructing a gloss-like constraint from text sentences and propose a Visual Alignment Pre-training (VAP) scheme.

3.2 Visual Alignment Pre-training

Recent gloss-based methods often leverage CTC [15] to provide supervision for ψ_V , which considers all monotonic alignments between $\mathcal{V}$ and g :

$$\mathcal{L}_{\text{CTC}}(g, \mathcal{V}) = -\log(p(g|\mathcal{V}; \theta_{\psi_V})) = -\log\left(\sum_{\pi} p(\pi|\mathcal{V}; \theta_{\psi_V})\right), \quad (3)$$

where π is an alignment between $\mathcal{V}$ and g . We assume that **the approximate meaning of each sign in the video can be found in the text sentence**, and try to leverage the alignment between $\mathcal{V}$ and $\mathbf{s}$ to supervise ψ_V . However, the existence of 2^{TU} potential alignments between $\mathcal{V}$ and $\mathbf{s}$ poses challenges in identifying the optimal alignment without appropriate constraints, such as the monotonic relationship gloss annotation has. Motivated by recent advancements in vision-language pre-training [51], we propose a VAP scheme to simplify the alignment, which can approximate the optimal alignment in a greedy manner and provide frame-wise labels for the visual encoder.

Specifically, given a text sentence $\mathbf{s}$, we first use a text encoder ψ_E to obtain the corresponding textual features $\psi_E(\mathbf{s})$, and then two simple linear layers to project $\mathcal{V}$ and $\psi_E(\mathbf{s})$ into a joint feature space:

$$\tilde{\mathbf{v}}_t = \text{Linear}(\mathbf{v}_t), \tilde{\mathbf{s}}_u = \text{Linear}(\psi_E(\mathbf{s}_u)). \quad (4)$$

Then, the alignment $\mathbf{A}^{\text{v2s}} \in \mathbb{R}^{T \times U}$ can be calculated by identifying the most similar textual feature for each visual feature, which can be formulated as:

$$\mathbf{A}_{t,u}^{\text{v2s}} = \begin{cases} 1, & \text{for } u = \arg \max_{u'} f(\tilde{\mathbf{v}}_t, \tilde{\mathbf{s}}_{u'}) \\ 0, & \text{otherwise} \end{cases}, \quad (5)$$

where $f(\cdot, \cdot)$ is implemented as the cosine similarity. To ensure the rationality of $\mathbf{A}^{\text{v2s}}$, we calculate the video-text similarity $\rho(\mathcal{X}, \mathbf{s})$ through:

$$\rho(\mathcal{X}, \mathbf{s}) = \frac{1}{T} \sum_{t=1}^T f(\tilde{\mathbf{v}}_t, \mathbf{A}_t^{\text{v2s}} \tilde{\mathbf{s}}), \quad (6)$$

and employ a contrastive learning pattern to optimize $\mathbf{A}^{\text{v2s}}$. However, the outputs of ψ_V may collapse to a constant value if solely relying on $\rho(\mathcal{X}, \mathbf{s})$ during pre-training. To prevent the collapse, we also calculate the text-video similarity $\rho(\mathbf{s}, \mathcal{X})$ with $\mathbf{A}^{\text{s2v}}$, which is calculated by identifying the most similar visual feature for each textual feature.

After that, for a mini-batch containing N video-text pairs $\{\mathcal{X}^i, \mathbf{s}^i\}_{i=1}^N$, we can compute the contrastive loss through:

$$\mathcal{L}_{\text{Align}} = -\frac{1}{2N} \left(\sum_{i=1}^N \log \frac{\exp(\rho(\mathcal{X}^i, \mathbf{s}^i)/\sigma)}{\sum_{j=1}^N \exp(\rho(\mathcal{X}^i, \mathbf{s}^j)/\sigma)} + \sum_{i=1}^N \log \frac{\exp(\rho(\mathbf{s}^i, \mathcal{X}^i)/\sigma)}{\sum_{j=1}^N \exp(\rho(\mathbf{s}^i, \mathcal{X}^j)/\sigma)} \right), \quad (7)$$

where σ is a pre-defined temperature to scale the logits. With the proposed $\mathcal{L}_{\text{Align}}$, we enlarge the similarity between paired visual and textual feature sequences, which will encourage each feature to find its most relevant paired feature and pull them closer.

In addition, the calculation of alignment only relies on word-level information and there are no constraints for the temporal consistency of the alignment. The decomposition in Eq. (2) reveals that the translation loss also contributes significantly to enhancing the quality of alignment. As we will demonstrate in experiments (Tab. 3), simply re-initializing the translation module can improve the performance, which indicates the importance of optimizing the alignment process in tandem with refining translation capabilities. Therefore, we incorporate $\mathcal{L}_{\text{SLT}}$ to ensure the temporal consistency of the alignment with a shallow translation module employed. The final supervision for the pre-training stage is:

$$\mathcal{L}_{\text{pre-train}} = \mathcal{L}_{\text{Align}} + \mathcal{L}_{\text{SLT}}. \quad (8)$$

Through the above designs, we can calculate an approximate alignment between visual and textual tokens and obtain a semantic-aware visual encoder.

3.3 End-to-end Fine-tuning

Compared with general machine translation datasets [3, 47], public SLT datasets [6, 13, 60] remain limited in scale. Thus, we adopt the translation module pre-trained on large-scale corpora (*e.g.* mBART [28], T5 [39]) to replace the shallow translation module during fine-tuning in order to alleviate the large demand for training data. Moreover, similar to the G2T task in [8, 9], we design a Pseudo-gloss2Text (P2T) task to fine-tune the pre-trained translation module for better adaptation.

Specifically, we can generate a pseudo gloss sequence $\tilde{g}$ based on the alignment $\mathbf{A}^{\text{v2s}}$ for each sign video. As $\mathbf{A}^{\text{v2s}}$ identifies the most similar textual token in the sentence for each visual feature, features within a local window may refer to the same token due to the continuity of the video. Thus, we simply merge the repeated tokens and obtain $\tilde{g}$ by:

$$\tilde{g} = \mathcal{B}(\mathbf{A}^{\text{v2s}} \mathbf{s}), \quad (9)$$

where $\mathcal{B}$ denotes the merging operation. The training objective of P2T can be formulated as:

$$\mathcal{L}_{\text{P2T}} = -\log p(\mathbf{s}|\tilde{g}). \quad (10)$$

So far, we have the well pre-trained visual encoder and the well fine-tuned translation module. Next is the end-to-end fine-tuning of the whole model. We adopt a two-layer MLP as the vision-language mapper (V-L mapper) as previous work [8] does, which is responsible for projecting visual features into the textual space. During fine-tuning, we still include the $\mathcal{L}_{\text{Align}}$ to continuously improve the visual encoder, and the total loss can be formulated as:

$$\mathcal{L}_{\text{fine-tune}} = \mathcal{L}_{\text{Align}} + \mathcal{L}_{\text{SLT}}. \quad (11)$$

3.4 Alignment Quality Measurement

In Sec. 3.2, we make an assumption that the approximate meaning of each sign in the video can be found in the text sentence, and calculate the alignment by identifying the most similar textual token. Based on this assumption, the generated pseudo gloss $\tilde{g}$ should also be monotonically aligned with the sign video. However, we can not directly evaluate the quality of generated alignments with gloss annotation due to the different vocabularies. Here, we propose two approaches to implicitly evaluate the quality of generated alignments.

If $\tilde{g}$ is monotonically aligned with the sign video, it should be able to guide the learning of the SLT network as gloss annotation does. Thus, we train a SLT network from scratch like most gloss-based SLT methods do but replace the gloss annotations with the pseudo glosses $\tilde{g}$, which can be formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{SLT}} + \mathcal{L}_{\text{CTC}}(\tilde{g}, \mathcal{V}), \quad (12)$$

the performance gap between such a pseudo-gloss-based model and VAP can implicitly reflect the quality of generated alignments.

Compared to evaluating the quality of alignments through translation performance, it will be more intuitive to evaluate through alignment performance, such as Word Error Rate (WER). Therefore, we propose to evaluate **the quality of alignment method** by replacing the text sentences in pre-training with gloss sequences, and calculating the WER between the generated pseudo glosses and the corresponding gloss annotations. In other words, this approach is designed to check whether an alignment method can find the correct order from an unordered sequence.

In a nutshell, the proposed approaches can evaluate the quality of alignments from both the generated results and the alignment method itself.

3.5 Training and Inference

Training. As shown in Fig. 3, the training pipeline of VAP includes three stages. The visual encoder is pre-trained in stage 1 using Eq. (8). The translation module pre-trained on large-scale corpora is fine-tuned in stage 2 using pseudo glosses generated by Eq. (9). Finally, the whole model is end-to-end fine-tuned in stage 3 using Eq. (11). Note that VAP is a gloss-free method and gloss annotations are only used for evaluating the quality of pseudo glosses as mentioned in Sec. 3.4.

Inference. The whole model fine-tuned after stage 3 is used for inference. Specifically, the visual encoder first extracts visual features $\mathcal{V}$ and the V-L mapper

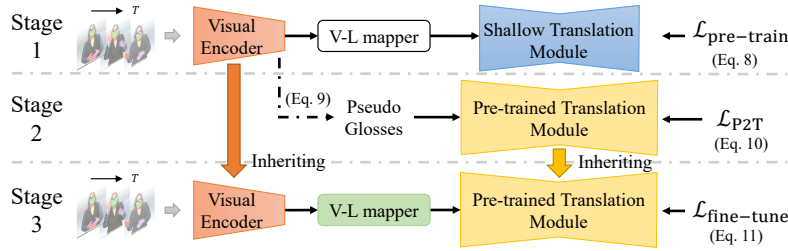


Fig. 3: Illustration of the training pipeline of VAP. The parameter inheriting between stages is indicated by the identical color.

projects them into the textual space. After that, the translation module generates text sentences in an autoregressive manner.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate the proposed method on four prominent SLT datasets. PHOENIX14T [6] is a German sign language dataset, which contains 8,257 samples. The vocabulary size is 2,887 for German words. CSL-Daily [60] is a Chinese sign language dataset and contains 20,654 samples. The vocabulary size is 2,343 for Chinese characters. How2Sign [13] and OpenASL [40] are large-scale American Sign Language datasets. How2Sign contains 79 hours of sign videos with a vocabulary size of 16k English words, while OpenASL includes 288 hours of sign videos with a vocabulary size of 33,549 English words.

Evaluation Metrics. BLEU [35] and ROUGE [26] are adopted (denoted as B, R) as the evaluation metric, and a higher BLEU or ROUGE score indicates better translation performance. WER is used to assess the quality of alignment and a lower WER indicates a better alignment.

Baseline. Considering the training efficiency, **we only use skeleton sequence as input** and employ CoSign-1s [21] as the visual encoder without the use of complementary regularization. The translation module is a three-layered transformer using 1024 hidden units on 8 heads in each layer. For the training process, we employ a similar setting as [59], involving the use of SGD for 200 epochs with the learning rate decay through a cosine schedule from 0.04 to 0.

Implementation Details. The pre-training strategies are the same as baseline, which can provide a better initialization. For the end-to-end fine-tuning and P2T, we use Adam optimizer and train 80 epochs with a cosine annealing schedule. The maximum learning rate is set to $2e-5$. We only use random temporal scaling ($\pm 20\%$) for augmentation, the label smoothing of translation loss is set to 0.2. The temperature σ used in $\mathcal{L}_{\text{Align}}$ is set to 0.07. During inference, we use beam search with a beam size of 4 and the length penalty is set to 1.

For network architecture, the shallow translation module is a three-layered transformer same as that used in baseline. As for the pre-trained translation

Table 1: Performance comparison on PHOENIX14T and CSL-Daily. ‡ denotes results reproduced by GASLT [52]. * denotes results without the inconsistent punctuation bug [31]. † denotes the results on CSL-Daily are reproduced by GFSLT-VLP [59]. The highest performance is highlighted in **bold**.

Gloss-based Method	PHOENIX14T						CSL-Daily					
	Dev			Test			Dev			Test		
	R	B1	B4	R	B1	B4	R	B1	B4	R	B1	B4
MMTLB [8]	53.10	53.95	27.61	52.65	53.97	28.39	53.38	53.81	24.42	53.25	53.31	23.92
IP-SLT [50]	54.43	54.10	28.22	53.72	54.25	27.97	44.33	45.26	16.74	44.09	44.85	16.72
2S-SLT [9]	54.08	54.32	28.66	53.48	54.90	28.95	55.10	55.21	25.76	55.72	55.44	25.79
2S-SLT* [9]							57.81	58.24	29.18	58.62	58.64	29.55
Gloss-free Method	Dev			Test			Dev			Test		
	R	B1	B4	R	B1	B4	R	B1	B4	R	B1	B4
NSLRT [6]	31.00	28.10	9.12	29.70	27.10	8.35	-	-	-	-	-	-
+ Bahdanau Attn. [14]	31.80	31.87	9.94	31.80	32.24	9.58	-	-	-	-	-	-
+ Luong Attn. [29]	32.60	31.58	10.00	30.70	29.86	9.00	-	-	-	-	-	-
SLRT† [7]	-	-	-	31.10	30.88	10.19	20.51	21.03	4.04	19.67	20.00	3.03
TSPNet† [25]	-	-	-	34.96	36.10	13.41	-	-	-	18.38	17.09	2.97
CSGCR [56]	38.96	35.85	15.08	38.85	36.71	15.18	-	-	-	-	-	-
GASLT [52]	-	-	-	39.86	39.07	15.74	-	-	-	20.35	19.90	4.07
GFSLT-VLP [59]	43.72	44.08	22.12	42.49	43.71	21.44	36.70	39.20	11.07	36.44	39.37	11.00
Sign2GPT [49]	-	-	-	48.90	49.54	22.52	-	-	-	42.36	41.75	15.40
Baseline	44.01	43.47	20.15	42.33	43.27	19.72	21.10	21.92	5.00	20.23	20.98	4.28
VAP	51.47	52.78	26.62	51.28	53.07	26.16	51.19	53.31	23.84	51.09	52.98	23.65
VAP*												

module, following [8], we utilize the official release of mBART-large-cc25³ pre-trained on CC25⁴. We use the shallow one for pre-training in order to prevent the large learning rate from destroying the learned features of mBART. For the text encoder ψ_E , we utilize the word embedding layer inherited from mBART and keep it frozen in order to facilitate the effective transfer of rich semantic knowledge to the visual encoder. The embedding layer and tokenizer are inherited from mBART and do not change among the experiments. For OpenASL, we select the person with maximal hand movements when there are multiple persons appearing in the video. The whole training process on the largest dataset OpenASL takes 4.5 days using a single NVIDIA RTX A800 GPU.

4.2 Comparison with State-of-the-art Methods

To verify the effectiveness of VAP, we first compare it with both gloss-based and gloss-free methods on PHOENIX14T and CSL-Daily. As shown in Tab. 1, the proposed VAP can bring substantial improvements over previous gloss-free methods. Specifically, VAP improves the BLEU-4 score by approximately 4.50/4.72 on the Dev/Test sets of PHOENIX14T and 12.77/12.65 on those of CSL-Daily than GFSLT-VLP [59], emphasizing the effects of aligning visual and textual tokens during pre-training. Although we only use skeleton sequences as input, VAP achieves competitive results to gloss-based methods. The performance gaps

³ <https://huggingface.co/facebook/mbart-large-cc25>

⁴ <https://commoncrawl.org/>

Table 2: Performance comparison on How2Sign and OpenASL. † denotes the results reproduced by OpenASL [40]. YT→H2S denotes training on YouTube-ASL and then fine-tuning on How2Sign [44]. The highest performance is highlighted in **bold**.

Method	How2Sign						OpenASL					
	R	Dev B1	B4	R	Test B1	B4	R	Dev B1	B4	R	Test B1	B4
NSLT† [6]	-	-	-	-	-	-	16.25	16.72	4.82	16.10	16.11	4.58
OpenASL [40]	-	-	-	-	-	-	20.43	20.10	6.57	21.02	20.92	6.72
GloFE-N [27]	-	-	-	-	-	-	21.63	21.78	7.51	21.23	20.49	6.82
Álvarez [36]	-	17.73	2.24	-	17.40	2.21	-	-	-	-	-	-
GloFE-VN [27]	12.98	15.21	2.37	12.61	14.94	2.24	21.37	21.06	6.68	21.75	21.56	7.06
Tarrés [42]	-	35.2	8.89	-	34.01	8.03	-	-	-	-	-	-
YT-ASL [44]	-	-	-	-	14.96	1.22	-	-	-	-	-	-
+YT→H2S [44]	-	-	-	-	37.82	12.39	-	-	-	-	-	-
Baseline	8.81	17.29	1.44	8.34	16.71	1.08	31.22	31.08	10.41	30.43	30.51	10.52
VAP	30.27	42.34	14.91	27.77	39.22	12.87	42.47	46.04	21.54	41.38	45.92	21.23

in BLEU-4 score between VAP and gloss-based SOTA [9] narrow to 2.04/2.79 on the Dev/Test sets of PHOENIX14T and 5.34/5.90 on those of CSL-Daily. These results indicate that the alignment identified by VAP can approximate the role of gloss annotation in gloss-based methods. For the larger performance improvements on CSL-Daily, we attribute them to the utilization of mBART [28], which greatly improves fluency on CSL-Daily. Similar observation can be obtained from the comparison between IP-SLT [50] and MMTLB [8] under gloss-based setting.

To further demonstrate the scalability of our approach, we conduct experiments on more challenging How2Sign and OpenASL datasets, where gloss-based methods are inapplicable due to the absence of gloss annotations. As shown in Tab. 2, our VAP continues to maintain its leading position. In detail, VAP achieves better BLEU-4 score (+0.48) than YT-ASL [44] on Test set of How2Sign without pre-training on extra data. These improvements highlight the ability of VAP to leverage data more effectively. Besides, the adopted baseline method achieves superior performance on OpenASL dataset and we attribute it to the usage of skeleton data, which enables the network to better focus on the key signer in the video. Building upon such a strong baseline, our VAP can further improve the BLEU-4 score by 11.13/10.71 on the Dev/Test sets.

4.3 Ablation Study

Effects of Translation Loss in gloss-free SLT. We first illustrate the effects of translation loss in gloss-free SLT by iteratively re-initializing the weights of the translation module in our baseline method. As shown in Tab. 3, the simple act of re-initializing yields significant improvements (+2.17 BLEU-4 after the first iteration), and achieves better performance than previous SOTA gloss-free method (23.75 *vs.* 22.52 of Sign2GPT [49]) without using any pre-trained translation modules. These results indicate the translation loss is helpful to improve the quality of alignment and provide supervision for ψ_V with proper constraints.

Table 3: Ablation results of pre-training objectives. Iter n denotes re-initializing the translation module n times.

$\mathcal{L}_{\text{Align}}$	$\mathcal{L}_{\text{SLT}}$	Dev		Test	
		R	B4	R	B4
		44.01	20.15	42.33	19.72
	iter 1	47.48	22.32	48.12	23.04
	iter 2	48.67	23.75	47.83	22.57
	iter 3	48.12	22.61	48.45	22.65
✓		48.50	23.32	48.80	22.94
✓	✓	50.02	24.38	50.07	24.62

Table 5: Ablation results of gloss-based VAP.

Stage	Dev				Test			
	WER	sub/del/ins	R	B4	WER	sub/del/ins	R	B4
Pre-training	18.6	3.3/8.3/7.1	42.75	19.61	19.5	3.1/7.9/8.5	42.28	19.00
Fine-tuning	24.0	4.0/9.0/11.1	50.98	25.58	23.7	3.6/8.8/11.2	50.10	25.13
+ $\mathcal{L}_{\text{P2T}}$	22.2	3.3/6.3/12.6	50.74	25.69	22.1	3.2/6.5/12.4	50.65	25.61

Table 4: Ablation results of fine-tuning and pseudo-gloss-based training.

mBART	$\mathcal{L}_{\text{P2T}}$	$\mathcal{L}_{\text{Align}}$	Dev		Test	
			R	B4	R	B4
			50.02	24.38	50.07	24.62
✓			51.45	26.23	50.06	25.54
✓	✓		50.94	26.48	50.51	25.50
✓	✓	✓	51.47	26.62	51.28	26.16
Pseudo-gloss-based Training			R	B4	R	B4
$\mathcal{L}_{\text{CTC}}$ (Random init)			49.47	25.04	49.00	24.39
$\mathcal{L}_{\text{CTC}}$ (Embedding init)			52.48	26.15	52.16	26.14
Gloss-based Training			53.14	28.01	51.65	27.00

In detail, we can observe the trend of performance characterized by an initial improvement followed by a subsequent decline with increasing iterations. For the improvement, re-initializing releases the over-fitting of translation module, thereby enabling translation loss to provide more adequate supervision for the visual encoder. The decline of performance can be ascribed to the diminishing of translation module, rendering it incapable of accurately modeling the alignment.

Ablation on Pre-training Objectives. We then evaluate the effects of different pre-training objectives in VAP. To intuitively show the quality of the pre-trained visual encoder, we only use $\mathcal{L}_{\text{SLT}}$ to fine-tune a shallow translation module from scratch. As shown in Tab. 3, utilizing $\mathcal{L}_{\text{Align}}$ to pre-train the visual encoder yields remarkable performance gains compared with directly training from scratch. The improvements can be attributed to the better utilization of visual information through aligning visual and textual tokens. Furthermore, the introduction of $\mathcal{L}_{\text{SLT}}$ can bring additional performance gains, which indicates its ability to ensure the temporal consistency of the established alignment.

Ablation on Fine-tuning. Compared to fine-tuning from scratch, fine-tuning with a translation module pre-trained large-scale corpora is anticipated to achieve better performance, which can be verified from Tab. 4 (+1.85 BLEU-4). Due to the distinct grammar and vocabularies between sign and spoken languages, further improvements are observed by fine-tuning mBART with the generated pseudo gloss sequences $\tilde{g}$ (+0.25 BLEU-4). Moreover, the incorporation of $\mathcal{L}_{\text{Align}}$ results in additional improvements, which underscores the effects continuously enhancing the visual encoder (+0.14 BLEU-4).

Quality of Alignment. As shown in Tab. 4, there is only a slight performance gap (26.15 *vs.* 26.62 BLEU-4) between pseudo-gloss-based model and VAP, which indicates that $\tilde{g}$ aligns well with the sign video and can provide adequate supervision for the visual encoder. We can also observe a small performance gap between

Table 6: Comparison with SOTA retrieval method on PHOENIX14T test set. R@k means recall at rank k. **Table 7:** Ablation results of aggregation method.

Method	T2V			V2T		
	R@1	R@5	R@10	R@1	R@5	R@10
CiCo [10]	69.5	86.6	92.1	70.2	88.0	92.8
VAP (encoder)	62.2	84.1	90.7	69.8	89.7	94.4
VAP (embedding)	66.4	87.7	93.0	56.9	81.8	89.4

Aggregation Method	Dev		Test	
	R	B4	R	B4
Average Pooling	44.04	18.57	42.28	17.13
Attention-Based Pooling	46.20	21.43	46.09	21.06
Max Pooling (VAP)	51.45	26.23	50.06	25.54

pseudo-gloss-based and gloss-based models (26.15 *vs.* 28.01 BLEU-4), indicating there is still room for improvement in alignment. Besides, initializing the classifier with the embedding layer inherited from mBART yields performance gains, which demonstrates the effective transfer of rich semantic knowledge to ψ_V .

To evaluate the quality of alignment method, as shown in Tab. 5, the prediction of gloss-based VAP achieves 18.6/19.5 WER on PHOENIX14T Dev/Test sets, which illustrates VAP can establish reasonable alignments between visual tokens and glosses. Notably, the prediction is **sampled from the ground-truth gloss sequence**, rendering the WER results **incomparable with** those reported in SLR works. Besides, directly fine-tuning with a pre-trained translation module (mBART) will harm the visual encoder in capturing sign words from videos (a higher WER), while fine-tuning the pre-trained translation module with $\mathcal{L}_{P2T}$ firstly prevents disturbing the alignments established during pre-training. For the inferior performance of using actual gloss in VAP, we attribute it to the usage of average gloss embedding and the incompatibility with mBART.

Ablation on Text Encoder. For the text encoder ψ_E , we utilize a simple frozen word embedding layer inherited from mBART, which is different from the Transformer encoder used in [10]. To elucidate the difference between different choices of ψ_E , we employ the alignments established by VAP to conduct sign language retrieval and show the comparison with SOTA method CiCo [10] in Tab. 6. Specifically, $\rho(\mathcal{X}, \mathbf{s})$ in Eq. (6) is used for sign-video-to-text (V2T) retrieval and $\rho(\mathbf{s}, \mathcal{X})$ is used for text-to-sign-video (T2V) retrieval. The variations of performance on T2V and V2T when switching from an embedding layer to an encoder indicate that, rich contextual information can not provide effective supervision for the visual encoder. However, CiCo opts for an encoder layer owing to the benefits of contextual information on the retrieval task. As shown in Tab C in the supplementary material, the performance gap (26.62 *vs.* 25.52 BLEU-4) also indicates the significance of utilizing the pre-trained embedding layer.

Ablation on Aggregation Method. When calculating the video-text similarity, VAP utilizes max pooling to aggregate the token-wise similarity, which is different from the attention-based pooling used in [49] and the average pooling used in [59]. To verify its effects, we compare different aggregation methods and show the results in Tab. 7. These results indicate VAP can better facilitate the utilization of visual information. In contrary, the average pooling and attention-based pooling methods provide insufficient supervision due to the flexible alignment between visual and textual tokens.

PHOENIXLAT	
Text:	am sonntag nur noch <u>schnee</u> im norden zeigt sich teilweise die <u>sonne</u> .
Translation:	am sonntag schneit es kaum noch vor allem richtung norden zeigt sich auch die <u>sonne</u> .
Gloss:	SONNTAG: <u>SCHNEE</u> WENIG: NORD REGION MOEGLICH SONNE
Pseudo Gloss:	am sonntag <u>schnee</u> <u>wenig</u> norden im sich zeigt <u>sonne</u>
CSL-Daily	
Text:	帮我看一下，这件衣服怎么样？
Translation:	你帮帮我看看这件衣服怎么样？
Gloss:	帮帮 帮 帮 这件 衣服 怎么样
Pseudo Gloss:	帮帮 帮 帮帮 怎么样
Text:	<u>我不去爬山</u> ，我有事。
Translation:	我不去爬山，我有事。
Gloss:	<u>山</u> 我不去 <u>爬山</u> 我有事
Pseudo Gloss:	<u>山</u> 我不去 <u>爬山</u> 我有事

Fig. 4: Qualitative results of translation and alignment. The correct parts of translation and pseudo glosses are marked by green and blue, respectively. Gloss sequences and text sentences with varying word orders are highlighted for clarity.



Fig. 5: Visualization of the retrieval results of two German words.

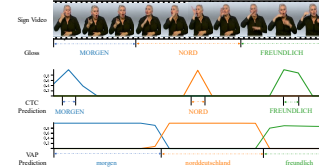


Fig. 6: Visualization of the CTC and VAP predictions.

Qualitative Results. Several results are shown in Fig. 4 to qualitatively evaluate the alignments established by VAP. Although we simply find the corresponding textual token for each visual token based on the similarity, the pseudo gloss sequence generated by VAP aligns well with the ground-truth gloss sequence. Moreover, a sign video with different predictions is visualized in Fig. 6 to show the localization ability of VAP. Different from the CTC prediction, VAP can better find the boundary between two signs, suggesting its potential as a valuable tool for localization. Besides, we utilize VAP to find the video clips corresponding to a given word and visualize the results in Fig. 5. As we can see, VAP can discover sign clips with similar actions among different signers, indicating it can cluster meaningful visual information within sign videos. While these initial findings are interesting, the optimization of alignment and translation quality by leveraging textual information is an ongoing challenge.

5 Conclusion

In this paper, we focus on the utilization of visual information in SLT when gloss annotations are unavailable. We propose a Visual Alignment Pre-training (VAP) scheme to enhance the visual encoder by constructing a gloss-like constraint from text sentences. The VAP approximates the optimal alignment between visual and textual tokens in a greedy manner, while providing frame-wise labels for the visual encoder and improving the compatibility with translation modules pre-trained on large-scale corpora. Experimental results demonstrate the efficacy of the proposed VAP scheme in guiding the learning of visual encoder and establishing meaningful alignment between visual and textual tokens. Moreover, the alignments generated by VAP can be used as pseudo gloss to provide intermediate supervision for a SLT model. We anticipate that the proposed VAP will catalyze future research endeavors, particularly in scaling SLT to large-scale datasets, thereby contributing to advancements in the field of SLT.

Acknowledgments

This work is partially supported by the National Science and Technology Major Project (No. 2021ZD0111901). We also thank the ICT computing platform for providing additional computing resources.

References

1. Albanie, S., Varol, G., Momeni, L., Afouras, T., Chung, J.S., Fox, N., Zisserman, A.: Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. In: European Conference Computer Vision. pp. 35–53. Springer (2020)
2. Bahdanau, D., Cho, K.H., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015 (2015)
3. Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M.L., Kamran, A., Kirefu, F., Koehn, P., et al.: Paracrawl: Web-scale acquisition of parallel corpora. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4555–4567 (2020)
4. Basri, R., Roth, D., Jacobs, D.: Clustering appearances of 3d objects. In: Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 414–420. IEEE (1998)
5. Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* **19**(2), 263–311 (1993), <https://aclanthology.org/J93-2003>
6. Camgoz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7784–7793 (2018)
7. Camgoz, N.C., Koller, O., Hadfield, S., Bowden, R.: 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. pp. 10023–10033 (2020)
8. Chen, Y., Wei, F., Sun, X., Wu, Z., Lin, S.: 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)
9. Chen, Y., Zuo, R., Wei, F., Wu, Y., Liu, S., Mak, B.: Two-stream network for sign language recognition and translation. *Advances in Neural Information Processing Systems* **35**, 17043–17056 (2022)
10. Cheng, Y., Wei, F., Bao, J., Chen, D., Zhang, W.: Cico: Domain-aware sign language retrieval via cross-lingual contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19016–19026 (2023)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houtsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), <https://openreview.net/forum?id=YicbFdNTTy>
12. Dreuw, P., Rybach, D., Deselaers, T., Zahedi, M., Ney, H.: Speech recognition techniques for a sign language recognition system. *hand* **60**, 80 (2007)

13. Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., Torres, J., Giro-i Nieto, X.: How2sign: a large-scale multimodal dataset for continuous american sign language. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 2735–2744 (2021)
14. Garg, S., Peitz, S., Nallasamy, U., Paulik, M.: Jointly learning to align and translate with transformer models. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. pp. 4453–4462 (2019)
15. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: *International Conference on Machine Learning*. pp. 369–376 (2006)
16. Guo, L., Xue, W., Guo, Q., Liu, B., Zhang, K., Yuan, T., Chen, S.: Distilling cross-temporal contexts for continuous sign language recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10771–10780 (2023)
17. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 16000–16009 (2022)
18. Hu, H., Zhao, W., Zhou, W., Li, H.: Signbert+: Hand-model-aware self-supervised pre-training for sign language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023)
19. Hu, L., Gao, L., Liu, Z., Feng, W.: Continuous sign language recognition with correlation network. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 2529–2539 (2023)
20. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: *International Conference on Machine Learning*. pp. 4904–4916. PMLR (2021)
21. Jiao, P., Min, Y., Li, Y., Wang, X., Lei, L., Chen, X.: Cosign: Exploring co-occurrence signals in skeleton-based continuous sign language recognition. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 20676–20686 (2023)
22. Jin, T., Zhao, Z.: Contrastive disentangled meta-learning for signer-independent sign language translation. In: *Proceedings of the 29th ACM International Conference on Multimedia*. pp. 5065–5073 (2021)
23. Kenton, J.D.M.W.C., Toutanova, L.K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of NAACL-HLT*. pp. 4171–4186 (2019)
24. Koller, O., Camgoz, N.C., Ney, H., Bowden, R.: Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **42**(9), 2306–2320 (2019)
25. Li, D., Xu, C., Yu, X., Zhang, K., Swift, B., Suominen, H., Li, H.: Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. *Advances in Neural Information Processing Systems* **33**, 12034–12045 (2020)
26. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: *Text summarization branches out*. pp. 74–81 (2004)
27. Lin, K., Wang, X., Zhu, L., Sun, K., Zhang, B., Yang, Y.: Gloss-free end-to-end sign language translation. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 12904–12916 (2023)

28. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L.: Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics* **8**, 726–742 (2020)
29. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. pp. 1412–1421 (2015)
30. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* **26** (2013)
31. Min, Y., CHEN, X.: Towards faithful sign language translation (2023), <https://openreview.net/forum?id=mWMJNOvbDF>
32. Misra, I., Maaten, L.v.d.: Self-supervised learning of pretext-invariant representations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6707–6717 (2020)
33. Müller, M., Jiang, Z., Moryossef, A., Rios, A., Ebling, S.: Considerations for meaningful sign language machine translation based on glosses. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. pp. 682–693 (2023)
34. Orbay, A., Akarun, L.: Neural sign language translation by learning tokenization. In: *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition*. pp. 222–228. IEEE (2020)
35. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: 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)
36. Patricia, Cabot, Á., Xavier, Giró, N., Laia, Tarrés, B.: Sign language translation based on transformers for the How2Sign dataset (2022), <https://imatge.upc.edu/web/publications/sign-language-translation-based-transformers-how2sign-dataset>
37. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: *Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing*. pp. 1532–1543 (2014)
38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International Conference on Machine Learning*. pp. 8748–8763. PMLR (2021)
39. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* **21**(140), 1–67 (2020)
40. Shi, B., Brentari, D., Shakhnarovich, G., Livescu, K.: Open-domain sign language translation learned from online video. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. pp. 6365–6379 (2022)
41. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: Videobert: A joint model for video and language representation learning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 7464–7473 (2019)
42. Tarrés, L., Gállego, G.I., Duarte, A., Torres, J., Giró-i Nieto, X.: Sign language translation from instructional videos. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. pp. 5625–5635 (June 2023)
43. Tenenbaum, J.B., Silva, V.d., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. *science* **290**(5500), 2319–2323 (2000)

44. Uthus, D., Tanzer, G., Georg, M.: Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023)
45. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., et al.: Image as a foreign language: Beit pre-training for vision and vision-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19175–19186 (2023)
46. Wei, F., Chen, Y.: Improving continuous sign language recognition with cross-lingual signs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23612–23621 (2023)
47. Wenzek, G., Lachaux, M.A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., Grave, É.: Ccnet: Extracting high quality monolingual datasets from web crawl data. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 4003–4012 (2020)
48. Wong, R., Camgoz, N.C., Bowden, R.: Learnt contrastive concept embeddings for sign recognition. In: 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 1937–1946. IEEE Computer Society (2023)
49. Wong, R., Camgoz, N.C., Bowden, R.: Sign2GPT: Leveraging large language models for gloss-free sign language translation. In: The Twelfth International Conference on Learning Representations (2024), <https://openreview.net/forum?id=LqaEEs3UxU>
50. Yao, H., Zhou, W., Feng, H., Hu, H., Zhou, H., Li, H.: Sign language translation with iterative prototype. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15592–15601 (2023)
51. Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., Xu, C.: Filip: Fine-grained interactive language-image pre-training. In: International Conference on Learning Representations (2021)
52. Yin, A., Zhong, T., Tang, L., Jin, W., Jin, T., Zhao, Z.: Gloss attention for gloss-free sign language translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2551–2562 (2023)
53. Yin, K., Read, J.: Better sign language translation with stmc-transformer. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 5975–5989 (2020)
54. Zhang, B., Müller, M., Sennrich, R.: Sltunet: A simple unified model for sign language translation. In: International Conference on Learning Representations (2023)
55. Zhang, H., Guo, Z., Yang, Y., Liu, X., Hu, D.: C2st: Cross-modal contextualized sequence transduction for continuous sign language recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21053–21062 (2023)
56. Zhao, J., Qi, W., Zhou, W., Duan, N., Zhou, M., Li, H.: Conditional sentence generation and cross-modal reranking for sign language translation. *IEEE Transactions on Multimedia* **24**, 2662–2672 (2021)
57. Zheng, J., Wang, Y., Tan, C., Li, S., Wang, G., Xia, J., Chen, Y., Li, S.Z.: Cvt-slr: Contrastive visual-textual transformation for sign language recognition with variational alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23141–23150 (2023)
58. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16793–16803 (2022)

- 59. Zhou, B., Chen, Z., Clapés, A., Wan, J., Liang, Y., Escalera, S., Lei, Z., Zhang, D.: Gloss-free sign language translation: Improving from visual-language pretraining. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20871–20881 (2023)
- 60. Zhou, H., Zhou, W., Qi, W., Pu, J., Li, H.: Improving sign language translation with monolingual data by sign back-translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1316–1325 (2021)
- 61. Zhou, H., Zhou, W., Zhou, Y., Li, H.: Spatial-temporal multi-cue network for sign language recognition and translation. *IEEE Transactions on Multimedia* **24**, 768–779 (2021)