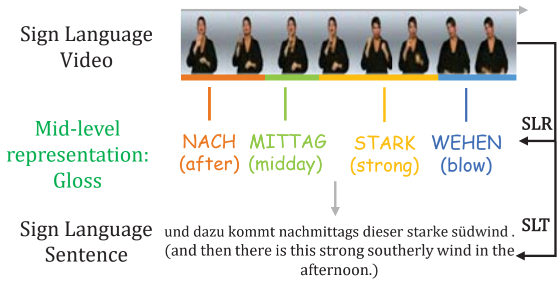
Gloss-free Sign Language Translation: Improving from Visual-Language Pretraining

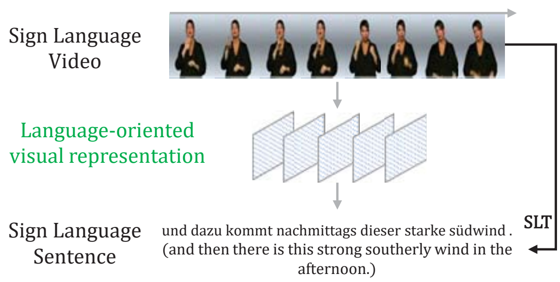
Benjia Zhou[1], Zhigang Chen[2,3]*, Albert Clapés[4,5], Jun Wan[1,2,3]†, Yanyan Liang[1],
Sergio Escalera[4,5,6], Zhen Lei[2,3,7], Du Zhang[1]
[1]MUST, Macau, China; [2]UCAS, China; [3]MAIS, CASIA, China; [4]Universitat de Barcelona, Spain;
[5]Computer Vision Center, Spain; [6]AAU, Aalborg, Denmark; [7]CAIR, HKISI, CAS, Hong Kong, China

Abstract

Sign Language Translation (SLT) is a challenging task due to its cross-domain nature, involving the translation of visual-gestural language to text. Many previous methods employ an intermediate representation, i.e., *gloss sequences, to facilitate SLT, thus transforming it into a two-stage task of sign language recognition (SLR) followed by sign language translation (SLT). However, the scarcity of gloss-annotated sign language data, combined with the information bottleneck in the mid-level gloss representation, has hindered the further development of the SLT task. To address this challenge, we propose a novel* ***G**loss-**F**ree **SLT** based on **V**isual-**L**anguage **P**retraining (**GFSLT-VLP**), which improves SLT by inheriting language-oriented prior knowledge from pre-trained models, without any gloss annotation assistance. Our approach involves two stages: (i) integrating Contrastive Language-Image Pretraining (CLIP) with masked self-supervised learning to create pre-tasks that bridge the semantic gap between visual and textual representations and restore masked sentences, and (ii) constructing an end-to-end architecture with an encoder-decoder-like structure that inherits the parameters of the pre-trained Visual Encoder and Text Decoder from the first stage. The seamless combination of these novel designs forms a robust sign language representation and significantly improves gloss-free sign language translation. In particular, we have achieved unprecedented improvements in terms of BLEU-4 score on the PHOENIX14T dataset (≥+5) and the CSL-Daily dataset (≥+3) compared to state-of-the-art gloss-free SLT methods. Furthermore, our approach also achieves competitive results on the PHOENIX14T dataset when compared with most of the gloss-based methods*[1].



(a) Gloss-based approach.



(b) Gloss-free approach (ours).

Figure 1: **Comparison of Two SLT Approaches:** (a) incorporating gloss sequences as intermediates, *e.g.*, Sign2Gloss2Text (directly) and Sign2Text (indirectly), and (b) excluding gloss information throughout the training and inference process.

1. Introduction

Sign language is the main medium of communication among deaf people. To facilitate effective communication with hard-of-hearing people, developing Sign Language Translation (SLT) techniques is a promising direction. SLT refers to translating sign language into fluent spoken language sentences, which is more challenging than traditional Natural Machine Translation (NMT) due to its

*Benjia Zhou and Zhigang Chen contributed equally to this paper.
†Corresponding author.
[1]https://github.com/zhoubenjia/GFSLT-VLP

cross-domain translation nature and the scarcity of annotated data.

Recently, a growing body of literature [4, 41, 11, 6, 5] has promoted the SLT by directly or indirectly employing the intermediate representations, namely sign glosses. Gloss is a simplified representation of each sign language in continuous video as illustrated in Figure 1a. Although gloss-based methods have significantly improved the SLT performance compared to end-to-end gloss-free approaches (as illustrated in Figure 1b), the former still suffers from the following problems: (i) annotating glosses is a labor-intensive task, which requires fine-grained alignment and labeled by specialists, significantly constraining the scalability of gloss-based SLT methods and (ii) the gloss-based approach introduces an information bottleneck in the mid-level gloss representation [4], which limits the network's ability to understand sign language as the translation model can only be as good as the sign gloss annotations it was trained from.

Drawing inspiration from CLIP [29], which utilizes natural language supervision for image representation learning, we discovered that learning language-indicated visual representation from sign language videos is an effective pre-training task for SLT as it establishes a potential connection between visual signs and language context. However, directly applying CLIP to SLT faces two challenges: (i) it lacks the ability to jointly pretrain the Visual Encoder and Text Decoder for SLT, and (ii) sufficient SLT data is necessary for this pretraining task. To address these obstacles, we must answer two key questions: (i) how can we efficiently perform joint pretraining using the limited SLT dataset? and (ii) how can we ensure that the pretrained model offers the most effective assistance for the downstream SLT task?

To overcome the first challenge, we present a solution that operates at both the algorithmic and data levels. Algorithmically, we introduce a fresh pre-training approach known as VLP (Visual-Language Pretraining), shown in Figure 2a, which combines masked self-supervised learning with CLIP. Specifically, we design a *pretext task* that aligns visual and textual representations in a joint multimodal semantic space, guiding the Visual Encoder to learn language-indicated visual representations. Meanwhile, we incorporate masked self-supervised learning into the pre-training process to help the Text Decoder to capture the syntactic and semantic properties of sign language sentences. At the data level, we investigate a set of strong data augmentation techniques for sign videos to increase the diversity of visual data. This aspect has often been overlooked in previous SLT techniques.

To cope with the second aspect, as depicted in Figure 2b, we design an end-to-end **G**loss-**F**ree **SLT** architecture referred to as GFSLT. This architecture takes the form of an encoder-decoder structure and inherits parameters from the pre-trained Visual Encoder and Text Decoder of the initial phase. GFSLT allows us to directly transform visual representations into spoken sentences without needing any intermediate steps or guidance. Furthermore, unlike alternative methods [4, 42, 5] that solely fine-tune the spatial feature extractor (visual embedding module) in the Visual Encoder, we fine-tune both the spatial feature extractor and the temporal relationship modeling network (Transformer encoder) as a unified whole.

In summary, the main contributions are listed:

- In this work, we have achieved unprecedented improvements in the BLEU-4 score for SLT without using gloss annotations. Specifically, compared with state-of-the-art gloss-free SLT methods, our method has got $\geq$+5 and $\geq$+3 improvements on the PHOENIX14T dataset and CSL-Daily dataset, respectively. We believe that these improvements represent a significant breakthrough in the task of gloss-free SLT.
- To the best of our knowledge, this is the first attempt to introduce the VLP strategy to align visual and textual representations in a joint semantic space in the gloss-free SLT task.
- We propose a novel pre-training paradigm that incorporates masked self-supervised learning together with contrastive language-image pre-training to facilitate the gloss-free SLT task. This approach represents a significant improvement over previous methods and has the potential to greatly enhance the accuracy and efficiency of SLT systems.

2. Related Works

Generally speaking, there are two methods for Sign Language Translation (SLT), namely, gloss-based and gloss-free. Before briefly surveying works along these two directions, we first introduce the Sign Language Recognition (SLR) task as it is an essential step for gloss-based SLT methods.

2.1. Sign Language Recognition

Sign Language Recognition (SLR) consists of two different tasks: Isolated Sign Language Recognition (ISLR) and Continuous Sign Language Recognition (CSLR). The goal of ISLR is to translate an isolated sign into a corresponding single sign language word [14, 15, 20, 22], which is somewhat similar to the isolated gesture recognition task [33, 18, 9, 38, 40, 39, 36]. CSLR is a more challenging task, which is dedicated to recognizing a continuous video of sign language into ordered sign language words, referred to as gloss sequences [4, 6, 11, 13, 17, 27, 42, 43]. Previous SLT work often utilized CSLR as a pre-task to predict gloss

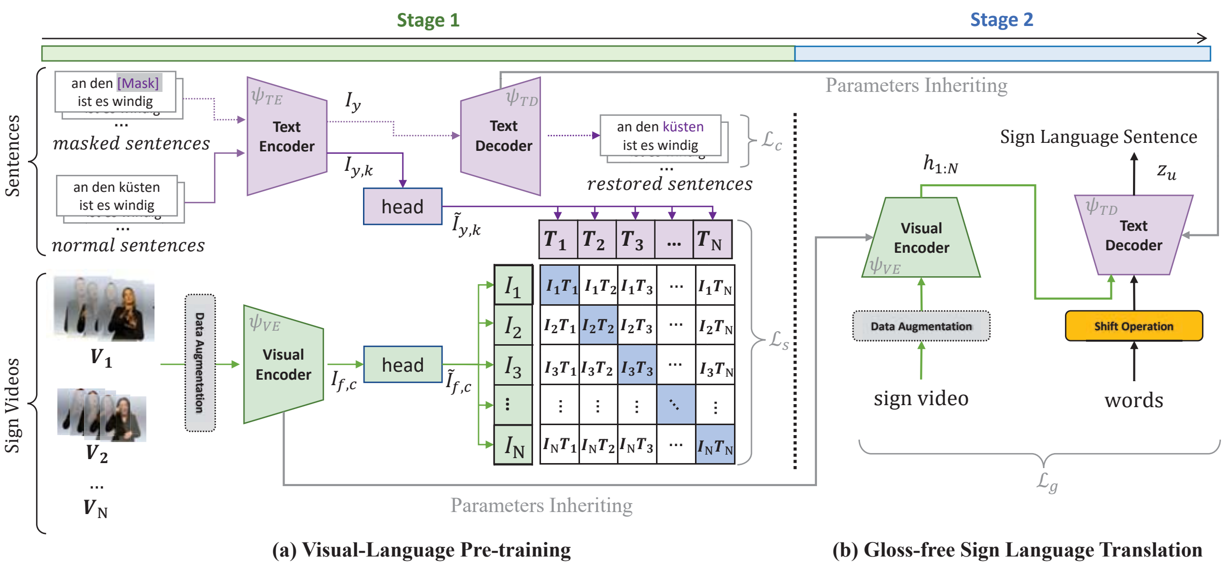


Figure 2: **Method Overview.** GFSLT-VLP improves the SLT by (a) performing Visual-Language Pretraining in stage 1 first, and then (b) transferring parameters of the pre-trained Visual Encoder and Text Decoder in stage 2. Initially, N video-language pairs are individually processed by a visual and text encoder to extract hidden visual and textual features (solid line in the figure). These features are subsequently projected onto a shared multimodal space through dedicated mapping heads to calculate the similarity. Simultaneously, the masked text is fed into a shared text encoder and a text decoder equipped with a causal mask to reconstruct the masked content (dotted line in the figure). Ultimately, the trained visual encoder and text decoder are utilized to train the downstream SLT model.

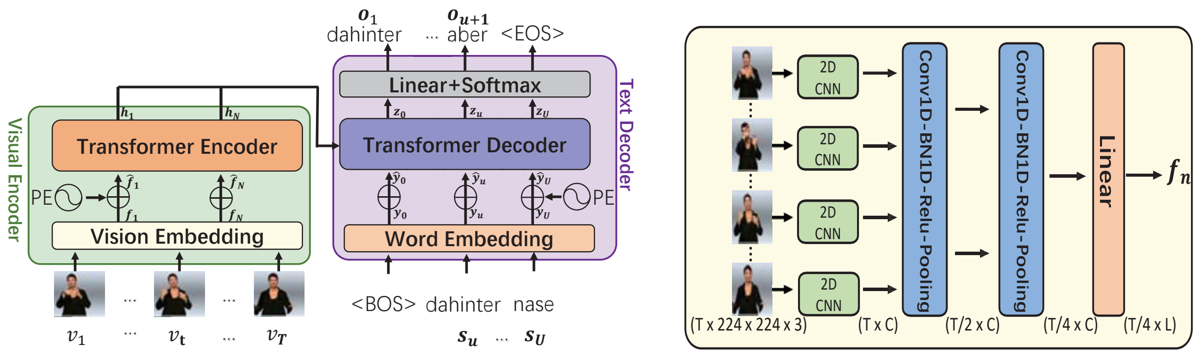
or obtain better visual representations [4, 42, 41, 5, 6]. Such methods often have high requirements on the accuracy of CSLR. In this work, however, we abandon gloss sequences entirely and explore a new gloss-free SLT approach.

2.2. Gloss-based Sign Language Translation

To improve the Sign Language Translation (SLT), several works have employed the mid-level representation of sign glosses. SLRT [4] first introduces a Transformer-based encoder-decoder framework to perform end-to-end SLT. This approach improves performance by using a Connectionist Temporal Classification (CTC) loss to soft-match sign representations and gloss sequences. STMC-T[42] approaches sign language understanding with multi-cue learning. It models sequence information by introducing intra-cue and inter-cue CTC [10] losses. SignBack[41] attempts to introduce advanced machine translation techniques such as back-translation [41] into SLT. Moreover, thanks to the successful application of transfer learning on NMT, Chen *et al.* [5, 6] made the first attempt to introduce large language models into SLT. All the above methods used the Gloss annotation directly or indirectly in SLT model training. However, in our case, we completely abandon the Gloss annotation because its existence limits the scale of sign language datasets. Instead, a more general design of sign language pre-training is introduced in this paper.

2.3. Gloss-free Sign Language Translation

"Gloss-free SLT" refers to the absence of gloss supervision throughout the training and testing, including pre-training and fine-tuning stages. NSLT [3] utilized CNN+RNN for end-to-end SLT, where CNN learned visual features of sign language, and RNN with attention [2, 26] managed sequence and text modeling. TSPNet [21] employed inter-scale and intra-scale attention to enhance visual feature learning by capturing local and global context in sign language videos. GASLT [35] revealed the essential role of gloss annotations for SLT and introduced the gloss-attention to leverage gloss-related benefits. In contrast, CS-GCR [37] aimed to improve the accuracy and fluency of SLT by proposing three modules: word existence verification, conditional sentence generation, and cross-modal re-ranking to learn better grammatical features. However, aligning modalities without gloss guidance is tough due to the significant difference in sign language video and spoken language order. Consequently, gloss-free SLT underperforms gloss-based SLT. In this paper, we adopt a VLP-based approach to obtain better cross-modal representations, effectively reducing the performance gap between gloss-free and gloss-based SLT.



(a) GFSLT Model

(b) Vision Embedding

Figure 3: (a) The framework of the Gloss-free SLT Model, where PE means Positional Encoding. (b) The structure of the Vision Embedding layer.

3. Method

In this paper, we suggest that language-indicated visual representations enjoy both low-redundancy and high-abstract properties of language information that can improve SLT. To this end, we introduce a new pre-training paradigm for SLT that combines masked self-supervised learning with CLIP, allowing us to jointly pre-train the Visual Encoder $\psi_{VE}(\cdot)$ and Text Decoder $\psi_{TD}(\cdot)$ for the downstream GFSLT model (Section 3.1). Subsequently, we transfer the parameters of the pre-trained Visual Encoder $\psi^*_{VE}(\cdot)$ and Text Decoder $\psi^*_{TD}(\cdot)$ to the GFSLT model $\psi_{GFSLT}(\cdot)$ meticulously to enhance its translation capabilities (Section 3.2). Algorithm 1 elaborates our entire algorithm flow.

3.1. Visual-Language Pretraining

In order to learn language-indicated visual representations from sign videos, two crucial issues need to be considered: (i) **how to design a *pretext task* that can effectively reduce the semantic gap between visual and textual representations?** and (ii) **how to achieve jointly pretraining on the limited SLT dataset?**

To cope with the first issue, we draw inspiration from CLIP [29] in the field of zero-shot transfer learning, which developed the "image-to-text" as a standardized input-output interface, allowing for transferable visual models from natural language supervision. CLIP has highlighted the advantages of learning from natural language over other task-agnostic pretraining methods, making it particularly suitable for sign language translation tasks. In other words, learning visual representations through language supervision is a straightforward yet effective pretext task for SLT, given that SLT data inherently has an image-text pair structure. From this insight, we present a new Visual-Language Pre-training scheme, termed VLP, as illustrated in Figure 2a. It jointly trains a Visual Encoder $\psi_{VE}(\cdot)$ and a Text Encoder $\psi_{TE}(\cdot)$ to predict the correct pairings of a batch of (sign video, language sentence) training examples. Formally, the video-text pair is first input into $\psi_{VE}(\cdot)$ and $\psi_{TE}(\cdot)$ in parallel to obtain corresponding high-dimensional semantic features:

$$I_f = \psi_{VE}(V), \quad V = (v_1, ..., v_T);$$
$$I_y = \psi_{TE}(S), \quad S = (s_1, ...s_U); \tag{1}$$

where V is a sign language video with T frames and S is a spoken language sentence with U words.

Visual Encoder: The Visual Encoder consists of a Visual

Algorithm 1 Two-Stage Gloss-free SLT.

Stage1: *Visual-Language Pre-training (VLP)*

Input: Dataset $\mathcal{D} = \{V^{(n)}, S^{(n)}\}_{n=1}^N$
Initialize the parameters Θ_* of $\psi_{VE}(\cdot)$, $\psi_{TD}(\cdot)$ and $\psi_{TE}(\cdot)$
while not converged **do**
  for $V^{(i)}, S^{(i)}$ in $\mathcal{D}$ **do**
    Update the $\psi_{VE}(\cdot)$ by descending $\bigtriangledown\mathcal{L}_s(\Theta_{VE}, V^{(i)}) + \mathcal{L}_s(\Theta_{TE}, S^{(i)})$
    Obtain the masked sentences $\tilde{S}^{(i)}$
    Update the $\psi_{TD}(\cdot)$ by descending $\bigtriangledown\mathcal{L}_c(\Theta_{TD}, \tilde{S}^{(i)})$
  end for
end while
Output: $\psi^*_{VE}(\cdot)$ and $\psi^*_{TD}(\cdot)$

Stage2: *Gloss-Free Sign Language Translation (GFSLT)*

Initialize the parameters Θ_{GFSLT} of $\psi_{GFSLT}(\cdot)$ with $\psi^*_{VE}(\cdot), \psi^*_{TD}(\cdot)$
while not converged **do**
  for $V^{(i)}, S^{(i)}$ in $\mathcal{D}$ **do**
    Update the $\psi_{GFSLT}(\cdot)$ by descending $\bigtriangledown\mathcal{L}_g(\Theta_{GFSLT}, \tilde{S}^{(i)})$
  end for
end while
Output: $\psi^*_{GFSLT}(\cdot)$

Embedding layer, as shown in Figure 3b, followed by a Transformer Encoder with multiple layers. Each frame of the video is first encoded by the weight-sharing 2D CNN layers; The resulting visual encoding is then fed through two temporal blocks, which use a combination of Conv1D-BN-Relu-Maxpooling to capture short-term dependencies. Finally, the features are passed through the Transformer Encoder to capture long-term dependencies in the video.

Text Encoder: To encode text data effectively, it is crucial to have a strong Text Encoder. As a result, we have opted to use the parameters initialized encoder with 12 layers in mBART [24]. This is an NMT model that has been pre-trained on CC25 [24], a multilingual corpus that covers 25 languages.

Then, the captured visual features I_f and textual features I_y are input to the corresponding heads to be linearly projected to the joint multimodal semantic space for similarity computation. Here, both heads are composed of a simple Linear layer. Formally, we can express the process as follows:

$$\tilde{I}_{f,c} = \text{Linear}(I_{f,c}), \tilde{I}_{y,k} = \text{Linear}(I_{y,k}); \quad (2)$$

where $I_{f,c}$ denotes the activation of the last layer of the Visual Encoder at the [CLS] token[2], and $I_{y,k}$ denotes the activation of the last layer of the Text Encoder at the <EOS>token[3]. Subsequently, similar to CLIP [29], $\tilde{I}_{f,c}$ and $\tilde{I}_{y,k}$ are layer-normalized and pairwise-scaled, and then used to calculate the loss value via a symmetric Cross-entropy loss function:

$$\mathcal{L}_s = -\frac{1}{2}\big(\sum V \log(\tilde{I}_{f,c}) + \sum S \log(\tilde{I}_{y,k})\big) \quad (3)$$

To tackle the second question, we take a dual approach at both the algorithm and data levels. At the algorithm level, as illustrated in Figure 2a, we introduce an additional auxiliary supervision stream - masked self-supervised learning (indicated by the dotted line in the figure) - to enable joint pre-training. Wherein the Text Decoder $\psi_{TD}(\cdot)$ takes as input the linguistic features extracted from the masked sentences via the weight-sharing Text Encoder, to predict the words in the masked regions.

Text Decoder: The Text Decoder is actually a standard Transformer decoder [31] with the causal mask. Its optimization goal can be formulated as follows:

$$\min_{\Theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_c\big(\psi_{TD}(\psi^*_{TE}(\tilde{S}^{(i)})), S^{(i)}\big) \quad (4)$$

where $\tilde{S}$ denotes the masked sentences, N is the number of training samples, $\mathcal{L}_c$ is the loss function.

[2]It is a special token that is added to make a global representation of the whole sequence.

[3]The text sequence is bracketed with <BOS>and <EOS>tokens.

This consideration of joint pre-training stems from the fact that different pre-training paradigms can capture different aspects of the data, and combining them can provide a more comprehensive representation of the data [16, 19].

At the data level, we introduce strong data augmentation (implemented by VIDAUG library [7]) for input videos in SLT, including geometric transformation, color space transformation, and temporal transformation. During training, we randomly combine these three augmentation methods to enlarge the data space.

In fact, this paradigm facilitates the Visual Encoder to acquire potent language representation skills that are similar to those of the Text Encoder, resulting in the generation of more robust and representative visual features. This is the reason why acquiring language-indicated visual features for SLT is feasible. Hence, after the Visual Encoder and Text Decoder have established such modeling capability, we employ them to perform the SLT task in the second stage.

3.2. Gloss-free Sign Language Translation

In this section, we present our Gloss-Free SLT (GFSLT) network, which can generate the corresponding sentence S from the given sign video V without any gloss annotation assistance. To achieve this, as illustrated in Figure 3a, we utilize Transformer [31] as the main framework of the model, as it has shown superior performance in Neural Machine Translation (NMT). Initially, the sign video is passed through the Visual Encoder ψ^*_{VE} pretained in the VLP stage to get the hidden semantic vectors:

$$h_{1:M} = \psi^*_{VE}(v_{1:T}) \quad (5)$$

where $M = T/4$. Meanwhile, Text Decoder ψ^*_{TD} pretained in the VLP stage takes the corresponding sentence $S = (s_1, .., s_U)$ along with the last encoder hidden state as input to generate one word at a time:

$$z_u = \psi^*_{TD}(s_{1:u-1}, h_{1:M}) \quad (6)$$

where the first word of a sentence is artificially set to a special flag word <BOS>, and the Transformer Decoder will end the generation until the flag word <EOS>. Finally, we calculate the conditional probability $p(S|V)$ after a Linear and a Softmax layer, and optimize the whole network by minimizing the video-to-sentence cross-entropy loss:

$$p(S|V) = \prod_{u=1}^{U} p(s_u|o_u), \quad o_u = softmax(Wz_u + b) \quad (7)$$

$$\mathcal{L}_g = -\log p(S|V) \quad (8)$$

4. Experiments

4.1. Datasets and Evaluation Metrics

Datasets. We evaluated our proposed method on two widely used SLT datasets: RWTH-PHOENIX-Weather

Method	Publisher	Dev					Test				
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE
Gloss-based											
SLRT [4]	CVPR20	47.26	34.40	27.05	22.38	-	46.61	33.73	26.19	21.32	-
STN-SLT [32]	ICCV21	49.12	36.29	28.34	23.23	-	48.61	35.97	28.37	23.65	-
STMC-T [42]	TMM21	47.60	36.43	29.18	24.09	48.24	46.98	36.09	28.70	23.65	46.65
BN-TIN-Transf.+SignBT [41]	CVPR21	51.11	37.90	29.80	24.45	50.29	50.80	37.75	29.72	24.32	49.54
MMTLB [5]	CVPR22	53.95	41.12	33.14	27.61	53.10	53.97	41.75	33.84	28.39	52.65
TS-SLT [6]	NeurIPS22	**54.32**	**41.99**	**34.15**	**28.66**	**54.08**	**54.90**	**42.43**	**34.46**	**28.95**	**53.48**
Gloss-free											
NSLT [3]	CVPR18	28.10	16.81	11.82	9.12	31.00	27.10	15.61	10.82	8.35	29.70
NSLT [3] + Bahdanau [2]	CVPR18	31.87	19.11	13.16	9.94	31.80	32.24	19.03	12.83	9.58	31.80
NSLT [3] + Luong [26]	CVPR18	31.58	18.98	13.22	10.00	32.60	29.86	17.52	11.96	9.00	30.70
SLRT-GF‡ [4]	CVPR20	-	-	-	-	-	30.88	18.57	13.12	10.19	31.10
TSPNet [21]	NeurIPS20	-	-	-	-	-	36.10	23.12	16.88	13.41	34.96
CSGCR [37]	TMM21	35.85	24.77	18.65	15.08	38.96	36.71	25.40	18.86	15.18	38.85
GASLT [35]	CVPR23	-	-	-	-	-	39.07	26.74	21.86	15.74	39.86
GFSLT (ours)	-	41.97	31.04	24.30	19.84	40.70	41.39	31.00	24.20	19.66	40.93
GFSLT-VLP (ours)	-	**44.08**	**33.56**	**26.74**	**22.12**	**43.72**	**43.71**	**33.18**	**26.11**	**21.44**	**42.49**
Improvement	-	+8.23	+8.79	+8.09	+7.04	+4.76	+4.64	+6.44	+4.25	+5.70	+2.63

Table 1: Experimental results on PHOENIX14T dataset. Wherein ‡ denotes results reproduced by [35]. We **bold** the best results in the gloss-based setting and gloss-free setting. Improvement represents the result of comparison with the latest gloss-free methods.

Method	Publisher	Dev					Test				
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE
Gloss-based											
SLRT* [4]	CVPR20	37.47	24.67	16.86	11.88	37.96	37.38	24.36	16.55	11.79	36.74
BN-TIN-Transf. [41]	CVPR21	40.66	26.56	18.06	12.73	37.29	40.74	26.96	18.48	13.19	37.67
BN-TIN-Transf.+SignBT [41]	CVPR21	51.46	37.23	27.51	20.80	49.49	51.42	37.26	27.76	21.34	49.31
MMTLB [5]	CVPR22	53.81	40.84	31.29	24.42	53.38	53.31	40.41	30.87	23.92	53.25
TS-SLT [6]	NeurIPS22	**55.21**	**42.31**	**32.71**	**25.76**	**55.10**	**55.44**	**42.59**	**32.87**	**25.79**	**55.72**
Gloss-free											
TSPNet‡ [21]	NeurIPS20	-	-	-	-	-	17.09	8.98	5.07	2.97	18.38
SLRT† [4]	CVPR20	21.03	9.97	5.96	4.04	20.51	20.00	9.11	4.93	3.03	19.67
GASLT [35]	CVPR23	-	-	-	-	-	19.90	9.94	5.98	4.07	20.35
NSLT [3] + Luong* [26]	CVPR18	34.22	19.72	12.24	7.96	34.28	34.16	19.57	11.84	7.56	34.54
GFSLT (ours)	-	37.60	23.30	14.89	9.92	35.42	37.69	23.28	14.93	9.88	35.16
GFSLT-VLP (ours)	-	**39.20**	**25.02**	**16.35**	**11.07**	**36.70**	**39.37**	**24.93**	**16.26**	**11.00**	**36.44**
Improvement	-	+4.98	+5.30	+4.11	+3.11	+2.42	+5.21	+5.36	+4.42	+3.44	+1.90

Table 2: Experimental results on CSL-Daily dataset. $\star$ denotes results reproduced by [41]; ‡ denotes results reproduced by [35]; and † denotes our reproduced result under the gloss-free setting.

2014T [3] and CSL-Daily [41]. PHOENIX-2014T contains 8257 parallel German sign language (DGS) videos with German translations from weather forecast programs, split into train, dev, and test sets of sizes 7096, 519, and 642 respectively. The German translations have a vocabulary size of 2887. CSL-Daily focuses on daily topics in Chinese sign language, containing 20654 parallel CSL videos with Chinese translations. The dataset is split into train, dev, and test sets of sizes 18401, 1077, and 1176, respectively, and the Chinese translations have a vocabulary size of 2343.

Evaluation Metrics. Following previous works [4, 41, 5, 6], we adopt BLEU [28] and ROUGE [23] to evaluate SLT. Higher BLEU and ROUGE-L indicate better translation performance.

4.2. Implementation details

GFSLT Model. We used ResNet18 [12] pre-trained on ImageNet [8] as our 2D-CNN. For the temporal blocks, we followed the configuration of [41], using a stride size of 1/2 and a kernel size of 5/2 for the Conv1D/Maxpooling layers. Our Transformer encoder and decoder both have 3 layers, with a hidden size of 1024 and a feed-forward size of 4096. Each layer has 8 attention heads, and we set the dropout to 0.1 to avoid overfitting.

VLP	Aug-S1	Aug-S2	Dev BLEU-1	Dev BLEU-2	Dev BLEU-3	Dev BLEU-4	Test BLEU-1	Test BLEU-2	Test BLEU-3	Test BLEU-4
✗	✗	✗	41.97	31.04	24.30	19.84	41.39	31.00	24.20	19.66
✓	✗	✗	41.50	31.26	24.64	20.12	41.81	31.34	24.40	19.77
✓	✓	✗	42.19	32.49	26.28	22.05	42.09	32.01	25.60	21.23
✗	✗	✓	41.84	30.98	24.12	19.65	40.57	29.59	22.73	18.41
✓	✓	✓	**44.08**	**33.56**	**26.74**	**22.12**	**43.71**	**33.18**	**26.11**	**21.44**

Table 3: Effect of VLP and data augmentation strategies. VLP: Visual-Language Pre-training, Aug-S1: strong data augmentation employed during stage 1 for sign video, Aug-S2: strong data augmentation employed during stage 2 for sign video.

Visual-Language Pretraining. We conduct respective pretraining tasks on the training sets of the two sign language datasets. The mini-batch size is set to 16 (we use AMP [1] technology to expand the batch size). The input sequences are first resized into 256×256, and then randomly/centrally cropped into 224×224 during training/inference. We employ SGD with 0.9 momentum as the optimizer and the learning rate is decayed with a cosine schedule [25] from 0.01 (maximum) to 1e-5 (minimum). The training lasts for 80 epochs.

SLT Training and Inference. The GFSLT network is trained end-to-end using cross-entropy loss with label smoothing of 0.2 and a mini-batch size of 8. We used SGD optimizer [30] with 0.9 momentum and initialized the learning rate to 0.01 with the cosine annealing scheduler. The network is trained for 200 epochs. During inference, decoding is performed using the beam search strategy with a length penalty [34] of 1, and a beam size of 5 is employed.

4.3. Comparison with State-of-the-art Methods

Results on PHOENIX14T dataset. Table 1 presents a comparison of our approach with state-of-the-art gloss-based and gloss-free methods for sign language translation. Our method achieves a significant performance gain when compared to other gloss-free approaches, such as CSGCR [37]. Specifically, our method improves the BLEU-4 score by approximately $+7.0$ (on Dev set) and $+5.7$ (on Test set), and improves the ROUGE score by approximately $+4.8$ (on Dev set) and $+2.6$ (on Test set). Moreover, our results are highly competitive when compared to most gloss-based methods. Notably, our method achieves competitive performance with SLRT [4] (22.38 $vs.$ 22.12) and STMC-T [42] (24.09 $vs.$ 22.12), highlighting its potential.

Results on CSL-Daily dataset. Table 2 compares our method with the state-of-the-art approaches on the CSL-Daily dataset. CSL-Daily is a large Chinese sign language dataset released in 2021 by [41] and there are therefore only a few methods that tested on it, especially gloss-free ones. Note that the result of SLRT[4] and NSLT+Loung[3, 26] are reproduced by [41]. As it can be seen, our method surpasses the gloss-free method NSLT+Loung[3, 26] in all metrics, especially improving the BLEU-4 score about $3.2_{\pm 0.1}$ and ROUGE score about $2.2_{\pm 0.2}$ on this dataset. Furthermore, compared with gloss-based methods, we are close to the SLRT[4] and BN-TIN-Transf[41] which did not use semi-supervised back-translation auxiliary training unlike BN-TIN-Transf+SignBT[41], large model transfer training such as MMTLB[5] or a multi-stream model like the one from TS-SLT[6].

4.4. Ablation Studies

The ablation studies were conducted mainly on the PHOENIX14T dataset, with a primary focus on improving the BLEU-4 score as it is the most reliable measure of SLT accuracy. Additionally, unless stated otherwise, we utilized the configuration outlined in Section 4.2 as the baseline settings for our network.

Visual-Language Pretraining. In our investigation of VLP, we delved into the key factors that affect its efficacy and discovered several phenomena. Firstly, from Table 3, we observed that data augmentation on sign videos plays a significant role in the success of VLP. Specifically, when utilizing lightweight data augmentation such as random cropping, the improvement of VLP for SLT is limited, only increasing the BLEU-4 score by about $+0.3$ on the Dev set and $+0.1$ on the Test set. However, when combined with strong data augmentation, VLP significantly enhances the SLT task, improving the BLEU-4 score from 19.84 to 22.05 (+2.2). This emphasizes the data-hungry nature of VLP. Secondly, we observed that without VLP, relying solely on strong data augmentation in stage 2 does not provide much benefit for SLT and may even impair performance slightly. But when combined with VLP, SLT performance can be continuously improved. This is because aggressive data augmentation methods may introduce excessive variations or distortions to the training data, which may pose challenges for the SLT model to adapt to the distribution of the augmented data. However, the VLP stage leverages the LLM to encourage the Visual Encoder to adapt to the distribution differences introduced by the augmented

Visual Encoder		T-Decoder	Dev				Test			
V-Embedding	T-Encoder		BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEU-1	BLEU-2	BLEU-3	BLEU-4
✗	✗	✗	41.97	31.04	24.30	19.84	41.39	31.00	24.20	19.66
✓	✗	✗	41.31	30.88	24.25	19.83	40.12	30.03	23.34	18.93
✗	✓	✗	42.75	32.31	25.65	21.23	42.94	32.68	25.83	21.25
✓	✓	✗	43.30	32.52	26.31	22.07	43.29	32.74	25.96	21.43
✗	✗	✓	42.27	31.68	25.14	20.70	41.55	31.11	24.56	20.27
✓	✓	✓	**44.08**	**33.56**	**26.74**	**22.12**	**43.71**	**33.18**	**26.11**	**21.44**

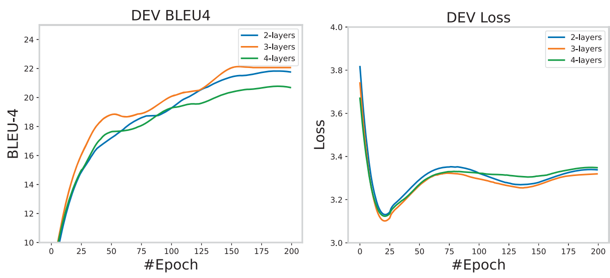
Table 4: Investigating the impact of fine-tuning individual components within the Visual-Language-Pretrain (VLP) framework. V-Embedding: Visual Embedding module; T-Encoder: Transformer Encoder; T-Decoder: Transformer Decoder. Notations ✗ and ✓ denote the initialization of corresponding layers with random parameters and pre-trained parameters, respectively.

period	40 epoch	80 epoch	160 epoch	200 epoch
fixing the training time (80 epochs) of stage 1.				
stage 2	18.23	20.42	21.13	**22.12**
fixing the training time (200 epochs) of stage 2.				
stage 1	20.62	22.12	**22.13**	22.07

Table 5: Effect of longer training regimes. We explore the optimal training time for both stages by fixing the training time of stage 1 to investigate the optimal training time of stage 2 and vice versa in this table.

data, helping the downstream SLT model develop the ability to generalize from the augmented data. Additionally, from Table 4, we find that fine-tuning the Visual Embedding module and Transformer Encoder as a unified whole can result in significant performance gains (+2.23) compared to fine-tuning them separately (-0.01 and +1.39, respectively). Finally, we observe that fine-tuning the Text Decoder can also bring some gains, but it seems limited ($\leq$ 1). These results confirm that a good visual feature is critical to Gloss-Free SLT. The VLP strategy facilitates the learning of low-redundancy and high-abstract features present in language representations by the Visual Encoder, which makes it a crucial component of the system.

Investigation of Training Time. The training time of gloss-based SLT models typically does not exceed 100 epochs. However, as shown in Table 5, with a fixed pre-training time, the gloss-free SLT model requires a longer training regime ($>$ 100 epochs) to achieve satisfactory performance. This is because without the aid of intermediate representation, the convergence speed of the network is reduced, necessitating more training time to make the model fit the desired effect. Moreover, we investigated the influence of pre-training duration on model performance. As observed, it doesn't seem necessary to have an extended pre-training duration. 80 epochs appears to be a trade-off



(a) BLEU-4 score on Dev data. (b) Loss value on Dev data.

Figure 4: Analysis of the impact of model parameter size. Increasing the network depth (to 4 layers) did not yield any positive results, but instead exacerbated model overfitting.

between the two sign language datasets.

Impact of Model Parameter Size. It is widely acknowledged that the size of the network parameters has a significant effect on the ultimate performance of the model, and a more intuitive perception is that the deeper the network, the better the performance. Nevertheless, for the GFSLT network, we noticed that adding network layers would cause more severe overfitting as shown by the green line in Figure 4. We attribute this to the limited scale of SLT data, suggesting that a sufficiently large SLT dataset may be able to alleviate this issue.

Impact of Freezing the Text Encoder. Considering that the Text Encoder is derived from the pre-trained Mbart [24], so in this experiment, we attempted to freeze its parameters and use it as a teacher model to supervise the learning of the Visual Encoder. Contrary to our expectations, this pre-training strategy did not produce satisfactory results, as shown in Table 6. We hypothesize that the reason for this may be that the text and visual features have fundamentally different underlying representations, and they must be optimized to a common representation for meaningful compar-

V-Encoder	T-Encoder	Dev BLEU-1	Dev BLEU-2	Dev BLEU-3	Dev BLEU-4	Test BLEU-1	Test BLEU-2	Test BLEU-3	Test BLEU-4
update	*freeze*	40.28	29.83	22.13	18.32	40.81	29.32	21.24	16.93
update	*update*	**44.08**	**33.56**	**26.74**	**22.12**	**43.71**	**33.18**	**26.11**	**21.44**

Table 6: Analyze the impact of freezing the Text Encoder during the pretraining stage. *update* means updating the network parameters, and *freeze* means freezing the network parameters.

isons and analysis. As a result, directly freezing the parameters of the Text Encoder may not provide sufficient guidance for the Visual Encoder to learn robust representations.

5. Qualitative Results

Reference:	sonst regnet es teilweise kräftig (Otherwise it rains heavily at times)
GFSLT:	sonst regnet es hier und da (Otherwise it rains here and there)
GFSLT-VLP:	sonst regnet es teilweise kräftig (Otherwise it rains heavily at times)
Reference:	am tag wechseln sonne und wolken einander ab teilweise ist es auch längere zeit sonnig (During the day sun and clouds alternate partly it is sunny for a long time)
GFSLT:	am tag sonne und wolken im wechsel nur vereinzelt schauer (During the day sun and clouds alternate only sporadic showers)
GFSLT-VLP:	am tag wechseln sonne und wolken einander ab es bilden sich längere zeit viel sonnenschein (During the day, sun and clouds alternate There is a lot of sunshine for a long time)
Reference:	am tag nur hier und da einige sonnige momente vor allem an den alpen (During the day only here and there some sunny moments, especially in the Alps)
GFSLT:	gebietsweise zeigt sich morgen häufig die sonne (In some areas, the sun will often show up tomorrow)
GFSLT-VLP:	morgen zeigt sich mal die sonne wenn dann vor allem an den alpen (Tomorrow the sun will show up , especially in the Alps)
Reference:	und nun die wettervorhersage für morgen sonntag den zwölften dezember (And now the weather forecast for tomorrow Sunday the twelfth of December)
GFSLT:	und nun die wettervorhersage für morgen sonntag den zwölften november (And now the weather forecast for tomorrow Sunday the twelfth of November)
GFSLT-VLP:	und nun die wettervorhersage für morgen sonntag den zwölften dezember (And now the weather forecast for tomorrow Sunday the twelfth of December)

Table 7: Qualitative results of PHOENIX14T. We highlight the difference between sentences. Green means totally same as the reference. Yellow means correct but different words. Red means totally wrong.

We visually demonstrate our model's performance on several sign language videos from PHOENIX14T test set in Table 7. While both models understand the general meaning of sign language videos and produce complete sentences, the baseline model is more error-prone on some keywords, resulting in drastically different translations (first and second rows). Additionally, the VLP model outperforms the baseline in recognizing named entities, accurately translating place names and months (third and fourth rows).

6. Conclusion and Future work

In this work, we propose a new perspective for the gloss-free SLT task by reducing the semantic gap between visual and textual representations, which enables us to learn language-indicated visual representations from sign videos. To achieve this, we introduce a novel pre-training paradigm that combines masked self-supervised learning with visual-language supervision learning. Our experiments reveal that both data scale and model parameters have a significant impact on the performance of this method. While our proposed pre-training paradigm is a crucial step towards gloss-free SLT, we acknowledge that further research is needed, especially in pre-training on a large-scale SLT dataset (without gloss annotations). We hope that our work will inspire future research in this area.

Acknowledgement

This work was supported by the National Key Research and Development Plan under Grant 2021YFE0205700, the External cooperation key project of Chinese Academy Sciences 173211KYSB20200002, the Science and Technology Development Fund of Macau Project (0004/2020/A1, 0123/2022/A3, 0070/2020/AMJ), Guangdong Provincial Key R&D Programme: 2019B010148001, CCF-Zhipu AI Large Model OF 202219, Open Research Projects of Zhejiang Lab No. 2021KH0AB07, InnoHK program, and has been partially supported by the Spanish project PID2022-136436NB-I00 and by ICREA under the ICREA Academia programme.

References

[1] Marc Baboulin, Alfredo Buttari, Jack Dongarra, Jakub Kurzak, Julie Langou, Julien Langou, Piotr Luszczek, and Stanimire Tomov. Accelerating scientific computations with mixed precision algorithms. *Computer Physics Communications*, 180(12):2526–2533, 2009. 7

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2015. 3, 6

[3] 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*, pages 7784–7793, 2018. 3, 6, 7

[4] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 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*, pages 10023–10033, 2020. 2, 3, 6, 7

[5] 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*, pages 5120–5130, 2022. 2, 3, 6, 7

[6] Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie LIU, and Brian Mak. Two-stream network for sign language recognition and translation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 2, 3, 6, 7

[7] darpa-sail on. Video augmentation techniques for deep learning. https://github.com/darpa-sail-on/videoaug, 2021. 5

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[9] Sergio Escalera, Jordi Gonzàlez, Xavier Baró, and Jamie Shotton. Guest editors' introduction to the special issue on multimodal human pose recovery and behavior analysis. 38(8):1489–1491, 2016. 2

[10] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006. 3

[11] Aiming Hao, Yuecong Min, and Xilin Chen. Self-mutual distillation learning for continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11303–11312, 2021. 2

[12] 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*, pages 770–778, 2016. 6

[13] Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Temporal lift pooling for continuous sign language recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 511–527. Springer, 2022. 2

[14] Alfarabi Imashev, Medet Mukushev, Vadim Kimmelman, and Anara Sandygulova. A dataset for linguistic understanding, visual evaluation, and recognition of sign languages: The k-rsl. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 631–640, 2020. 2

[15] Hamid Reza Vaezi Joze and Oscar Koller. Ms-asl: A large-scale data set and benchmark for understanding american sign language. In *British Machine Vision Conference*, 2019. 2

[16] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018. 5

[17] Oscar Koller, Necati Cihan Camgoz, Hermann Ney, and Richard Bowden. 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. 2

[18] Jakub Konečnỳ and Michal Hagara. One-shot-learning gesture recognition using hog-hof features. *The Journal of Machine Learning Research*, 15(1):2513–2532, 2014. 2

[19] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671, 2019. 5

[20] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1459–1469, 2020. 2

[21] Dongxu Li, Chenchen Xu, Xin Yu, Kaihao Zhang, Benjamin Swift, Hanna Suominen, and Hongdong Li. Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. *Advances in Neural Information Processing Systems*, 33:12034–12045, 2020. 3, 6

[22] Dongxu Li, Xin Yu, Chenchen Xu, Lars Petersson, and Hongdong Li. Transferring cross-domain knowledge for video sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6205–6214, 2020. 2

[23] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6

[24] 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*, 8:726–742, 2020. 5, 8

[25] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 7

[26] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. In *Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, 2015. 3, 6, 7

[27] Yuecong Min, Aiming Hao, Xiujuan Chai, and Xilin Chen. Visual alignment constraint for continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11542–11551, 2021. 2

[28] 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*, pages 311–318, 2002. 6

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 4, 5

[30] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951. 7

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[32] Andreas Voskou, Konstantinos P Panousis, Dimitrios Kosmopoulos, Dimitris N Metaxas, and Sotirios Chatzis. 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, 2021. 6

[33] Jun Wan, Qiuqi Ruan, Wei Li, and Shuang Deng. One-shot learning gesture recognition from RGB-D data using bag of features. 14(1):2549–2582, 2013. 2

[34] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016. 7

[35] Aoxiong Yin, Tianyun Zhong, Li Tang, Weike Jin, Tao Jin, and Zhou Zhao. Gloss attention for gloss-free sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2551–2562, 2023. 3, 6

[36] Zitong Yu, Benjia Zhou, Jun Wan, Pichao Wang, Haoyu Chen, Xin Liu, Stan Z Li, and Guoying Zhao. Searching multi-rate and multi-modal temporal enhanced networks for gesture recognition. *IEEE Transactions on Image Processing*, 2021. 2

[37] Jian Zhao, Weizhen Qi, Wengang Zhou, Nan Duan, Ming Zhou, and Houqiang Li. Conditional sentence generation and cross-modal reranking for sign language translation. *IEEE Transactions on Multimedia*, 24:2662–2672, 2021. 3, 6, 7

[38] Benjia Zhou, Yunan Li, and Jun Wan. Regional attention with architecture-rebuilt 3d network for rgb-d gesture recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3563–3571, May 2021. 2

[39] Benjia Zhou, Pichao Wang, Jun Wan, Yanyan Liang, and Fan Wang. A unified multimodal de- and re-coupling framework for rgb-d motion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–15, 2023. 2

[40] Benjia Zhou, Pichao Wang, Jun Wan, Yanyan Liang, Fan Wang, Du Zhang, Zhen Lei, Hao Li, and Rong Jin. Decoupling and recoupling spatiotemporal representation for rgb-d-based motion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20154–20163, June 2022. 2

[41] Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1316–1325, 2021. 2, 3, 6, 7

[42] Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. Spatial-temporal multi-cue network for sign language recognition and translation. *IEEE Transactions on Multimedia*, 24:768–779, 2021. 2, 3, 6, 7

[43] Ronglai Zuo and Brian Mak. C2slr: Consistency-enhanced continuous sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5131–5140, 2022. 2