A Simple Multi-Modality Transfer Learning Baseline for Sign Language Translation

Yutong Chen[1*]   Fangyun Wei[2]   Xiao Sun[2]   Zhirong Wu[2]   Stephen Lin[2]

[1]Tsinghua University   [2]Microsoft Research Asia

chenytjudy@gmail.com   {fawe, xias, wuzhiron, stevelin}@microsoft.com

Abstract

This paper proposes a simple transfer learning baseline for sign language translation. Existing sign language datasets (e.g. PHOENIX-2014T, CSL-Daily) contain only about 10K-20K pairs of sign videos, gloss annotations and texts, which are an order of magnitude smaller than typical parallel data for training spoken language translation models. Data is thus a bottleneck for training effective sign language translation models. To mitigate this problem, we propose to progressively pretrain the model from general-domain datasets that include a large amount of external supervision to within-domain datasets. Concretely, we pretrain the sign-to-gloss visual network on the general domain of human actions and the within-domain of a sign-to-gloss dataset, and pretrain the gloss-to-text translation network on the general domain of a multilingual corpus and the within-domain of a gloss-to-text corpus. The joint model is fine-tuned with an additional module named the visual-language mapper that connects the two networks. This simple baseline surpasses the previous state-of-the-art results on two sign language translation benchmarks, demonstrating the effectiveness of transfer learning. With its simplicity and strong performance, this approach can serve as a solid baseline for future research.

1. Introduction

Sign languages are visual signals for communication among the deaf and hard of hearing. These languages are primarily expressed through manual articulations, but are also greatly aided by the movement of body, head, mouth, eyes and eyebrows. While technology for automatic machine translation of spoken languages have successfully been deployed in production [8,25,33,44], research on sign language translation (SLT) lags behind and is still in early-stage development. An effective system for automatic sign language translation may help to build a bridge between
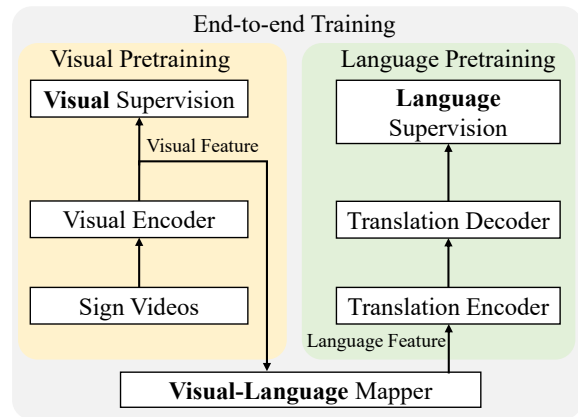

Figure 1. We decouple sign language translation into a visual task (left part) and a language task (right part), and propose a visual-language mapper (V-L Mapper) to bridge the connection between them. The decoupling allows both the visual and language networks to be effectively and independently pretrained before joint training. Both spatio-temporal information from sign videos and semantic knowledge from text transcriptions are encoded through VL-Mapper.

hearing-impaired and unimpaired people.

Existing sign language translation methods follow the framework of neural machine translation (NMT) originally developed for spoken languages [4–6, 46, 49, 50], with the distinction that the source language is represented as spatio-temporal pixels instead of discrete tokens. To be concrete, sign videos are first fed into a video backbone network to extract an intermediate representation, which is then mapped to the target language text via NMT. The intermediate representation is usually supervised by glosses[1] [6, 49,50], where each gloss corresponds to the semantic meaning of a single sign (e.g. happy, sad) in the continuous video input.

Despite adopting the formulation of advanced neural machine translation, the current results are far from sat-

---

*Accomplished during Yutong Chen's internship at MSRA.

[1]Glosses are the word-for-word transcription of sign language where each gloss is a unique label for a sign. Typically, we identify each gloss by a capitalized word which is loosely associated with the sign's meaning.

isfactory. The best reported sign language translation performance [49] on the PHOENIX-Weather-2014T test dataset [4] is 24.32 in terms of BLEU-4, while a baseline transformer achieves a 30.9 BLEU-4 score for English to German translation [33]. We hypothesize that the key factor that hinders the progress of sign language translation is the scale of the training data. To effectively train a typical NMT model, it usually requires a corpus of 1M paralleled samples [42]. However, existing sign language datasets are an order of magnitude smaller, containing only fewer than 20K paralleled samples [4, 49].

In this paper, we study a multi-modal pretraining approach to cope with the data scarcity issue for sign language translation. While pretraining and transfer learning has greatly improved performance in tasks of vision [11, 17, 34], language [10, 25, 27, 33, 40] and cross-modality [26, 31, 35, 39, 47], they are still under explored in SLT. Our work aims to exploit their strength in SLT.

SLT can be broken down into two disjoint tasks: a visual action recognition task that converts sign videos into semantic glosses (Sign2Gloss), and a language translation task that maps glosses into spoken language texts (Gloss2Text). Our transfer learning approach progressively pretrains each task separately and then finetunes the joint model. For Sign2Gloss, we first pretrain the visual model on a general domain to learn generic human actions [21, 28], and then we transfer it to within the domain to learn fine-grained glosses. Similarly for Gloss2Text, we adopt mBART [33], a denoising auto-encoder pretrained on a large-scale general-domain multilingual corpus, and transfer it to the within-domain task of gloss-to-text translation. By leveraging existing datasets and supervisions that can effectively transfer to sign language translation, the necessity of gathering large parallel data is lessened.

With well-trained Sign2Gloss and Gloss2Text modules, we can build a two-staged pipeline known as Sign2Gloss2Text to generate a gloss sequence from the video and then translate the predicted gloss sequence into text. This two-staged pipeline is also implemented in [4, 6, 46, 49] and shows promising results. However, glosses are discrete representations of the language modality, without encoding any spatio-temporal visual information from sign videos such as facial expressions[2], which may lead to degraded translation performance. For example, hearing-impaired individuals use exaggerated facial expressions to convey the adverb 'Extremely', but this kind of information is ignored in gloss annotation. In contrast, labelers and linguists have to take into account these adverbs to produce translated sentences that are complete and semantically accurate. Incorporation of both visual and language modalities is thus needed.

To this end, we introduce a visual-language mapper which connects the visual features before gloss classification in the visual model to the gloss embedding in the translation model. With this mapper, the full model is jointly optimized and the discrete gloss representation is circumvented in joint training. The mapper is simply implemented as a fully connected MLP with two hidden layers. Figure 1 shows our design.

In contrast to previous works which attempt to improve translation performance by integrating multiple cues from mouthing or pose in a handcrafted manner [5, 50] or by adopting advanced machine translation techniques such as back-translation [49], our overall framework is extremely simple, resulting in a transfer learning approach on top of a standard NMT model. Some previous works conduct transfer learning for SLT by pretraining visual backbone on human action recognition [29] or loading pretrained word embeddings [29, 46], while we are the first to adopt both general-domain and within-domain pretraining in a progressive manner and incorporate pretrained spoken language model into SLT. Our experimental results demonstrate that this progressive pretraining of visual and translation models greatly boosts performance. Our simple approach surpasses all existing methods by large margins, including those that employ semi-supervised learning, on PHOENIX-2014T [4] and CSL-Daily [49].

2. Related Work

Sign Language Recognition. A fundamental task in sign language understanding is Isolated Sign Language Recognition (ISLR), which aims to identify a single gloss word label for a short video clip [2, 19, 20, 28, 30, 43]. The more challenging task of Continuous Sign Language Recognition (CSLR) seeks to convert a continuous sign video into a gloss sequence using only weak sentence-level annotations [9, 22, 23, 37, 50]. Our work fully exploits gloss annotations for SLT by transferring within-domain knowledge from ISLR to CSLR and SLT.

Sign Language Translation. Sign Language Translation (SLT) aims to translate a raw video sequence to a spoken language sentence [4–6, 12, 29, 46, 49, 50]. Existing works attempt to formulate this task as a neural machine translation (NMT) problem. However, unlike NMT which benefits from a large-scale parallel corpus, SLT greatly suffers from data scarcity. To tackle this issue, [6] jointly trains SLR and SLT to enforce regularization on the translation encoder; [49] proposes a data augmentation strategy of back-translating text to visual features using glosses as the pivot. Moreover, [5, 50] manually design sophisticated multi-cue channels to model the collaboration of multiple visual cues in sign language, and [29] introduces a temporal semantic pyramid network to capture multiple levels of temporal granularity in sign videos. Compared to these ef-

---

[2]In sign languages, facial expressions are used to express both linguistic information and emotions.
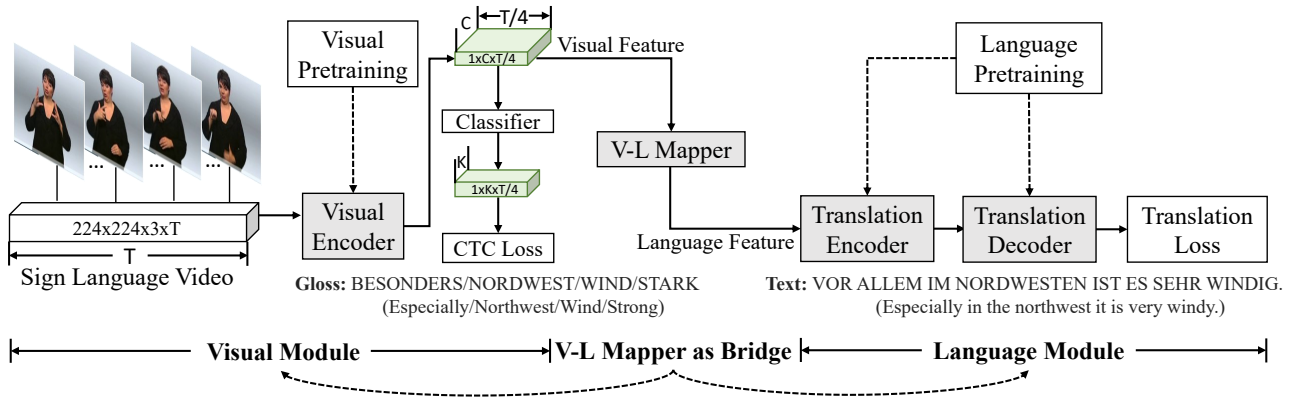
Figure 2. Overview of our framework. We decouple sign language translation into a visual task and a language task. The proposed visual-language mapper (V-L Mapper) establishes a bridge between features of the visual modality and language modality for end-to-end training. The decoupling allows both visual and language networks to be progressively and independently pretrained.

forts, our method is simple yet more effective by utilizing a large amount of external supervision through progressive pretraining.

Action Recognition. A related research field that may facilitate visual modeling of sign language is action recognition, where many works focus on network architecture [7, 13, 14, 38, 45] and large-scale dataset construction [15, 18, 21]. As fine-grained gesture understanding is a special case of human action recognition, some works for ISLR [2, 28, 30] and SLT [29] initialize their visual network with weights pretrained on action classification task. We employ general-domain pretraining on action recognition together with within-domain pretraining on Sign2Gloss in a progressive manner.

Pretraining for Text Generation. Recently, the NLP community has seen rapid progress in large-scale self-supervised pretraining [8, 10, 27, 40, 41], which brings significant gains on downstream tasks. In particular, pretraining a language model on a large-scale monolingual corpus brings large improvements in low-resource NMT [1, 3, 8, 33]. Some multi-modality tasks such as image caption and VQA also leverage pretrained language models as initialization for bi-modal transformers [31, 35, 48]. As sign language is a full-fledged language system, powerful NLP techniques can likely be extended into SLT to help address the data-scarcity issue. We are the first to apply a pretrained language model for spoken language in SLT.

Transfer learning in SLT. Some previous works attempt to transfer external vision or language knowledge to SLT. For visual pretraining, [29] pretrains the visual backbone on Kinetics-400 [21] and two ISLR datasets [20, 28]. [4, 6, 49, 50] pretrain their visual backbones on within-domain Sign2Gloss task with gloss annotations. We adopt both general-domain and within-domain pretraining in a progressive manner. For language pretraining, [29, 46] load pretrained word embeddings into the decoder embedding layer

but fail to demonstrate their effectiveness. We are the first to leverage powerful pretrained language models, which brings significant improvement.

3. Method

In this section, we introduce our simple method for sign language translation. Given an input sign video $\mathcal{V} = (v_1, ..., v_T)$ with T frames, our goal is to learn a neural network $N_\theta(\cdot)$ that can predict the associated spoken language sentence $\mathcal{S} = (s_1, ..., s_U)$ with U words directly from the sign video $\mathcal{V}$:

$$\mathcal{S} = N_\theta(\mathcal{V}). \tag{1}$$

In order to transfer knowledge from general domains of action recognition and machine translation, we break down the SLT framework into two disjoint tasks: a visual action recognition task that converts the sign videos to semantic glosses (Sign2Gloss), and a language translation task that maps glosses to spoken language texts (Gloss2Text). This allows us to pretrain each task separately and then finetune the joint model.

In our approach, the overall network $N_\theta(.)$ is composed of three sub-networks: 1) a visual encoder network $\mathcal{E}$ that transforms raw video into visual features; 2) a sequence-to-sequence translation network $\mathcal{D}$ that translates language features into spoken language text; 3) a visual-language mapper $\mathcal{M}$ that bridges between features of the visual modality and language modality for joint training. The framework is illustrated in Figure 2.

In this work, we demonstrate that using such a simple, no-frills framework can achieve high sign language translation performance. Besides its simplicity and high performance, we reveal that the bottleneck of current SLT systems mainly lies in the lack of training data, so that a more flexible architecture that can leverage as much training data as possible via pretraining, from both the vision and language sides, is superior.
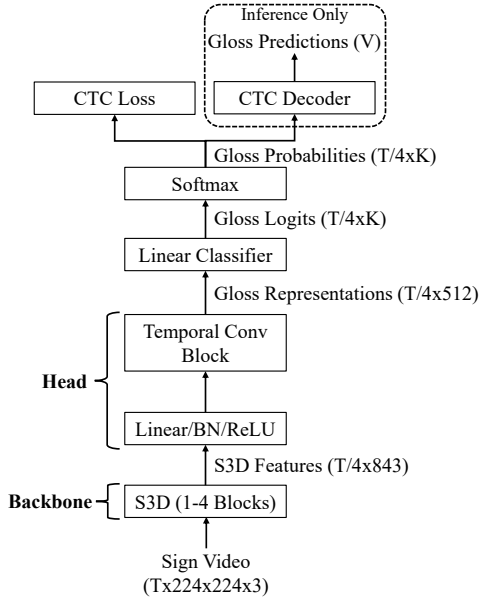
Figure 3. Architecture of our visual encoder network.

3.1. Visual Encoder Network and Pretraining

The visual encoder network $\mathcal{E}$ transforms the raw video input into a visual feature. The visual feature in this stage is mainly used to predict gloss labels, which is essentially a fine-grained action recognition task. Figure 3 shows the network architecture, which consists of a video backbone and a lightweight head to further encode temporal information.

Video Backbone. We use S3D [45] as our backbone due to its excellent trade-off between performance and inference speed. We feed each $T \times 224 \times 224 \times 3$ video into the backbone. Only the first four blocks of S3D are used since our goal is to extract a dense representation for gloss sequence prediction, and thus the extracted S3D features are of size $T/4 \times 832$ after spatial pooling. The extracted features then serve as the input of our head network.

Head Network. As shown in Figure 3, our lightweight head network consists of a projection block containing a temporal linear layer, a batch normalization layer, and a ReLU layer, as well as a temporal convolutional block which contains two temporal convolutional layers with a temporal kernel size of 3 and a stride of 1, a linear translation layer, and a ReLU layer. We feed the S3D features into the projection block and the following temporal convolutional block to generate $\mathcal{Z} \in \mathcal{R}^{T/4 \times 512}$. We call it a gloss representation since it represents gloss categories in the high dimensional space. Then a linear classifier and a Softmax function are applied to extract frame-level gloss probabilities $\mathcal{P} \in \mathcal{R}^{T/4 \times K}$, where K is the size of the gloss vocabulary.

Progressive Pretraining. We progressively pretrain the visual encoder $\mathcal{E}$ by first pretraining it in a general domain to learn generic human actions and then transferring it

to the within-domain task of learning fine-grained glosses. Specifically, for general-domain pretraining, we pretrain our S3D backbone on Kinetics-400, an action recognition dataset [21] and then WLASL, an isolated sign recognition dataset [28]. Next, for within-domain pretraining, we train our visual encoder under the Sign2Gloss task supervised by the continuous gloss annotations provided in SLT datasets.

Unlike spoken language texts, the continuous gloss annotations are chronologically consistent with the sign signals. We utilize the well-known connectionist temporal classification (CTC) loss [16] for within-domain pretraining under the supervision of gloss annotations. The CTC loss considers all possible alignments between two sequences while minimizing the error. Concretely, for an input video $\mathcal{V}$ and the corresponding ground truth gloss sequence $\mathcal{G}$, we use CTC to compute $p(\mathcal{G}|\mathcal{V})$ by marginalizing over all possible $\mathcal{V}$ to $\mathcal{G}$ alignments:

$$p(\mathcal{G}|\mathcal{V}) = \sum_{\pi \in \mathcal{B}} p(\pi|\mathcal{V}), \qquad (2)$$

where π denotes a path and $\mathcal{B}$ is the set of all viable paths that correspond to $\mathcal{G}$. The probability $p(\pi|\mathcal{V})$ is computed by the visual encoder $\mathcal{E}$. The CTC loss is then formulated as

$$\mathcal{L} = -\ln p(\mathcal{G}|\mathcal{V}). \qquad (3)$$

Gloss Sequence Prediction. Once the pretraining is finished, our visual encoder network can be used to predict gloss sequences given sign videos. As shown in Figure 3, we first use our visual encoder to extract gloss probabilities, then CTC decoding is utilized to generate the predicted gloss sequence. Details of the CTC decoding can be found in the supplementary materials.

3.2. Translation Network and Pretraining

Now we introduce the translation network $\mathcal{D}$, which learns a mapping between gloss sequences and spoken language texts, and present the corresponding progressive pretraining procedure.

Translation Network. Inspired by the recent progress of neural machine translation and multilingual denoising pretraining, we use mBART [33], a sequence-to-sequence denoising auto-encoder pretrained on large-scale multi-lingual corpus, as initialization for our translation network. The architecture is a standard sequence-to-sequence Transformer [44] with 12 layers for the encoder, 12 layers for the decoder, and a model dimension of 1024 on 16 heads.

Progressive Pretraining. With mBART initialization, our translation network is already pretrained in the general language domain. We further conduct within-domain pretraining on the Gloss2Text task to transfer mBART to the specific domain of gloss-to-text translation. Our goal is to train a translation network that can predict a text sentence $\mathcal{S}$ from a given gloss sequence $\mathcal{G}$. Concretely, we split

both $\mathcal{G}$ and $\mathcal{S}$ into sub-word units using mBART's SentencePiece tokenizer [24] and project one-hot vectors into dense embeddings via mBART's pretrained word embedding layer. Then, we add positional embeddings to the word embeddings as inputs to the bottoms of the encoder and decoder stacks. We train mBART on the Gloss2Text corpus to minimize the sequence-to-sequence cross-entropy loss $\mathcal{L} = -\log P(\mathcal{S}|\mathcal{G})$. After obtaining a well-trained translation model, we can predict spoken language sentences given gloss sequences. Translating from *ground-truth* sign gloss sequences to spoken language texts (Gloss2Text) is regarded as a virtual upper bound for performance on the SLT task [4, 6]. The two-stage translation task that first utilizes a Sign2Gloss model (our visual encoder) to generate a gloss sequence and then feeds the *predicted* gloss sequence to a well-trained Gloss2Text pipeline is known as Sign2Gloss2Text. However, using glosses as the intermediate representations may be sub-optimal since glosses cannot fully encode spatio-temporal visual information. To overcome this limitation, we bridge vision and language modalities via our V-L mapper for joint training.

3.3. End-to-end Sign Language Translation

So far, we have described the architectures and the pretraining processes of our visual encoder and translation network. Now we introduce the Visual-Language Mapper (V-L Mapper), which builds a connection between the two networks modeling different modalities for the purpose of joint training. Our V-L Mapper is simply implemented as a fully-connected MLP with two hidden layers. As shown in Figure 2, it converts visual features extracted by the visual encoder to language features, which are subsequently taken as the input of the translation encoder. We study the effects of feeding different visual features into the V-L Mapper in Section 4.4.3, and use gloss representations (see Figure 3) as our default setting. Thanks to the V-L Mapper, our framework can be trained in an end-to-end manner, under the joint supervision of the CTC loss and translation loss. Surprisingly, our framework even outperforms the acknowledged upper bound, i.e., translating from ground-truth sign gloss sequences to spoken language texts by using a well-trained Gloss2Text model, on the RWTH-PHOENIX-Weather-2014T test set. This is because our framework encodes both spatio-temporal information from sign videos and semantic knowledge from text transcriptions, providing more clues compared with the Gloss2Text model of only the language modality.

4. Experiments

4.1. Datasets and Evaluation Metrics

RWTH-PHOENIX-Weather 2014T. PHOENIX-2014T [4] is the most widely used benchmark for SLT in recent years [4, 6, 46, 49, 50]. The parallel corpus is collected from weather forecast news of the German public TV station PHOENIX over three years, including 8k triplets of RGB sign language videos of nine signers performing German Sign Language (DGS), sentence-level gloss annotations, and German translations transcribed from the news speaker. It contains 7096, 519 and 642 video segments in train, dev and test splits, respectively. The vocabulary size is 1066 for sign glosses and 2887 for German text. We compare our method with state-of-the-art methods on both the dev set and the test set. All ablation studies are conducted on this dataset.

CSL-Daily. CSL-Daily [49] is a recently published Chinese sign language (CSL) translation dataset recorded in a studio. It contains 20k triplets of (video, gloss, text) performed by ten different signers. The content contains topics such as family life, medical care, and school life. CSL-Daily contains 18401, 1077 and 1176 segments in train, dev and test splits. The vocabulary size is 2000 for sign glosses and 2343 for Chinese text. We compare our approach with state-of-the-art methods on both the dev set and test set.

Evaluation Tasks. We examine performance on the following tasks:

- **Sign2Gloss**: Predict the gloss sequence given raw video input. This task is also known as CSLR (Continuous Sign Language Recognition). This task is mainly used to evaluate our visual encoder.

- **Gloss2Text**: Translate a *ground-truth* gloss sequence to text. Its results are generally regarded as an upper bound for the sign language translation task. We also use this task to evaluate our translation model.

- **Sign2Gloss2Text**: A two-stage pipeline where we first adopt a Sign2Gloss module to predict a gloss sequence and then translate the *predicted* glosses to text by a Gloss2Text module. We use this to evaluate pipelines in which the visual encoder and the translation model are connected by the predicted gloss sequence.

- **Sign2Text**: Directly translate sign language video into text, which is our goal.

Following previous work [4, 6, 49, 50], we use Word Error Rate (WER) to evaluate Sign2Gloss, and ROUGE [32] and BLEU [36] to evaluate the other three tasks.

4.2. Implementation details

Our model is implemented in PyTorch. Details about all hyperparameters are given in the supplementary materials.

Visual Encoder Pretraining. We progressively pretrain the visual encoder from general domain to within domain. First we pretrain the S3D backbone on two action recognition datasets sequentially, namely Kinetics-400 [21], the most popular human action recognition dataset containing 400 action classes, and WLASL [28], a large-scale Word-Level American Sign Language video datasets containing 2000 isolated sign classes. The training procedure follows [45].

Sign2Gloss2Text (two-stage)	Dev					Test				
	R	B1	B2	B3	B4	R	B1	B2	B3	B4
SL-Luong [4]	44.14	42.88	30.30	23.02	18.40	43.80	43.29	30.39	22.82	18.13
SL-Transf [6]	-	47.73	34.82	27.11	22.11	-	48.47	35.35	27.57	22.45
BN-TIN-Transf [49]	47.83	47.72	34.78	26.94	21.86	47.98	47.74	35.27	27.59	22.54
BN-TIN-Transf + BT* [49]	49.53	49.33	36.43	28.66	23.51	49.35	48.55	36.13	28.47	23.51
STMC-Transf [46]	46.31	48.27	35.20	27.47	22.47	46.77	48.73	36.53	29.03	24.00
Ours	**50.23**	**50.36**	**37.50**	**29.69**	**24.63**	**49.59**	**49.94**	**37.28**	**29.67**	**24.60**
Sign2Text (end-to-end)	R	B1	B2	B3	B4	R	B1	B2	B3	B4
SL-Luong[†] [4]	31.80	31.87	19.11	13.16	9.94	31.80	32.24	19.03	12.83	9.58
TSPNet-Joint[†] [29]	-	-	-	-	-	34.96	36.10	23.12	16.88	13.41
SL-Transf [6]	-	47.26	34.40	27.05	22.38	-	46.61	33.73	26.19	21.32
STMC-T [50]	48.24	47.60	36.43	29.18	24.09	46.65	46.98	36.09	28.70	23.65
BN-TIN-Transf + BT* [49]	50.29	51.11	37.90	29.80	24.45	49.54	50.80	37.75	29.72	24.32
Ours	**53.10**	**53.95**	**41.12**	**33.14**	**27.61**	**52.65**	**53.97**	**41.75**	**33.84**	**28.39**

Table 1. Comparison with state-of-the-art methods on PHOENIX-2014T. [†] denotes methods without using gloss annotations. * denotes methods with semi-supervised learning. 'R' represents ROUGE, and 'B1' denotes BLEU-1, with the same for 'B2-B4'. Our framework outperforms all methods by large margins.

Video clips are fed through five blocks in S3D backbone followed by a 3D average pooling layer and a linear classification layer to predict the action class. Next we conduct within-domain pretraining on Sign2Gloss task using the CTC loss (Eq. 3), where we only use the first four blocks of the pretrained S3D and spatially pool the S3D features to the size of $T/4 \times 832$ as inputs of our head network.

Translation Pretraining. For general-domain pretraining, we initialize our language model with the official release of mBART-large-cc25[3] which is pretrained on CC25, a multilingual corpus of size 1300GB from Common Crawl[4] that covers 25 languages. We also try GPT2 [40] pretrained on 16GB German monolingual corpus. Unless otherwise specific, we use mBART by default.

Joint Training. We load the two independently pretrained modules as the initialization for joint training. The features before the linear classifier, i.e. gloss representation, are projected into vectors of 1024 dimension by the V-L Mapper and position embeddings are added to them to form inputs to the translation encoder. The whole network is trained under the joint supervision of the CTC loss and cross-entropy loss with both weights set to 1.0.

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

We compare our approach to state-of-the-art methods on PHOENIX-2014T and CSL-Daily, as shown in Table 1 and Table 2. Without integrating multi-cue features [5, 50] nor advanced data augmentation strategies such as back translation [49], our simple method significantly surpasses all

---
[3]https://huggingface.co/facebook/mbart-large-cc25
[4]https://commoncrawl.org/

counterparts on PHOENIX-2014T and CSL-Daily.

4.4. Ablation Study

4.4.1 Pretraining of Visual Encoder

Our visual encoder is pretrained in a progressive manner. We first study the effects of using different general-domain pretraining strategies:

- **Scratch.** No general-domain pretraining is conducted. The S3D backbone is trained from scratch.
- **K-400.** General-domain pretraining is done on Kinetics-400 [21], a large-scale action recognition set.
- **K-400→WLASL.** We further pretrain the K-400 pretrained S3D backbone on WLASL [28], a large-scale word-level sign language recognition dataset.

We conduct within-domain Sign2Gloss pretraining on these pretrained models and report the effects on both the Sign2Gloss and Sign2Text tasks in Table 3. The performance of Sign2Gloss directly reflects the effects of different general-domain pretrained models. Although K-400 is an action classification dataset, using the model pretrained on it as initialization nevertheless improves Sign2Gloss performance compared to the model trained from scratch, reducing the WER from 28.06 to 23.50 on the test set. Using K-400→WLASL as the initialization further boosts performance, achieving 22.45 WER on the test set. Though there exist differences between WLASL and PHOENIX-2014T, e.g., the former is proposed to solve isolated sign language recognition for American sign language while the latter aims to solve continuous sign language recognition for German sign language, general-domain pretraining on WLASL still learns relevant representations, e.g., low-level

Sign2Gloss2Text (two-stage)	Dev					Test				
	R	B1	B2	B3	B4	R	B1	B2	B3	B4
SL-Luong [4]	40.18	41.46	25.71	16.57	11.06	40.05	41.55	25.73	16.54	11.03
SL-Transf [6]	44.18	46.82	32.22	22.49	15.94	44.81	47.09	32.49	22.61	16.24
BN-TIN-Transf [49]	44.21	46.61	32.11	22.44	15.93	44.78	46.85	32.37	22.57	16.25
BN-TIN-Transf + BT* [49]	48.38	**50.97**	36.16	26.26	19.53	48.21	**50.68**	36.00	26.20	19.67
Ours	**51.35**	50.89	**37.96**	**28.53**	**21.88**	**51.43**	50.33	**37.44**	**28.08**	**21.46**
Sign2Text (end-to-end)	R	B1	B2	B3	B4	R	B1	B2	B3	B4
SL-Luong[†] [4]	34.28	34.22	19.72	12.24	7.96	34.54	34.16	19.57	11.84	7.56
SL-Transf [6]	37.06	37.47	24.67	16.86	11.88	36.74	37.38	24.36	16.55	11.79
BN-TIN-Transf [49]	37.29	40.66	26.56	18.06	12.73	37.67	40.74	26.96	18.48	13.19
BN-TIN-Transf + BT* [49]	49.49	51.46	37.23	27.51	20.80	49.31	51.42	37.26	27.76	21.34
Ours	**53.38**	**53.81**	**40.84**	**31.29**	**24.42**	**53.25**	**53.31**	**40.41**	**30.87**	**23.92**

Table 2. Comparison with state-of-the-art methods on CSL-Daily. † denotes methods without using gloss annotations. * denotes methods with semi-supervised learning. Without any tricks nor using manually generated extra data, our method surpasses the recent semi-supervised method BN-TIN-Transf + BT [49].

Pretraining			Sign2Gloss		Sign2Text									
			Dev	Test	Dev					Test				
K	WL	S2G	WER	WER	R	B1	B2	B3	B4	R	B1	B2	B3	B4
		✓	27.25	28.06	51.97	52.99	40.42	32.59	27.27	51.82	52.53	40.18	32.37	27.01
✓		✓	23.05	23.50	**53.24**	**53.99**	**41.47**	**33.63**	**28.19**	52.42	53.66	41.27	33.36	27.91
✓	✓	✓	**21.90**	**22.45**	53.10	53.95	41.12	33.14	27.61	**52.64**	**53.97**	**41.75**	**33.84**	**28.39**
✓	✓		-	-	45.84	47.31	33.64	25.83	20.76	45.93	47.40	34.30	26.47	21.44

Table 3. Ablation study of visual encoder with different pretraining settings on the **PHOENIX Sign2Gloss** and **PHOENIX Sign2Text** tasks. K, WL, S2G denote pretraining on Kinetics-400, WLASL and Sign2Gloss, respectively.

gesture features. As for Sign2Text, the gains of visual pretraining become narrowed, which suggests that learning favorable visual features is not the only determining factor for Sign2Text. For example, the translation model provides complementary information. In addition, to verify the importance of within-domain Sign2Gloss pretraining for Sign2Text, we load the visual encoder only pretrained on K-400 and WLASL into Sign2Text joint training. As the last column in Table 3 shows, skipping within-domain pretraining considerably hurts performance, reducing BLEU-4 by nearly 7 on both sets. We conclude that both general-domain and within-domain pretraining contribute to our method's high performance.

4.4.2 Pretraining of Translation Model

SLT greatly suffers from the data scarcity issue. Recently, language pretraining has shown promising results in low-resource NMT [1, 8, 25, 33], which inspires us to introduce language pretraining to SLT.

General-domain Pretraining Improves Gloss2Text. We first experiment with two popular pretrained language models, namely mBART [33] and GPT2 [40], to verify the effects of using different architectures and different large-

scale general-domain corpus, through direct evaluation on the PHOENIX Gloss2Text task. Table 4 shows the results. As baselines, we train two translation networks with the same architecture as mBART or GPT2 but with random initializations. mBART outperforms GPT2, suggesting that the encoder-decoder architecture and bidirectional attention of mBART makes it more suitable for Gloss2Text than GPT2 which only has a decoder with unidirectional attention. However, general-domain pretraining on large corpus improves both mBART and GPT2 on Gloss2Text and mBART pretrained on CC25 achieves the best performance. We use mBART for further experiments. Additionally, mBART is pretrained on the multilingual corpus and thus can be used as a generic pretraining model for various sign languages.

Progressive Pretraining Improves Sign2Text. We examine the effects of progressive pretraining of the translation model on the Sign2Text task, which is our final goal. Four pretraining settings are studied: 1) without pretraining; 2) pretraining on the Gloss2Text task; 3) pretraining on the CC25 corpus; 4) progressive pretraining, i.e., the translation model is first pretrained on CC25, then a further within-domain pretraining is conducted on the Gloss2Text task.

General-domain Pretraining	Dev					Test				
	R	B1	B2	B3	B4	R	B1	B2	B3	B4
GPT2 Scratch	44.45	45.93	31.96	24.17	19.40	43.57	44.62	31.27	23.65	18.96
mBart Scratch	46.56	48.42	35.22	27.55	22.71	47.03	48.16	35.32	27.63	22.57
GPT2 w/ German Corpus	49.99	50.86	38.07	30.32	25.11	48.61	48.98	36.55	28.88	23.82
mBART w/ CC25	**53.79**	**54.01**	**41.41**	**33.50**	**28.19**	**52.54**	**52.65**	**39.99**	**32.07**	**26.70**

Table 4. Ablation study of general-domain language pretraining on the **PHOENIX Gloss2Text** task. We conduct within-domain pretraining on the Gloss2Text task to transfer from the general domain to the specific domain of gloss-to-text translation.

CC25	Gloss2Text	Dev					Test				
		R	B1	B2	B3	B4	R	B1	B2	B3	B4
		46.47	46.99	33.75	25.83	20.70	46.67	47.45	34.39	26.47	21.36
	✓	48.77	49.46	35.98	27.98	22.89	47.65	49.01	36.16	28.33	23.28
✓		52.30	52.39	39.97	32.28	26.99	51.83	53.05	40.34	32.28	26.95
✓	✓	**53.10**	**53.95**	**41.12**	**33.14**	**27.61**	**52.65**	**53.97**	**41.75**	**33.84**	**28.39**

Table 5. Ablation study of mBART with different pretraining settings on the **PHOENIX Sign2Text** task. 'CC25' denotes transformer with mBART initialization, 'Gloss2Text' represents transformer pretrained on the Gloss2Text task.

Method	Dev					Test				
	R	B1	B2	B3	B4	R	B1	B2	B3	B4
Gloss2Text	53.79	54.01	41.41	33.50	28.19	52.54	52.65	39.99	32.07	26.70
Sign2Gloss2Text	50.23	50.36	37.50	29.69	24.63	49.59	49.94	37.28	29.67	24.60
Sign2Text w/ Gloss Logits	52.45	53.23	40.55	32.66	27.23	**52.71**	53.70	41.20	33.22	27.78
Sign2Text w/ Gloss Reps	**53.10**	**53.95**	**41.12**	**33.14**	**27.61**	52.65	**53.97**	**41.75**	**33.84**	**28.39**
Sign2Text w/ S3D Features	43.53	44.11	31.46	24.60	20.24	43.77	44.68	32.02	25.02	20.62

Table 6. Ablations on different visual features as the V-L Mapper input on **PHOENIX Sign2Text**.

For all settings, we use the same joint training process for Sign2Text. The results are shown in Table 5. We use transformer of the same architecture without pretraining as our baseline. From the table we can observe that the translation model pretrained on the Gloss2Text task yields a slight improvement (+1.92 BLEU-4 on the test set). When pretraining on CC25, our method achieves 26.95 BLEU-4 on the test set, which demonstrates the importance of language pretraining on the large-scale corpus. The best result is achieved by progressive pretraining, which can be attributed to both general-domain pretraining on the large-scale corpus and domain alignment through within-domain pretraining (Gloss2Text) and the downstream task (Sign2Text).

4.4.3 Joint Multi-modality Training

At last, we study the effectiveness of our joint multi-modality Sign2Text training by bridging the two modalities via the V-L Mapper. The most straightforward approach is to build a two-stage translation pipeline, i.e. Sign2Gloss2Text, where predicted glosses serve as the intermediate state. As the discrete glosses cannot fully capture the semantics in sign video, here we study using different visual features as the inputs to the V-L Mapper. In an ablation study, we examine the three features shown in Figure 3, namely gloss logits, gloss representations, and S3D features. Table 6 shows the results. Translating from a ground-truth gloss sequence to text (Gloss2Text) is generally regarded as an upper bound in SLT. Surprisingly, our joint Sign2Text training with gloss representations and with gloss logits outperforms not only Sign2Gloss2Text, but also the Gloss2Text upper bound, which demonstrates the effectiveness of our progressive pretraining and the proposed multi-modality transfer learning.

5. Conclusion

We present a simple yet effective multi-modality transfer learning baseline for sign language translation. To alleviate the data scarcity issue, we exploit large-scale external knowledge from human action and spoken language by progressively pretraining visual and language modules from general domains to within the target domains. The two individually pretrained modules are then bridged via the Visual-Language Mapper for joint SLT training. Experiments on two SLT datasets show that our approach outperforms all state-of-the-art methods. Our method can be applied to various sign languages. In future work, we would like to use this framework to transfer more external knowledge into SLT for further improvement. We hope our simple baseline can facilitate future research in SLT and motivate more researchers to engage in this field.

References

[1] Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT)*, 2019. 3, 7

[2] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *European Conference on Computer Vision*, 2020. 2, 3

[3] Christos Baziotis, Barry Haddow, and Alexandra Birch. Language model prior for low-resource neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. 3

[4] 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*, 2018. 1, 2, 3, 5, 6, 7

[5] Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. Multi-channel transformers for multi-articulatory sign language translation. In *ECCV 2020 Workshops*, 2020. 1, 2, 6

[6] 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*, 2020. 1, 2, 3, 5, 6, 7

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

[8] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. 1, 3, 7

[9] Runpeng Cui, Hu Liu, and Changshui Zhang. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, 2019. 2, 3

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2

[12] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2sign: A large-scale multi-modal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[13] Haoqi Fan, Yanghao Li, Bo Xiong, Wan-Yen Lo, and Christoph Feichtenhofer. Pyslowfast. https://github.com/facebookresearch/slowfast, 2020. 3

[14] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[15] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *IEEE International Conference on Computer Vision, ICCV*, 2017. 3

[16] 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*, 2006. 4

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[18] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2015. 3

[19] 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*, 2020. 2

[20] Hamid Reza Vaezi Joze and Oscar Koller. MS-ASL: A large-scale data set and benchmark for understanding american sign language. In *30th British Machine Vision Conference, BMVC*, 2019. 2, 3

[21] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, 2017. 2, 3, 4, 5, 6

[22] 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*, 2019. 2

[23] Oscar Koller, Sepehr Zargaran, and Hermann Ney. Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[24] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and deto-

kenizer for neural text processing. In Eduardo Blanco and Wei Lu, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP: System Demonstrations*, 2018. 5

[25] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 1, 2, 7

[26] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learningvia sparse sampling. In *CVPR*, 2021. 2

[27] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. 2, 3

[28] 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 *The IEEE Winter Conference on Applications of Computer Vision*, 2020. 2, 3, 4, 5, 6

[29] 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. In *Advances in Neural Information Processing Systems*, 2020. 2, 3, 6

[30] 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*, 2020. 2, 3

[31] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV*, 2020. 2, 3

[32] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. 5

[33] 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*, 2020. 1, 2, 3, 4, 7

[34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *International Conference on Computer Vision (ICCV)*, 2021. 2

[35] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 2, 3

[36] 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*, 2002. 5

[37] Junfu Pu, Wengang Zhou, and Houqiang Li. Iterative alignment network for continuous sign language recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2019. 2

[38] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatiotemporal representation with pseudo-3d residual networks. In *IEEE International Conference on Computer Vision, ICCV*, 2017. 3

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML*, 2021. 2

[40] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 2, 3, 6, 7

[41] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 2020. 3

[42] Rico Sennrich and Biao Zhang. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. 2

[43] Ozge Mercanoglu Sincan and Hacer Yalim Keles. Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access*, 2020. 2

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, 2017. 1, 4

[45] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. 3, 4, 5

[46] Kayo Yin and Jesse Read. Better sign language translation with STMC-transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, 2020. 1, 2, 3, 5, 6

[47] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Making visual representations matter in vision-language models. *CVPR*, 2021. 2

[48] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2021. 3

[49] 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*, 2021. 1, 2, 3, 5, 6, 7

[50] Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. Spatial-temporal multi-cue network for sign language recognition and translation. *IEEE Transactions on Multimedia*, 2022. 1, 2, 3, 5, 6