StepNet: Spatial-temporal Part-aware Network for Isolated Sign Language Recognition

XIAOLONG SHEN, Zhejiang University, Zhejiang, China
ZHEDONG ZHENG, University of Macau, Macau SAR, China
YI YANG, Zhejiang University, Zhejiang, China

The goal of sign language recognition (SLR) is to help those who are hard of hearing or deaf overcome the communication barrier. Most existing approaches can be typically divided into two lines, *i.e.*, Skeleton-based and RGB-based methods, but both the two lines of methods have their limitations. Skeleton-based methods do not consider facial expressions, while RGB-based approaches usually ignore the fine-grained hand structure. To overcome both limitations, we propose a new framework called Spatial-temporal Part-aware network (StepNet), based on RGB parts. As its name suggests, it is made up of two modules: Part-level Spatial Modeling and Part-level Temporal Modeling. Part-level Spatial Modeling, in particular, automatically captures the appearance-based properties, such as hands and faces, in the feature space without the use of any keypoint-level annotations. On the other hand, Part-level Temporal Modeling implicitly mines the long-short term context to capture the relevant attributes over time. Extensive experiments demonstrate that our StepNet, thanks to spatial-temporal modules, achieves competitive Top-1 Per-instance accuracy on three commonly-used SLR benchmarks, *i.e.*, 56.89% on WLASL, 77.2% on NMFs-CSL, and 77.1% on BOBSL. Additionally, the proposed method is compatible with the optical flow input and can produce superior performance if fused. For those who are hard of hearing, we hope that our work can act as a preliminary step.

CCS Concepts: • **Computing methodologies** → **Activity recognition and understanding**; **Video summarization**.

Additional Key Words and Phrases: Sign Language Recognition, Video Analysis.

ACM Reference Format:
Xiaolong Shen, Zhedong Zheng, and Yi Yang. 2023. StepNet: Spatial-temporal Part-aware Network for Isolated Sign Language Recognition. *ACM Trans. Multimedia Comput. Commun. Appl.* 5, 8, Article 39 (October 2023), 19 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

As the primary communication tool for deaf and hard of hearing, sign language employs the dynamic movement of the hands and body as well as facial expressions to convey meaning. Sign language has complex rules and is challenging to learn because it is independent of spoken language. Moreover, sign languages are not universal although there are some similarities among different

Xiaolong Shen and Yi Yang are with the College of Computer Science and Technology, Zhejiang University, China 310027. E-mail: sxlongcs@zju.edu.com, yangyics@zju.edu.cn. Zhedong Zheng is with Faculty of Science and Technology, and Institute of Collaborative Innovation, University of Macau, China 999078. E-mail: zhedongzheng@um.edu.mo.
This work was supported by the National Natural Science Foundation of China (U2336212) and the Fundamental Research Funds for the Central Universities (No. 226-2022-00051).
Authors' addresses: Xiaolong Shen, Zhejiang University, 310013, Zhejiang, China; Zhedong Zheng, University of Macau, 999078, Macau SAR, China; Yi Yang, Zhejiang University, 310013, Zhejiang, China.

1551-6857/2023/10-ART39 $15.00
https://doi.org/0000001.0000001

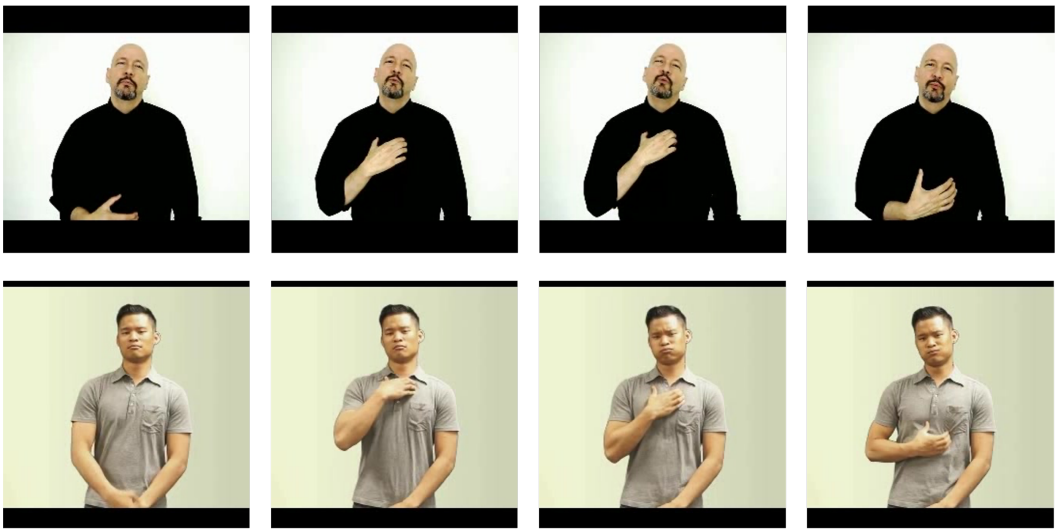

Fig. 1. Selected challenging examples from the widely-used sign language WLASL benchmark [33]. We could observe that there are some similar-appearance hand gestures, *e.g.*, "wish" (top) and "hungry" (bottom). Such cases demand the learned model to mine more fine-grained details, such as facial expressions.

sign languages [8]. It makes a communication barrier between deaf-mute people and others, as well as among deaf-mute people. In an attempt to break such barriers, more researchers [24, 27, 33] have started to explore Sign language recognition (SLR), which aims to predict the word class of the sign video. SLR is of importance, since it is the fundamental prerequisite of sign language translation [5, 20] and other related tasks, such as sign language generation [62].

Therefore, in this work, we study word-level sign language recognition, called Isolated SLR. Isolated SLR is generally regarded as a fine-grained action recognition task. Previously, SLR approaches [18, 28, 63, 68] encode hand gestures with hand-crafted features. Recently, owing to the development of deep learning, many methods [24, 26, 27] have emerged for learning a sign video representation in an end-to-end manner. These works generally involve two families based on the input format: Skeleton-based (using the keypoint sequence) and RGB-based (using the RGB video) methods. Skeleton-based methods [27, 33, 50, 59] mainly utilize Recurrent neural networks (RNNs) or Graph neural networks (GNNs) to model the spatial and temporal representation. However, the Skeleton-based approaches have two primary drawbacks. First, appearance attributes are totally discarded in the keypoint inputs, which is sub-optimal. The facial expressions in SLR convey the emotion of the signer. The emotion can serve as an additional clue to discriminate the sign words with similar hand or limb movements (see Figure 1). Second, keypoint annotations are not stable and usually noisy. For instance, the keypoints [27, 33] are extracted by an offline pose estimator. Due to the gap between different training and testing sets, the offline keypoint model could introduce incorrect predictions, especially in the video scenario, *e.g.*, the occlusion on the fingers and the motion blur. In contrast, another line of works, *i.e.*, RGB-based methods [1, 32, 33], mainly fine-tune the action recognition methods. Some works [33] deploy 2D Convolutional neural networks (CNNs) with RNNs to build spatial and temporal relations respectively, while others [1, 33] also apply 3D CNNs to capture spatial as well as temporal information simultaneously. However, directly borrowing the structure of the general action recognition framework ignores the geometric characteristics of sign language videos, such as fine-grained movement and human structure. As shown

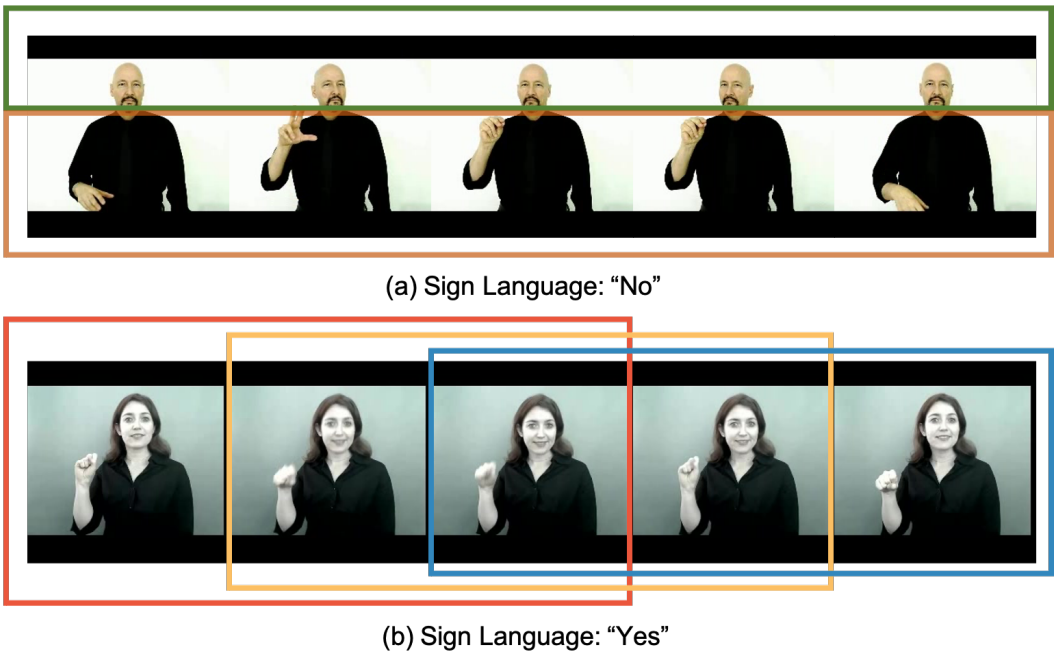

Fig. 2. Our motivation for designing the part-aware modules. (a) Part-level Spatial Partition: We can infer what the sign video means by looking at just one part, either the green part (shaking head) or the brown part (closing three fingers) (b) Part-level Temporal Partition: Similarly, we observe that any three short video clips, *i.e.*, red, yellow, and blue, can represent the gloss of the sign video. These two kinds of partition inspire us to harness RGB parts to mine fine-grained features for sign language recognition.

in Figure 1, most pixels in sign videos are static, while the discriminative parts only take a few spaces in the frame. Therefore, the non-trivial slight hand movement and facial expression need ad-hoc optimization.

Based on the above observation, we propose a new SLR network, called Spatial-temporal Part-aware network (StepNet), which is based on RGB parts instead of commonly-used skeleton data. It is because RGB parts well preserve the local human structure movement as well as the fine-grained texture changes, which is in line with the demands of the SLR task. As the name implies, StepNet contains two different part modeling branches. One branch is Part-level Spatial Modeling, and the other is Part-level Temporal Modeling. Part-level Spatial Modeling is to capture the appearance-based properties, such as hands and faces, in the feature space, while Part-level Temporal Modeling aims at capturing the changes along with the time (see Figure 2). When inference, we fuse the spatial and temporal representations for the final prediction. Since we involve two kinds of fine-grained branches, the learned model is robust to similar classes with subtle differences and achieves competitive performance on several public benchmarks [2, 26, 33]. The ablation studies also show that our model effectively exploits the Spatial-temporal cues in sign videos. Our main contributions are as follows:

- Inspired by the inherent structure of sign language, we design a new RGB-based network, called Spatial-temporal Part-aware network (StepNet). As the name implies, StepNet contains the Part-level Spatial Modeling that learns the symmetric and top-bottom association, and the Part-level Temporal Modeling that captures bottom-up long-short term changes.

- Extensive experiments verify the effectiveness of the proposed method, yielding competitive performance on public benchmarks, *i.e.*, 77.1% on BOBSL [2], 56.89% on WLASL-2000 [33] and 77.2% on NMFs-CSL [26]. Moreover, StepNet is compatible with the optical flow inputs for the further ensemble.

2 RELATED WORK

Sign language recognition (SLR). Depending on the input format, SLR can be classified into RGB-based and Skeleton-based approaches. Previously, pioneering RGB-based methods [4, 9, 43, 77] mainly use hand-crafted features, *e.g.*, HOG-based features, and SIFT-based features, to model spatio-temporal representation. Recently, deep learning has shown great potential in various computer vision tasks. Some approaches [29, 30] utilize 2D CNNs for spatial feature extraction instead of hand-crafted features and Hidden Markov Models (HMMs) for temporal modeling. In addition, [1, 22, 33] deploy 3D CNNs for spatio-temporal modeling, which achieves higher recognition accuracy. It also reflects the powerful characterization capability of 3D CNNs. Skeleton-based methods [3, 33, 67] have attracted much interest from researchers thanks to the fact that skeleton data is not affected by background and appearance. Similar to the method of processing RGB data, some methods [6, 12, 31, 60] utilize CNNs or RNNs to process skeleton data. Besides, there are other processing methods for skeleton data. For instance, one line of works [3, 27, 33] utilizes GNNs to process skeleton data. Taking one step further, SAM [27] employs the multi-stream strategy [57] that separates the skeleton data into joint, bone, joint motion, and bone motion, and then applies the designed SL-GCN to model these representations. Recently, the pretraining and finetuning schema has been widely used in computer vision tasks. For sign language recognition, there are some works[23, 24, 55, 80] that focus on this schema. Owing to this self-reconstruction schema in the pretraining stage, the model will mine the inherent relations as much as possible. Pretraining a model on a large dataset will learn a more powerful representation, improving the robustness of the model, particularly in some challenging cases like self-occlusion among joints and motion blur. Thus when finetuning this model on other downstream tasks, the model can leverage more dense information to make the choice. However, the Skeleton-based approaches have two primary drawbacks. First, appearance attributes are totally discarded in the keypoints input. Second, keypoint annotations are not stable and usually noisy. As for the existing RGB-based methods, they do not fully exploit the relation between face and hand, except that borrowing the structure of the general action recognition framework.

RGB-based action recognition. The methods in this task mainly deploy three types of neural networks: 3D CNNs, 2D CNNs, and the mixtures of 3D and 2D. [7] proposes I3D to learn spatio-temporal features from videos directly. TEINet [41] and TEA [35] design more powerful temporal modules that can be directly incorporated into 2D CNNs for efficient information interaction between neighboring frames. For example, TSM [36] shifts the channels along the temporal dimension, which enhances the ability of temporal modeling for 2D CNNs. R(2+1)D [65] and S3D [71] decompose the 3D convolution into a 2D spatial convolution and 1D temporal convolution to improve the efficiency of 3D convolution. Two-path networks are also effective 3D designs. For instance, SlowFast [15] applies a slow branch and a fast branch to capture long-short term temporal information by feeding video with different frame rates.

Multi-cues methods. To improve the effectiveness of models for sign language recognition, researchers explore various fusion methods [16, 21] that utilize the combination of diverse cues, including multiple modalities and multiple local patterns. As one of the pioneering works, the two-stream method [58] uses two kinds of inputs, *i.e.*, RGB and optical flow, to model appearance and motion information in videos separately and then fuse them with the late-fusion strategy. Similarly, SAM [27] designs a multi-branch framework that fuses RGB and depth-based modalities by the

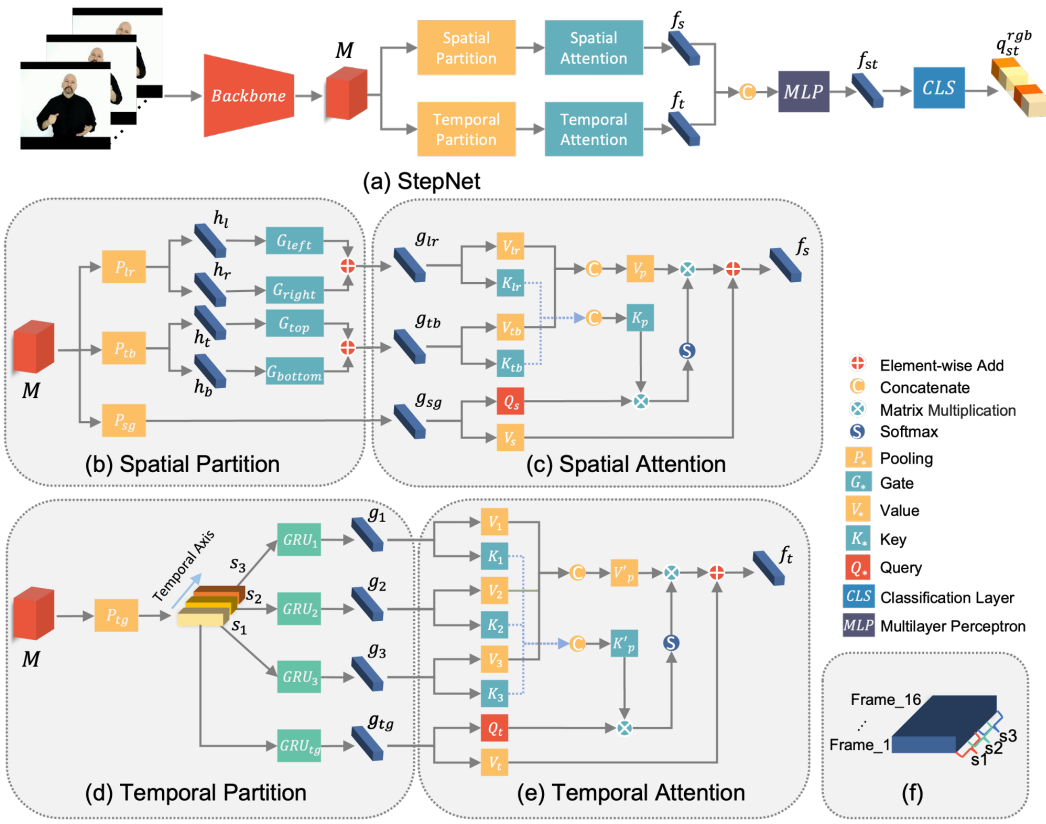

Fig. 3. (a) A schematic of our framework. StepNet consists of four parts, *i.e.*, backbone, Part-level Spatial Modeling, Part-level Temporal Modeling, and classification layer. The backbone is utilized to extract feature maps of inputs **M**. The feature maps **M** are then processed in parallel using a Part-level Spatial Modeling and a Part-level Temporal Modeling. Finally, outputs of two part-level modeling are fused and put into the classification layer to obtain classification logit vectors q_{st}^{rgb}. Sub-figure (b),(c). Details of the spatial partition and attention. Spatial partition includes the local guidance, *i.e.*, left-right(lr) and top-bottom (tb), and the global guidance (sg). The lr and tb are composed of Pooling (P) and Gate (G) operators, which build the channel relationships between the lr or tb. Global guidance(sg) guides the model to obtain a coarse but global representation g_{sg}. After that, we introduce spatial attention in Sub-figure (c) to compute how to aggregate these features g_{lr}, g_{tb}, g_{sg}. Sub-figure (d),(e). Details of the temporal partition and attention. We first pool the feature maps and then split them along the temporal dimension into three segments, *i.e.*, s_1, s_2, s_3. Each segment is utilized to explore the short clip context by GRU. In addition, we also apply the GRU to the long clip (before partition), which models the long-term representation g_{tg}. After that, temporal attention learns how to complement the long-term representation g_{tg} from the short-term context. Sub-figure (f). An example of partition with overlaps along the temporal axis.

proposed GEM, yielding competitive performance. For sign language translation, [20] proposes a temporal encoder to capture RGB and skeleton clues adaptively. Another line of work leverages local patterns within the input to mine the fine-grained feature. Hu *et al.* propose a model [25] that refines the representation of the hand by the hand prior, which enhances the model interpretability, and then fuses the hand representation with RGB-based or Skeleton-based methods by late fusion to improve

the model accuracy. Zhou *et al.* propose a spatial-temporal multi-cue network [84] for continuous SLR and sign language translation which involves multiple customized modules and optimization strategies for multi-cue fusion. Moreover, some methods [11, 13, 14, 17, 40, 51, 53, 74, 76, 81] in other fine-grained tasks also adopt a similar multi-cues strategy for learning discriminative representation. For instance, PCB [61] horizontally splits the feature maps, while LPN [70] cuts the local attention parts in a circle format. Yu *et al.* propose a method [78] that searches multi-rate and multi-modal architectures for RGB-D gesture recognition, and designs a 3D-CDC block to capture temporal context. Liu *et al.* propose a video dataset [40] for micro-gesture understanding and emotion analysis and design an unsupervised framework for MG Recognition. Some works [19, 83] introduce spatial transform on feature maps. In this work, we follow a similar multi-cue spirit. Differently, we focus on the symmetric and top-bottom association, and learning temporal modeling in an end-to-end manner. Therefore, the proposed method achieves competitive performance. For sign language recognition, using external tools to detect faces or hands and then feeding into the model separately to learn relations incur expensive computational overhead for Sign Language tasks.

3 METHODOLOGY

3.1 Overview

As shown in Figure 3, StepNet consists of three stages: the first stage is to extract the coarse spatiotemporal representation from the input RGB frames. Second, the coarse representation is refined by two parallel modules, *i.e.*, Part-level Spatial Modeling and Part-level Temporal Modeling. Part-level Spatial Modeling contains spatial partition and spatial attention, while Part-level Temporal Modeling consists of temporal partition and temporal attention. The two modeling process is complementary. Part-level Spatial Modeling focuses on capturing the appearance-based properties in the spatial space, while Part-level Temporal Modeling is to capture the motion changes from long-short term contexts along the temporal dimension. Finally, we fuse the refined spatial and temporal features with Multilayer Perceptron (MLP) and one linear classification layer for predicting the sign word. It is worth noting that we do not require extra key-point annotations throughout the whole training process.

3.2 Part-level Spatial Modeling

Current skeleton-based approaches [37] suffer from the inaccuracy of the keypoints, such as motion blur and self-occlusion, and neglect facial expressions. Therefore, we resort to Part-aware RGB-based methods. Unlike most existing RGB-based methods, we explicitly draw the network's attention to local patterns, which can capture the fine-grained positions between two hand gestures, and the subtle facial expression changes. Specifically, we design a module named Part-level Spatial Modeling that can explicitly capture the hand relationships and the hand-face correlation without extra pose estimators. Figure 3 (b) (c) show two components of Part-level Spatial Modeling: **spatial partition** and **spatial attention**.

Spatial Partition. We first introduce the spatial partition strategy in the Part-level Spatial Modeling. Given a sign video $\mathbf{X}$ with T frames, $\mathbf{X} \in \mathbb{R}^{T\times C\times H\times W}$, where C, H, and W are the channel, height, and weight of a single frame, we put the frames of the sign video into the backbone, a spatial-temporal extraction model, and then obtain feature maps $\mathbf{M}$. Here we have two types of operations for feature maps. One is the global guidance, which averagely pools the feature maps $\mathbf{M}$ along the channel dimension and outputs the global feature g_{sg} of size $T \times C$. The other processing method is called local guidance, which splits $\mathbf{M}$ into two stripes and averages the inner vectors of stripes into a single part-level vector. For the left-right partition, the splitting direction is vertical, and the

part-level vectors are named h_l, h_r. In addition, we deploy a top-bottom partition that horizontally splits $\mathbf{M}$, and the part-level vectors are termed as h_t, h_b. After that, Considering the different scales of local features, the Gate function is introduced to normalize these part-level features. Finally, we add up these re-scaled features after gating as the intermediate feature g. The process can be formulated as follows,

$$\begin{aligned} g_{lr} &= G_{left}(h_l) + G_{right}(h_r), \\ g_{tb} &= G_{top}(h_t) + G_{bottom}(h_b), \end{aligned} \tag{1}$$

where $G(h) = h \cdot Sigmoid(MLP(h))$, and MLP denotes Multilayer Perceptron.

Spatial Attention. Spatial attention focus on complementing information capturing the relationship of hands and faces from refined features g_{lr}, g_{tb} to the global feature g_{sg}. Specifically, we deploy linear layers for the refined features g_{lr}, g_{tb} to generate key-value pairs, termed as $K_{lr}, V_{lr}, K_{tb}, V_{tb}$, which encodes the local patterns about hands and faces. Similarly, we generate a query-value pair from the global feature g_{sg}, named Q_s, V_s. Then we concatenate the key and value of the refined features g_{lr}, g_{tb} as a large key-value pool and further compute the attention maps with the query of the global feature g_{sg}. In this way, we can extract complementary information from the refined features g_{lr}, g_{tb} by the attention maps. Finally, we add the complementing features to the value of the global feature V_s. The operations are defined as follows,

$$f_s = Softmax(Q_s K_p^T) V_p + V_s, \tag{2}$$

where K_p is the concatenated feature of K_{lr}, K_{tb}. V_p is the concatenated feature of V_{lr}, V_{tb}, and f_s is the final representation of the Part-level Spatial Modeling.

Discussion. In this section, we propose a Partition strategy that adaptively mines the location of face and hands by the pooling operation and fuse the normalized part-level features through the Gate mechanism. Instead of cropping hands or face to mine discriminative representation, this method harnesses the spatial properties of pooling that obtain larger receptive fields. It promotes the bottom-top alignment to build the relationship between the hands and face, and left-right alignment to build the relationship between hands. It is also efficient because of no need to preprocess the input, such as the pose estimator. Besides, we also propose an Attention method, which further aggregates the local and global clues. In summary, mutual modeling of the Part-level Spatial Modeling, especially face and hand in an end-to-end way, helps to capture discriminative fine-grained attention.

3.3 Part-level Temporal Modeling

Current SLR methods overlook the bottom-up property that the sign sequences can be viewed as a combination of multiple sub-actions. This property motivates us to explore the short-segment contexts. In Figure 2(b), combining more part sequences as sub-actions can fully exploit the part-level temporal property. Hence, we introduce Part-level Temporal Modeling for mining the long-short term action change for sign language. The idea is that we view the sign sequence as a combination of multiple decomposed sub-actions. Specifically, this module learns how to handle small segments to complement the long segment by explicitly segmenting the long segment in the temporal dimension. This module involves two stages: **temporal partition** and **temporal attention**.

Temporal Partition. The temporal partition is similar to the spatial partition. The main difference is that we split feature maps along the temporal dimension. Given feature maps $\mathbf{M}$, we first pool the (H, W) dimension of feature maps to $(1, 1)$ because we mainly model the temporal cues in this branch, which also reduces computing costs. Then we segment the temporal dimension T to $N = 3$ parts with overlaps as shown in Figure 3 (f), $\mathbf{S} = \{s_n\}_{n=1}^{N}$.

Temporal Attention. To mine the temporal motion changes, we deploy Gated Recurrent Neural Networks (*GRUs*). We apply an independent *GRU* on the global feature that captures the long-term change in the whole video. Other GRUs are used for every short segment s_n as follows:

$$g_n = GRU_n(s_n), \tag{3}$$

where GRU_n does not share weights, considering the original short sequence orders. Next, we calculate the key-value pairs, termed as K_n, V_n, through a linear function and concatenate these keys anD values as K'_p, V'_p, respectively. Similarly, we compute the query-value pair of the global feature g_t, named as Q_t, V_t. Then we deploy the temporal attention function to calculate the relationship between short segments and the long sequence. The attention of this part aims to explore the long-short-term problem. This operation enhances the capacity to model long-short-term variations of hands or faces as follows,

$$f_t = Softmax(Q_t {K'_p}^T) V'_p + V_t. \tag{4}$$

After achieving the spatial and temporal refined features, *i.e.*, f_s, f_t, we apply an MLP on the concatenated feature of f_s and f_t to derive the final representation f_{st}.

Discussion. In Figure 2, we show that the short video clips alone can facilitate the network to predict the correct sign language class. Inspired by this observation, we explicitly extract short-term representations from short video clips and then fuse multiple short-term predictions as one long-short-term prediction, which facilitates the robust inference process. Following the spirit, we propose a simple fusion method based on splitting and aggregation, where the splitting allows our model to fully exploit the fine-grained short-term representation and leverage these features to conduct the attention mechanism. The aggregation allows our model to explore the complementary way between the context of short and long segments and handle the long-short range dependencies and make full use of local and global cues.

Discussion about partition-attention mechanism. Sign language recognition tasks need to consider both hand movement and facial appearance. Thus mining the fine-grained relations between them both in spatial and temporal dimensions will help the model to learn more discriminate features. Based on the above observations, we design a spatial partition method that targets to model the local relations between hands and faces separately. After that, the mined fine-grained local feature will complement the global feature by an attention mechanism, yielding a more robust global feature. Moreover, simply modeling the spatial relations is not enough to discriminate challenging cases. Thus we further design a temporal modeling branch that summarises the long-short term changes in temporal context, capturing the hands and faces changes. By coupling all the spatial-temporal cues, the model achieves significant improvements.

3.4 Optimization Objectives

Given the class predictions from different spatial and temporal parts, we optimize the classification error following the existing works [25, 27, 33], which can be formulated as cross-entropy loss: $\mathcal{L}_{ce}(q, y) = -\log(\frac{exp(q_y)}{\sum_{c=1}^{C} exp(q_c)})$, where q is the predicted logits, and y is the index of the ground-truth category. Different from previous works, we accumulate the cross-entropy losses on all partial and

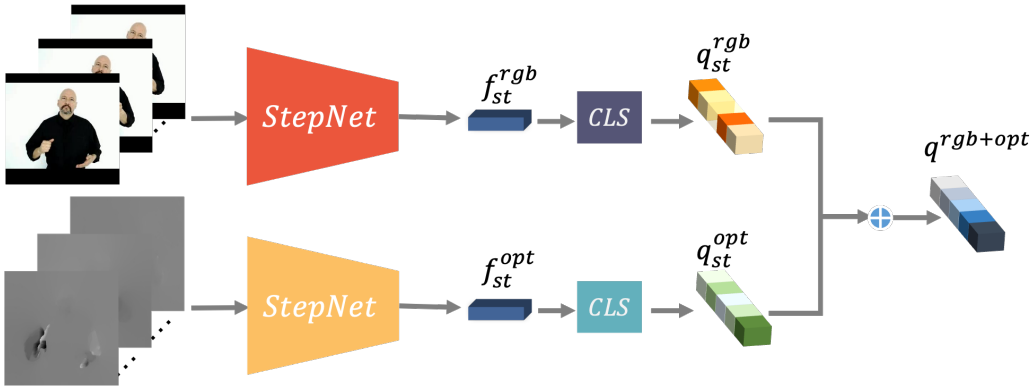

Fig. 4. A schematic of our two-stream fusion method. We fuse the classification logits ($q_{st}^{rgb}, q_{st}^{opt}$) processed from two StepNets to obtain the final logits (q_{r+o}). **Notably, two StepNets do not share weights.**

global observations, and thus the total loss can be defined as follows,

$$\begin{aligned}\mathcal{L}_{total} = &\overbrace{\mathcal{L}_{ce}(q_l, y) + \mathcal{L}_{ce}(q_r, y) + \mathcal{L}_{ce}(q_t, y) + \mathcal{L}_{ce}(q_b, y)}^{\mathcal{L}_{spatial}} \\ &+ \overbrace{\mathcal{L}_{ce}(q_{lr}, y) + \mathcal{L}_{ce}(q_{tb}, y) + \mathcal{L}_{ce}(q_{sg}, y) + \mathcal{L}_{ce}(q_s, y)}^{\mathcal{L}_{spatial}} \\ &+ \underbrace{\mathcal{L}_{ce}(q_t, y)}_{\mathcal{L}_{temporal}} + \underbrace{\mathcal{L}_{ce}(q_{st}, y)}_{\mathcal{L}_{fuse}}\end{aligned} \tag{5}$$

where $\mathcal{L}_{spatial}$ is utilized to supervise the predictions from spatial features including q_l, q_r, q_t, q_b as well as the aggregated features, *e.g.*, q_{lr}, q_{tb}. Applying loss on q_l, q_r, q_t, q_b can help the model to mine the local discriminate feature so that only seeing one part of these can yield relatively robust results. Moreover, applying loss on the aggregated features q_{lr}, q_{tb} can help the model learn the relations and correlations between the local discriminate feature. The classification loss motivates the model to discriminate the video with different sign language meanings, and implicitly encourages the model to learn the symmetric relationship, *e.g.*, between two hands, and the top-down correlation, *e.g.*, between hands and faces. Besides, it also aggregates the features representing global-local human structure by attention mechanism. Similarly, for another temporal branch, we deploy $\mathcal{L}_{temporal}$ on the prediction q_t of aggregated feature f_t to implicitly capture the long-short term variation. Finally, $\mathcal{L}_{fuse}$ motivates the model to learn adaptive weights between the spatial and the temporal feature f_s and f_t. During inference, we only deploy the top prediction q_{st}^{rgb} based on the final feature as the predicted sign word.

3.5 Two-Stream StepNet

StepNet is flexible and compatible with inputs of different modalities. We can equip multiple StepNets to form a simple late-fusion framework for the multi-modality input. For instance, as shown in Figure 4, we introduce another modality, *i.e.*, optical flow, which captures the motion between frames. We separately train the two StepNets for RGB and optical flow, of which weights are not shared. We adopt the late-fusion strategy to combine the category predictions. Specifically,

Table 1. A statistical summary of SLR datasets.

Datasets	# Signs	# Signers	# Samples	Languages	Type
WLASL [33]	2,000	119	21K	American	Isolated
NMFs-CSL [26]	1,067	10	32K	Chinese	Isolated
BOBSL [2]	2,281	39	452K	British	Co-articulated

we sum up the two modality logits ($q_{st}^{rgb}, q_{st}^{opt}$), generated from two StepNets with weights to obtain the final prediction:

$$q^{rgb+opt} = q_{st}^{rgb} + \alpha q_{st}^{opt}, \tag{6}$$

where q denotes the prediction logits, α is a weight hyper-parameter. For instance, we empirically set $\alpha = 0.4$ for the WLASL. Albeit simple, we show that the late-fusion strategy can arrive at a competitive sign language recognition accuracy in the experiment.

4 EXPERIMENTS

4.1 Datasets

We evaluate the proposed StepNet on three widely-used public datasets, including WLASL [33], NMFs-CSL [26], BOBSL [2], which span different sign languages. The datasets are briefly summarized in TABLE 1.

WLASL [33] is an American Sign Language (ASL) dataset, which includes four subsets consisting of different vocabulary sizes, *i.e.*, WLASL-100, WLASL-300, WLASL-1000, WLASL-2000. For instance, WLASL contains 21,083 videos performed by 119 signers and originated from unrestricted videos on the web. WLASL-2000 is one of the most challenging subsets because of more infrequent words.

NMFs-CSL [26] is a Chinese Sign Language (CSL) dataset, which includes 25,608 and 6,402 samples with 1,067 words for training and testing, respectively. It is also a challenging dataset due to a large variety of confusing words.

BOBSL [2] is a large-scale video collection of British Sign Language (BSL), which contains 1,962 episodes spanning a total of 426 differently named TV shows. It has 452K samples performed by 39 different signers for Co-articulated sign language recognition. This dataset differs from the above datasets because it is under the Co-articulated signing setting, which signs in context. It benefits to build robust models for understanding sign language "in the wild".

Note that all these datasets employ the signer-independent setting, which means no signer overlap between the training and testing splits.

4.2 Implementation Details

We apply FFmpeg [64] to extract all frames from RGB videos and resize all frames to 320×256. During training, we uniformly split the video into 16 splits and then randomly sample 16 frames from each split. The frames are then randomly cropped to 256×256. When testing, we center sample 16 frames and then center crop the frames to 256×256. We also apply the random horizontal flip augmentation to frames with a probability of 0.5. StepNet is implemented by PyTorch [49] and trained on one NVIDIA TESLA V100 with a batch size of 8. We train our framework using the AdamW [44] optimizer. For WLASL [33] and NMFs-CSL [26], the weight decay is set to 0.1 and the epoch is set to 100. In the training phase, we use linear warm-up in the first five epochs. The initial learning rate is 1e-4, and we reduce it to 1e-5 by CosineAnnealing [45] learning rate scheduler. For BOBSL [1], we follow the existing work [1], and deploy SGD optimizer with 0.9 momentum. The learning rate and the epoch are set to 0.03 and 30. We apply a multistep scheduler to decay the

Table 2. Comparison with the state-of-the-art methods in terms of Top-1, Top-2, and Top-5 accuracy on the NMFs-CSL [26]. The best results are in **bold**. *H* denotes the reconstructed hand representation by MANO [54]. *H+P* and *H+R* mean fusing the reconstructed hand representation (*H*) to the features of the Skeleton-based (*P*) and the RGB-based (*R*) method, respectively. *R+F* denotes using RGB data and optical flow data as inputs.

Method	Total			Confusing			Normal		
	Top-1	Top-2	Top-5	Top-1	Top-2	Top-5	Top-1	Top-2	Top-5
Skeleton-based									
ST-GCN [73]	59.9	74.7	86.8	42.2	62.3	79.4	83.4	91.3	96.7
Signbert (*H*) [24]	67.0	86.8	95.3	46.4	78.2	92.1	94.5	98.1	99.6
BEST [80]	68.5	-	94.4	49.0	-	90.3	94.6	-	99.7
RGB-based									
3D-R50 [52]	62.1	73.2	82.9	43.1	57.9	72.4	87.4	93.4	97.0
DNF [10]	55.8	69.5	82.4	33.1	51.9	71.4	86.3	93.1	97.0
I3D [7]	64.4	77.9	88.0	47.3	65.7	81.8	87.1	94.3	97.3
TSM [36]	64.5	79.5	88.7	42.9	66.0	81.0	93.3	97.5	99.0
Slowfast [15]	66.3	77.8	86.6	47.0	63.7	77.4	92.0	96.7	98.9
GLE-Net [26]	69.0	79.9	88.1	50.6	66.7	79.6	93.6	97.6	99.3
Ours	**77.2** ↑ 8.2	**86.2** ↑ 6.3	**92.5** ↑ 3.8	**62.4** ↑ 11.8	**76.1** ↑ 9.4	**86.9** ↑ 5.9	**96.9** ↑ 3.3	**99.7** ↑ 2.1	**99.9** ↑ 0.6
Fusion-based									
Signbert (*H* + *P*) [24]	74.9	**93.2**	**98.2**	58.6	**88.6**	**96.9**	96.7	99.3	99.9
Signbert (*H* + *R*) [24]	78.4	92.0	97.3	64.3	86.5	95.4	97.4	99.3	99.9
BEST (*H* + *R*) [80]	79.2	-	97.1	65.5	-	95.0	97.5	-	99.9
Ours (*R+F*)	**83.6** ↑ 4.4	92.7 ↓ 0.5	97.0 ↓ 1.2	**72.3** ↑ 6.8	87.2 ↓ 1.4	94.8 ↓ 2.1	**98.7** ↑ 1.2	**99.9** ↑ 0.6	**100.0** ↑ 0.1

learning rate. The decay parameter γ and step are set to 0.1 and [15, 25], respectively. For multi-modality fusion, we follow [36] to extract optical flow data, and adopt the TVL1 algorithm [79] implemented by Openmmlab Denseflow API [69]. The model parameter is 194M. We use TSM (R50) as our backbone, and the spatial and temporal attentions are composed of several linear layers, activation layers, and normalization layers. The running time is 0.0346s per batch (8 bach size) on Tesla V100 (16G). The shape of the tensor is listed as shown in Table 3.

Table 3. The shape of tensors.

name	M	f_s	f_t	f_{st}	h_l	h_r	h_t	h_b	g_{lr}	g_{tb}	g_{sg}	g_1	g_2	g_3	g_t
size	$16 \times 2048 \times 16 \times 16$	16×1024	16×2048									8×1024			16×2048

4.3 Quantitative Results

To validate the effectiveness of our model, we use Per-instance and Per-class accuracy metrics, which mean the average accuracy over each instance and each class, respectively. For NMFs-CSL [26], we follow [26] and report the Per-instance accuracy including Top-1, Top-2, Top-5 accuracy. For WLASL [33] and BOBSL [2], we show Per-instance and Per-class metrics with the Top-1 and Top-5 accuracy.

Comparison with state-of-the-art methods. We compare the proposed method with several competitive SLR methods on three benchmark datasets, *i.e.*, NMFs-CSL [26], WLASL [33] and BOBSL [2].

Evaluation on NMFs-CSL [26]. As shown in Table 2, we obtain the best accuracy on total, normal and confusing words, which validates the effectiveness of our proposed method. Specifically, when only one modality is employed, our RGB-based StepNet surpasses the best Skeletion-based method,

Table 4. Comparison with the state-of-the-art methods in terms of Top-1 and Top-5 accuracy on the WLASL [33]. The best results are in **bold**. Per-i denotes Per-instance, and Per-c represents Per-class. *H* denotes the reconstructed hand representation by MANO [54]. *P* means using skeleton data as input.

Method	WLASL100 [33]				WLASL300 [33]				WLASL1000 [33]				WLASL2000 [33]			
	Per-i		Per-c		Per-i		Per-c		Per-i		Per-c		Per-i		Per-c	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Skeleton-based																
ST-GCN [73]	50.78	79.07	51.62	79.47	44.46	73.05	45.29	73.16	-	-	-	-	34.40	66.57	32.53	65.45
Pose-TGCN [33]	55.43	78.68	-	-	38.32	67.51	-	-	-	-	-	-	23.65	51.75	-	-
PSLR [66]	60.15	83.98	-	-	42.18	71.71	-	-	-	-	-	-	-	-	-	-
Signbert (*H*) [24]	76.36	91.09	77.68	91.67	62.72	85.18	63.43	85.71	-	-	-	-	39.40	73.35	36.74	72.38
BEST [80]	77.91	91.47	77.83	92.50	67.66	89.22	68.31	89.57	-	-	-	-	46.25	79.33	43.52	77.65
Signbert+ [23]	79.84	2.64	80.72	93.08	73.20	90.42	73.77	90.58	-	-	-	-	48.85	82.48	46.37	81.33
SAM (*P*) [27]	-	-	-	-	-	-	-	-	-	-	-	-	51.50	84.94	48.87	84.02
RGB-based																
MEN [42]	-	-	-	-	-	-	-	-	-	-	-	-	25.54	53.72	-	-
I3D [33]	65.89	84.11	67.01	84.58	56.14	79.94	56.24	78.38	47.33	76.44	-	-	32.48	57.31	-	-
TCK [32]	77.52	91.08	77.55	91.42	68.56	89.52	68.75	89.41	-	-	-	-	-	-	-	-
BSL [1]	-	-	-	-	-	-	-	-	-	-	-	-	46.82	79.36	44.72	78.47
Ours	**78.29** ↑ 0.77	**92.25** ↑ 1.17	**78.77** ↑ 1.22	**92.63** ↑ 1.21	**74.70** ↑ 6.14	**91.02** ↑ 1.50	**75.32** ↑ 6.57	**91.17** ↑ 1.76	**67.91**	**91.10**	**67.76**	**91.33**	**56.89** ↑ 10.07	**88.64** ↑ 9.28	**54.54** ↑ 9.82	**87.97** ↑ 9.5

Table 5. Comparison with the state-of-the-art fusion-based methods in terms of Top-1 and Top-5 accuracy on WLASL-2000 [33]. The best results are in **bold**. *C* denotes the five-clip ensemble method. *H* is the reconstructed hand representation by MANO [54]. *H+P* or *H+R* in the Signbert mean fusing the reconstructed hand representation (*H*) to the features of the Skeleton-based or RGB-based method. *R+F* denotes using RGB data and optical flow data as inputs.

Model	Per-instance		Per-class	
	Top-1	Top-5	Top-1	Top-5
Signbert (*H+P*) [24]	47.46	83.32	45.17	82.32
Signbert (*H+R*) [24]	54.69	87.49	52.08	86.93
BEST [80]	54.59	88.08	52.12	87.28
Signbert+ [23]	55.59	89.37	53.33	88.82
SAM (7 Modalities + *C*) [27]	59.39	91.48	56.63	90.89
Ours (*R+F*)	**61.17** ↑ 1.78	**91.94** ↑ 0.46	**58.43** ↑ 1.8	**91.43** ↑ 0.54

i.e., Signbert [24], by a large margin, *i.e.*, +10.2% Top-1 increment. Meanwhile, compared with the best RGB-based GLE-Net [26], our method acquires an +8.2% improvement in Top-1 accuracy. For comparison with the fusion-based method, we implement the two-stream framework using RGB and optical flow data, termed as *R+F* network (See Section 3.5 for more details). We also achieve very competitive performance compared with the fusion-based method.

Evaluation on WLASL [33]. WLASL has four subsets, *i.e.*, WLASL-100, WLASL-300, WLASL-1000, and WLASL-2000. The number in the subset represents the number of words provided within the subset. TABLE 4 shows that our proposed method outperforms the previous state-of-the-art method on all subsets by a clear margin. In particular, our method surpasses the recent Skeleton-based SAM [27] with +5.39% Top-1 Per-instance improvement and +5.67% Top-1 Per-class improvement. We also implement StepNet (Optical flow) mentioned above. As shown in TABLE 5, when fusing with the optical flow network, we achieve 61.17% Top-1 er-instance accuracy and 91.94% Top-5 Per-instance accuracy, which outperform the SAM [27] (7 modalities + 5 clips). It is worth noting that we only use two modalities (RGB + Optical flow) to improve our model and only use one clip during inference. Our proposed method surpasses the previous state-of-the-art method with fewer modality inputs.

Evaluation on BOBSL [2]. BOBSL is a new dataset containing more words and sign videos than WLASL-2000 [33]. We observe similar performance improvement. As shown in Table 6, our implemented StepNet (I3D) achieves 77.1% Top-1 accuracy and 92.7% Top-5 accuracy in Per-instance metrics, which boosts the baseline(I3D) by 1.3% Top-1 and 0.3% Top-5 accuracy. In Per-class metrics,

Table 6. Comparison with the state-of-the-art methods in terms of Top-1 and Top-5 accuracy on the BOBSL [2]. The best results are in **bold**.

Model	Per-instance		Per-class	
	Top-1	Top-5	Top-1	Top-5
2D pose-Sign [2]	61.8	82.1	30.6	56.6
Flow-I3D [2]	52.1	75.7	19.2	41.7
RGB-I3D [2]	75.8	92.4	50.5	77.6
Ours (I3D)	**77.1**	**92.7**	**51.3**	**78.2**

Table 7. Ablation studies on WLASL-2000 [33].

(a) Two proposed branches. *w* denotes that we harness with such a component.

Backbone	*w/ Spatial*	*w/ Temporal*	Top-1
I3D + Ours			43.38
	✓		46.06 (+2.68)
		✓	48.52 (+5.14)
	✓	✓	49.46 (+6.08)
TSM + Ours			54.01
	✓		55.89 (+1.88)
		✓	55.05 (+1.04)
	✓	✓	56.89 (+2.88)

(b) Part-level Spatial Modeling. *tb* means top-bottom partition, while *lr* denotes left-right partition.

tb	*lr*	*attention*	*concatenate*	Top-1
				54.01
✓				54.15 (+0.14)
	✓			54.19 (+0.18)
✓	✓		✓	55.05 (+1.04)
✓	✓	✓		55.89 (+1.88)

we also obtain the improvement in Top-1 and Top-5 accuracy. Specifically, Our network arrives at 51.3% Top-1 and 78.2% Top-5 accuracy.

4.4 Ablation Studies

To validate the effectiveness of the proposed method, we design the following ablation studies, *i.e.*, spatial and temporal clues, changing backbones and the fusion ratio. If not specified, we mainly evaluate the model on the widely-used WLASL-2000.

Effect of spatial and temporal clues. As shown in Table 7a, we have several observations: 1) our model based on TSM [36] obtains +1.88% Top-1 accuracy improvement when we merely use the proposed spatial clues. 2) when only using temporal cues, our model also obtains 1.04% Top-1 improvement compared with the baseline (TSM). 3) when fusing spatial and temporal clues, our model obtains 56.89% Top-1 per-instance accuracy, which brings 2.88% accuracy improvement. It also shows that spatial and temporal cues are complementary.

Different backbones. To validate the scalability of our proposed Part-level Spatial and Temporal modeling methods, we explore the influence of the different backbones. We re-implement the I3D backbone for adapting our proposed method and report the re-implemented I3D performance. As shown in Table 7a, our method also yields about +6.08% accuracy improvement on the I3D backbone.

Effect of Spatial and Temporal submodules. In the Part-level Spatial Modeling, we ablate the spatial partition and attention. 1) As shown in Table 7b, we study the left-right and top-bottom partition strategies. The result shows that left-right and top-bottom clues improve the baseline by 0.14% and 0.18%, respectively. 2) To validate the effectiveness of spatial attention, we implement a simple concatenation method for fusing the left-right and top-bottom features. We observe that the

Table 8. Ablation studies of our proposed Part-level Temporal Modeling.

(a) Lengths of the segmented frames and segments.

Temporal	2 Segments	3 Segments	4 Segments
4 Frames	-	-	55.20 (+1.19)
6 Frames	-	55.19 (+1.18)	**55.68 (+1.67)**
8 Frames	55.30 (+1.29)	55.08 (+1.07)	55.33 (+1.32)

(b) Effect of GRUs.

Type	Top-1	Top-5
w/o GRUs	55.23	87.08
w/ GRUs	**56.89**	**88.66**

attention method surpasses the concatenate method by 0.84% accuracy in Table 7b. It shows that the attention method builds the relationship between the global and local clues. 3) As shown in Table 8a, applying any number of segments or length of segments can improve the baseline (54.01% Top-1 accuracy). 4) Four splits generally obtain better performance, and setting the appropriate length of sub-clips, such as 6, also can lead to better accuracy.

Effect of GRUs. As shown in Table 8b, applying GRUs on temporal split features improves the performance before using temporal attention. We consider that this operation helps the model to learn the temporal information that represents the sub-action and global-action of signers.

Effect of different parts. In TABLE 5, we observe 1) The classifier on local parts also achieves a competitive performance. 2) The global representation, based on local parts, usually achieves a consistent improvement. For instance, cls_{lr} is better than either cls_l or cls_r. 3) The final classifier on the fused spatial and temporal representation has arrived at the best Top-1 and Top-5 accuracy. Moreover, we conduct experiments on the local temporal features. Our model achieves 56.93 in top-1, and 87.98 in top-5 accuracy. We can see that introducing supervision of short-term temporal features is also helpful to the model.

StepNet (TSM [36])	Top-1	Top-5
cls_t	48.45	82.18
cls_b	49.84	83.88
cls_{tb}	54.01	87.95
cls_l	51.23	85.59
cls_r	48.73	83.81
cls_{lr}	53.94	87.95
cls_{sg}	54.22	87.81
cls_s	55.96	88.64
cls_t	56.69	88.57
cls_{st}	56.89	88.64

Fig. 5. The Per-instance accuracy of each classifier on WLASL-2000 [33].

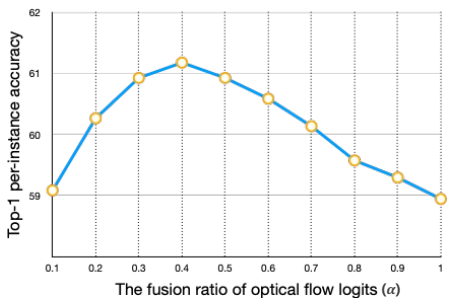

Fig. 6. The performance of our StepNet (*R+F*) when fusing different ratios on WLASL-2000 [33]. The horizontal axis donates the ratio α in Eq.(6).

Effect of fusing the optical flow. Optical flow is able to complement the motion changes between frames for RGB data. To compare with the other fusion-based method, we follow [36] and introduce a StepNet trained on optical flow, which simply replaces the 3-channel RGB inputs with the 10-channel optical flow as discussed in Section 3.5. StepNet (Optical Flow) alone reaches a competitive performance 51.2% Top-1 per-instance accuracy on WLASL-2000 [33]. We apply the late-fusion strategy to fuse predictions of StepNet (RGB) and StepNet (Optical flow). As shown in Figure 6, the final prediction is not sensitive to the fusion rate. Fusing optical flow can always improve our RGB

network (56.89%). Therefore, when applying to an unseen situation, RGB and Optical StepNet with a 1:0.3 to 1:0.5 ratio can be a good initial option.

5 CONCLUSION

We identify the challenge of underexplored spatial and temporal clues in sign language recognition. As an attempt to fill the gap, we propose a new framework, named as StepNet, by constructing two branches, *i.e.*, Part-level Spatial Modeling and Part-level Temporal Modeling. The Part-level Spatial Modeling learns the symmetric association, *e.g.*, between hands and the top-bottom relationship, *e.g.*, between the hands and faces, while the Part-level Temporal Modeling implicitly captures the long-short-term changes. As a result, we achieve competitive performance on three large-scale WLASL, BOBSL, and NMFs-CSL benchmarks, which verifies the effectiveness of the proposed method. In the future, we will continue to study the potential of the proposed method in other fields, such as 3D person re-identification [82], cooking videos with instruction [48], first-view action recognition [46], and video question answering [39, 72].

Limitations. Some of the used networks are out of date, thus the performance may not be better than based on new technology, like transformers. Although the model can handle the optical flow data, the performance may not be better than others designed for optical flow.

Future Enhancements. Recently, transformers [34, 47, 56] have dominated large numbers of domains of computer vision. Thus designing a transformer-based architecture may enhance the vision representation and further improve the performance. Understanding sign language using LLMs [38, 75] may be a new research area. Moreover, designing a pretraining and finetuning schema is also a good solution like BEST [80], Signbert [24]. Our current fusing strategy is based on late fusion which is relatively simple. Designing an effective fusing module is an enhancement for sign language recognition.

REFERENCES

[1] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. 2020. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *ECCV*. 35–53.

[2] Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, et al. 2021. BBC-Oxford British Sign Language Dataset. *arXiv:2111.03635* (2021).

[3] Matyáš Boháček and Marek Hrúz. 2022. Sign Pose-based Transformer for Word-level Sign Language Recognition. In *WACV*. 182–191.

[4] Patrick Buehler, Andrew Zisserman, and Mark Everingham. 2009. Learning sign language by watching TV (using weakly aligned subtitles). In *CVPR*. 2961–2968.

[5] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *CVPR*. 10023–10033.

[6] Congqi Cao, Cuiling Lan, Yifan Zhang, Wenjun Zeng, Hanqing Lu, and Yanning Zhang. 2019. Skeleton-Based Action Recognition With Gated Convolutional Neural Networks. *IEEE TCSVT* 29, 11 (2019), 3247–3257. https://doi.org/10.1109/TCSVT.2018.2879913

[7] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? A new model and the Kinetics dataset. In *CVPR*. 6299–6308.

[8] Wikipedia contributors. 2004. Sign language — Wikipedia The Free Encyclopedia. https://en.wikipedia.org/wiki/Sign_language [Online; accessed 22-July-2004].

[9] Helen Cooper, Eng-Jon Ong, Nicolas Pugeault, and Richard Bowden. 2012. Sign language recognition using sub-units. *JMLR* 13, Jul (2012), 2205–2231.

[10] Runpeng Cui, Hu Liu, and Changshui Zhang. 2019. A Deep Neural Framework for Continuous Sign Language Recognition by Iterative Training. *IEEE TMM* 21, 7 (2019), 1880–1891. https://doi.org/10.1109/TMM.2018.2889563

[11] Yuhang Ding, Xin Yu, and Yi Yang. 2021. RFNet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In *ICCV*. 3975–3984.

[12] Yong Du, Wei Wang, and Liang Wang. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*. 1110–1118.

[13] Jiali Duan, Jun Wan, Shuai Zhou, Xiaoyuan Guo, and Stan Z. Li. 2018. A Unified Framework for Multi-Modal Isolated Gesture Recognition. *ACM TOMM* 14, 1s, Article 21 (feb 2018), 16 pages. https://doi.org/10.1145/3131343

[14] Hehe Fan, Xin Yu, Yuhang Ding, Yi Yang, and Mohan Kankanhalli. 2022. Pstnet: Point spatio-temporal convolution on point cloud sequences. *ICLR* (2022).

[15] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *ICCV*. 6202–6211.

[16] Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. 2021. TMMF: Temporal Multi-Modal Fusion for Single-Stage Continuous Gesture Recognition. *IEEE TIP* 30 (2021), 7689–7701. https://doi.org/10.1109/TIP.2021.3108349

[17] Zan Gao, Leming Guo, Tongwei Ren, An-An Liu, Zhi-Yong Cheng, and Shengyong Chen. 2022. Pairwise Two-Stream ConvNets for Cross-Domain Action Recognition With Small Data. *IEEE Transactions on Neural Networks and Learning Systems* 33, 3 (2022), 1147–1161. https://doi.org/10.1109/TNNLS.2020.3041018

[18] Kirsti Grobel and Marcell Assan. 1997. Isolated sign language recognition using hidden Markov models. In *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, Vol. 1. IEEE, 162–167.

[19] Qingji Guan, Yaping Huang, Zhun Zhong, Zhedong Zheng, Liang Zheng, and Yi Yang. 2020. Thorax disease classification with attention guided convolutional neural network. *Pattern Recognition Letters* 131 (2020), 38–45.

[20] Dan Guo, Wengang Zhou, Anyang Li, Houqiang Li, and Meng Wang. 2019. Hierarchical recurrent deep fusion using adaptive clip summarization for sign language translation. *IEEE Transactions on Image Processing* 29 (2019), 1575–1590.

[21] Dan Guo, Wengang Zhou, Houqiang Li, and Meng Wang. 2017. Online early-late fusion based on adaptive hmm for sign language recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14, 1 (2017), 1–18.

[22] Al Amin Hosain, Panneer Selvam Santhalingam, Parth Pathak, Huzefa Rangwala, and Jana Kosecka. 2021. Hand pose guided 3d pooling for word-level sign language recognition. In *WACV*. 3429–3439.

[23] Hezhen Hu, Weichao Zhao, Wengang Zhou, and Houqiang Li. 2023. SignBERT+: Hand-Model-Aware Self-Supervised Pre-Training for Sign Language Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 9 (Sept. 2023), 11221–11239. https://doi.org/10.1109/tpami.2023.3269220

[24] Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. 2021. SignBERT: Pre-Training of Hand-Model-Aware Representation for Sign Language Recognition. In *ICCV*. 11087–11096.

[25] Hezhen Hu, Wengang Zhou, and Houqiang Li. 2021. Hand-model-aware sign language recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 1558–1566.

[26] Hezhen Hu, Wengang Zhou, Junfu Pu, and Houqiang Li. 2021. Global-local enhancement network for NMF-aware sign language recognition. *ACM transactions on multimedia computing, communications, and applications (TOMM)* 17, 3 (2021), 1–19.

[27] Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. 2021. Sign Language Recognition via Skeleton-Aware Multi-Model Ensemble. *arXiv:2110.06161* (2021).

[28] Oscar Koller, Jens Forster, and Hermann Ney. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding* 141 (2015), 108–125.

[29] Oscar Koller, O Zargaran, Hermann Ney, and Richard Bowden. 2016. Deep sign: Hybrid CNN-HMM for continuous sign language recognition. In *Proceedings of the British Machine Vision Conference*. Article 136, 136.1-136.12 pages.

[30] Oscar Koller, Sepehr Zargaran, Hermann Ney, and Richard Bowden. 2018. Deep sign: Enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs. *IJCV* 126, 12 (2018), 1311–1325.

[31] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. 2018. Co-occurrence Feature Learning from Skeleton Data for Action Recognition and Detection with Hierarchical Aggregation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, 786–792. https://doi.org/10.24963/ijcai.2018/109

[32] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. Transferring cross-domain knowledge for video sign language recognition. In *CVPR*. 6205–6214.

[33] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *WACV*. 1459–1469.

[34] Kexin Li, Zongxin Yang, Lei Chen, Yi Yang, and Jun Xiao. 2023. Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 1485–1494.

[35] Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. 2020. Tea: Temporal excitation and aggregation for action recognition. In *CVPR*. 909–918.

[36] Ji Lin, Chuang Gan, and Song Han. 2019. TSM: Temporal shift module for efficient video understanding. In *ICCV*. 7083–7093.
[37] Jinliang Lin, Zhedong Zheng, Zhun Zhong, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. 2022. Joint Representation Learning and Keypoint Detection for Cross-view Geo-localization. *IEEE Transactions on Image Processing* (2022).
[38] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning.
[39] Rui Liu and Yahong Han. 2022. Instance-sequence reasoning for video question answering. *Frontiers of Computer Science* 16, 6 (2022), 1–9. doi: https://doi.org/10.1007/s11704-021-1248-1.
[40] Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. 2021. iMiGUE: An Identity-free Video Dataset for Micro-Gesture Understanding and Emotion Analysis. arXiv:2107.00285 [cs.CV]
[41] Zhaoyang Liu, Donghao Luo, Yabiao Wang, Limin Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Tong Lu. 2020. Teinet: Towards an efficient architecture for video recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11669–11676.
[42] Zhengzhe Liu, Lei Pang, and Xiaojuan Qi. 2022. MEN: Mutual Enhancement Networks for Sign Language Recognition and Education. *IEEE Transactions on Neural Networks and Learning Systems* (2022), 1–15. https://doi.org/10.1109/TNNLS.2022.3174031
[43] Stephan Liwicki and Mark Everingham. 2009. Automatic recognition of fingerspelled words in British sign language. In *CVPR Workshops*. 50–57.
[44] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv:1711.05101* (2017).
[45] Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Skq89Scxx
[46] Minlong Lu, Ze-Nian Li, Yueming Wang, and Gang Pan. 2019. Deep attention network for egocentric action recognition. *IEEE Transactions on Image Processing* 28, 8 (2019), 3703–3713.
[47] Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. 2024. Vista-LLaMA: Reliable Video Narrator via Equal Distance to Visual Tokens. In *CVPR*.
[48] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*.
[49] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. (2017).
[50] Lionel Pigou, Aäron Van Den Oord, Sander Dieleman, Mieke Van Herreweghe, and Joni Dambre. 2018. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. *International Journal of Computer Vision* 126, 2 (2018), 430–439.
[51] Zhenyue Qin, Yang Liu, Pan Ji, Dongwoo Kim, Lei Wang, R. I. McKay, Saeed Anwar, and Tom Gedeon. 2022. Fusing Higher-Order Features in Graph Neural Networks for Skeleton-Based Action Recognition. *IEEE Transactions on Neural Networks and Learning Systems* (2022), 1–15. https://doi.org/10.1109/TNNLS.2022.3201518
[52] Zhaofan Qiu, Ting Yao, and Tao Mei. 2017. Learning spatio-temporal representation with pseudo-3D residual networks. In *ICCV*. 5533–5541.
[53] Ruijie Quan, Wenguan Wang, Zhibo Tian, Fan Ma, and Yi Yang. 2024. Psychometry: An Omnifit Model for Image Reconstruction from Human Brain Activity. In *CVPR*.
[54] Javier Romero, Dimitrios Tzionas, and Michael J. Black. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36, 6 (Nov. 2017).
[55] Prem Selvaraj, Gokul NC, Pratyush Kumar, and Mitesh Khapra. 2021. OpenHands: Making Sign Language Recognition Accessible with Pose-based Pretrained Models across Languages. arXiv:2110.05877 [cs.CL]
[56] Xiaolong Shen, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, and Yi Yang. 2023. Global-to-local modeling for video-based 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8887–8896.
[57] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2020. Skeleton-Based Action Recognition With Multi-Stream Adaptive Graph Convolutional Networks. *IEEE Transactions on Image Processing* 29 (2020), 9532–9545. https://doi.org/10.1109/TIP.2020.3028207
[58] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems* 27 (2014).
[59] Ozge Mercanoglu Sincan, Anil Osman Tur, and Hacer Yalim Keles. 2019. Isolated sign language recognition with multi-scale features using LSTM. In *2019 27th Signal Processing and Communications Applications Conference (SIU)*. IEEE, 1–4.
[60] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. 2017. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *AAAI*. 4263–4270.
[61] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*.

480–496.
[62] Yucheng Suo, Zhedong Zheng, Xiaohan Wang, Bang Zhang, and Yi Yang. 2024. Jointly Harnessing Prior Structures and Temporal Consistency for Sign Language Video Generation. *ACM Trans. Multimedia Comput. Commun. Appl.* 20, 6, Article 185 (mar 2024), 18 pages. https://doi.org/10.1145/3648368
[63] Alaa Tharwat, Tarek Gaber, Aboul Ella Hassanien, Mohamed K Shahin, and Basma Refaat. 2015. SIFT-based Arabic sign language recognition system. In *Proceedings of the Afro-European Conference for Industrial Advancement*. 359–370.
[64] Suramya Tomar. 2006. Converting video formats with FFmpeg. *Linux Journal* 2006, 146 (2006), 10.
[65] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*. 6450–6459.
[66] Anirudh Tunga, Sai Vidyaranya Nuthalapati, and Juan Wachs. 2020. Pose-based Sign Language Recognition using GCN and BERT. In *WACV Workshop*. 31–40.
[67] Anirudh Tunga, Sai Vidyaranya Nuthalapati, and Juan Wachs. 2021. Pose-based sign language recognition using gcn and bert. In *WACV*. 31–40.
[68] Li-Chun Wang, Ru Wang, De-Hui Kong, and Bao-Cai Yin. 2014. Similarity Assessment Model for Chinese Sign Language Videos. *IEEE Transactions on Multimedia* 16, 3 (2014), 751–761. https://doi.org/10.1109/TMM.2014.2298382
[69] Shiguang* Wang, Zhizhong* Li, Yue Zhao, Yuanjun Xiong, Limin Wang, and Dahua Lin. 2020. denseflow. https://github.com/open-mmlab/denseflow.
[70] Tingyu Wang, Zhedong Zheng, Chenggang Yan, jiyong Zhang, Yaoqi Sun, Bolun Zheng, and Yi Yang. 2021. Each Part Matters: Local Patterns Facilitate Cross-view Geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology* (2021). doi: 10.1109/TCSVT.2021.3061265.
[71] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*. 305–321.
[72] Hongyang Xue, Zhou Zhao, and Deng Cai. 2017. Unifying the Video and Question Attentions for Open-Ended Video Question Answering. *IEEE Transactions on Image Processing* 26, 12 (2017), 5656–5666. https://doi.org/10.1109/TIP.2017.2746267
[73] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*. 7444–7452.
[74] Hao Yang, Chunfeng Yuan, Li Zhang, Yunda Sun, Weiming Hu, and Stephen J. Maybank. 2020. STA-CNN: Convolutional Spatial-Temporal Attention Learning for Action Recognition. *IEEE Transactions on Image Processing* 29 (2020), 5783–5793. https://doi.org/10.1109/TIP.2020.2984904
[75] Zongxin Yang, Guikun Chen, Xiaodi Li, Wenguan Wang, and Yi Yang. 2024. Doraemongpt: Toward understanding dynamic scenes with large language models. *arXiv preprint arXiv:2401.08392* (2024).
[76] Zongxin Yang, Yunchao Wei, and Yi Yang. 2021. Collaborative video object segmentation by multi-scale foreground-background integration. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 9 (2021), 4701–4712.
[77] Farhad Yasir, PW Chandana Prasad, Abeer Alsadoon, and Amr Elchouemi. 2015. SIFT based approach on bangla sign language recognition. In *IWCIA*. 35–39.
[78] Zitong Yu, Benjia Zhou, Jun Wan, Pichao Wang, Haoyu Chen, Xin Liu, Stan Z. Li, and Guoying Zhao. 2021. Searching Multi-Rate and Multi-Modal Temporal Enhanced Networks for Gesture Recognition. *IEEE Transactions on Image Processing* 30 (2021), 5626–5640. https://doi.org/10.1109/tip.2021.3087348
[79] Christopher Zach, Thomas Pock, and Horst Bischof. 2007. A duality based approach for realtime TV-L1 optical flow. In *Proceedings of Joint Pattern Recognition Symposium*. Springer, 214–223.
[80] Weichao Zhao, Hezhen Hu, Wengang Zhou, Jiaxin Shi, and Houqiang Li. 2023. BEST: BERT Pre-Training for Sign Language Recognition with Coupling Tokenization. arXiv:2302.05075 [cs.CV]
[81] Zengqun Zhao, Qingshan Liu, and Shanmin Wang. 2021. Learning Deep Global Multi-Scale and Local Attention Features for Facial Expression Recognition in the Wild. *IEEE Transactions on Image Processing* 30 (2021), 6544–6556. https://doi.org/10.1109/TIP.2021.3093397
[82] Zhedong Zheng, Xiaohan Wang, Nenggan Zheng, and Yi Yang. 2022. Parameter-Efficient Person Re-identification in the 3D Space. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)* (2022). https://doi.org/10.1109/TNNLS.2022.3214834
[83] Zhedong Zheng, Liang Zheng, and Yi Yang. 2018. Pedestrian alignment network for large-scale person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 10 (2018), 3037–3045. doi:10.1109/TCSVT.2018.2873599.
[84] Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. 2022. Spatial-Temporal Multi-Cue Network for Sign Language Recognition and Translation. *IEEE Transactions on Multimedia* 24 (2022), 768–779. https://doi.org/10.1109/TMM.2021.3059098

Received October 2023; revised xx xxxx; accepted xx xxxx