Sign Language Recognition Based on R(2+1)D With Spatial–Temporal–Channel Attention

Xiangzu Han, Fei Lu, Jianqin Yin, *Member, IEEE*, Guohui Tian, *Member, IEEE*, and Jun Liu, *Member, IEEE*

***Abstract*—Previous work utilized three-dimensional (3-D) convolutional neural networks (CNNs) tomodel the spatial appearance and temporal evolution concurrently for sign language recognition (SLR) and exhibited impressive performance. However, there are still challenges for 3-D CNN-based methods. First, motion information plays a more significant role than spatial content in sign language. Therefore, it is still questionable whether to treat space and time equally and model them jointly by heavy 3-D convolutions in a unified approach. Second, because of the interference from the highly redundant information in sign videos, it is still nontrivial to effectively extract discriminative spatiotemporal features related to sign language. In this study, deep R(2+1)D was adopted for separate spatial and temporal modeling and demonstrated that decomposing 3-D convolution filters into independent spatial and temporal convolutions facilitates the optimization process in SLR. A lightweight spatial–temporal–channel attention module, including two submodules called channel–temporal attention and spatial–temporal attention, was proposed to make the network concentrate on the significant information along spatial, temporal, and channel dimensions by combining squeeze and excitation attention with self-attention. By embedding this module into R(2+1)D, superior or comparable results to the state-of-the-art methods on the CSL-500, Jester, and EgoGesture datasets were obtained, which demonstrated the effectiveness of the proposed method.**

***Index Terms*—Attention mechanism, R(2+1)D, sign language recognition (SLR).**

I. Introduction

Sign language recognition (SLR) aims to convert sign language, a nonverbal communication form for deaf-mute people, into speech or text to bridge the gap between deaf–mute communities and hearing people. SLR can also be applied as a human–computer interaction (HCI) tool and promote communication between humans and machines. In recent years, SLR has made significant progress due to the increasing development of deep learning [1] but has not been entirely resolved.

The first challenge is spatiotemporal modeling. The two-dimensional (2-D) convolutional neural networks (CNNs) have been well demonstrated in various vision tasks, such as object detection [2], [3] and image classification [4]–[8]. A simple way to extend 2-D CNNs is to use off-the-shelf 2-D CNNs as frame-level feature extractors and then perform a later fusion to predict the video category [9]–[11]. However, this method ignores the temporal evolution between adjacent frames, and the performance is even worse than that of handcrafted feature-based methods. Some [12]–[17] have proposed adopting recurrent neural networks (RNNs) over the top of 2-D CNNs to model the temporal connection. However, this kind of method only builds the temporal relationship of high-level features and is difficult to optimize. Recently, 3-D CNNs have demonstrated the capacity of spatiotemporal modeling simultaneously [18]–[22] from low-level to high-level benefit from the hierarchical architecture, but the disadvantage is that 3-D CNNs are of high computational cost and thus challenging to optimize. In addition, with the directly extended temporal dimension of the original 2-D convolution, 3-D convolution treats spatial and temporal dimensions equally. However, space and time are not always equivalent, and jointly modeling them may also lead to optimization difficulties. Although the emergence of large-scale datasets and pretrained models alleviates the difficulty of training a deep 3-D CNN to some extent [23], whether it is necessary to model spatial appearance and temporal evolution jointly by 3-D convolution filters is still a question in SLR, especially considering that most of the video samples cannot be distinguished by similar spatial content, while the temporal dynamics are of more vital significance in sign language.

Second, video-based SLR is still a challenge due to the highly redundant information in space and time, which may distract a video model from extracting discriminative spatiotemporal features and thus limit the performance improvement. A sign gloss consists of manual features such as fine-grained hand gestures and even nonmanual features, including facial expressions and upper-body posture [24]. These regions of interest are relatively smaller than the whole video frame, so the cluttered background easily misguides the network. Furthermore, videos are redundant along the temporal dimension, and there is usually no distinct difference between consecutive video frames. Therefore, it is a challenge to focus on the exact spatial regions

Manuscript received January 10, 2021; revised May 23, 2021, July 15, 2021, and November 15, 2021; accepted January 9, 2022. Date of publication February 3, 2022; date of current version July 14, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61973187 and Grant 61773239, in part by the Guangdong Province Basic and Applied Basic Research Fund Project under Grant 2019A1515110175, and in part by the Research Grants Council of Hong Kong under Grant 21212720. This article was recommended by Associate Editor H. Zhou. *(Corresponding authors: Fei Lu; Jun Liu.)*

Xiangzu Han, Fei Lu, and Guohui Tian are with the School of Control Science and Engineering, Shandong University, Jinan 250061, China (e-mail: xiangzu@mail.sdu.edu.cn; lawyerlf@sdu.edu.cn; g.h.tian@sdu.edu.cn).

Jianqin Yin is with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100088, China (e-mail: jqyin@bupt.edu.cn).

Jun Liu is with the Department of Mechanical Engineering,City University of Hong Kong, Hong Kong (e-mail: jun.liu@cityu.edu.hk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/THMS.2022.3144000.

Digital Object Identifier 10.1109/THMS.2022.3144000

and temporal moments where and when motion occurs. Some works used external tools, such as hand detection [11], [14], [25], [26], to focus on the hand, but the performance depends heavily on these tools and ignores nonmanual features. Some relied heavily on multistream inputs such as depth, skeleton, and optical flow [27]–[36] to make the network extract more discriminative features related to sign motions. However, the repetitive feature extraction of multistream inputs inevitably increases the demand for memory and computing resources.

In contrast, human vision systems have effective attention mechanisms to concentrate on the relevant objects and motion patterns of regions of interest and significant moments without external tools or multichannel inputs, which is difficult for deep neural networks. In previous research, attention mechanisms have been investigated and demonstrated remarkable improvement [8], [37]–[39]. Despite some successful attempts at the attention mechanism in SLR [31], [40]–[43], few studies have examined the combination of spatial, temporal, and channel attention.

In this article, to effectively model the spatial appearance and temporal evolution, R(2+1)D [44]was adopted, which decomposes 3-D convolutions into 2-D convolutions for spatial modeling and 1-D convolutions for temporal learning as the backbone. In conducted experiments, it was demonstrated that (2+1)D decomposition greatly facilitates the difficulty of training a 3-D CNN and shows obvious performance improvement. To eliminate the interference from irrelevant information and extract more discriminative spatiotemporal features without adding heavy overhead, a lightweight SE-like attention module called spatial–temporal–channel attention (STCA) was proposed that decomposes the whole STCA into two submodules: channel–temporal attention (CTA) and spatial–temporal attention (STA). For CTA, the global information along spatial dimensions was first obtained to generate the channel-temporal descriptor (channel descriptor with temporal dimension). Then, a temporalwise multi-layer perceptron (MLP) network with a self-attention operation was adopted to model the channel relation and global temporal evolution. For STA, global information was utilized along the channel dimension, and multipath convolution layers were adopted for multiscale spatial modeling and local temporal learning. The two submodules were complementary for the whole STCA, especially both local and global temporal components. It was demonstrated that R(2+1)D with the STCA module improved the model performance and obtained superior or competitive results to state-of-the-art methods on the CSL-500, Jester, and EgoGesture datasets.

The rest of this article is organized as follows. First, we review the related work in Section II. Then, we describe the details of our method in Section III. Experiments and analyses on CSL-500 and Jester are shown in Section IV. Finally, Section V concludes this article.

II. Related Work

A. Spatiotemporal Modeling of CNNs

CNNs have proven to be effective methods of tackling various vision tasks, such as object detection [2], [3], image classification [4]–[8], and action recognition [9], [12], [13], [18]–[21], [44]–[50]. Karpathy *et al.* [9] proposed extending image-based 2-D CNN models into video tasks by later fusion to predict the action class of a video. However, the performance does not compare with the handcrafted feature methods without considering the motion changes among consecutive frames. Some [12], [13] proposed utilizing RNNs over the representations of frames to model the temporal evolution of videos. However, they only build a high-level temporal connection while neglecting the low-level component.

To incorporate temporal information, Simonyan *et al.* [45] introduced a two-stream network. This design consists of a spatial path to learn the spatial appearance and a temporal path to model the temporal evolution by optical flow. TSN [46] extended such a two-steam framework and proposed the sparse sampling strategy. Although the two-steam architecture has powerful spatiotemporal learning ability, there is an expensive computational cost for calculating the optical flow.

Recently, because of the pretraining on large-scale video datasets, 3-D CNNs have demonstrated their effectiveness over 2-D CNNs in video understanding tasks such as action recognition [18]–[21], [44], [47]. Tran *et al.* [18] proposed the first well-known but shallow 3-D CNN network C3D, which is designed based on VGG [4]. To make full use of off-the-shelf pretrained 2-D CNNs, I3D [19] proposed inflating 2-D Inception [5]. Hara *et al.* [21] presented very deep 3D-ResNet and its variants by replacing all the 2-D filters with 3-D convolution filters and demonstrated that pretrained models on Kinetics [23] contribute to significant progress in various video tasks. Feichtenhofer *et al.* [22] proposed a two-path design called SlowFast for capturing both static and dynamic information and showed impressive performance. To reduce the heavy computational cost of 3-D CNNs, Tran *et al.* [44] and Xie *et al.* [47] rethought spatial and temporal modeling, proposed factorizing 3-D convolutions into 2-D spatial convolutions and 1-D temporal convolutions, and demonstrated that this (2+1)D is also much easier to optimize.

B. Sign Language Recognition (SLR)

In addition to being a tool for HCI to promote communication between humans and machines, SLR can also convert videos of hand signs into text or speech, which is helpful in real-world scenarios to assist people with speech or hearing impairments [51].

SLR generally includes two tasks, i.e., isolated SLR (or isolated hand gesture recognition) [10], [14]–[17], [27]–[36], [40], [41], [52]–[59], which aims to recognize sign glosses independently, and continuous SLR [24], [42], [43], [60]–[63], which recognizes a longer sign video into ordered glosses. In this article, we focus on the basic isolated SLR.

Conventional approaches usually extract hand-crafted features [25], [61], [64] and utilize hidden Markov (HMM) [33], [52], [60] or dynamic time warping [53], [61] to model the temporal relationships in sign videos. With the great success of 2-D CNNs, Pigou *et al.* [10] adopted a 2-D CNN as the backbone to extract frame-level features and then some fusion strategies [9] for the classification. Mohammed *et al.* [11] proposed using hand detection and lightweight 2-D CNN for gesture recognition. However, such an approach cannot effectively capture motion information and results in temporal information loss because the framewise feature is isolated with adjacent frames. Due to the great success of RNNs for temporal modeling, others [14]–[17] adopted this sequence model on top of a 2-D CNN to model the temporal relation. Nevertheless, the frame-level feature is of a high level and may also lose some valuable low-level information. Kopuklu *et al.* [59] introduced a novel method to fuse the data and recognize hand gestures based on TSN [46].

Multimodality has been shown to be helpful in SLR. Kumar *et al.* [33] proposed an isolated SLR method based on multiple classifiers, including HMM and long short-term memory (LSTM), for multimodal inputs. Xiao *et al.* [34] also proposed a similar approach for isolated Chinese SLR. Chansri *et al.* [35] adopted a later fusion for the depth and color information from the captured images. Similar research by Bird *et al.* [36] also confirmed that the fusion-based method was effective for multimodal American SLR.

The 3-D CNNs have proven to be effective in spatiotemporal modeling simultaneously; therefore, many 3-D CNN-based networks have been applied to SLR [27]–[31], [54], [55], [58] and can further be improved by multimodal data. Molchanov *et al.* [54] proposed a shallow 3-D CNN to recognize hand gestures using depth video and intensity data. Huang *et al.* [27] also proposed a shallow 3-D CNN-based method using multimodal inputs for SLR. Molchanov *et al.* [55] used both RNNs and 3-D CNNs for hand detection and recognition. Wu *et al.* [28] proposed applying 3-D CNN for multimodal gesture recognition. Li *et al.* [29] proposed a large-scale gesture recognition approach using C3D. Li *et al.* [30] proposed using C3D and multimodal data for gesture recognition. Although multimodal approaches can provide significant performance gains, the additional memory and computational costs cannot be ignored. Zhang *et al.* [58] proposed deep deformable 3-D CNNs for gesture recognition but only utilized the RGB modality for efficiency.

C. Attention Mechanism

Attention is of vital importance in human perception [37]. The attention mechanism is characterized by concentrating on discriminative information, which is more critical to the task. Recently, there have been many attempts [8], [37], [38] to adopt the attention mechanism to improve CNNs. Hu *et al.* [8] introduced a lightweight squeeze and excitation (SE) module to exploit the channel relationship based on global average-pooling information. Woo *et al.* [37] further exploited spatial and channel attentions and proposed a convolutional block attention module (CBAM). Perez *et al.* [38] also proposed an SE-like attention module that decomposes the process of generating the STCA into two subprocesses. In addition, in natural language processing (NLP), Vaswani *et al.* [39] proposed a novel self-attention operation for global modeling, and their transformer dominated many NLP tasks.

D. Applications of Attention in SLR

In SLR, attention mechanisms, including spatial attention, temporal attention, and self-attention, have been used to improve performance. Huang *et al.* [31] proposed an attention-based C3D [18] using multimodal inputs for large-vocabulary isolated SLR. In this article, an untrainable spatial attention mask was applied for selected joints, and a temporal attention module based on LSTM was integrated to model the attention of different clips for the transformation of clip-level predictions into video-level predictions. In other studies, DE *et al.* [40] utilized a multi-head attention-based transformer encoder for SLR, and Slimane *et al.* [41] adopted a similar method to consider the context impact. Camgoz *et al.* [42], [43] adopted the transformer-based encoder–decoder framework and multichannel inputs to achieve high reliability for SLR.

III. Methodology

An overview of the proposed approach is illustrated in Fig. 1. First, RGB videos were randomly sampled into long-range clips using the sparse sampling strategy [46]. Then, these clips were forwarded into R(2+1)D with the STCA block. Finally, the prediction of a sign video was obtained. The details of the approach and discussion are as follows.

A. Channel–Temporal Attention

Each channel of the deep network was recognized as a feature extractor [37], so the CTA module computes the significance of different object-motion patterns and their temporal evolution. Given a 4-D feature map $F \in \mathbb{R}^{C\times T\times W\times H}$, the global information was utilized by average pooling and max pooling, as in [37] and [38], along the spatial dimension ($W \times H$) to yield two compact channel-temporal descriptors $d_{\text{ct-max}}$ and $d_{\text{ct-avg}}$ ($d \in \mathbb{R}^{C\times T}$), each of which can be recognized as a channel descriptor of size C evolving T moments.

To model the channel relation, $d_{\text{ct-max}}$ and $d_{\text{ct-avg}}$ were forwarded into a temporalwise MLP network with one hidden layer as in [8], [37], and [38]. The design of the reduction ratio r was also followed, which reduces the number of neurons in the hidden layer to the size of C/r to limit the complexity. To model the temporal dynamic, different from [38], which uses local temporal convolution, a self-attention layer with position encoding [(1) and (2) were adopted, and (5) was proposed in [39]) was used for global temporal attention. Considering the computational complexity, the self-attention operation was inserted into the middle of the bottleneck for efficiency, which was also different from W3 [38]. A sigmoid function was finally applied to generate a single CTA descriptor D_{CT}. This process is summarized as

$$\text{PE}(\text{pos}, 2i) = \sin(\text{pos}/10\,000^{2i/c}) \tag{1}$$

$$\text{PE}(\text{pos}, 2i+1) = \cos(\text{pos}/10\,000^{2i/c}) \tag{2}$$

$$d_{\text{ct}} = relu\left(w_1\left(d_{\text{ct-max}} + d_{\text{ct-avg}}\right)\right)) \tag{3}$$

$$[q_{ct}, k_{ct}, v_{\text{ct}}] = [w_q, w_k, w_v] \cdot (\text{PE}({d_{\text{ct}}}^T) + {d_{\text{ct}}}^T) \tag{4}$$

$$d'_{\text{ct}} = \text{ softmax}\left(\frac{q_{\text{ct}} \cdot k_{\text{ct}}^T}{\sqrt{d_k}}\right) \cdot v_{\text{ct}} \tag{5}$$

$$D_{\text{CT}} = \sigma(w_2\left({d'_{\text{ct}}}^T\right)) \tag{6}$$

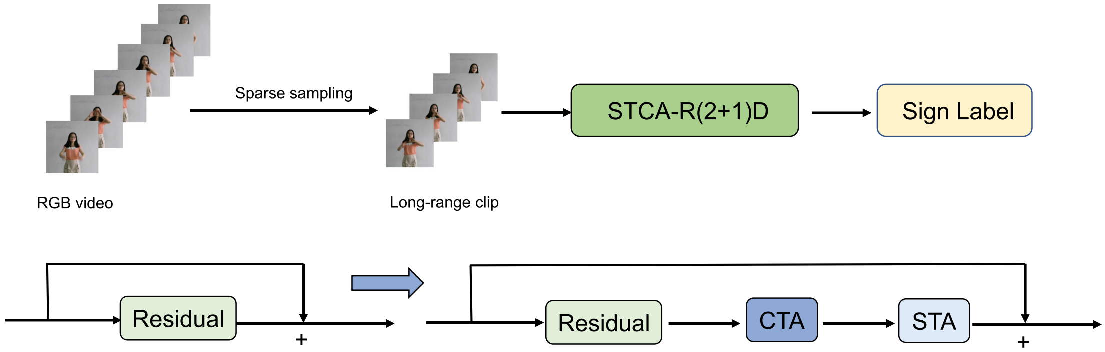

Fig. 1. Top: An overview of our method for isolated SLR; the sparse sampling strategy is utilized to generate a long-range clip of T frames. Bottom: The CTA and STA are inserted into each (2+1)D residual block.

$$c = d_k = C/r \tag{7}$$

where PE denotes the position encoding, $w_1 \in \mathbb{R}^{C\times C/r}$ and $w_2 \in \mathbb{R}^{C/r\times C}$ denote the components of MLP, $w_q \in \mathbb{R}^{C/r\times C/r}$, $w_k \in \mathbb{R}^{C/r\times C/r}$, and $w_v \in \mathbb{R}^{C/r\times C/r}$ denote the linear projection for the self-attention as in [39] to generate the query (q_{ct}), key (k_{ct}), and value (v_{ct}) vector, d_k denotes the dimension of k_{ct} and is equal to both c and C/r, and σ denotes a sigmoid function.

In addition, a residual connection was adopted inside CTA when refining the feature map to preserve the original feature map. The final process is summarized as

$$F_{\text{CT}} = D_{\text{CT}} \bigotimes F + F \tag{8}$$

where $\bigotimes$ denotes the elementwise multiplication with broadcasting, $+$ denotes the inner residual connection in CTA, and F_{CT} denotes the refined feature map after the CTA submodule.

B. Spatial–Temporal Attention

A sign consists of the temporal evolution of hands, body, head, eyes, and faces, so STA modeling was crucial to concentrate on these regions of interest evolving over time. Given a 4-D feature map $F \in \mathbb{R}^{C\times T\times W\times H}$, global information was first utilized by both max-pooling and average-pooling along the channel dimension to obtain two spatial–temporal descriptors $d_{\text{st-max}} \in \mathbb{R}^{1\times T\times W\times H}$and $d_{\text{st-avg}} \in \mathbb{R}^{1\times T\times W\times H}$. Then, the descriptors were directly concatenated along the channel dimension into $d_{\text{st}} \in \mathbb{R}^{2\times T\times W\times H}$ as in [37] and [38]. Different from CBAM and W3, which adopt fixed filter size for both spatial convolution and temporal convolution filters, inspired by Szegedy *et al.* [5], the d_{st} was then forwarded into a multipath 2-D convolution layer (filter size of each path is set to 3, 5, and 7) to exploit the multiscale spatial relation and 1-D temporal convolution layer (filter size is set to 3, 5, and 7) to model the multiscale local temporal interdependency. Finally, a sigmoid function was applied to yield the final STA descriptor D_{ST}. In summary, the STA descriptor is computed as

$$d_{\text{st}} = \text{concat}\left[d_{\text{st-max}}, d_{\text{st-avg}}\right] \tag{9}$$

$$D_{\text{ST}} = \sigma(\text{Conv1}d\left(relu(\text{Conv2}d\left(D_{\text{st}}\right))\right)) \tag{10}$$

where σ denotes a sigmoid function, Conv2d represents the multiscale 2-D spatial convolution layer, and Conv1d denotes the multipath 1-D temporal convolution layer.

Similarly, a residual connection was also adopted inside the STA to preserve the original information (the last plus sign at the bottom of Fig. 2). The last process of STA to refine the feature map is summarized as

$$F_{\text{STC}} = D_{\text{ST}} \bigotimes F_{\text{CT}} + F_{\text{CT}} \tag{11}$$

where F_{CT} denotes the refined feature map after CTA and F_{STC} denotes the final refined feature map.

C. Integration With Residual Blocks

R(2+1)D is a temporal extension of standard ResNet [6]. In the conducted experiments, the same expanding strategy as [44] was first adopted, which used one downsampling along the spatial dimension at conv1 and three downsampling at conv3_1, conv4_1, and conv5_1 along spatial and temporal dimensions to obtain R3D. Then, the spatial and temporal modeling was decomposed by factorizing 3-D convolution filters and constructing R(2+1)D by replacing the 3-D residual block of R3D with a (2+1)D block. Tran *et al.* [44] provide more specifications of the R(2+1)D architectures considered in conducted experiments.

As illustrated in Fig. 3, our STCA-R(2+1)D was built by inserting CTA and STA into each residual block. Therefore, there were two types of residual connections: the adopted residual connections inside CTA and STA and the original residual connection in ResNet. CTA and STA were combined to simulate human attention mechanisms to concentrate on discriminative features (channel) of regions of interest (spatial) and significant moments (temporal). As illustrated in Fig. 2, given a 4-D feature map $F \in \mathbb{R}^{C\times T\times W\times H}$ in the residual block as input, CTA first computed a 2-D CTA descriptor $D_{\text{CT}} \in \mathbb{R}^{C\times T}$ using global spatial information and then generated the refined feature map F_{CT}; STA first generated a 3-D STA descriptor $D_{\text{ST}} \in \mathbb{R}^{T\times W\times H}$ using global channel information and then generated the final feature map of the convolution path of the residual block, which was refined along spatial, temporal, and channel dimensions.

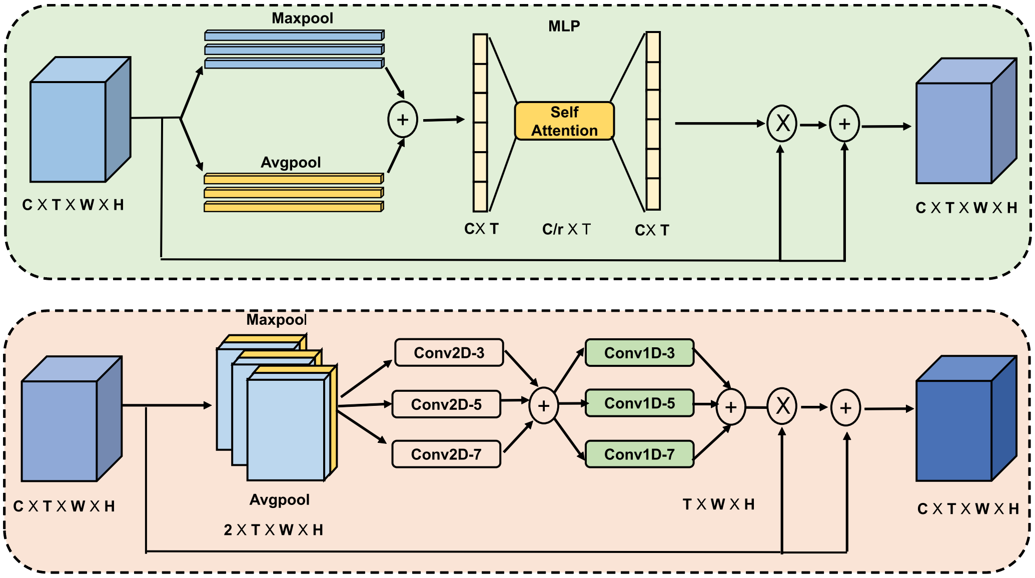

Fig. 2. Overview of the proposed STCA module. Top: CTA. Bottom: STA.

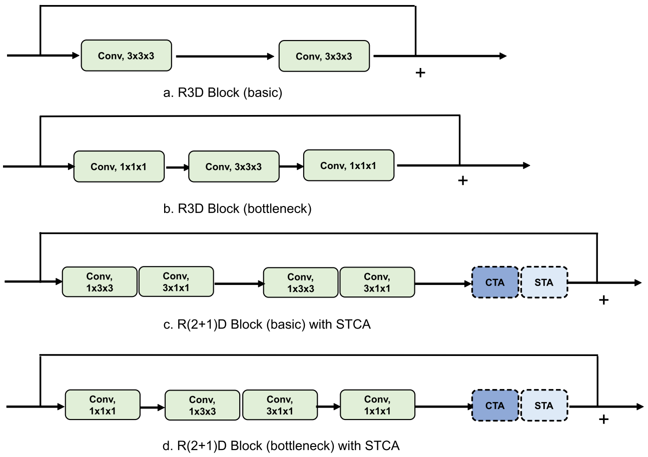

Fig. 3. Altered residual blocks in our experiments.

D. Comparisons With SE, CBAM, and W3

Inspired by SE [8] and CBAM [37], global information was utilized and temporalwise MLP with a bottleneck design for modeling the channel relation in the STCA was used. Following W3 [38], which shows strong performance in video understanding, the static SE and CBAM were modified by introducing temporal attention, and the whole STCA module was decomposed into two submodules.

The first difference between STCA and W3 was the temporal modeling in CTA. Global temporal modeling was adopted by a self-attention operation in this article, while W3 was focused on local modeling using temporal convolution layers. Instead of simply replacing the temporal convolution, the temporal modeling component was inside the bottleneck, making the module lighter. However, the temporal convolution layers of W3 introduce too many parameters (see Tables II and VII).

The second improvement was that multiscale convolution filters were adopted that increased the receptive field and generalization performance for generating the spatial–temporal descriptor in STA, while W3 only utilizes convolution filters of fixed size. Although it was proven in CTA that the self-attention operation performs better than temporal convolution and even multiscale temporal convolution layers (see Table II), considering the rapidly increasing module complexity and the risk of overfitting (the complexity of self-attention after the spatial convolution of STA is O ($T^2\times$ H × W)), self-attention was not utilized in STA (see Table IV). Instead, multiscale local convolutional operations were adopted for temporal learning.

Finally, the W3 module in [38] did not use inner residual connections that were critical in the proposed method. Without the inner residual connection, the background noise or unrelated features in the current layers will be permanently suppressed. In the proposed STCA, the inner residual connection layer can preserve the original feature map (the output of the main branch), which is helpful for subsequent layers (see Table VI).

IV. Experimental Details

We performed our experiments on three datasets, i.e., CSL-500 [31], Jester [65], and EgoGesture [66]. First, we introduce the three datasets. Then, the implementation details of our experiments are described. Finally, the results and analyses are discussed.

A. Datasets

1) CSL-500 Dataset: The CSL-500 [31] dataset consists of a single vocabulary of 500 signs in daily life, and each is recorded five times by 50 signers. Therefore, there were 125 000 instances in total. Each sample includes RGB, depth, and skeleton data. In this study, only RGB was utilized.

TABLE I
COMPARISON RESULTS ON CSL-500 [(2+1)D VERSUS 3-D]

Network	Params(M)	GFLOPs	Pretraining dataset	Test Top1-Acc(%)
R3D-18	33.46	55.68	Kinetics	88.69
R3D-18+STCA	33.56	55.72	Kinetics	90.15
R(2+1)D-18	33.42	54.36	Kinetics	95.26
R(2+1)D-18+STCA	33.52	54.40	Kinetics	96.22
R3D-34	63.77	101.00	Kinetics	90.58
R3D-34+STCA	63.95	101.01	Kinetics	91.74
R(2+1)D-34	63.74	99.76	Kinetics	96.08
R(2+1)D-34+STCA	63.93	99.82	Kinetics	96.94
R3D-50	47.22	74.17	Kinetics	93.26
R3D-50+STCA	50.06	74.30	Kinetics	94.23
R(2+1)D-50	47.19	75.71	Kinetics	96.55
R(2+1)D-50+STCA	50.02	75.83	Kinetics	**97.45**
R3D-152	118.43	163.57	Kinetics	91.82
R(2+1)D-152	118.43	165.16	Kinetics	95.07

TABLE II
EFFECT OF SELF-ATTENTION IN CTA

Setting	Position	Test Top1	Params(M)	GFLOPs
1D CNN(k=3)	M	97.20	49.9	75.832
1D CNN(k=5)	M	97.24	50.1	75.832
1D CNN(k=7)	M	97.24	50.3	75.833
1D CNN(k=3,5,7)	M	97.32	50.9	75.834
Self-attention	M	97.45	50.0	75.832
Self-attention	L	97.50	110.0	76.020

TABLE III
IMPACT OF r IN CTA

r	Test Top1	Params(M)	GFLOPs
8	97.53	53.48	75.832
16	97.45	50.02	75.832
32	97.17	48.53	75.832

TABLE IV
IMPACT OF MULTISCALE SETTING IN STA

Setting	Top1	Params(M)	GFLOPs
k=3	97.31	50	75.79
k=5	97.35	50	75.80
k=7	97.37	50	75.81
Multi-scale(k=3,5,7)	**97.45**	50	75.83
self-attention	96.36	215	78.34

TABLE V
COMPONENT ANALYSIS OF STCA

Setting	Top1	Params(M)	GFLOPs
R(2+1)D-50	96.55	47.19	75.71
R(2+1)D-50+STA	96.87	47.19	75.75
R(2+1)D-50+CTA	97.26	50.02	75.79
R(2+1)D-50+STA+CTA	97.42	50.02	75.83
R(2+1)D-50+CTA+STA	**97.45**	50.02	75.83

TABLE VI
IMPACT OF INNER RESIDUAL CONNECTION IN STCA

Setting	Top1
R(2+1)D-50	96.55
R(2+1)D-50+STCA(w/o residual connection)	97.36
R(2+1)D-50+STCA(w/ residual connection)	**97.45**

TABLE VII
COMPARISON WITH SE-LIKE ATTENTION MODULES

Setting	Top1	Params(M)	GFLOPs
R(2+1)D-50	96.55	47.19	75.71
R(2+1)D-50+SE	97.12	49.70	75.78
R(2+1)D-50+CBAM	97.15	49.71	75.81
R(2+1)D-50+W3	97.30	110.21	75.93
R(2+1)D-50+STCA	**97.45**	50.02	75.83

For a fair evaluation, as in [31], 36 signers were selected for training (90 000 videos) and the rest for testing (35 000 videos), and no overlap exists between the signers of the two subsets.

2) Jester Dataset: The Jester dataset [65] contains 27 kinds of predefined hand gestures. There are 148 092 instances in total and are officially segmented into a training set (118 562 samples), validation set (14 787 samples), and test set (14 743 samples). Although sign language, which also involves nonmanual components, is not equal to hand gestures, the performance of the proposed method can also be evaluated, especially the performance of temporal modeling.

3) EgoGesture Dataset: EgoGesture [66] is also a large-scale egocentric hand gesture dataset containing 2081 RGB and depth videos and 24 161 gesture samples involving 83 dynamic gesture classes. In the experiment, the trained model using the validation subset and test subset was evaluated.

B. Implementation Details

R(2+1)D was utilized as the backbone and inserted the STCA into each residual block. In CTA, the reduction ratio r of CTA was initially set to 16. Pretrained models were utilized on Kinetics [23] for parameter initialization.

For all experiments without additional explanation, RGB only was adopted as input, and the sparse sampling strategy [46] was adopted to produce long-range clips ($T = 16$ in the experiments without other explanation). First, a video of N frames was evenly divided into T segments. Then, one frame was randomly selected from each segment to generate a long-range clip with T frames. Due to the sparse strategy, the long-range clips can represent the whole temporal evolution of sign videos. Because of random selection, the sparse sampling strategy can be recognized as a temporal augmentation method.

During the training process, the short side of the frames was scaled to s, and the original aspect ratio was maintained. s was randomly selected between 144 and 170. Then, random cropping was utilized, and the cropped size of each frame was 128×128. All networks were trained with SGD, and the batch size, initial learning rate, weight decay, and momentum were 16, 0.001, 1e-4, and 0.9, respectively.

During the test process, each frame of the long-range clip generated by the sparse sampling strategy was first resized to 144×144. The prediction of a sign video was then obtained by averaging multiple clips (ten clips per video unless otherwise specified) of 128×128 generated by center cropping.

C. Ablation Study

1) (2+1)D Versus 3-D: The deep R(2+1)D with R3D was first compared on the CSL-500 dataset to evaluate whether independent spatial and temporal modeling was suitable for SLR. For a fair comparison, pretrained models on Kinetics were used for both R3D and R(2+1)D. As given in Table I, even the shallowest R3D-18 can obtain over 0.88 test accuracy on CSL-500, which demonstrates the spatiotemporal modeling capacity of 3-D convolutions. However, each R3D of different depths in the experiments was inferior to its (2+1)D variant, and R(2+1)D-50 achieved the best performance among basic architectures in the conducted experiments. It was argued that modeling spatial appearance and temporal evolution jointly by 3-D convolutions was not necessary for SLR, and (2+1)D decomposition greatly facilitates the optimization process (factorizing the spatiotemporal modeling process) by decomposing the heavy 3-D convolution filters into 2-D filters and 1-D filters. In addition, R(2+1)D doubles the nonlinearity because of the extra RELU in each (2+1)D residual block, resulting in the improved model capacity without introducing additional parameters.

As the model depth increases, the performances of R3D and R(2+1)D are improved significantly. However, as illustrated in the last block of Table I, the very deep R3D-152 and R(2+1)D-152 declined markedly because of overfitting. It was argued that the number of video samples of CSL-500 was not sufficient for optimizing the two heavy networks.

2) Ablation Study of STCA: The impact of the self-attention layer in CTA was first analyzed by comparing the self-attention with both multiscale and single-scale temporal convolution settings. As given in Table II, the self-attention operation for global temporal modeling performs the best among all settings. It was believed that the global temporal modeling of self-attention was more effective than the local modeling of temporal convolution in SLR, which is more related to long-range motion patterns. Although the absolute improvement of the self-attention method was minimal (+0.21), the increased accuracy compared to the incremental benefits from the conventional multiscale convolution and single-scale convolution was appreciable (i.e., 0.21 versus 0.58). Accordingly, over 36% of the overall improvement of CTA (see Table V) could be beneficial for future investigations of similar problems.

Then, the effects of different positions of the self-attention layer in CTA were compared. As given in the last two rows of Table II, placing the self-attention operation in the middle (M) of the MLP to form a bottleneck structure was more efficient (50 M versus 110 M) with slight performance loss than placing it at the last (L). Therefore, the self-attention operation was placed in the middle of the MLP in CTA for subsequent experiments.

The impact of the reduction ratio r was also analyzed, which was the reduction ratio of CTA. It can allow for the control of the number of parameters of the temporalwise MLP and self-attention operation. Table III reveals that the performance of SCTA does not improve obviously when the network complexity is further increased, which is consistent with [8]. This was likely because of the overfitting of the channel-temporal relation. Therefore, we set r=16 for our subsequent experiments for the tradeoff between speed and accuracy.

The adopted multipath convolution layer of STA was also compared for generating multiscale STA descriptors with the single-scale setting and self-attention. As given in Table IV, the multipath setting surpasses the single path (+0.08). It was argued that multiscale convolution filters enhance the generalization performance by combining spatiotemporal information at different scales. The advantage of the multiscale convolution over the single-scale convolution method was not very obvious, but it is worth noting that the multiscale convolution complements the local temporal component, and the relative improvement is 42% (0.08/0.19) of the overall performance improvement of STA (see Table V). In addition, although self-attention performs well in CTA, its use in STA brings about a sharp increase in complexity (from 50 M to 215 M) and, therefore, results in overfitting (the performance of 96.36 is even lower than the original R(2+1)D). Therefore, self-attention was excluded in STA.

Three settings of the STCA based on R(2+1)D-50 were then compared: STA only, CTA only, and their combination in a different order. In Table V, both the STA and CTA can improve test accuracy on CSL-500. It can also be seen that incorporating both submodules performs better than STA or CTA only. It was argued that this was because the two components were combined to capture different attention descriptors along spatial, temporal, and channel dimensions, especially the complimentary local and global temporal attention operations in CTA and STA. In addition, it can also be seen that the order of these two submodules did not count (such a slight performance difference is negligible).

The effect of the inner residual connection was also explored in the STCA, and Table VI lists the performance improvement. It was argued that this was because it facilitates the optimization process and preserves the original features, which was not useful for the current layer but may be helpful for the next layers.

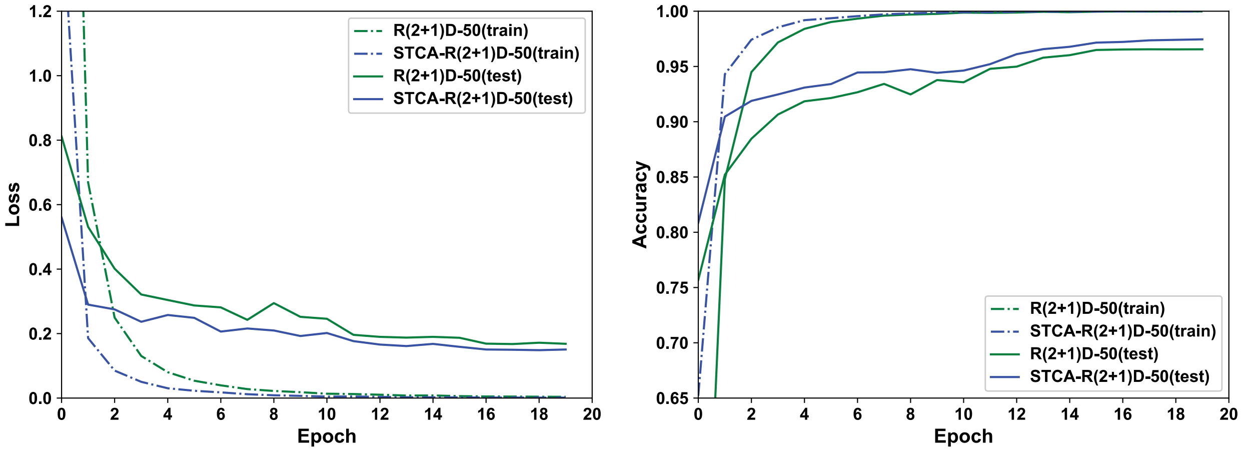

Fig. 4. Training process of R(2+1)D-50 with and without STCA.

Finally, the generalization performance of STCA with different backbones was explored. In Table I, one can also discover the consistent and noticeable performance improvement of all R3D and R(2+1)D networks of different depths with STCA. Among them, R(2+1)D-18 incorporated with STCA can obtain comparable performance to deeper R(2+1)D-34 with negligible additional overheads, and a similar phenomenon also existed in other pairs. As illustrated in Fig. 4, with the whole STCA, both training loss and test loss are reduced faster, and the test loss converges to a lower value when R(2+1)D-50 was embedded with STCA, which demonstrated that the network can be optimized to be faster and better. It was argued that this was because STCA can more capture discriminative spatiotemporal features related to sign language and further improve the representation capacity.

3) Comparison With Other SE-Like Attention Modules: We first compare our STCA with two famous attention modules called SE [8] and CBAM [37], which are first designed for image-based tasks. For a fair comparison, we insert these two lightweight modules into each residual block of R(2+1)D-50 using the same strategy as STCA. As given in Table VII, SE and CBAM can also improve the performance on CSL-500. We argue that the channel attention of both SE and CBAM can make the network concentrate on the discriminative motion patterns, and the spatial attention in CBAM can further make the network focus on the relevant spatial regions. However, their performance hit a bottleneck without exploiting temporal attention reasoning. In contrast, we propose to generate temporal attention by introducing self-attention operations for global temporal attention and temporal convolution layers for local temporal attention in our STCA.

According to the description in [38], we implement W3 by ourselves. To examine the pure effect of attention modules, we exclude the proposed mature feature guided regularization (MFR) in [38]. As given in Table VII, our STCA performs better than W3 (97.45 versus 97.30), with other settings being the same. We argue that the complementary global temporal modeling of self-attention in CTA, multiscale local temporal modeling in STA, and inner residual connections account for the performance improvement. In addition, our STCA is of lighter weight than W3 when combined with R(2+1)D (50 M versus 110 M), which also demonstrates the efficiency of placing the self-attention layer in the middle of the MLP to form a bottleneck structure.

D. Comparison With the State-of-the-Art

1) Comparison With the State-of-the-Art on CSL-500: First, we compare our proposed method with previously reported methods on CSL-500 published in [31]. Hand-crafted features [64], [67] are implemented in [31]. C3D [18] support vector machine (SVM) first splits the video into short-range clips via the sliding window and then forward the feature vectors of these clips into an SVM. Attention-based C3D [31] adopts a spatial attention mask for selected joints and a temporal attention module to model the attention of different clips when combining clip-level predictions into a video-level prediction. The implementation details of the above methods are described in [31].

Other 3-D and (2+1)D CNN-based approaches were also implemented. For the 3-D baselines, we included deformable 3D-ResNeXt-101 [58], which achieved state-of-the-art in the Jester dataset, I3D [19], which inflated the pretrained 2-D convolution filters in the framework of Inception [5], and SlowFast [22]. For the (2+1)D method, we implemented S3D [47], a separate spatial and temporal modeling version of I3D.

As given in Table VIII, STCA-R(2+1)D achieved the best result on CSL-500. STCA-R(2+1)D not only exceeded hand-crafted feature-based methods but also surpassed the above state-of-the-art 3-D and (2+1)D methods. This result demonstrated the effectiveness of R(2+1)D, which decomposed spatial and temporal modeling, and STCA, which made the network extract discriminative spatiotemporal features for SLR.

2) Comparison With the State-of-the-Art on Jester: Although sign language, which involves both manual and nonmanual components, is not the same as hand gestures, we can also use the Jester dataset to evaluate the generalization performance of our proposed method. We compared our proposed STCA-R(2+1)D with the state-of-the-art methods on the Jester dataset.

TABLE VIII
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON CSL-500

Method	Modality	Frames	Params(M)	GFLOPs×Clips	Test Top1	Test Top5
STIP-FV [64]-SVM	RGB	-	-	-	61.8	-
iDT-FV [67]-SVM	RGB+optical flow	-	-	-	68.5	-
C3D [18]-SVM	RGB+depth	-	79	-	74.7	-
Attention-C3D [31]	RGB+depth+skeleton	-	79	-	88.7	-
I3D [19]	RGB	64	13	111×10	96.47	99.4
S3D [47]	RGB	16	8.4	181×10	96.51	99.4
SlowFast [22]	RGB	32	34.8	51×30	96.88	99.5
Deformable-3D-ResNeXt101 [58]	RGB	32	55.3	140×10	97.22	99.5
STCA-R(2+1)D(ours)	**RGB**	16	50.0	75.8×10	**97.45**	99.5

TABLE IX
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON JESTER VALIDATION SET

Method	Modality	Frames	Params(M)	GFLOPs×Clips	Val Top1	Val Top5
Modified C3D [58]	RGB	32	36.2	-	92.20	-
3D-ResNet-101 [58]	RGB	32	81.7	-	94.40	-
Multi-scale TRN [48]	RGB	8	31.8	33×N/A	95.31	-
TSM [49]	RGB	16	24.3	65×N/A	95.30	99.8
Motion Fused Frames [59]	RGB + optical flow	32	-	-	96.33	-
3D-ResNeXt101 [58]	RGB	32	45.8	-	96.40	-
S3D [47]	RGB	32	8.4	-	96.60	-
STM [50]	RGB	16	24	67×30	96.70	99.9
Deformable-3D-ResNeXt-101 [58]	RGB	32	52.2	-	97.10	-
SlowFast [22]	RGB	32	34.8	51×30	96.91	99.9
R(2+1)D [44]	RGB	16	47.2	75.7×10	96.28	99.8
STCA-R(2+1)D(ours)	**RGB**	16	50.0	75.8×10	**97.05**	99.9

TABLE X
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE EGOGESTURE DATASET

Method	Modality	Frames	GFLOPs×Clips	Val Top1	Val Top5	Test Top1	Test Top5
C3D [18]	RGB	16	-	-	-	86.4	-
I3D [19]	RGB	32	153×1	-	-	90.3	-
TSM [49]	RGB	8	33×1	79.7	96.9	80.5	97.8
TSM+W3+**MFR** [38]	RGB	8	33.5×1	93.9	98.7	94.3	99.2
3D-ResNeXt-101 [68]	RGB	32	-	-	-	93.8	-
3D-ResNeXt-101 [68]	depth	32	-	-	-	94.0	-
SlowFast [22]	RGB	32	51×1	92.2	98.6	92.8	99.1
R(2+1)D [44]	RGB	16	76×1	93.2	98.6	93.4	99.1
STCA-R(2+1)D(ours)	**RGB**	16	76×1	**94.0**	98.7	94.3	99.2

In testing, one clip per video is used for fair comparison.

The implementation details were similar to those of CSL-500. As given in Table IX, the modified C3D [18] that removes the last three FC layers achieves 92.2% accuracy. Deeper 3-D or (2+1)D CNNs, including 3D-ResNet101 [21], 3D-ResNeXt101 [21], S3D [47], and SlowFast [22] (SlowFast was implemented by ourselves) based methods, also obtain excellent results without overfitting. TRN [48], TSM [49], and STM [50], which proposed temporal reasoning modules, also achieve impressive performance. Zhang *et al.* [58] proposed utilizing deformable convolution in the framework of 3-D-ResNeXt101 and achieved state-of-the-art, i.e., 97.10% accuracy. Our STCA promotes the R(2+1)D and achieves 97.05% accuracy, which was also a competitive result.

The confusion matrix using STCA-R(2+1)D is shown in Fig. 5, and most of the gestures were recognized correctly. However, the pair "turning hand clockwise" and "turning hand counterclockwise" was much more confused than other pairs because there were many incorrect labels [58] .

3) Comparison With the State-of-the-Art on EgoGesture: We finally compared our proposed STCA-R(2+1)D with the state-of-the-art methods on the EgoGesture dataset to further evaluate the generalization performance of our proposed method. As given in Table X, TSM [49] with W3 and MFR [38] surpasses not only reported methods, including C3D [18], I3D [19], and 3D-ResNeXt [68] but also results implemented by ourselves, including R(2+1)D [44] and SlowFast [22], which demonstrates the effectiveness of spatial, temporal, and channel attention.

In contrast, our STCA further improved the performance of R(2+1)D and achieved comparable results of 94.0% Top1 validation accuracy and 94.3 % Top1 test accuracy. Although

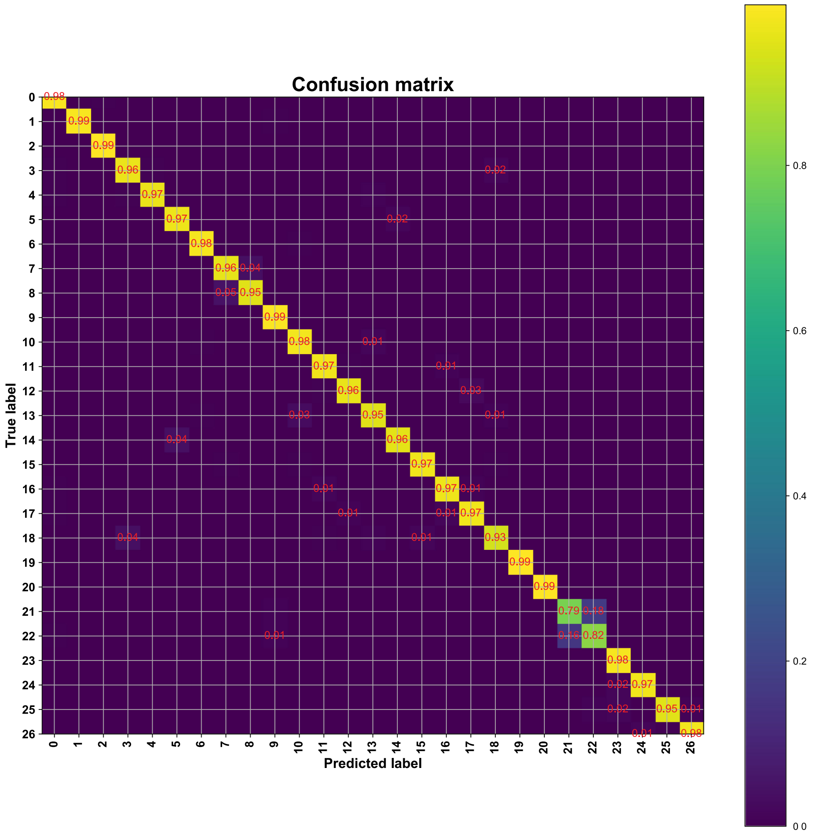

Fig. 5. Confusion matrix on Jester validation set.

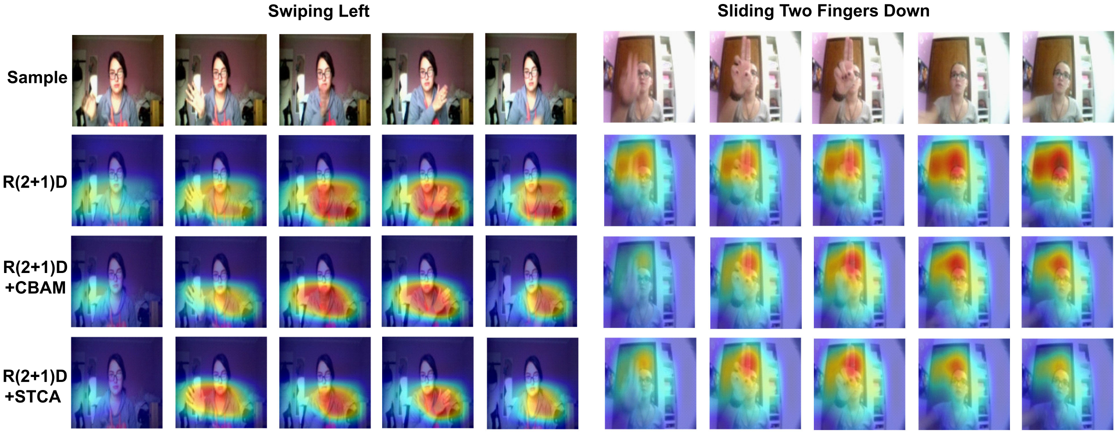

Fig. 6. Visualization of attention map.

our best STCA-R(2+1)D required more computational resources than [38] during the inference process (76 G versus 36 G), our training process was one stage directly without the extra complicated MFR stage in [38].

E. Visualization and Interpretation

We finally visualized our STCA-R(2+1)D using saliency tubes [69] to present whether our network focuses on the main points in frames (spatial) over time (temporal). As illustrated in Fig. 6, although all three R(2+1)Ds can focus on the regions of interest evolving over time, the spatial regions when R(2+1)D is incorporated with CBAM or STCA were not only smaller than the plain R(2+1)D but also more related to hands and arms. We argue that this is because of the spatial attention in both CBAM and STCA to alleviate the interference from the redundant spatial information and further make the useful regions attended.

We also observed that our STCA tended to attend hands and arms with more temporal dynamics than CBAM. We argue that this is because of both local and global temporal modeling in

STCA. For instance, in the hand gesture "Swiping left," our STCA was not activated in the beginning when the gesture was not happening but was rapidly activated when the gesture was ongoing.

V. Conclusion

In this article, we propose to adopt R(2+1)D with STCA for isolated SLR. We demonstrate that R(2+1)D, which separately models spatial appearance and temporal evolution, greatly exceeds joint spatiotemporal modeling in SLR. Our STCA combines SE-like attention with self-attention and simulates the perception of human vision to concentrate on useful regions and extract the most discriminative motion patterns evolving over time. By inserting STCA into R(2+1)D, we alleviate the interference from the redundant information in sign videos and achieve the state-of-the-art performance on the CSL-500 dataset and competitive performance on the Jester and EgoGesture datasets.

In combination with other methods, such as model compression, the proposed method can be run not only on high-performance computers but also on mobile devices. In the future, we will focus on real-world applications of our STCA on embedded devices with limited computational capacity. In addition, the fusion of multimodality for SLR without sacrificing efficiency will also be involved.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.

[5] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[7] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1492–1500.

[8] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[9] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1725–1732.

[10] L. Pigou, S. Dieleman, P.-J. Kindermans, and B. Schrauwen, "Sign language recognition using convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 572–578.

[11] A. A. Q. Mohammed, J. Lv, and M. Islam, "A deep learning-based end-to-end composite system for hand detection and gesture recognition," *Sensors*, vol. 19, no. 23, 2019, Art. no. 5282.

[12] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2625–2634.

[13] J. Yue-Hei *et al.*, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4694–4702.

[14] R. Rastgoo, K. Kiani, and S. Escalera, "Video-based isolated hand sign language recognition using a deep cascaded model," *Multimedia Tools Appl.*, vol. 79, pp. 22965–22987, 2020.

[15] L. Pigou, A. van Den Oord, S. Dieleman, M. van Herreweghe, and J. Dambre, "Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video," *Int. J. Comput. Vis.*, vol. 126, no. 2/4, pp. 430–439, 2018.

[16] C. Mao, S. Huang, X. Li, and Z. Ye, "Chinese sign language recognition with sequence to sequence learning," in *Proc. CCF Chin. Conf. Comput. Vis.*, 2017, pp. 180–191.

[17] K. Bantupalli and Y. Xie, "American sign language recognition using deep learning and computer vision," in *Proc. IEEE Int. Conf. Big Data*, 2018, pp. 4896–4899.

[18] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4489–4497.

[19] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6299–6308.

[20] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[21] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imagenet," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6546–6555.

[22] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6202–6211.

[23] W. Kay *et al.*, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*.

[24] H. Zhou, W. Zhou, Y. Zhou, and H. Li, "Spatial–temporal multi-cue network for continuous sign language recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 13009–13016.

[25] P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman, "Upper body detection and tracking in extended signing sequences," *Int. J. Comput. Vis.*, vol. 95, no. 2, 2011, Art. no. 180.

[26] Y. Li, X. Wang, W. Liu, and B. Feng, "Pose anchor: A single-stage hand keypoint detection network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 7, pp. 2104–2113, Jul. 2020.

[27] J. Huang, W. Zhou, H. Li, and W. Li, "Sign language recognition using 3D convolutional neural networks," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2015, pp. 1–6.

[28] D. Wu *et al.*, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1583–1597, Aug. 2016.

[29] Y. Li *et al.*, "Large-scale gesture recognition with a fusion of RGB-D data based on saliency theory and C3D model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2956–2964, Oct. 2018.

[30] Y. Li *et al.*, "Large-scale gesture recognition with a fusion of RGB-D data based on optical flow and the C3D model," *Pattern Recognit. Lett.*, vol. 119, pp. 187–194, 2019.

[31] J. Huang, W. Zhou, H. Li, and W. Li, "Attention-based 3D-CNNs for large-vocabulary sign language recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2822–2832, Sep. 2018.

[32] S. Ravi *et al.*, "Multi modal spatio temporal co-trained CNNs with single modal testing on rgb-d based sign language gesture recognition," *J. Comput. Lang.*, vol. 52, pp. 88–102, 2019.

[33] P. Kumar, H. Gauba, P. P. Roy, and D. P. Dogra, "A multimodal framework for sensor based sign language recognition," *Neurocomputing*, vol. 259, pp. 21–38, 2017.

[34] Q. Xiao, M. Qin, P. Guo, and Y. Zhao, "Multimodal fusion based on LSTM and a couple conditional hidden Markov model for Chinese sign language recognition," *IEEE Access*, vol. 7, pp. 112258–112268, 2019.

[35] C. Chansri and J. Srinonchat, "Hand gesture recognition for thai sign language in complex background using fusion of depth and color video," *Procedia Comput. Sci.*, vol. 86, pp. 257–260, 2016.

[36] J. J. Bird, A. Ekárt, and D. R. Faria, "British sign language recognition via late fusion of computer vision and leap motion with transfer learning to American sign language," *Sensors*, vol. 20, no. 18, 2020, Art. no. 5151.

[37] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[38] J.-M. Perez-Rua, B. Martinez, X. Zhu, A. Toisoul, V. Escorcia, and T. Xiang, "Knowing what, where and when to look: Efficient video action modeling with attention," 2020, *arXiv:2004.01278*.

[39] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[40] M. De Coster, M. van Herreweghe, and J. Dambre, "Sign language recognition with transformer networks," in *Proc. 12th Int. Conf. Lang. Resour. Eval.*, 2020, pp. 6018–6024.

[41] F. B. Slimane and M. Bouguessa, "Context matters: Self-attention for sign language recognition," in *Proc. Int. Conf. Pattern Recognit.*, 2021, pp. 7884–7891.

[42] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10023–10033.

[43] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Multi-channel transformers for multi-articulatory sign language translation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 301–319.

[44] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6450–6459.

[45] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[46] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 20–36.

[47] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 305–321.

[48] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 803–818.

[49] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 7083–7093.

[50] B. Jiang, M. Wang, W. Gan, W. Wu, and J. Yan, "STM: Spatiotemporal and motion encoding for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 2000–2009.

[51] L. Quesada, G. López, and L. Guerrero, "Automatic recognition of the American sign language fingerspelling alphabet to assist people living with speech or hearing impairments," *J. Ambient Intell. Humanized Comput.*, vol. 8, no. 4, pp. 625–635, 2017.

[52] K. Grobel and M. Assan, "Isolated sign language recognition using hidden Markov models," in *Proc. IEEE Int. Conf. Syst., Man, Cybern., Comput. Cybern. Simul.*, 1997, pp. 162–167.

[53] L.-C. Wang, R. Wang, D.-H. Kong, and B.-C. Yin, "Similarity assessment model for Chinese sign language videos," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 751–761, Apr. 2014.

[54] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3D convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 1–7.

[55] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4207–4215.

[56] L. Pigou, M. van Herreweghe, and J. Dambre, "Gesture and sign language recognition with temporal residual networks," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 3086–3093.

[57] K. M. Lim, A. W. C. Tan, C. P. Lee, and S. C. Tan, "Isolated sign language recognition using convolutional neural network hand modelling and hand energy image," *Multimedia Tools Appl.*, vol. 78, no. 14, pp. 19917–19944, 2019.

[58] Y. Zhang, L. Shi, Y. Wu, K. Cheng, J. Cheng, and H. Lu, "Gesture recognition based on deep deformable 3D convolutional neural networks," *Pattern Recognit.*, vol. 107, 2020, Art. no. 107416.

[59] O. Kopuklu, N. Kose, and G. Rigoll, "Motion fused frames: Data level fusion strategy for hand gesture recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 2103–2111.

[60] R.-H. Liang and M. Ouhyoung, "A real-time continuous gesture recognition system for sign language," in *Proc. IEEE Int. Conf. Automat. Face Gesture Recognit.*, 1998, pp. 558–567.

[61] G. Fang, W. Gao, and D. Zhao, "Large-vocabulary continuous sign language recognition based on transition-movement models," *IEEE Trans. Syst., Man, Cybern. A, Syst. Human*, vol. 37, no. 1, pp. 1–9, Jan. 2007.

[62] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation," in *Proc. 32nd AAAI Conf. Artif. Intell., 30th Innov. Appl. Artif. Intell. Conf., 8th AAAI Symp. Educ. Adv. Artif. Intell.*, 2018, pp. 2257–2264.

[63] R. Cui, H. Liu, and C. Zhang, "A deep neural framework for continuous sign language recognition by iterative training," *IEEE Trans. Multimedia*, vol. 21, no. 7, pp. 1880–1891, Jul. 2019.

[64] I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.*, vol. 64, no. 2/3, pp. 107–123, 2005.

[65] J. Materzynska, G. Berger, I. Bax, and R. Memisevic, "The jester dataset: A large-scale video dataset of human gestures," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2019, pp. 2874–2882.

[66] Y. Zhang, C. Cao, J. Cheng, and H. Lu, "Egogesture: A new dataset and benchmark for egocentric hand gesture recognition," *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1038–1050, May 2018.

[67] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 3169–3176.

[68] O. Köpüklü, A. Gunduz, N. Kose, and G. Rigoll, "Real-time hand gesture detection and classification using convolutional neural networks," in *Proc. 14th IEEE Int. Conf. Automat. Face Gesture Recognit.*, 2019, pp. 1–8.

[69] A. Stergiou, G. Kapidis, G. Kalliatakis, C. Chrysoulas, R. Veltkamp, and R. Poppe, "Saliency tubes: Visual explanations for spatio-temporal convolutions," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 1830–1834.

Xiangzu Han was born in 1996. He is currently working toward the Graduate degree with the School of Control Science and Engineering, Shandong University, Jinan, China.

His research interests include deep learning, video understanding, and sign language recognition.

Fei Lu received the M.S. degree in system engineering from Northeastern University, Shenyang, China, in 1997, and the Ph.D. degree in control theory from Shandong University, Jinan, China, in 2007.

She is a Professor with the School of Control Science and Engineering, Shandong University, Jinan, China. Her research interests include service robots, intelligent space, robot cognition, deep learning, and computer vision.

Jianqin Yin (Member, IEEE) was born in 1978. She received the Ph.D. degree in control science and control engineering from Shandong University, Jinan, China, in 2013.

She is currently a Professor with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China. Her research interests include service robots, pattern recognition, machine learning, and image processing.

Guohui Tian (Member, IEEE) was born in 1969. He received the Ph.D. degree in automatic control theory and application from the School of Automation, Northeastern University, Shenyang, China, in 1997.

He was a Lecturer from 1997 to 1998 and an Associate Professor from 1998 to 2002 with Shandong University, Jinan, China, and also from 1999 to 2001, a Postdoctoral Researcher with the School of Mechanical Engineering. From 2003 to 2005, he was a Visiting Professor with the Graduate School of Engineering, Tokyo University, Tokyo, Japan. He is currently a Professor with the School of Control Science and Engineering, Shandong University, Jinan, China. His current research interests include service robots, intelligent space, cloud robotics, and brain-inspired intelligent robotics.

Jun Liu (Member, IEEE) received the Ph.D. degree in mechanical engineering from the University of Toronto, Toronto, ON, Canada, in 2016.

From 2017 to 2019, he was a Postdoctoral Fellow with Weill Cornell Medical College, Cornell University. He is currently an Assistant Professor with the Department of Mechanical Engineering, City University of Hong Kong, Hong Kong. His research interests include micronanorobotics, medical robotics, and medical image analysis.

Dr. Liu was a recipient of multiple awards, including the Best Student Paper Award and the Best Medical Robotics Paper Finalist Award from the IEEE International Conference on Robotics and Automation in 2014 and the IEEE Transactions on Automation Science and Engineering Best New Application Paper Award in 2018 in the field of robotics and automation.