

Received 16 December 2022, accepted 25 December 2022, date of publication 5 January 2023, date of current version 16 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3234743

RESEARCH ARTICLE

STFE-Net: A Spatial-Temporal Feature Extraction Network for Continuous Sign Language Translation

JIWEI HU¹, YUNFEI LIU¹, KIN-MAN LAM², AND PING LOU¹

¹School of Information Engineering, Wuhan University of Technology, Wuhan, Hubei 430070, China

²Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong, China

Corresponding author: Ping Lou (pinglou@whut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 52075404, and in part by the Application Foundation Frontier Special Project of Wuhan Science and Technology Bureau under Grant 2020010601012176.

ABSTRACT The main challenge of continuous sign language translation (CSLT) lies in the extraction of both discriminative spatial features and temporal features. In this paper, a spatial-temporal feature extraction network (STFE-Net) is proposed for CSLT, which optimally fuses spatial and temporal features, extracted by the spatial feature extraction network (SFE-Net) and the temporal feature extraction network (TFE-Net), respectively. SFE-Net performs pose estimation for the presenters in sign-language videos. Based on COCO-WholeBody, 133 key points are abbreviated to 53 key points, according to the characteristics of the sign language. High-resolution pose estimation is performed on the hands, along with the whole-body pose estimation, to obtain finer-grained hand features. The spatial features extracted by SFE-Net and the sign language words are then fed to TFE-Net, which is based on Transformer with relative position encoding. In this paper, a dataset for Chinese continuous sign language was created and used for evaluation. STFE-Net achieves Bilingual Evaluation Understudy (BLEU-1, BLEU-2, BLEU-3, BLEU-4) scores of 77.59, 75.62, 74.25, 72.14, respectively. Furthermore, our proposed STFE-Net was also evaluated on two public datasets, RWTH-Phoenix-Weather 2014T and CLS. The BLEU-1, BLEU-2, BLEU-3 and BLEU-4 scores achieved by our method on the former dataset are 48.22, 33.59, 26.41 and 22.45, respectively, and the corresponding scores are 61.54, 58.76, 57.93 and 57.52, respectively, on the latter dataset. Experiment results show that our model can achieve promising performance. If any reader needs the code or dataset, please email lunfee@whut.edu.cn.


INDEX TERMS Continuous sign language translation, pose estimation, transformer, relative position encoding.

I. INTRODUCTION

As the main communication channel between the people having total or partial hearing loss and hearing people, sign languages have played a very important role in daily life. The rapid development in the fields of computer vision and deep learning has opened up new opportunities for sign language recognition. Sign language uses different parts of the body, such as fingers, arms, hand movement trajectories, head

and facial expressions, to convey information [1]. In sign languages, each gesture has a specific meaning, and strong contextual information and grammatical rules are factors that should be considered in continuous sign language translation (CSLT).

In recent years, research on video-based sign language recognition, translation, and generation has received increasing and wide attentions. Sign language recognition (SLR) refers to the use of algorithms and techniques to recognize the resulting gesture sequences and elaborate their meaning in the form of text or speech [2]. SLR is a typical

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval .

interdisciplinary problem involving several fields, such as computer vision, natural language processing, image recognition, human-computer interaction, and pattern recognition [3]. Challenges and difficulties in sign language recognition still exist because of the large set of vocabularies, rich and diverse expressions, and complex semantic-grammar structures.

With the availability of large-scale CSLT datasets, research on sign language translation (SLT) in an end-to-end manner has been emerging. Forster et al. [4] constructed the RWTH-PHOENIX-Weather German sign language dataset. Then, Camgoz et al. [5] expanded the dataset to form the RWTH PHOENIX-Weather 2014T dataset, and introduced the first end-to-end CSLT model. The model uses convolutional neural networks (CNNs) to extract the spatial features of sign language actions. An attention-based architecture was also employed to learn the mapping relationship between the sign language from a video and the reference translated text. Huang et al. [6] constructed a large-scale Chinese continuous sign language dataset CSL, and proposed the Hierarchical Attention Network with Latent Space (LS-HAN).

CSLT is a generalized sequence-to-sequence problem, and one of the difficulties is the recognition of visual information from videos. CSLT not only considers the information from the current image frame, but also relates it to the complex dynamic relationship between the consecutive frames. The information of the current frame is represented as spatial features, and the key points of the person concerned can be extracted by pose estimation methods. For CSLT, the key points on the body, face and hands are required to be located. A recent method for pose estimation is OpenPose [7], which combines multiple deep neural networks (DNNs) on different datasets, one DNN for body pose estimation on COCO [8], one DNN for locating facial key points trained with multiple datasets (i.e., FRGC [9] and i-bug [10]), and one DNN for locating hand's key points based on Panoptic dataset [11]. The COCO-WholeBody dataset [12] has 133 key points for the human body, but some of the key points are redundant or unnecessary for CSLT.

With the emergence of Transformer [13], the performance for CSLT has been further improved. Camgoz et al. [14] proposed an advanced architecture based on Transformer, namely Sign Language Transformer (SLT). SLT was trained using the Connectionist Temporal Classification (CTC) loss function and cross-entropy loss function. In a recent study, Camgoz et al. [10] used additional modal and cross-modal attention to synchronize the flow of different information. Kim et al. [15] proposed a key point normalization method and built a Korean sign language translation framework based on Transformer. Yin et al. [11] proposed an STMC-Transformer model, where the SMC module decomposes the input video into spatial features of multiple visual cues (face, hands, full frame, and pose), and the TMC module computes temporal correlations for different time steps.

In this paper, because of the scarcity of Chinese continuous sign language datasets, we construct a Chinese continuous

sign language teaching dataset for real translation scenarios. In addition, the selection of the sign-language key points, pose estimation method, and the Transformer network will be studied in depth for CSLT. A spatial-temporal feature extraction network (STFE-Net) is proposed, which combines the spatial features and temporal features for the CSLT task.

The main contributions of this work are summarized as follows:

1. In this work, a Chinese continuous sign language dataset was constructed for real translation scenarios. The dataset provides useful data to support the study of Chinese continuous sign language translation.
2. Based on the fused spatial-temporal information, an end-to-end Chinese continuous sign language translation network (STFE-Net) is proposed, which is composed of the spatial feature extraction network (SFE-Net) and the temporal feature extraction network (TFE-Net).
3. For SFE-Net, 53 key points related to a sign language are selected from the 133 key points in the COCO-WholeBody dataset. The selected key points can result in achieving better pose estimation performance than using all the 133 key points. In addition, high-resolution pose estimation is performed on the hands so as to obtain fine-grained sign language information.
4. For TFE-Net, Transformer is used to implement temporal feature extraction, in which relative position encoding and position-aware self-attention optimization are adopted.
5. Combining SFE-Net and TFE-Net realizes an end-to-end network, i.e., STFE-Net, for CSLT, which achieves excellent performance on our created dataset and multiple public datasets. Moreover, STFE-Net outperforms many state-of-art methods.

II. RELATED WORKS

This section will focus on the work related to the various techniques used in SFET-Net, i.e., the whole-body pose estimation method and Transformer network.

A. POSE ESTIMATION

The COCO-WholeBody dataset contains 133 key points for whole-body pose estimation. However, for sign language recognition, 133 key points are more than necessary and many of them are redundant. Selecting appropriate key points for sign language recognition can lead to better performance.

OpenPose [16], [17], [18], Single-Network (SN) [19], HPRNet [20], and HRNet [21] are the current state-of-the-art methods for pose estimation. OpenPose is a two-step pose estimation network, which requires separate training for hands and body of different scales. SN, HPRNet and HRNet are one-step pose estimation networks. HPRNet is pose estimation. HRNet is a high-resolution network, which iteratively exchanges information between parallel multi-resolution sub-networks to perform multi-scale repetitive fusion. In this paper, we employ HRNet to construct a parallel structure for performing global body pose estimation and fine-grained hand pose estimation.

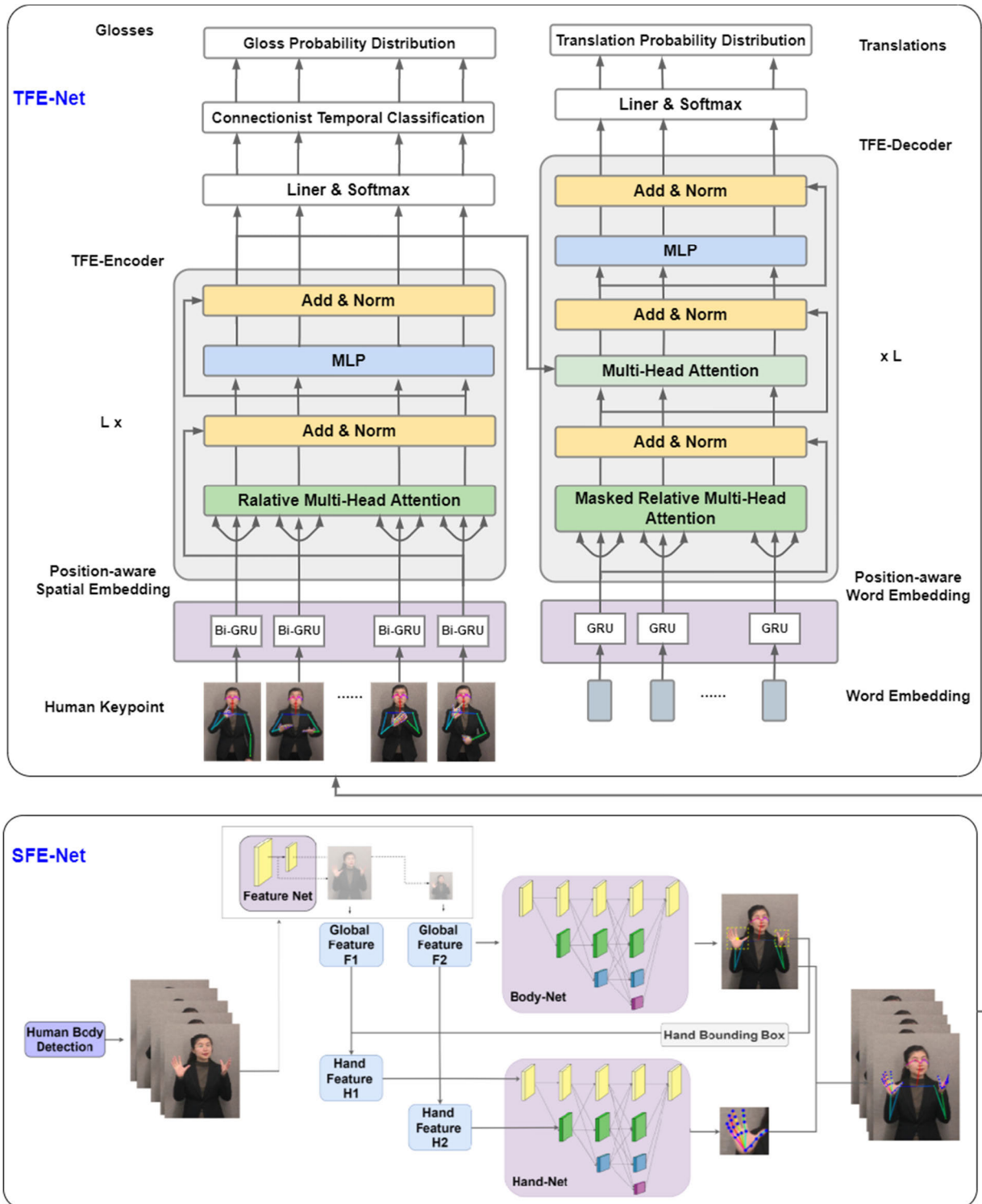


FIGURE 1. The entire structure of STFE-Net. STFE-Net contains SFE-Net and TFE-Net. SFE-Net is used to extract key points in video. First the input image is processed by Feature-Net to extract Global Features (F1 and F2), using the features extracted from Feature-Net, Body-Net predicts body key points and hand bounding boxes at the same time. Using hand bounding boxes predicted, then Hand Features (H1 and H2) are applied to Hand-Net to predict the heatmaps of hand key points. Next, TFE-Net processes the key points obtained by SFE-Net. Transformer is used to extract the sign-language sequence features. To extract the temporal features, the embeddings of the spatial features and the sign-language words are first computed, and then input to the encoders and decoders of the Transformer, respectively. In the feature embedding, GRUs are used to achieve position awareness of the sequences. The Position-aware Spatial Embedding uses Bi-GRU, while the Position-aware Word Embedding uses unidirectional GRU to mask future inputs during prediction.

For hand detection in sign language recognition, we use CornerNet [22], which treats the target to be detected as an envelope composed of two corner points. Previous target detection models, such as Faster R-CNN [23], use RoI Pooling for coordinate mapping and scale transformation. RoI Align, proposed in Mask R-CNN [24], uses bilinear interpolation to avoid generating bias in coordinate mapping and pooling. Thus, in our model, RoI Align is adopted for hand-box detection.

B. TRANSFORMER WITH RELATIVE POSITION ENCODING

Transformer is an encoder proposed by Vaswani et al. [13], which utilizes self-attention mechanism with an encoding layer consisting of two sub-layers, i.e., the self-attention layer and the fully connected layer. In contrast to RNN structure-based encoders, Transformer is not structured to directly obtain absolute or relative position information [13], [25]. Therefore, in Transformer, in addition to the input embedding, a position vector based on sinusoidal functions of different frequencies is produced to embed the position information. Then, the input embedding and position embedding are added to form the input to the self-attention layer. To further improve the performance for long sequences, Shaw et al. [26] proposed the relative position embedding. The relative position embedding takes into account the positional relationships of the words in the sequence and allows for better modelling of the semantic information contained in the sequence.

1) SELF-ATTENTION LAYER

The encoder in Transformer mainly relies on the Multi-Head Self-Attention mechanism. For each self-attention head, denote the input and output sequences $H = \{h_1, \dots, h_n\}$, $h_i \in R^{d_H}$ and $Z = \{z_1, \dots, z_n\}$, $z_i \in R^{d_Z}$, respectively, where n is the length of the sequences, d_H is the dimension of the input sequence, and d_Z is the dimension of the output sequence. The computations in the self-attention layer are described in the following.

First, each input h_i is mapped to three different spaces to obtain query $q_i \in R^{d_Z}$, key $k_i \in R^{d_Z}$ and value $v_i \in R^{d_Z}$, i.e.,

$$\begin{cases} q_i = h_i W^Q \\ k_i = h_i W^K \\ v_i = h_i W^V \end{cases} \quad (1)$$

where $W^Q, W^K, W^V \in R^{d_H \times d_Z}$ are trainable parameters. Then, for inputs h_j and h_l , the attention score A_{ij} between them is computed as follows:

$$A_{ij} = q_j k_i^T \quad (2)$$

After that, A_{ij} is scaled and normalized to a_{jl} , as follows:

$$a_{jl} = \frac{\exp(A_{jl}/\sqrt{d_Z})}{\sum \exp(A_{jm}/\sqrt{d_Z})} \quad (3)$$

Finally, the output z_j is obtained by weighted summation as follows:

$$a_{jl} = \sum_{l=1}^n a_{jl} v_j \quad (4)$$

where $j, l = [1, n]$ are the positions of the vectors in the input sequence.

2) RELATIVE POSITION ENCODING

Absolute positional encoding does not perceive the sequential information of the sequences. To better learn the position letters of the sign language sequences and reference translation word sequences, relative position encoding is adopted in our model. The relative position information is incorporated in the Multi-Head self-Attention layer of the encoder and decoder. In this work, bidirectional GRU is used for relative position encoding in the encoder, while GRU is used for relative position encoding in the decoder.

In [27], recurrent neural networks (RNNs) [28] were used for relative position encoding. Instead of adding position information, RNNs can output feature embeddings with position information. In [29], the sinusoidal position embeddings are replaced by learned two-dimensional convolutional layers. The latest Transformer variants, such as Reformers [30] and Transformer-XL [31], also employ the relative localization schemes.

III. THE PROPOSED MODEL

In this paper, a spatial-temporal feature extraction network (STFE-Net) is proposed for CSLT, which is implemented and trained for Continuous Sign Language Recognition (CSLR) in an end-to-end manner. The detailed architecture of our proposed model is shown in Figure 1. This deep model is mainly composed of two modules, namely the spatial feature extraction network (SFE-Net) and the temporal feature extraction network (TFE-Net). The model first learns the spatial features of 53 key points, selected for sign language recognition, from sign language videos. A global pose-estimation network is used to generate the spatial features. In order to obtain fine-grained hand pose information, high-resolution pose estimation is performed on the hands. After that, temporal features are extracted from the consecutive frames. TFE-Net is based on Transformer, which includes an encoding module and a decoding module. Firstly, spatial features and the corresponding words are converted into embeddings. The feature embeddings are then combined with relative position embedding based on GRU. The encoded feature and word embeddings are then input to the encoders and decoders of Transformer, respectively, for learning. The details of our method will be described in the following.

A. SPATIAL FEATURE EXTRACTION NETWORK

1) KEY POINTS LOCALIZATION METHOD

COCO-WholeBody [12] is the first dataset for whole-body pose estimation. Each whole body is represented by four

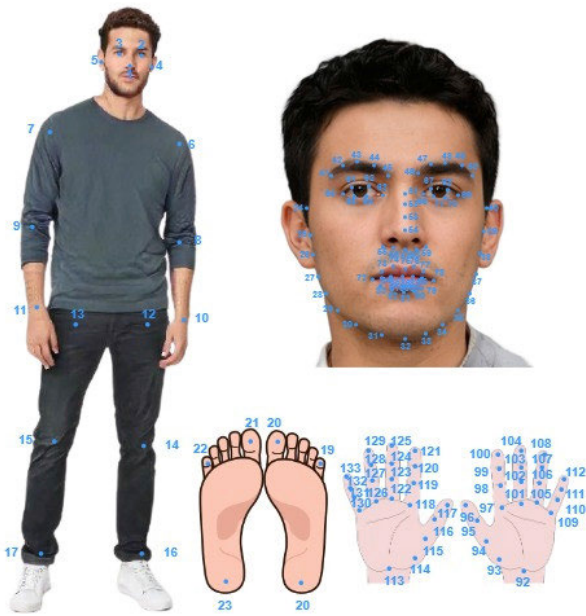


FIGURE 2. The 133 key points locations in COCO-WholeBody.

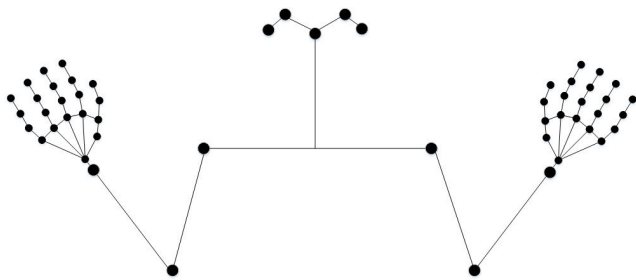


FIGURE 3. The 53 key points selected in this work.

bounding boxes: the person box, face box, left-hand box, and right-hand box. In addition, 133 key points are located over the whole body: 17 for the body, 6 for the two legs, 68 for the face, and 42 for the two arms. The locations of these key points are illustrated in Figure 2. However, a large number of key points results in a lot of redundancy for sign language recognition. Consequently, using all the key points will reduce the learning efficiency of the model. Therefore, in this paper, only 53 of the key points are selected, according to the amount of motions in a sign language, to simplify feature learning and extraction.

The key points are selected based on the following method. All the body movements, expressed in the sign language, are in the upper body, so the key points 12 to 23 are ignored. On the other hand, the face movements in the sign language do not vary significantly, so the key points 24 to 91 on the face are also ignored. Finally, only 53 key points (11 key points for the upper body and 21 key points for the left and right hands) are selected. These selected key points are connected to form a human skeleton, as illustrated in Figure 3.

2) STRUCTURE FOR GLOBAL POSE ESTIMATION

Sign language movements are characterized by the interaction of the hands and the body to collaboratively express semantic information. However, the body and hand regions have significant scale differences. Considering the human body hierarchy, this work performs feature extraction for the global body, while performing high-resolution feature extraction for the hands. As shown in Figure 1 (SFE-Net), this structure, which contains three parts: Feature-Net, Body-Net and Hands-Net, can efficiently acquire fine-grained hand pose information. Furthermore, it enables multi-scale sign language action feature extraction by fusing features for the body and hand postures.

a: FEATURE-NET

This network includes two convolutional layers. Each convolutional layer reduces the input to half of its resolution, and the corresponding features are denoted as F1 and F2, which are used as low-level features for Body-Net and Hand-Net.

b: BODY-NET

Inspired by CornerNet [22], Body-Net predicts both the body box and the hand box, based on the shared feature maps F1 and F2. Using the shared maps F1 and F2, fine-grained hand and facial features will be extracted. Then BottleNeck and BasicBlock, proposed in ResNet [32], were utilized. Thus, shallow features can be preserved in the forward path. At the end of the network, the highest resolution feature maps obtained are used for representation prediction. HRNet-w32 [21] was chosen for the body-pose estimation network in this paper. Table 2-1 lists the parameters of the HRNet-w32 network. Using Body-Net, 21 global key-point (11 for body, 5 for left hand, and 5 for right hand) features can be obtained.

c: HAND-NET

By using the hand boxes predicted by Body-Net, the features corresponding to the hand regions in F1 and F2 are cropped. To obtain finer hand feature, HRNetV2p-w18 [21] was used. Compared to HRNet-w32 used in Body-Net, the number of channels corresponding to its highest resolution feature map is halved. This method is based on RoI Align in Faster R-CNN [23]. Hand-Net performs high-resolution pose estimation for the hands, which facilitates learning fine-grained sign-language information.

B. TEMPORAL FEATURE EXTRATION NETWORK

At this point, the spatial features have been extracted from the current frame. To perform continuous sign-language translation, it is necessary to handle the temporal relationships between the spatial features from consecutive frames. In TFE-Net, Transformer is used to extract the sign-language sequence features. To extract the temporal features, the embeddings of the spatial features and the sign-language words are first computed, and then input to the encoders

and decoders of the Transformer, respectively. GRU, which is a variant of RNN, is used to perform relative position encoding for the embeddings. Figure 1 shows the structure of the entire Transformer network, i.e., TFE-Net, which is described below in detail.

1) SPATIAL FEATURE AND WORD EMBEDDINGS

Transformer networks are based on self-attention mechanisms and lack sequential or positional information about sequences. Similar translation results can be obtained, even if the order of the utterances is disrupted, and such translations may be ambiguous in conveying the true message of the semantics. Therefore, sequential Position Embedding (PE) is introduced to the input embeddings. Both spatial features and word vectors need to be fused with Position Embedding, which is generated by using sine and cosine functions of different frequencies in the original Transformer. The input embeddings are added with the corresponding position embedding, so the dimension of the position vector must be the same as that of the spatial feature/word embedding. The position embedding can be computed as follows:

$$PE(t, 2i) = \sin(t/10000^{2i/d}) \quad (5)$$

$$PE(t, 2i + 1) = \cos(t/10000^{2i/d}) \quad (6)$$

where t denotes the temporal order of the absolute position of the spatial features/words in a sequence, d denotes the dimension of the corresponding embedding, and i refers to the position in an embedding. The position embeddings at even and odd positions, i.e., $2i$ and $2i + 1$, are calculated using (5) and (6), respectively.

Bi-GRU is used in the relative position encoder to generate position-aware spatial embeddings, while GRU is used to generate position-aware word embeddings in the decoder. As the decoder stack needs to mask future inputs during predicting, so GRU, instead Bi-GRU, is used for decoding. With the spatial feature embedding, Bi-GRU is used to calculate the state h_i of the current layer, as follows:

$$h_i = GRU(s_k, h_{i-1}) \quad (7)$$

$$s'_k = h_i \quad (8)$$

where s_k and s'_k are the spatial embedding and position-aware spatial embedding, respectively, and h_{i-1} is the state of the hidden layer before the current layer.

Similarly, the position-aware word embedding is calculated as follows:

$$h_i = GRU(w_k, h_{i-1}) \quad (9)$$

$$w'_k = h_i \quad (10)$$

The position-aware spatial embeddings, s'_k , are input to the decoder stack, while the position-aware word embeddings, w'_k , are input to the decoder stack of Transformer.

2) ENCODER STACK

Transformer, with relative position encoding, can optimize the self-attention layer in Multi-Head Attention. The input

embedding is treated as a directed fully connected graph. Denote the edges between input elements x_i and x_j as r_{ij}^V and r_{ij}^K , which represent the relative position information between two vectors. Here K , V are the query, key and value matrices generated from the input set of vectors. Then, by substituting x_i and x_j into the formulation for Multi-Head Attention, we have:

$$z_i = \sum_{j=1}^n \alpha_{ij}(x_j W^V + r_{ij}^V) \quad (11)$$

where α_{ij} is calculated as follows:

$$\alpha_{ij} = \frac{\exp e_{ij}}{\sum_{t=1}^n \exp e_{it}} \quad (12)$$

and e_{ij} is calculated as follows:

$$e_{ij} = \frac{(x_i W^Q)(x_j W^K + r_{ij}^K)^T}{\sqrt{d_z}} \quad (13)$$

From the above equations, it can be seen that, for Multi-Headed Attention, the values of r_{ij}^V and r_{ij}^K are shared among the multiple attention heads. Furthermore, there is no linear transformation, and the position relationship is not decimated. The distance $|j - i|$ between the input x_i and x_j is restricted to be within a fixed range, i.e., $|j - i| < k$. If the distance between x_i and x_j is greater than this range k , the distance is truncated to k .

To train the encoders, weakly supervised learning is adopted with the Connectionist Temporal Classification (CTC) loss function [33]. The sequence of spatial features extracted from T consecutive frames is denoted as $S = \{s_1, s_2, \dots, s_T\}$. The annotated sequence of N sign-language isolated words is denoted as $G = \{g_1, g_2, \dots, g_N\}$. Then, the conditional probability $p(G/S)$ is modeled as follows:

$$p(G/S) = \sum_{\delta \in \beta} p(\delta/S) \quad (14)$$

where δ represents the path and β represents the set of all allowed paths that matches G .

The CSLR loss is calculated as follows:

$$L_{CSLR} = 1 - p(G/S) \quad (15)$$

3) DECODER STACK

Unlike the Multi Head Attention in encoder, the decoding module uses a sequence mask, which is designed to mask the future inputs during prediction. In addition, different from the Multi-Head Attention layer in the encoding module, the decoding module incorporates attention rather than self-attention.

Denote the translated word of the m^{th} decoding step as t_m , and the translation representation learned by the decoding module at the $m-1^{\text{st}}$ step as r_{m-1} . The conditional probability

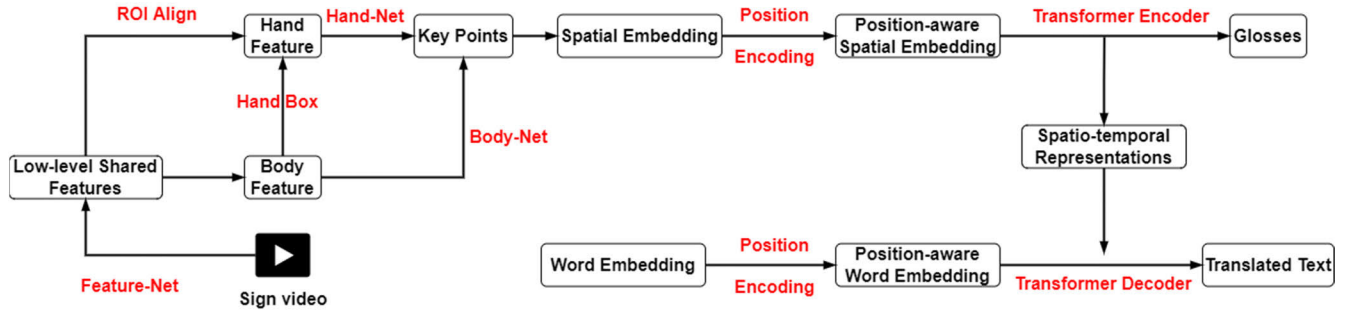


FIGURE 4. Workflow of STFE-Net. Feature-Net: extracts Low-level Shared Features from Sign video. ROI Align: crops corresponding Hand Features and up-scales them to a higher resolution. Hand Box: represents a box with four corner points and a center point for locating the hands. Body-Net: extracts body key points and Hand Box. Hand-Net: extracts hand key points. Position Encoding: encodes the extracted key points/words into Position-aware Spatial Embedding. Transformer Encoder/Decoder: outputs Glosses in the intermediate stage and finally the Translated Text.

distribution $p(T/S)$ of the sequence $T = \{t_1, t_2, \dots, t_M\}$ after consecutive sign language translation is:

$$p(T/S) = \prod_{m=1}^M p(t_m/r_{m-1}) \quad (16)$$

The CSLT loss can be calculated as follows:

$$L_{CSLT} = 1 - \prod_{m=1}^M \sum_{d=1}^D p(\tilde{r}_m^d) p(t_m^d/r_{m-1}) \quad (17)$$

The CSLR loss and CSLT loss are combined to train the encoder and decoder jointly as follows.

$$L_{total} = w_1 L_{CSLR} + w_2 L_{CSLT} \quad (18)$$

where w_1 and w_2 are the weights assigned for the CSLR loss and CSLT loss, respectively.

C. SPATIAL-TEMPORAL FEATURE EXTRACTION NETWORK

TFE-Net receives spatial features from SFT-Net to form STFE-Net. STFE-Net incorporates spatial and temporal features, to achieve continuous sign language translation in an end-to-end manner. The entire workflow is shown in Figure 4.

IV. EXPERIMENT

A. OUR DATASET

In this paper, we construct a RGB camera-based Chinese continuous sign language dataset. It provides data support for the study of Chinese continuous sign language translation in practical scenarios. The dataset contains 60 utterances categorized by scene units and 70 sign language presenters. Each utterance from each sign language demonstrator was recorded 5 times. In total, 21,000 sign language videos were recorded. The number of Chinese characters for each sign language utterance in each video ranges from 2 to 13. Each video is labeled with a corresponding translated text. The Chinese continuous sign language teaching dataset was annotated by 15 professional sign language interpreters and 55 school students having total or partial hearing loss. The recording process ensures that the sign language movements were carried out in strict accordance with the standards of

the "List of Words Commonly Used in National Sign Language". The recordings were made using a monocular camera with a frame rate of 30 frames per second, at a resolution of 1280×720 . The total size of the video data is 144G. These annotated reference sign language utterances were separated by spaces to obtain words with Chinese semantics. Then, manual checking and adjustment were performed. Finally, the word database was constructed based on the frequency of occurrence of words. A total of 171 meaningful words were obtained. These words are used for model learning. An example of the key actions and annotations of the recorded sign language utterance is illustrated in Figure 5.

Our dataset was divided into training set, validation set, and test set with the number of presenters in the ratio of 5:1:1. This means that 15,000 videos from 50 sign language presenters are used for training, 3,000 videos from 10 sign language presenters are used for validation, and 3,000 videos from the remaining 10 sign language presenters are used for testing.

B. PUBLIC DATASETS

1) DATASETS FOR SFE-NET

The COCO-WholeBody dataset is the first large-scale benchmark dataset for whole-body pose estimation. In this paper, the 133 key points in the COCO-WholeBody dataset are reduced to 53 key points. Experiments are conducted on this dataset to evaluate the performance of our deep models with different numbers of key points. In addition, other whole-body pose estimation methods are also compared on this dataset.

The CSL-500 dataset [34] consists of 50 sign language presenters, 25,000 labeled video samples, and 500 sign language vocabularies. Each vocabulary contains 50 corresponding sign language videos, depth videos, and 21 skeleton key-points coordinate sequences. The dataset was divided according to the sign language demonstrators, with 36 presenters in the training set and 14 presenters in the test set. Only the RGB video data is used. The depth information and key point annotation are ignored, replaced by the 53 key points as in COCO-WholeBody.

Key actions and annotations of sign language sentences

Daily expression: What's your name?



Translation: What is your name?

Family expression: My grandpa is 80 this year.



Translation: My grandpa is 80 years old this year

Campus expression: I can't do this problem, can you?



Translation: I can't solve this problem, can you?

FIGURE 5. Examples of the key actions and annotations of the recorded sign language utterance.

2) DATASETS FOR TFE-NET

RWTH-PHOENIX-Weather 2014T [5] was used for training, validation and testing according to the official division of 7096, 519 and 642 video samples, respectively.

3) DATASETS FOR STFE-NET

To validate our proposed model, the continuous sign language dataset RWTH-Phoenix-Weather 2014T [5] and the Chinese continuous sign language dataset CSL [6] were used. The CSL dataset was divided into training set, validation set, and test set according to the number of sign language presenters in the ratio of 36:7:7. Only the RGB data in CSL was used, ignoring the depth information and key point annotation. A statistical comparison of our dataset with the public datasets is shown in TABLE 1.

C. EXPERIMENT ENVIRONMENT

The evaluation metric used for the COCO-WholeBody dataset is Object Key point Similarity (OKS), which is a measure of the similarity between the true and predicted key points. This metric is derived from Intersection of Union (IoU), calculated as follows:

$$OKS = \frac{\sum_i \exp(-\frac{d_i^2}{2s^2\sigma_i^2})\delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \quad (19)$$

where i indexes the key point in the range (1, 133). Key points (1, 11) and (91, 133) are selected and used in our model. d_i is the Euclidean distance between the predicted key point and

TABLE 1. Comparison with other datasets.

Datasets	RWTH-PHOENIX-Weather 2014T	CSL	Ours
Languages	German	Chinese	Chinese
RGB Resolution	210*260	1920*1080	1280*720
Framerate (FPS)	25	25	30
No. of statements	1066	100	60
No. of presenters	9	50	70
No. of instances	8257	25000	21000
No. of words	2887	178	171

the ground truth of key point i . s is the scale factor for the samples, calculated as the square root of the area enclosed by the detection label box. σ_i is a normalization factor, which is the standard deviation of the key point i across all samples. v_i is the visibility flag, indicating the visibility of a key point. $v_i > 0$ indicates that the key point i is considered. By setting $v_i = 0$ for those unimportant or redundant key points, the number of key points is reduced from 133 to 53. $\delta(v)$ is the impulse function, i.e., $\delta(v) = 1$ if $v = 0$, otherwise $\delta(v) = 0$. The value of OKS is in the range of [0,1]. The closer OKS is to 1, the closer the predicted value is to the true value.

The quantitative analysis of the pose estimation results is based on Average Precision (AP) and Average Recall (AR):

$$AP = \frac{\sum_i \delta(OKS > T)}{n} \quad (20)$$

$$AR = \frac{\sum_{j=1}^m \delta(OKS > T)}{m} \quad (21)$$

where T is the OKS threshold, n is the number of key points used, and m is the number of samples.

To measure AP and AR, the OKS threshold is set in the range of 0.5-0.95, with a step size of 0.05, to obtain 10 AP and AR values, which are then averaged to compute the mean Average Precision (mAP) and mean Average Recall (mAR), respectively.

In addition, the highest probability accuracy (Top-1 Accuracy) is used as the accuracy metric for sign language recognition. That is, for an input sign language word, the sign language word is considered to be correctly recognized if the output word with the highest probability agrees with the reference word.

The Word Error Rate (WER) is used to evaluate performance on the sign language recognition task, based on the accuracy of the annotation of isolated words in sign language. WER is calculated as follows:

$$WER = \frac{O_s + O_i + O_d}{N_{word}} \quad (22)$$

where N_{word} is the number of words contained in the actual text, O_s , O_i , and O_d represent the number of substitution operations, insertion operations, and deletion operations, respectively, when transforming the recognized text into the actual text. The smaller the value of WER, the better the recognition performance.

The performance on the sign language translation task is also measured using the Bilingual Evaluation Understudy (BLEU) [35], which is the most commonly used translation metric in machine translation. BLEU is calculated as follows:

$$BLEU = BP \times \exp\left(\sum_{n=1}^N w_n \log p_n\right) \quad (23)$$

where w_n and p_n are the weight and precision of the N-gram, respectively. N-gram means matching all clauses of length N in a sentence. BP is the penalty factor. If the length of the translation is less than the reference translation, BP is set less than 1 to avoid losing information, as follows:

$$BP = \begin{cases} 1 & lt > lr \\ \exp(1 - lr/lt) & lt \leq lr \end{cases} \quad (24)$$

where lt and lr represent the lengths of the actual translated text and the reference translated text, respectively. In our experiments, we set $N = 1$, $N = 2$, $N = 3$, $N = 4$ and the translation quality is evaluated using BLEU-1, BLEU-2, BLEU-3, BLEU-4.

TABLE 2. Parameters of SFE-Net on the COCO-WholeBody dataset with 53 key points.

Parameters	size/type
Loss function	SSE
Batch size	32
epoch	120
Optimizer	Adam
β_1	0.9
β_2	0.999
Initial learning rate	0.001
Decay rate	0.1

TABLE 3. Results of COCO-WholeBody with 53 key points and 133 key points.

Key points	body pose		hand pose		global pose	
	mAP	mAR	mAP	mAR	mAP	mAR
133	70.2	72.2	38.1	49.5	50.1	60
53	72.9	77.6	40.3	52.8	52.7	63.4

D. EXPERIMENT ENVIRONMENT

All experiments were conducted on a hardware server with NVIDIA Tesla V100 32GB GPU and Intel(R) Xeon(R) CPU E5-2698 V4 @2.20GHz. Pytorch is the deep learning framework used in this paper.

E. EXPERIMENT RESULTS

Our spatial-temporal model consists of two main modules: the spatial feature extraction module and the temporal feature extraction module. We have conducted experiments on both modules and the whole network.

1) SFE-NET RESULTS

TABLE 2 presents the parameters of SFE-Net, with the COCO-WholeBody dataset used. To validate the 53 key points selected in this work, TABLE 3 shows the pose-estimation results on the COCO-WholeBody dataset with 53 key points and 133 key points. It can be seen that the mAP and mAR of body pose, hand pose, and global pose are improved by reducing the 133 key points to 53 key points. This demonstrates the positive effect of key point reduction for sign language spatial features extraction. Moreover, the OpenPose [16], [17], [18], Single-Network (SN) [19], HPRNet [20], and HRNet [21] are also compared with our pose estimation method SFE-Net, which are presented in TABLE 4. These methods are compared on the COCO-WholeBody dataset with original 133 key points and 53 key points. SFE-Net achieves the best estimation results for body-pose estimation and global-pose estimation. HPRNet outperforms SFE-Net in hand-pose estimation, while



FIGURE 6. Examples of CSL-500 dataset using SFE-Net.

TABLE 4. Comparison with other methods on COCO-WholeBody with 133 and 53 key points.

Methods	Kp	body pose		hand pose		global pose	
		mAP	mAR	mAP	mAR	mAP	mAR
OpenPose	133	56.3	61.2	19.8	34.2	33.8	44.9
[16-18]	53	56.5	61.7	19.8	34.2	34.2	45.1
SN [19]	133	28.0	33.6	13.8	33.6	16.1	20.9
	53	28.4	33.8	13.6	33.5	16.2	21.7
HPRNet	133	55.2	63.1	47.0	60.8	33.3	43.4
[20]	53	61.2	69.1	51.4	64.8	37.5	49.6
HRNet	133	65.9	70.9	30.0	36.3	43.2	52.0
[21]	53	67.8	71.3	30.1	36.5	45.4	52.9
SFE-	133	70.2	72.2	38.1	49.5	50.1	60.1
Net(ours)	53	72.9	77.6	40.3	52.8	52.7	63.4

SFE-Net is the second best among the methods compared. HRNet performs better than HPRNet on global estimation, which verifies its superiority of feature sharing. SN, HPRNet, and HRNet are all one-step pose estimation networks, which perform predictions for the given 53 key points simultaneously. HPRNet performs well for small-scale estimation, such as hand pose. Both OpenPose and SFE-Net belong to two-step pose estimation networks. This means that separate training is required for hand and body pose estimation. From the results, SFE-Net outperforms OpenPose for all pose estimations.

In addition, we also evaluate the different methods on the public dataset CSL-500. TABLE 5 shows the parameters of SFE-Net on the CSL-500 dataset. TABLE 6 compares the accuracy of SFE-Net with other methods on the CSL-500

TABLE 5. Parameters of SFE-Net on CSL-500.

Parameters	size/type
epoch	50
Batch size	16
Initial learning rate	0.1
Decay rate	0.1
Optimizer	SGD
Momentum	0.9

TABLE 6. Comparison with other methods on CSL-500.

Methods	accuracy
3D-CNN	86.9
HPRNet	92.4
SFE-Net(ours)	95.7

dataset, with the same temporal feature extraction model. The proposed SFE-Net achieves an accuracy of 95.7%, which is better than 3D-CNN and HPRNet. This proves that using the 53 key points can lead to better learning of spatial features. Figure 6 shows some results by SFE-Net on the CSL-500 dataset. It can be seen that the proposed SFE-Net can accurately extract the pose features for sign language actions.

2) TFE-NET RESULTS

In order to evaluate the performance of TFE-Net, the RWTH-PHOENIX-Weather2014 dataset was used. TABLE 7 tabulates the parameters of the Transformer for the dataset. TABLE 8 presents the comparison results based on different position embedding methods. The Transformer network based on absolute position encoding is used as the benchmark in the experiments. The methods based on LSTM and GRU perform relative position encoding for temporal feature embedding. Since the encoder input is fully

TABLE 7. Parameters of the transformer on RWTH-PHOENIX-Weather2014.

Parameters	size/type
Number of encoders	3
Number of decoders	3
Number of self-attention heads	8
Initialization method	Xavier
dropout	0.1
Batch size	32
Optimizer	Adam
β_1	0.9
β_2	0.999
Initial learning rate	0.001
Decay rate	0.1
Decoding method	Greedy search decoding

TABLE 8. Results with different position embedding methods on RWTH-PHOENIX-Weather2014.

Position embedding method	WER		BLEU-4	
	Validation set	Test set	Validation set	Test set
Absolute position coding	24.61	24.49	22.12	21.80
LSTM	24.07	25.02	22.23	21.76
GRU	23.85	24.28	22.57	22.31

context-aware, bi-directional Bi-LSTM and Bi-GRU are used for feature embedding. On the RWTH-Phoenix-Weather 2014T dataset, compared to absolute position encoding, Using Bi-GRU for relative position embedding achieves smaller WER, reduced by 0.76 and 0.21 on the validation and test sets, respectively. Furthermore, BLEU-4 is improved by 0.45 and 0.51 on the validation set and test set, respectively. Bi-GUR also outperforms Bi-LSTM for relative position encoding. This means that using Bi-GRU for relative position encoding can enrich the position information for the sign language recognition and translation task. We also present the results of optimizing the self-attention layer using position awareness in TABLE 9. Using position awareness can reduce WER by 1.41 and 0.13 and improves the BLEU-4 by 0.23 and 0.60 on the validation and test sets, respectively, compared to no position awareness.

In addition, TABLE 10 tabulates the results of our method and other Transformer models for sign language recognition and translation on the RWTH-Phoenix-Weather 2014T dataset. It can be seen that, compared with the Sign Language Transformer [14], using our method achieves a lower WER on the validation set and test set by 1.61 and 1.21, respectively, and a higher BLEU-4 by 1.05 and 0.94, respectively. NSLT [5] is used as a benchmark experiment for

TABLE 9. Results of different self-attention approaches on RWTH-PHOENIX-Weather2014T.

Self-attention position	WER		BLEU-4	
	Validation set	Test set	Validation set	Test set
aware				
No position awareness	24.61	24.49	22.12	21.80
With position awareness	23.20	24.36	22.35	22.40

TABLE 10. Results of different temporal models on RWTH-PHOENIX-Weather2014T.

Models	WER		BLEU-4	
	Valid set	Test set	Valid set	Test set
NSLT [5]	-	-	18.40	18.13
Sign Language Transformer [14]	24.61	24.49	22.12	21.80
Multi-channel [10]	-	-	19.52	19.61
STMC-Transformer [11]	-	-	22.23	21.65
TFE-Net(ours)	23.00	23.28	23.17	22.74

the dataset, it achieves BLEU-4 of 18.40 and 18.13 on the validation set and test set, respectively. The temporal model of NSLT employs GRU-based attention mechanisms. All other models using Transformer, such as Sign Language Transformer [14], Multi-channel [10], and STMC-Transformer [11], can achieve better BLEU-4. Our proposed method achieves the highest BLEU-4 scores, which are 23.17 and 22.74 on the validation set and test set, respectively. Then, the direct CSLT task, i.e., with no intermediate supervision, can be achieved by setting the weight w_2 in equation (18) to 0. The corresponding results are tabulated in TABLE 11. Our network still outperforms all other models.

3) STFE-NET RESULTS

In this section, we evaluate the performance of STFE-Net, which combines the spatial and temporal models for continuous sign language translation. Figure 7 illustrates some pose estimation results of STFE-Net on our developed dataset.

After extracting the 53 key points, these spatial features are encoded and input to Transformer. The parameters of the Transformer on our dataset are shown in TABLE 12.

TABLE 13 shows the results of STFE-Net on our dataset and the public datasets RWTH-PHOENIX-Weather2014 and CSL. The results indicate that our network can perform better for Chinese continuous sign language translation. We also compare our method with different methods on the RWTH-PHOENIX-Weather 2014T dataset, and the results

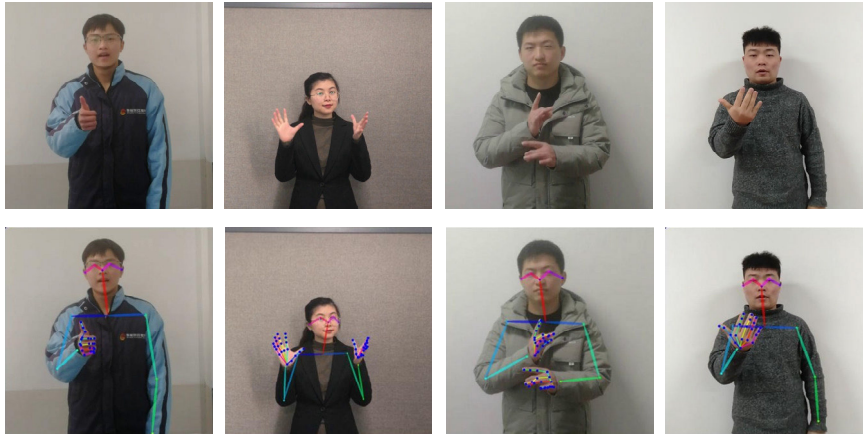


FIGURE 7. Examples of our dataset using SFE-Net.

TABLE 11. Results of different temporal models with no intermediate supervision on RWTH-PHOENIX-Weather2014T.

Models	BLEU-4	
	Valid set	Test set
NSLT [5]	9.94	9.58
Sign Language Transformer [14]	20.69	20.17
Multi-channel [10]	19.51	18.51
TFE-Net(ours)	21.48	21.29

TABLE 12. Parameters of TFE-Net on our dataset.

Parameters	size/type
Number of encoders	3
Number of decoders	3
Number of self-attention heads	8
Initialization method	Xavier
dropout	0.1
Batch size	32
Optimizer	Adam
β_1	0.9
β_2	0.998
Initial learning rate	0.001
Decay rate	0.1
Decoding method	Greedy search decoding

are shown in TABLE 14. Transformer [14] performs better than NSLT [5]. Multi-channel [10] was evaluated only for BLEU-4, and its result is inferior to that of Transformer. Our method achieves 22.57 and 22.45, in term of BLEU-4,

TABLE 13. Results of different datasets.

Datasets	Validation set				Test set			
	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4
RWTH	48.93	34.16	27.04	22.57	48.22	33.59	26.41	22.45
CSL	63.17	60.13	58.95	58.23	61.54	58.76	57.93	57.52
Our dataset	76.45	74.03	73.37	70.18	77.59	75.62	74.25	72.14

on the validation set and test set, respectively, and outperforms Transformer [14].

V. DISCUSSION

In this paper, we design a CSLT network based on spatial-temporal feature fusion. The model mainly consists of two networks, the spatial feature extraction network and the temporal feature extraction network. Our CSLT network is optimized for CSLT.

The spatial feature extraction network is an improvement on an existing pose estimation method. On the one hand, we ablate the 133 key points in COCO-WholeBody to 53 key points. The ablation of the key points allows our model to learn better sign language spatial features. The original numbers of key points for body pose, hand pose, and global pose estimation are 17, 42, and 133, respectively. After the reduction, the corresponding numbers of key points are 11, 42, and 53. As the results shown in TABLE 4, the mAP and mAR for the key point detection of the body, hand, and whole body are improved by using the ablated key points. This demonstrates the positive impact of the key point ablation operation for learning the sign language action features. While the key point ablation is to increase attention to the hands, this abatement may not be applicable to the scenarios involving lower limb posture, which is a limitation of this method. On the other hand, the hand pose expresses a large amount of sign language information, so a parallel structure, based on HRNet [21], is adopted. High-resolution pose

TABLE 14. Results of different methods on RWTH-PHOENIX-Weather 2014T.

Models	Validation set				Test set			
	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4
NSLT [5]	31.87	19.11	13.16	9.94	32.24	19.03	12.83	9.58
Transformer [14]	45.54	32.60	25.30	20.69	45.34	32.31	24.83	20.17
Multi-channel [10]	-	-	-	19.51	-	-	-	18.51
Ours	48.93	34.16	27.04	22.57	48.22	33.59	26.41	22.45

estimation is performed on the hand along with whole-body pose estimation. The proposed pose-estimation model was evaluated on the sign language recognition dataset CLS-500, and the recognition results were tabulated in TABLE 4. The results show that our model designed for pose estimation can extract better sign-language features compared to 3D-CNN [36] and HPRNet [20].

After extracting the continuous spatial features, the transformer network is utilized for temporal feature learning. To learn the sequential ordering information for the sign language in video sequences, relative position coding is used and is introduced into the self-attentive layer. The experimental results shown in TABLE 8 and TABLE 9 demonstrate that relative position encoding and self-attentive position perception positively affect CLSR. In addition, TABLE 10 compares the results of different temporal models on RWTH-Phoenix-Weather 2014T, and our approach outperforms four methods, including NSLT [5], Sign Language Transformer [14], Multi-channel [10], and STMC-Transformer [11].

The spatial feature extraction model and the temporal feature extraction model are combined to achieve end-to-end video sign language translation. TABLE 14 shows the results of our method on RWTH-PHOENIX-Weather 2014T, as well as the results of other methods, i.e., NSLT [5], Transformer [14], and Multi-channel [10]. Our proposed network outperforms the other methods. In addition, as can be seen in TABLE 13, the proposed model achieves BLUE-1=77.59, BLUE-2=75.62, BLUE-3=74.25, BLUE-4=72.14 on our Chinese continuous sign language dataset. The results for the CLS dataset are also promising. However, on RWTH-PHOENIX-Weather 2014T, our results are worse. This may be due to the fact that the RWTH-PHOENIX-Weather 2014T dataset has a larger amount of data and sentences, but a smaller number of samples per sentence. This makes the generalization of the model worse.

VI. CONCLUSION AND FUTURE WORKS

In this paper, a spatial-temporal feature extraction network was proposed for Chinese Sign Language Translation (CSLT) for real-world scenes. In the spatial feature extraction module, 53 key points related to sign language characteristics are selected. A parallel structure is designed to perform whole-body pose estimation, along with high-resolution hand pose estimation, to obtain finer-grained sign language features. In the temporal feature extraction module, temporal

features are learned based on Transformer. The designed Transformer employs relative position encoding for spatial features and sign language words, as well as incorporating relative position into Multi-Head Attention (MHA). Furthermore, a Chinese sign language dataset was created for this research, which enriches the research on Chinese sign language translation. Our proposed method can achieve promising performance on our dataset and multiple public datasets. This shows that our method is effective for continuous sign language recognition in practical scenarios.

The key points used in this paper are migrated from the whole-body pose estimation dataset COCO-WholeBody. Research on key points selected specifically for CSLT can further improve the translation accuracy. Furthermore, if a smaller number of key points are used, the model size and computational complexity can be further reduced. In addition, the dataset created in this work can be extended for the study of sign language recognition and translation with intermediate supervision. For a deep learning model, especially in real-time processing scenarios, accuracy and real-time performance (or availability) are two important evaluation metrics. The approach in this paper focuses more on providing a feasible approach for continuous sign language translation, more attention is paid to the recognition accuracy of pose estimation and sign language translation, while the real-time performance of the model is not quantified. Actually, key frame extraction will improve the real-time performance of video processing. Future work will be centered on efficient key frame extraction algorithms and real-time performance.

REFERENCES

- [1] M. J. Cheok, Z. Omar, and M. H. Jaward, "A review of hand gesture and sign language recognition techniques," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 1, pp. 131–153, Aug. 2019.
- [2] S. M. Kamal, Y. Chen, S. Li, X. Shi, and J. Zheng, "Technical approaches to Chinese sign language processing: A review," *IEEE Access*, vol. 7, pp. 96926–96935, 2019.
- [3] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.*, vol. 2018, Nov. 2018, Art. no. 7068349.
- [4] J. Forster, C. Schmidt, T. Hoyoux, O. Koller, U. Zelle, J. Piater, and H. Ney, "RWTH-PHOENIX-weather: A large vocabulary sign language recognition and translation corpus," in *Proc. 8th Int. Conf. Lang. Resour. Eval. (LREC)*, 2012, pp. 3785–3789.
- [5] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7784–7793.
- [6] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, Apr. 2018, pp. 1–8.
- [7] N. M. Kakoty and M. D. Sharma, "Recognition of sign language alphabets and numbers based on hand kinematics using a data glove," *Proc. Comput. Sci.*, vol. 133, pp. 55–62, 2018.
- [8] R. Rastgoo, K. Kiani, and S. Escalera, "Video-based isolated hand sign language recognition using a deep cascaded model," *Multimedia Tools Appl.*, vol. 79, nos. 31–32, pp. 22965–22987, Aug. 2020.
- [9] S.-K. Ko, C. J. Kim, H. Jung, and C. Cho, "Neural sign language translation based on human keypoint estimation," *Appl. Sci.*, vol. 9, no. 13, p. 2683, Jul. 2019.
- [10] 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.* Springer, 2020, pp. 301–319.

- [11] K. Yin and J. Read, "Better sign language translation with STMC-transformer," in *Proc. 28th Int. Conf. Comput. Linguistics*, Dec. 2020, pp. 5975–5989.
- [12] S. Jin et al., "Whole-body human pose estimation in the wild," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 196–214.
- [13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
- [14] N. Cihan Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10023–10033.
- [15] S. Kim, C. J. Kim, H.-M. Park, Y. Jeong, J. Y. Jang, and H. Jung, "Robust keypoint normalization method for Korean sign language translation using transformer," in *Proc. Int. Conf. Inf. Commun. Technol. Conver. (ICTC)*, 2020, pp. 1303–1305.
- [16] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7291–7299.
- [17] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1145–1153.
- [18] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4724–4732.
- [19] G. H. Martinez, Y. Raaj, H. Idrees, D. Xiang, H. Joo, T. Simon, and Y. Sheikh, "Single-network whole-body pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6982–6991.
- [20] N. Samet and E. Akbas, "HPRNet: Hierarchical point regression for whole-body human pose estimation," *Image Vis. Comput.*, vol. 115, Nov. 2021, Art. no. 104285.
- [21] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5693–5703.
- [22] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.
- [23] 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.*, vol. 28, 2015, pp. 1–9.
- [24] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.
- [25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
- [26] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," 2018, *arXiv:1803.02155*.
- [27] M. Neishi and N. Yoshinaga, "On the relation between position information and sentence length in neural machine translation," in *Proc. 23rd Conf. Comput. Natural Lang. Learn. (CoNLL)*, 2019, pp. 328–338.
- [28] L. R. Medsker and L. C. Jain, "Recurrent neural networks," *Design Appl.*, vol. 5, pp. 64–67, Dec. 2001.
- [29] A. Mohamed, D. Okhonko, and L. Zettlemoyer, "Transformers with convolutional context for ASR," 2019, *arXiv:1904.11660*.
- [30] N. Kitaev, Ł. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," 2020, *arXiv:2001.04451*.
- [31] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," 2019, *arXiv:1901.02860*.
- [32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
- [33] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, 2006, pp. 369–376.
- [34] J. Zhang, W. Zhou, C. Xie, J. Pu, and H. Li, "Chinese sign language recognition with adaptive HMM," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2016, pp. 1–6.
- [35] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2001, pp. 311–318.
- [36] M. C. Ariesta, F. Wiryana, Suharjito, and A. Zahra, "Sentence level Indonesian sign language recognition using 3D convolutional neural network and bidirectional recurrent neural network," in *Proc. Indonesian Assoc. Pattern Recognit. Int. Conf. (INAPR)*, Sep. 2018, pp. 16–22.



JIWEI HU received the B.E. degree in electronic information engineering from the Wuhan University of Technology, China, in 2007, and the Ph.D. degree in electronic and information engineering from The Hong Kong Polytechnic University, in 2013. He is currently an Associate Professor with the School of Information Engineering, Wuhan University of Technology. His research interests include computer vision, computer science, signal processing, and data mining.



YUNFEI LIU received the bachelor's degree in electronic information engineering from the Wuhan University of Technology, Wuhan, China, in 2020, where he is currently pursuing the master's degree with the Department of Information Engineering. His research interests include machine learning, 3-D reconstruction, natural language processing, and simultaneous localization and mapping.



KIN-MAN LAM received the M.Sc. degree in communication engineering from the Department of Electrical Engineering, Imperial College of Science, Technology and Medicine, London, U.K., in 1987, and the Ph.D. degree from the Department of Electrical Engineering, University of Sydney, Australia, in 1996. He is currently a Professor with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University. His research interests include human face recognition, image and video processing, and computer vision.



PING LOU received the M.S. and Ph.D. degrees in mechanical engineering from the Huazhong University of Science and Technology, Wuhan, China, in 1997 and 2004, respectively. She is currently a Professor with the School of Information Engineering, Wuhan University of Technology. Her research interests include network manufacturing, digital manufacturing, intelligent manufacturing, and supply chain.

...