IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 18 October 2022, accepted 1 December 2022, date of publication 6 December 2022, date of current version 9 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3227042

RESEARCH ARTICLE

Spatial–Temporal Graph Transformer With Sign Mesh Regression for Skinned-Based Sign Language Production

ZHENCHAO CUI[1,2], ZIANG CHEN[1,2], ZHAOXIN LI[3], AND ZHAOQI WANG[3]

[1]School of Cyber Security and Computer, Hebei University, Baoding 071002, China
[2]Hebei Machine Vision Engineering Research Center, Hebei University, Baoding 071002, China
[3]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

Corresponding author: Zhaoxin Li (cszli@hotmail.com)

This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFC1523302, in part by the Post-Graduate's Innovation Fund Project of Hebei University under Grant HBU2022ss014, in part by the National Natural Science Foundation of China under Grant 62172392, in part by the Scientific Research Foundation for Talented Scholars of Hebei University under Grant 521100221081, and in part by the Scientific Research Foundation of Colleges and Universities in Hebei Province under Grant QN2022107.

ABSTRACT Sign language production aims to automatically generate coordinated sign language videos from spoken language. As a typical sequence to sequence task, the existing methods are mostly to regard the skeletons as a whole sequence, however, those do not take the rich graph information among both joints and edges into consideration. In this paper, we propose a novel method named Spatial-Temporal Graph Transformer (STGT) to deal with this problem. Specifically, according to kinesiology, we first design a novel graph representation to achieve graph features from skeletons. Then the spatial-temporal graph self-attention utilizes graph topology to capture the intra-frame and inter-frame correlations, respectively. Our key innovation is that the attention maps are calculated on both spatial and temporal dimensions in turn, meanwhile, graph convolution is used to strengthen the short-term features of skeletal structure. Finally, due to the generated skeletons are based on the form of skeleton points and lines so far. In order to visualize the generated sign language videos, we design a sign mesh regression module to render the skeletons into skinned animations including body and hands posture. Comparing with states of art baseline on RWTH-PHONEIX Weather-2014T in Experiment Section, STGT can obtain the highest values on BLEU and ROUGE, which indicates our method produces most accurate and intuitive sign language videos.

INDEX TERMS Transformer, graph convolution, human mesh reconstruction, sign language production.

I. INTRODUCTION

As a useful language, sign language conveys information through gestures and spatial movements of limbs. It is the most natural way for hearing-impaired people to interact with the outside world. Sign language production (SLP) aims to automatically translate spoken language sentences into the corresponding sign language videos. Both accurate and vivid SLP can significantly improve the communication quality for the Deaf community. Sign glosses are intermediary words that match the meaning of spoken language. As shown in Fig. 1, our work can be divided into two parts: (1) Translating gloss sequences for spoken language sentences into the corresponding sign pose sequences. (2) Generating skinned-based animations from skeleton sequences.

Recently, Transformer-based methods [1], [2], [3], [4], [5] became the most widespread methods to produce skeletons for SLP. However, there is still a problem in these works: such architecture always ignores the structural relationships of the human skeletons, by which poor performance would be obtained. Thereupon, the existing SLP method [4] devises a spatial-temporal graph convolution (GCN) as pose generator which implemented from a standard 2D convolution. Skeletal graph self-attention [6] encodes the spatio-temporal

The associate editor coordinating the review of this manuscript and approving it for publication was Hossein Rahmani.

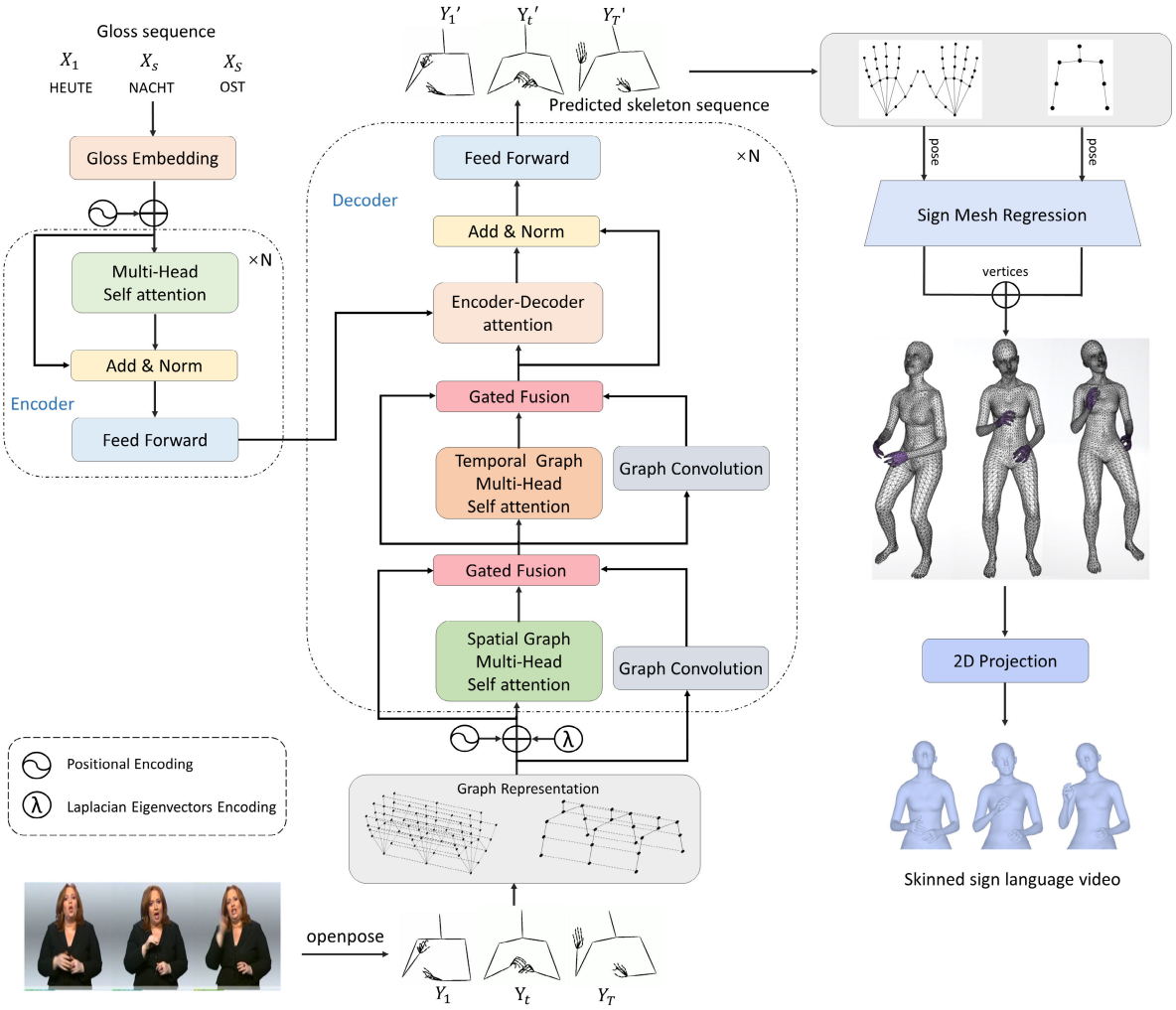


FIGURE 1. The overall architecture of STGT, which composes of gloss sequence encoder, skeleton sequence decoder and sign mesh regression module. The encoder learns semantic features from source sentences and the decoder captures both intra-frame and inter-frame correlations between dynamic skeletons. Sign mesh regression module takes the skeletons from the encoder-decoder network, and renders the final output into skinned sign language animation.

connectivity into the node features while calculating attention matrices. To deal with the problem, we conduct two self-attention layers with different dimensions in turn and equip with GCN to strengthen the short-term structure features lacked in attention results.

Due to the small motion ranges of hands and large range of motions in upper limbs, based on kinesiology, we propose a novel graph partition strategy with a combination of connectivity and motion relationship. The novel graph topology is characterized by a partitioned laplacian matrix, which makes the encoded representation more comprehensive.

To facilitate the analysis of sign language, we design a sign mesh regression module for animating the generated skeleton sequences. We employ SMPL [7] for generating skinned body shapes of the upper limb and MANO [8] model accurate reconstruction of hands. The skinned meshes of different parts are assembled by a fast Copy-and-Paste, providing a more comprehensive and graphic sign language video which can better reflect the real 3D structure of the human body.

The major contributions of our work are summarized as follows:

- We introduce a novel Spatial-Temporal Graph Transformer, STGT, considering both the intra-frame and inter-frame correlations. It is able to exploit the spatial displacements and temporal dynamics of skeletal data more effectively. Meanwhile, the gated fusion module is proposed for modeling both long-term and short-term dependencies of skeletal structure in an efficient way.
- In graph topology representation, an additional motion relation between fingers is combined with bone connectivity. Moreover, we explicitly utilize the novel graph representation inductive bias in self-attention layers, which further improve the model performance.

- To produce realistic and visual sign language videos, a sign mesh regression module is presented for rendering skinned sign language animations from skeletons. To our knowledge, this is the first work of skinned-based sign language production.

Experiments demonstrate the superior performance of our method to the competing methods on RWTH-PHOENIX-Weather-2014T dataset. We achieve BLEU-1 score of 36.01 and ROUGE score of 37.62 on STGT($\mathcal{C}\&\mathcal{M}$), which increases 1.49 BLEU score and 1.34 ROUGE score than reproduced Saunders' results [6] in a fair comparison.

The rest of this paper is organised as follows: In Section 2, we survey the related works in the field of SLP and human mesh reconstruction. In Section 3, we introduce the novel graph representation, gloss sequence encoder, sign sequence decoder and sign mesh regression module. We share our experimental details in Section 4, the quantitative results and the qualitative examples are also presented here. Finally, we draw conclusions in Section 5 and suggest future work.

II. RELATED WORK

A. GRAPH CONVOLUTIONAL NETWORK AND TRANSFORMER

Several works have considered Transformer [9] and Spatial-Temporal Graph Convolutional Network(ST-GCN)) [10] to process the spatial-temporal connectivity in non-Euclidean datasets. The original Transformer operates on fully connected graphs representing all connections between the tokens. So that it sticks to poor performance when the graph topology has not been encoded into the node features. While ST-GCN introduced high-level semantics such as both the spatial and temporal edges from data, it could provide strong complementarities to Transformer.

The combination of graph and transformer has many applications in other fields. Guo et al. [11] proposed a self-attention based graph neural network for traffic forecasting, which is specialized for capturing the temporal dynamics of traffic data by self-attention and using graph convolution module to capture the spatial correlations. Specific to skeleton-based human action recognition, a recent study by Plizzari et al. [12] model dependencies between joints by the self-attention operator and use a two-stream mechanism for conditionally building the natural human body structure. Dwivedi et al. [13] proposed a generalization of transformer neural network architecture for arbitrary graphs which is extended to both node and edge feature representation. Inspired by their works, our mining of sign language information is extended to the edge dimension, which represents the relative distance of joints during gesture movement.

B. SIGN LANGUAGE PRODUCTION

Sign language production is a fundamental problem in neural machine translation and has been widely attracting a lot of attentions in recent years [14], [15], [16], [17]. Since Transformer [9] adopts the self-attention mechanism without convolution, which has made great breakthroughs in the field of natural language processing. Saunders et al. [1] proposed the first Transformer-based SLP model to learn the mapping between spoken language sentences and sign pose sequences in an end-to-end manner. The above researches usually convert the sign pose sequences into Euclidean data which seriously ignores the original structure, semantics and other characteristics of the skeletal data.

Further, Saunders et al. [6] proposed a spatial-temporal skeletal graph attention layer that embeds a hierarchical body inductive bias into the self-attention mechanism. Huang et al. [4] developed spatial-temporal graph convolution layers into the pose generator which is able to capture both intra-frame and inter-frame information of sign language videos. However, all these methods disregard each joint has different contributions to gestures expression. Both motion relationship and the action amplitude will influence the sign language meaning. To make an efficient representation of non-Euclidean data, we define a novel graph partition strategy constructing the upper limb and hands respectively.

C. HUMAN MESH RECONSTRUCTION

In human mesh reconstruction, a skinned vertex-based model reconstructs the skin that is represented by 3D mesh and can be regarded as the modeling of real geometry. Skinned multi-person linear model (SMPL) [7] is used to parameterize the basic attributes of the human body model, such as a wide variety of body shapes in natural human poses. SMPL uses skeletons to drive meshes for deformation. It consists of 6890 vertices, 13776 triangular meshes and 24 joints, however the reconstructed 3D surface does not include hand details. Hand model with articulated and non-rigid deformations (MANO) [8] is an end-to-end learnable model which provides a compact mapping from hand poses to mesh blend shape corrections. Faces Learned with an articulated model and expressions (FLAME) [18] assumes a whole head mapping, captures the 3D rotation of the head, and also models the neck area. SMPLX [19] computes a 3D model of human body pose, hand pose, and facial expression. It combines SMPL with FLAME head model and MANO hand model, yielding in natural and expressive results.

III. METHODOLOGY

In this section, we introduce technical details of the proposed spatial-temporal graph transformer (STGT) with sign mesh regression for sign language production (SLP), and Fig. 1 shows its overall architecture. We first formulate the SLP task as a spatial-temporal translation problem. Given the source spoken language sentence $\mathcal{X} = (X_1, X_2, \ldots, X_S)$ with S glosses, we focus on translating it into the corresponding target sign poses sequence $\mathcal{Y} = (Y_1, Y_2, \ldots, Y_T)$ with T frames. Intra-frame skeletons are expressed as $Y_t = (y_1^t, y_2^t, \ldots, y_G^t) \in \mathbb{R}^G$ and contained G joints, where y_g^t denotes the position of joint g at time t. Our goal is to fit our model of maximizing the computation of conditional probability $\mathcal{P}(Y|X)$.

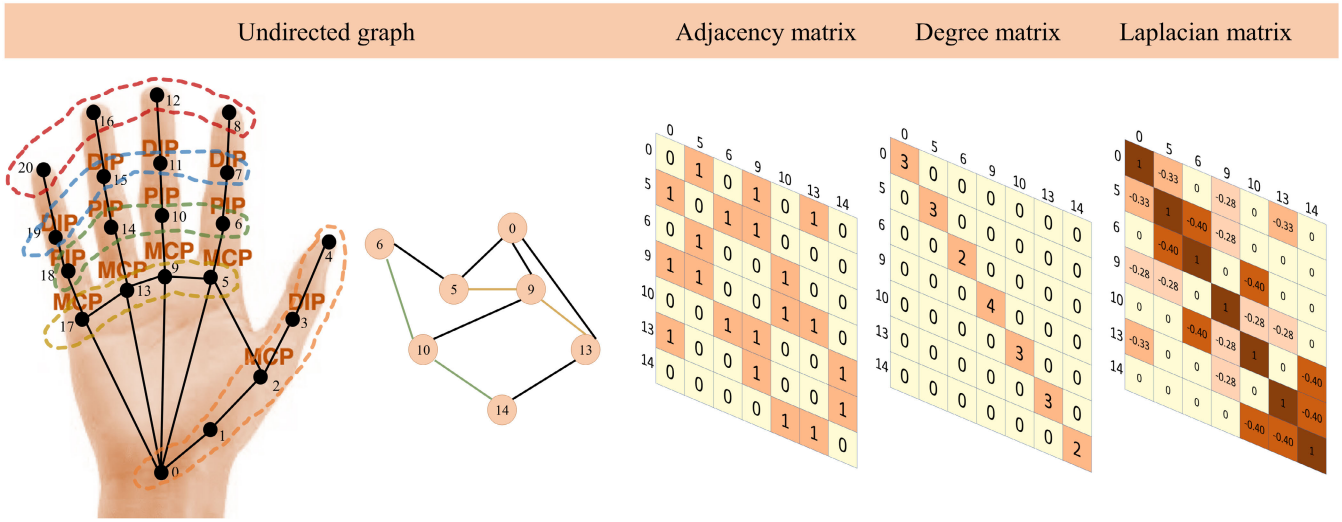


FIGURE 2. The graph topology representation and adjacency matrix of right hand. Exploiting Dynamic Information. As a demonstration, we choose seven finger joints, calculating the adjacency matrix, degree matrix and normalized laplacian matrix respectively.

Firstly, a novel graph representation method is introduced for understanding human skeletons in sign language action. Then we elaborate on the spatial-temporal graph transformer block, automatically capturing both intra-frame and inter-frame correlations between dynamic skeletons. Finally, the SLP results are displayed in the form of skinned-based animation by our proposed sign mesh regression module, which moves beyond the visual presentations of previous methods.

A. NOVEL GRAPH REPRESENTATION

1) NOVEL GRAPH PARTITION STRATEGY

In sign language gestures, the role of hands in semantic representation is the most obvious. The second is the upper limb, which collaborates with the hands to perform the elaborate sign language expression through the lifting or lowering action. Due to the motion amplitude between hands and upper limb being different, we divide the skeleton sequence into three parts: the upper limb, the left and right hands. The graph based on sign language skeleton can be constructed as a combination of sub-graphs corresponding to each part, where the adjacent sub-graphs have at least one common joint (the wrist joint). As shown in the Fig. 2, the right hand sub-graph representation can be formulated as a spatial adjacency matrix symbolizing the intra-frame relationship between each joint.

2) NOVEL SPATIAL ADJACENCY MATRIX

Previous work [6] builds the spatial adjacency matrix only by the skeleton structure. It ignores that, during sign language communication, joints in different fingers are often connected due to the motion relationship. On the basis of the skeleton graph, we give a supplementary motion relationship graph associated with finger movement. As shown in the Fig. 2, finger joints include metacarpophalangeal point (MCP), proximal interphalangeal point (PIP) and distal interphalangeal point (DIP). Because the thumb has high freedom and flexibility, we divide the thumb joints into a motion graph separately. The other four fingers have the same structure, and the same joint of different fingers will have a similar movement trend. Therefore, we have established the motion relationship graphs of four MCP joints, four PIP joints and four DIP joints.

Our spatial adjacency matrix $\mathcal{A} \in \mathbb{R}^{G\times G}$ takes both the connectivity and the motion relationship within a frame into consideration. The extended motion relationship strengthens the rationality and sensitivity of humans poses in sign language actions. The topology analysis of left hand $\mathcal{A}^l$ can be formalised as:

$$\mathcal{A}^l[i][j] = \begin{cases} 1 & if\ Con(i,j) = True \\ 1 & if\ Mot(i,j) = True \\ 0 & otherwise \end{cases} \tag{1}$$

where $Con(i,j)$ and $Mot(i,j)$ indicate the joints have the connectivity and the motion relationship respectively. Degree matrix $\mathcal{D} \in \mathbb{R}^{G\times G}$ is a diagonal matrix where $\mathcal{D}^l[i][j] = \sum_j \mathcal{A}^l[i][j]$, and the elements on the diagonal are the degrees of each joint. $\mathcal{I} \in \mathbb{R}^{G\times G}$ is an identity matrix, which represents self-connections. Due to imbalanced weights may undesirably affect the matrix spectrum, we use the symmetrically normalized laplacian matrix $\mathcal{L}^l_{sym}$ for undirected graph representation. Normalisation can be formulated as:

$$\mathcal{L}^l_{sym} = \mathcal{I}^l - \mathcal{D}^{-\frac{1}{2}}\mathcal{A}^l\mathcal{D}^{\frac{1}{2}} \tag{2}$$

It is noteworthy that the ultimate $\mathcal{L}_{sym}$ is actually a partitioned matrix which is composed of the left hand laplacian matrix $\mathcal{L}^l_{sym}$, right hand laplacian matrix $\mathcal{L}^r_{sym}$, and the upper

limb laplacian matrix $\mathcal{L}^u_{sym}$:

$$\mathcal{L}_{sym} = \begin{bmatrix} \mathcal{L}^l_{sym} & O & O \\ O & \mathcal{L}^r_{sym} & O \\ O & O & \mathcal{L}^u_{sym} \end{bmatrix} \tag{3}$$

B. IMPLEMENTATION DETAILS OF GLOSS SEQUENCE ENCODER

The encoder learns semantic features from an embedded gloss sequence $\tilde{\mathcal{X}} \in \mathbb{R}^{S\times d_{model}}$, where S denotes the number of sign glosses and d_{model} represents the dimension of embedded vectors. The order information plays an important role in neural machine translation tasks since it defines the syntax of sentences and the composition of videos. Hence, we equip with the positional encoding in (4) which aims at explicitly inducing the order bias into the gloss sequence.

$$\begin{aligned} PE(d, 2i) &= \sin\left(\frac{d}{10000^{\frac{2i}{d_{model}}}}\right) \\ PE(d, 2i+1) &= \cos\left(\frac{d}{10000^{\frac{2i}{d_{model}}}}\right) \end{aligned} \tag{4}$$

where d is the relative index of each gloss in the sequence and i is introduced to distinguish odd-even.

The architecture of gloss sequence encoder closely resembles the classical transformer [9], which is composed of N blocks with multi-head self attention (MHA) and feed forward network (FFN). MHA projects query vector Q, key vector K and value vector V through h different linear transformations, and finally concatenate h different attention results of the global gloss sequence. Then the MHA outputs pass into the FFN, which is a fully connected network with two linear layers. The basic operation in MHA is the scaled dot-product attention defined in (5):

$$\begin{aligned} Attention(Q, K, V) &= softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \\ Q = \tilde{\mathcal{X}}W^Q,\ K &= \tilde{\mathcal{X}}W^K,\ V = \tilde{\mathcal{X}}W^V \end{aligned} \tag{5}$$

where dividing by $\sqrt{d_k}$ is to prevent the saturation led by softmax function and the input $\tilde{\mathcal{X}}$ is projected by three matrices $W^Q \in \mathbb{R}^{d_Q\times d_{model}}$, $W^K \in \mathbb{R}^{d_K\times d_{model}}$, $W^V \in \mathbb{R}^{d_V\times d_{model}}$ to the corresponding representations Q, K, V. Since the encoder consists of a stack of identical layers, the residual connection and layer normalization are used inside the blocks to ensure stable training as the encoder goes deeper.

C. IMPLEMENTATION DETAILS OF SKELETON SEQUENCE DECODER

1) POSITIONAL LAPLACIAN EIGENVECTORS ENCODING

The embedded skeleton sequence after the positional encoding in (4) is symbolized by $\tilde{\mathcal{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_T) \in \mathbb{R}^{G\times T\times d_{model}}$, where G is the total number of joints, T is the total frames number and d_{model} represents the dimension of embedded vectors. The action information of skeleton sequence mainly exists inside a single frame (in spatial dimension), while the motion track information is contained between consecutive frames (in temporal dimension). We build the undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to represent spatial-temporal structure, where the node features $\mathcal{V} = \{y^t_g \mid g = 1, \ldots, G$ and $t = 1, \ldots, T\}$, t represents the frames in temporal domain, and g represents skeleton joints in spatial domain. The spatial edge features $\mathcal{E}_s = \{y^t_i y^t_j \mid (i, j) \in G\}$ contains both the connectivity of skeletons and the motion relationship, which expresses the relationship between joint i and joint j at frame t. We also tried to learn the representations of temporal edge features $\mathcal{E}_t = \{y^t_i y^{t+1}_i\}$ through connecting the joints between adjacent frames. Finally, we convert the spatial graph $\mathcal{G}^s$ and temporal graph $\mathcal{G}^t$ into laplacian matrices $L^s_{sym} \in \mathbb{R}^{G\times G}$ and $L^t_{sym} \in \mathbb{R}^{T\times T}$ followed by (1) (2) (3). In order to integrate $\tilde{\mathcal{Y}}$ with L^s_{sym} and L^t_{sym}, we expand them along the spatial and temporal axes to generate $\mathcal{L}^s \in \mathbb{R}^{G\times G\times T}$ and $\mathcal{L}^t \in \mathbb{R}^{T\times G\times T}$.

2) SPATIAL-TEMPORAL GRAPH SELF-ATTENTION

An advantage of Transformer is its global receptive field, which we use to capture long-term interactions of skeleton sequences. Instead of using the classical self attention in (5), we introduce the spatial graph self-attention (S-GSA) module to embed the intra-frame dependencies into the Query-Key product matrix M. The query $Q^s \in \mathbb{R}^{G\times T\times d_{model}}$, key $K^s \in \mathbb{R}^{G\times T\times d_{model}}$, value $V^s \in \mathbb{R}^{G\times T\times d_{model}}$ representations are projected into different subspaces by applying multiple trainable transformations to $\tilde{\mathcal{Y}}$. When calculating M, we use the Einstein summation convention to convert the inner product into the same dimension as $\mathcal{L}^s$. The calculation procedure of spatial graph self-attention is shown in (6).

$$\begin{aligned} M^s &= Q^s(K^s)^T \in \mathbb{R}^{G\times G\times T} \\ Attention(M^s, V^s, \mathcal{L}^s) &= softmax\left(\frac{M^s + \mathcal{L}^s}{\sqrt{d^s_k}}\right)V^s \end{aligned} \tag{6}$$

where the spatial graph matrix $\mathcal{L}^s \in \mathbb{R}^{G\times G\times T}$ will be added to M, and we set $\mathcal{L}^s$ as a learnable weight matrix during training. Note that we equip with the masked mechanism [9] in graph self-attention to prevent the leakage of future information when decoding target sequences.

Wu et al. [20] prove that convolutions can be incorporated into the Transformer to improve performance and robustness for datasets containing local structures. Considering the local information of skeleton sequences, we further use graph convolution (GCN) to make better initialize the weight of edge information and strengthen the short-term features lacking in S-GSA. Note that convolution can remember the position information, so the position embedding operation dropped here. The GCN operation in (7) calculates for each frame separately based on $L^s_{sym} \in \mathbb{R}^{G\times G}$, and concatenates all results at last.

$$GCN(\tilde{Y}, L^s_{sym}) = Concat^T_t(\sigma(W^{GCN}L^s_{sym}\tilde{Y}_t)) \tag{7}$$

where $\tilde{Y}_t \in \mathbb{R}^{G\times d_{model}}$, $W^{GCN} \in \mathbb{R}^{d_{model}\times d_{model}}$ and σ are the embedded skeletons at one time step, projection matrix and sigmoid nonlinear activation, respectively.

Similar to S-GSA, we build the temporal graph self-attention (T-GSA) module in (8) to capture the long-term inter-frame correlations based on $\mathcal{L}^t \in \mathbb{R}^{T\times G\times T}$. Note that S-GSA is performed for each frame, while T-GSA is performed for each joint. And we also use *GCN* to build short-term dependencies in temporal dimension based on $L^t_{sym} \in \mathbb{R}^{T\times T}$.

$$M^t = Q^t(K^t)^T \in \mathbb{R}^{T\times G\times T}$$
$$Attention(M^t, V^t, \mathcal{L}^t) = softmax\left(\frac{M^t + \mathcal{L}^t}{\sqrt{d_k^t}}\right)V^t \quad (8)$$

3) GATED FUSION MODULE

Especially, to combine with the local dependencies (GCN_{out}) and global dependencies (GSA_{out}) in an efficient way, we design a gated fusion module to control information flows with gates. in (9).

$$gate = \sigma(W^1\mathcal{Y} + W^2GCN_{out} + W^3GSA_{out} + b) \quad (9)$$

where W and b_1 are the weight matrix and bias vector of the fully connected layer. As a result, the output is obtained by weighting GCN_{out} and GSA_{out} with the gate:

$$out = (1 - gate) \odot GCN_{out} + gate \odot GSA_{out} \quad (10)$$

where $\odot$ is element-wise multiplication. Note that both S-GSA and T-GSA outputs use this gated weighting method to aggregate useful information with GCN outputs.

4) ENCODER-DECODER ATTENTION AND LOSS FUNCTION

After spatial-temporal graph self-attention layers, an Encoder-Decoder attention layer is used to focus on the appropriate alignment between gloss sequence and skeleton sequence. Therefore, it likes the classical self attention in (5), except creating the query matrix from the output of the previous layer (Gated Fusion of T-GSA and GCN). The Key and Value matrices come from the gloss sequence encoder actually.

The output of our skeleton sequence decoder is a vector of floats and we use a linear layer to turn that into the predicted skeleton sequence $\mathcal{Z} \in \mathbb{R}^{G\times T\times d_{model}}$. Mean square error loss $MSE(\mathcal{Z}, \tilde{\mathcal{Y}})$ is utilized to fit out model minimizing the error between predicted $\mathcal{Z}$ and the ground truth $\tilde{\mathcal{Y}}$.

D. SIGN MESH REGRESSION

The results of existing SLP approaches are mostly embodied in the form of 2D joint points and lines, which leads to abstract expression of the human body. Our sign mesh regression module provides both body mesh parameter map and hand mesh parameter map, which can jointly describe the skinned sign language videos based on the skeleton sequences $\mathcal{Z}$.

For an efficient integration of SMPL [7] and MANO [8] model, we refer to FrankMocap [21] to assemble body and hands by a fast Copy-and-Paste. The 2D coordinates in $\mathcal{Z}$ are converted into $\theta \in \mathbb{R}^{G\times 3}$, which symbolize 3D rotation of G body joints in Rodrigues representation. When transferring the corresponding joint angle parameters from the hands and body, the wrist joints that connect the two parts need to be treated independently. The 3D rotation parameters for the whole joints are denoted as $\theta^{whole} : \{\theta^{body} \cup \theta^{wrist} \cup \theta^{hand}\}$, where θ^{whole} is respectively composed of body, both wrists and both hands.

Sign mesh regression can be expressed as a differentiable mesh function M and mesh vertex position function T in (11):

$$M(\beta^{fix}, \theta^{whole}) = W(T(\beta^{fix}, \theta), J(\beta^{fix}), \theta^{whole}, \omega)$$
$$T(\beta^{fix}, \theta^{whole}) = \overline{T} + B_S(\beta^{fix}) + B_p(\theta^{whole}) \quad (11)$$

where W is a mixed skinning linear function, ω is the blended weight of each joint. Note that our purpose is to animate the generated sign language sequences, the sign language meaning is not related to the change of body shape. Hence we choose a set of fixed shape parameters β^{fix}, inputting it into the corresponding joint location function J and mesh vertex position function T. $\overline{T}$ is a uniform template, which represents the whole body mesh at rest. The shape mixing function B_S gets the blended shape of the whole body. B_p is posture mixing function, which inputs θ^{whole} and outputs the mesh deformation caused by posture change.

IV. EXPERIMENTS

A. IMPLEMENTATION DETAILS

1) DATASET AND PREPROCESSING

The training dataset of the proposed approach is RWTH-PHOENIX-Weather-2014T [22], which records the daily news and weather forecast airings of the German public tv-station PHOENIX featuring sign language interpretation. It contains 8257 video samples, and a total of 2887 words are combined into 5356 continuous sentences related to the weather forecast. In order to eliminate redundant information and reduce the amount of calculation, sign language videos are extract the 2D skeletons by Openpose [23] which contains 8 joints of the upper body and 21 joints of each hand. By observing imbalance data distribution of the skeleton sequences, we discard the abnormal joints and process the missing joints through weighted linear interpolation.

2) EVALUATION METRICS AND BASELINES

We evaluate STGT and benchmark sign language production methods through back-translation by continuous sign language translation (SLT) model [24]. Baselines include progressive transformer (PT) [1] and skeletal graph self-attention (Skeletal-GSA) [6]. According to the baseline methods, the input of SLT is changed from sign video frames to skeleton sequences. The score is presented with standard metrics including BLEU-1/4 and ROUGE. BLEU measures how much the frames in the machine generated sign language

video appeared in the Ground Truth. ROUGE measures how much the frames in the Ground Truth appeared in the machine generated sign language video.

3) EXPERIMENTAL SETTINGS

The proposed model is built by PyTorch deep learning framework, and a NVIDIA geforce RTX 3060 GPU is used for model training and inference. During the training phase, both our model and compared methods almost follow the batch size 32 and Adam [25] as the optimizer. We set the number of heads for multi-head attention to 8, the spatial-temporal graph transformer layers to 4, the embedding dimensions d_{model} to 512 and the feed forward dimensions in each layer to $4 \times d_{model}$. The cosine decay with warmup learning rate [26] is employed in the first 100 steps with the maximum $1e-3$ and the minimum $1e-4$.

B. ABLATION STUDIES

In this section, we will experimentally analyze STGT in detail from the following aspects.

TABLE 1. Comparison of precision and edge number in different graph relationships on RWTH-PHOENIX-Weather-2014T.

Graph relationship	Edges	BLEU-1	ROUGE
connectivity($\mathcal{C}$)	55	33.52	35.39
connectivity&motion($\mathcal{C}\&\mathcal{M}$)	73	**34.35**	**36.24**

1) THE EFFECTIVENESS OF COMBINING CONNECTIVITY AND MOTION RELATIONSHIP

We first vary the graph embedding of connectivity and motion relationship in the skeleton sequence. The results provide a fair comparison in the decoder configuration of S-GSA and GCN. Table 1 summarises BLEU-1 and ROUGE scores in different graph relationships. The basic connectivity relationship of sign language skeletons has 55 edges to represent bones that link joints, while our method incorporates additional 18 motion edges on this basis. According to the results, combining the connectivity and motion relationship will lift 0.83 BLEU score and 0.85 ROUGE score than the single connectivity relationship. The main reason is that adding motion relationships can make the model pay attention to the coordinate changes of joints with sign language actions on the basis of skeleton structure.

2) THE EFFECTIVENESS OF SPATIAL-TEMPORAL GRAPH SELF-ATTENTION

To verify the effectiveness of our proposed modules, we gradually embed S-GSA, T-GSA and both the two modules (ST-GSA) into the decoder configuration. The results are listed in Table 2. Note that all spatial graph representations consider connectivity and motion relationships.

- The effectiveness of S-GSA can be clearly seen in the cases of Transformer and S-GSA&GCN. When the decoder uses S-GSA to establish correlations in spatial dimension and GCN to capture temporal features, the BLEU-1 and ROUGE scores are improved by 2.10 and 3.15 respectively. It verifies the availability for the spatial graph self-attention on combined joints and bones information.
- We further validate that the proposed T-GSA is efficient to capture long-range temporal dependencies for each joint in temproal dimension. In comparison Transformer with our T-GSA, the BLEU-1 and ROUGE scores are slightly improved by 0.67 and 0.38.
- After simultaneously using both our S-GSA and T-GSA, ST-GSA shows that the BLEU score lifts 2.84 and ROUGE score lifts 3.92, confirming the effectiveness of the spatial-temporal graph self-attention on skeleton sequences.

TABLE 2. Comparison of precision in different decoder configurations on RWTH-PHOENIX-Weather-2014T.

Decoder configuration	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE
Transformer	11.38	14.59	20.35	32.25	33.09
S-GSA&GCN	13.87	17.25	23.20	34.35	36.24
GCN&T-GSA	12.21	15.50	21.20	32.92	33.47
ST-GSA	14.52	18.07	23.94	35.09	37.01
ST-GSA&GCN	**15.17**	**18.69**	**24.33**	**36.01**	**37.62**

3) THE EFFECTIVENESS OF ADDITIONAL GRAPH CONVOLUTION NETWORK

We further employ GCN and ST-GSA to capture both the global and local structure of skeleton sequences across spatial-temporal dimension. The patterns hidden in the graphs are compatible through gated fusion module. Compared with ST-GSA and ST-GSA&GCN in Table 2, the combination lifts 0.92 BLEU score and 0.61 ROUGE score. It proves that although Transformer has the advantages of global attention, it is not strong in extracting details and local features.

C. QUANTITATIVE EVALUATION

We compare our STGT with several other state-of-the-art models, including PT [1] with gaussian noise and skeletal-GSA($\mathcal{C}$) [6]. Table 3 summaries results of SLP on dataset RWTH-PHOENIX-Weather-2014T. Note that the pre-trained back-translation model in Saunders' work [1], [6] is not publicly available, we train the back-translation model based on SLT [24] by ourselves. Although the results presented in their papers are not comparable to ours, we reproduced their results as much as possible and made a relatively fair comparison in the same standard training settings.

To evaluate our method, we first reproduce the results of PT [1] and skeletal-GSA($\mathcal{C}$) [6] in Table 3. The decoder of PT is in classic Transformer structure. Although both our STGT and skeletal-GSA combine the spatial-temporal graph topology into self-attention, the performance results show that it is effective to use spatial graph self attention and temporal graph

TABLE 3. Comparison of the performance with state-of-the-art models on RWTH-PHOENIX-Weather-2014T. The results of PT [1] and Skeletal-GSA($\mathcal{C}$) [6] are reproduced with gaussian noise and translated by our back-translation model in a fair comparison.

Approach	DEV SET					TEST SET				
	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE
PT[1]	11.38	14.59	20.35	32.25	33.09	9.45	12.52	17.08	26.59	27.31
Skeletal-GSA($\mathcal{C}$)[6]	14.01	17.27	23.11	34.52	36.28	11.26	14.68	19.72	28.87	30.01
STGT($\mathcal{C}$)	14.60	18.11	23.88	35.07	37.00	12.30	15.71	20.62	29.95	31.00
STGT($\mathcal{C}\&\mathcal{M}$)	**15.17**	**18.69**	**24.33**	**36.01**	**37.62**	**12.49**	**16.06**	**21.27**	**30.45**	**31.37**

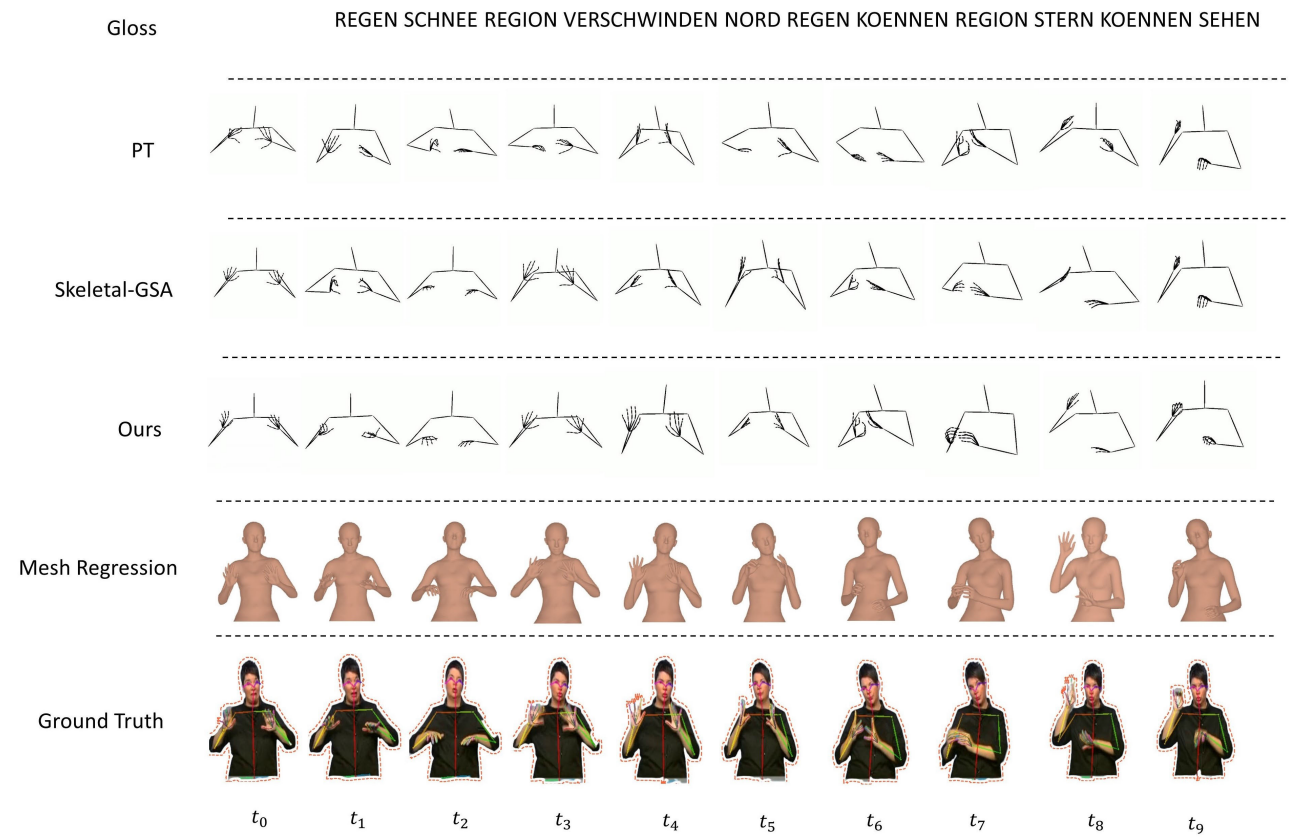


FIGURE 3. Qualitative results on DEV SET of RWTH-PHOENIX-Weather 2014 T dataset. The top row is the input glosses. The second row is the produced frames by PT [1]. The third row is the produced frames by Sleketal-GSA [6]. The fourth row is our method in STGT($\mathcal{C}\&\mathcal{M}$) configuration, we also render the generated skeleton sequence into skinned animation in the fifth row. The last row is the ground truth.

self attention on the skeleton sequence of different dimensions in turn. Our STGT($\mathcal{C}$) improves 1.08 BLEU-1 score, 0.99 ROUGE score than Sleketal-GSA($\mathcal{C}$) on TEST SET, and their graph representation only contains connectivity relationships. After combining connectivity and motion relationships in the skeleton graph, our STGT($\mathcal{C}\&\mathcal{M}$) achieves the best performance. Finally, compared with PT on TEST SET, the STGT($\mathcal{C}\&\mathcal{M}$) improves 3.86 BLEU score and 4.06 ROUGE score. Compared with Sleketal-GSA($\mathcal{C}$), the STGT($\mathcal{C}\&\mathcal{M}$) improves 1.58 BLEU score and 1.36 ROUGE score.

D. QUALITATIVE EVALUATION

In order to show the performance of STGT($\mathcal{C}\&\mathcal{M}$), we compare the generated skeleton sequences by different models both on DEV SET and TEST SET separately. To prevent errors caused by different proportions of human bones, we suggest normalization and alignment among skeletons from different signers. Due to the skeleton information being redundant, we refer to Saunders' [1] extraction method for RWTH-PHOENIX-Weather-2014T dataset. Each sign language video is processed into the corresponding skeleton sequence of 50 joints.

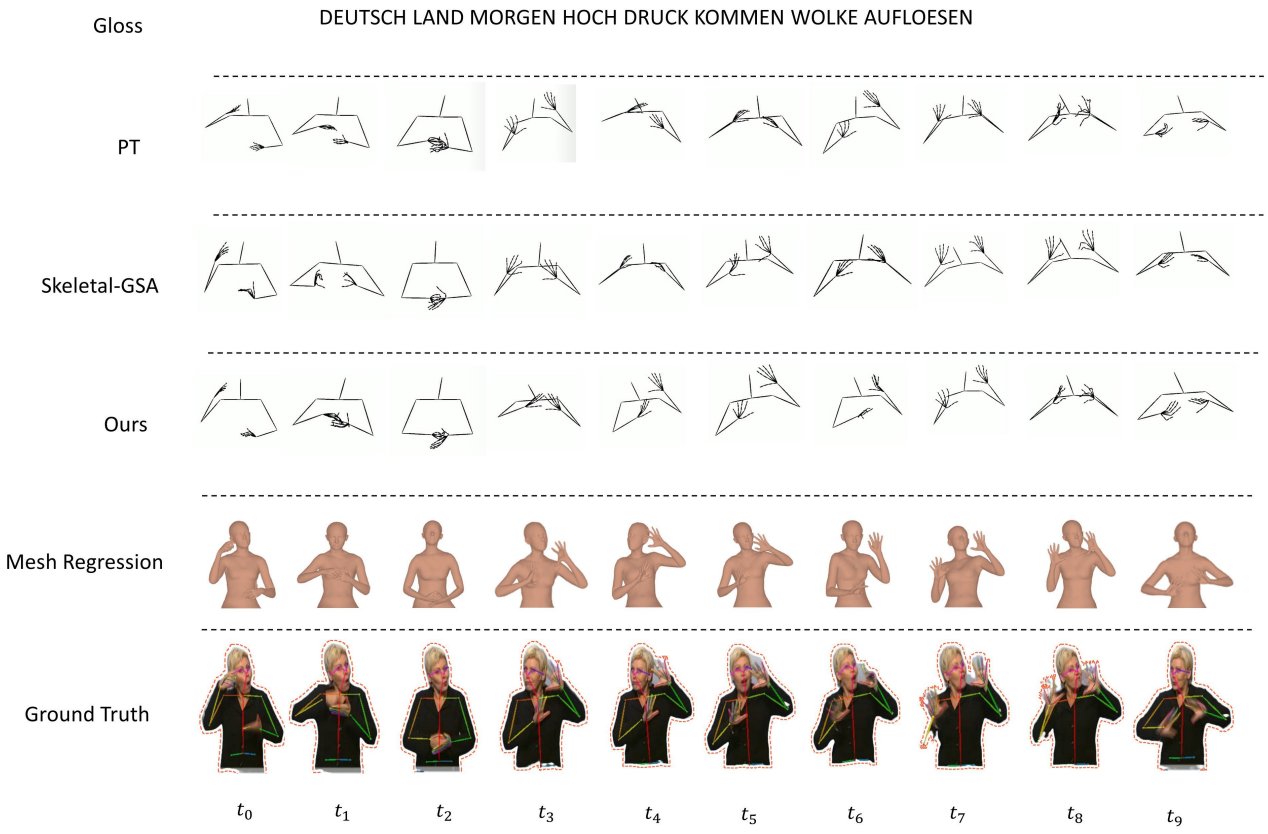


FIGURE 4. Qualitative results on TEST SET of RWTH-PHOENIX-Weather 2014 T dataset. The top row is the input glosses. The second row is the produced frames by PT [1]. The third row is the produced frames by Sleketal-GSA [6]. The fourth row is our method in STGT($\mathcal{C}\&\mathcal{M}$) configuration, we also render the generated skeleton sequence into skinned animation in the fifth row. The last row is the ground truth.

Fig. 3 and Fig. 4 are the visualization results on DEV SET and TEST SET respectively. From left to right, we sample every 10 frames of the predicted sequences for a fair comparison, where each column represents the frame generated by different models at a certain time. In comparison, the results in STGT($\mathcal{C}\&\mathcal{M}$) present more stable and accurate skeleton sequences. We also present more realistic and expressive skinned animations by the sign mesh regression module.

V. CONCLUSION

In this work, we propose a novel SLP method named STGT($\mathcal{C}\&\mathcal{M}$), which aims at producing realistic skinned sign language videos in the spatial-temporal dimensions. This is the first skinned-based SLP method which translates sign glosses into skinned animations. The spatial-temporal graph self-attention utilizes graph topology to capture the intra-frame and inter-frame correlations respectively, meanwhile, graph convolution is used to strengthen the short-term features of skeletal structure. Another significant finding is that motion relationships are important features and for the first time we use them to induct bias in the self-attention layer. The extensive experiments demonstrate the efficiency and effectiveness of STGT, which achieve the superior performance on the RWTH-PHOENIX-Weather-2014T dataset. In the future, we plan to conduct multimodal learning composed of sign poses, lip moving and head expressions in an efficient way.

ACKNOWLEDGMENT

(Zhenchao Cui and Ziang Chen contributed equally to this work.)

REFERENCES

[1] B. Saunders, N. C. Camgoz, and R. Bowden, "Progressive transformers for end-to-end sign language production," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 687–705.
[2] B. Saunders, N. C. Camgoz, and R. Bowden, "Continuous 3D multi-channel sign language production via progressive transformers and mixture density networks," *Int. J. Comput. Vis.*, vol. 129, no. 7, pp. 2113–2135, Jul. 2021.
[3] B. Saunders, N. C. Camgoz, and R. Bowden, "Mixed SIGNals: Sign language production via a mixture of motion primitives," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1919–1929.
[4] W. Huang, W. Pan, Z. Zhao, and Q. Tian, "Towards fast and high-quality sign language production," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 3172–3181.
[5] E. Hwang, J.-H. Kim, and J.-C. Park, "Non-autoregressive sign language production with Gaussian space," in *Proc. 32nd Brit. Mach. Vis. Conf. (BMVC)*, 2021, pp. 1–13. [Online]. Available: https://www.bmvc2021-virtualconference.com/assets/papers/1102.pdf
[6] B. Saunders, N. C. Camgoz, and R. Bowden, "Skeletal graph self-attention: Embedding a skeleton inductive bias into sign language production," 2021, *arXiv:2112.05277*.

[7] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.

[8] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid, "Learning joint reconstruction of hands and manipulated objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11799–11808.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[10] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," 2018, *arXiv:1801.07455*.

[11] S. Guo, Y. Lin, H. Wan, X. Li, and G. Cong, "Learning dynamics and heterogeneity of spatial-temporal graph data for traffic forecasting," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 11, pp. 5415–5428, Nov. 2022.

[12] Q. Zhang, T. Wang, M. Zhang, K. Liu, P. Shi, and H. Snoussi, "Spatial-temporal transformer for skeleton-based action recognition," in *Proc. China Autom. Congr. (CAC)*. Cham, Switzerland: Springer, Oct. 2021, pp. 694–701.

[13] V. Prakash Dwivedi and X. Bresson, "A generalization of transformer networks to graphs," 2020, *arXiv:2012.09699*.

[14] Q. Xiao, M. Qin, and Y. Yin, "Skeleton-based Chinese sign language recognition and generation for bidirectional communication between deaf and hearing people," *Neural Netw.*, vol. 125, pp. 41–55, May 2020.

[15] B. Saunders, N. Cihan Camgoz, and R. Bowden, "Adversarial training for multi-channel sign language production," 2020, *arXiv:2008.12405*.

[16] L. Ventura, A. Duarte, and X. Giró-i-Nieto, "Can everybody sign now? Exploring sign language video generation from 2D poses," 2020, *arXiv:2012.10941*.

[17] S. Stoll, N. C. Camgoz, S. Hadfield, and R. Bowden, "Text2Sign: Towards sign language production using neural machine translation and generative adversarial networks," *Int. J. Comput. Vis.*, vol. 128, no. 4, pp. 891–908, Apr. 2020.

[18] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4D scans," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 1–17, Nov. 2017.

[19] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3D hands, face, and body from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10975–10985.

[20] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 22–31.

[21] Y. Rong, T. Shiratori, and H. Joo, "FrankMocap: Fast monocular 3D hand and body motion capture by regression and integration," 2020, *arXiv:2008.08324*.

[22] J. Forster, C. Schmidt, O. Koller, M. Bellgardt, and H. Ney, "Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-weather," in *Proc. 9th Int. Conf. Lang. Resour. Eval. (LREC)*, 2014, pp. 1911–1916.

[23] W. Chen, Z. Jiang, H. Guo, and X. Ni, "Fall detection based on key points of human-skeleton using OpenPose," *Symmetry*, vol. 12, no. 5, p. 744, May 2020.

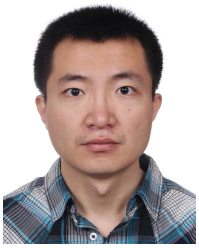
[24] N. Cihan Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10023–10033.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[26] A. Gotmare, N. S. Keskar, C. Xiong, and R. Socher, "A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation," 2018, *arXiv:1810.13243*.

ZHENCHAO CUI received the master's degree from Yanshan University, in 2010, and the Ph.D. degree from the Harbin Institute of Technology, in 2015. He is currently a Master Supervisor with the School of Cyber Security and Computer, Hebei University. He is also the Founder of the Machine Vision Engineering Research Center, Hebei. His research interests include computer vision and deep learning. He is a member of ACM and CCF.

ZIANG CHEN received the B.E. degree from the Chongqing University of Technology, in 2020. He is currently pursuing the master's degree with Hebei University. His research interests include deep learning and computer vision.



ZHAOXIN LI received the Ph.D. degree in computer application technology from the Harbin Institute of Technology, Harbin, China, in 2016. From September 2018 to March 2019, he worked as a Postdoctoral Fellow with the Department of Computing, The Hong Kong Polytechnic University. He is currently with the Institute of Computing Technology, Chinese Academy of Sciences, China. His research interests include 3D computer vision and 3D data processing.

ZHAOQI WANG received the Ph.D. degree from the State Key Laboratory of Virtual Reality, Beihang University, China, in 1999. He is currently a Professor and an Advisor of Ph.D. student's at the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include virtual reality and intelligent human–computer interaction. He is a Senior Member of the China Computer Federation.