IIAI Open Conference Publication Series
IIAI Letters on Informatics and Interdisciplinary Research
Vol.004, LIIR167
DOI: https://doi.org/10.52731/liir.v004.167

Spatial-Temporal Unfold Transformer for Skeleton-based Human Action Recognition

Hu Cui [*], Tessai Hayama [*]

Abstract

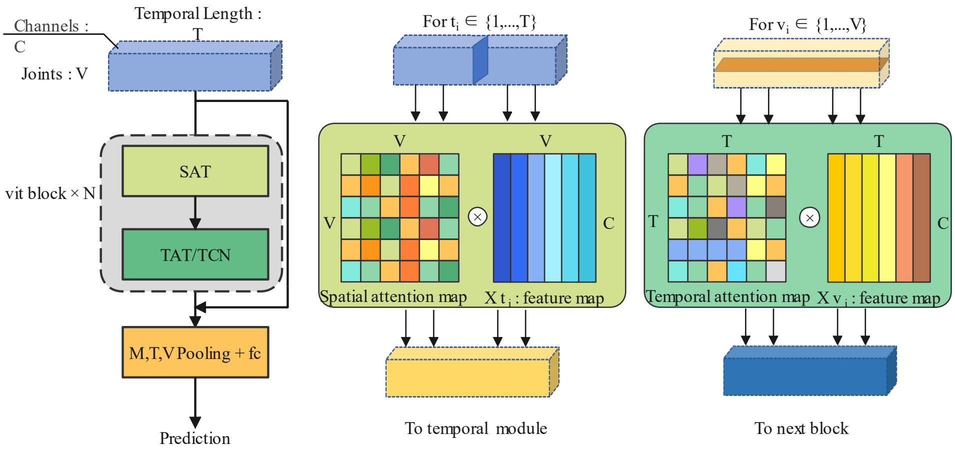
Transformer-based architecture has been proven to be effective for action and gesture recognition. In contrast to Graph Convolutional Networks (GCNs), it can automatically model joint relationships through attention mechanisms without any predefined topological graph. However, most of the previous approaches do attention to the spatial and temporal dimensions in a completely decoupled manner, ignoring the local dynamic features of the action and human body semantics. And the performance lag behind state-of-the-art GCN-based methods. To overcome the issues, we propose a Spatial-Temporal Unfold Attention Network (STUT). Firstly, it locally unfolds skeleton data in the temporal dimension such that all neighboring frames are included in each unfolded frame. Then, the human body structural semantics of actions are extracted by a hypergraph convolution used for guiding the local spatio-temporal attention operation in each unfolded frame. In addition, in order to distinguish the importance of different frames, we introduce temporal squeezing attention (TSE) for multi-scale global spatial-temporal modeling. Extensive experiments are conducted and our model achieves 96.4% on NW-UCLA and 96.91% / 94.88% on SHREC17 (14-gestures / 28-gestures).

Keywords: Action recognition, spatial temporal model, graph attention network, skeleton-based action recognition, gesture recognition.

1 Introduction

Every human action, no matter how insignificant, has a purpose. People use their hands, arms, legs, torso, body, etc. to perform actions of different meanings. They eat when they are hungry, drink when they are thirsty, run, and swim for their health. The goal of human action recognition is to apply machine vision methods to automatically detect and classify various human actions in order to analyze human activities more accurately and easily. In recent years, skeleton-based action recognition has gradually become a growing research hotspot. Compared to video action data, there are several advantages of using skeleton data for action recognition. First, the skeleton is almost a complete abstraction of the body's posture and movement information. One is able to identify the action category simply by

[*] Nagaoka University of Technology, Niigata, Japan



(a) Classic Spatial-Temporal Attention Network

(b) Spatial Attention Module

(c) Temporal Attention Module

Figure 1: The classic Spatial Temporal Transformer for skeleton-based action recognition. (a) A Transformer model consists of N vit blocks, each vit block consists of a spatial attention module (SAT) and a temporal module (TAT/TCN). (b) The spatial module performs self-attention operation across joints in every frame $t_i \in \{1,...,T\}$. (c) The temporal module learns temporal motion features by temporal attention (TAT) or temporal convolution (TCN).

observing the motion of the skeleton. Second, the skeleton data is easy to access by pose estimation algorithms [1] which is another essential technique for human behavior understanding. By utilizing pose estimation algorithms, we can effortlessly acquire an adequate dataset for action recognition and analysis. In addition, compared to RGB video, skeleton data is more robust to camera view, background, and lighting. The data set is smaller and more computationally efficient.

For skeleton-based action recognition, deep learning is widely utilized to model the spatial-temporal representation of skeletal sequences. Many different architectures of networks have been explored, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Graph Convolutional Networks (GCNs). Recently, GCNs have tended to become the de facto standard of choice for skeleton-based action recognition, which treats skeleton joints as nodes and bones as edges on a graph, applying graph theory [2, 3, 4] to represent the spatial semantics and applying temporal convolution to represent the temporal semantics of an action. To our best knowledge, ST-GCN [5] is the first study to utilize the GCN for skeleton-based human action recognition. The shortcoming is that they use a fixed predefined graph based on the natural connections of the human joints, which limits the interactions between the unconnected joints. To overcome this issue, some variants of ST-GCNs [6, 7, 8, 9, 10], use dynamically learnable graph structures to learn relationships between non-connected joints. In order to distinguish the importance of interactions between different joints, attention mechanisms are widely used.

The widespread use of attention mechanisms has motivated studies of attention-based Transformer models for skeleton-based action recognition. In contrast to GCNs, Transformer-based methods [11, 12, 13] can naturally model the relationships between local and global joints. DG-STA [11] applies graph-based spatial-temporal attention to calculates edge weights and learn joints features for gesture recognition. ST-TR [12] proposed a two-stream model which uses spatial attention and temporal convolution for the spatial stream and uses

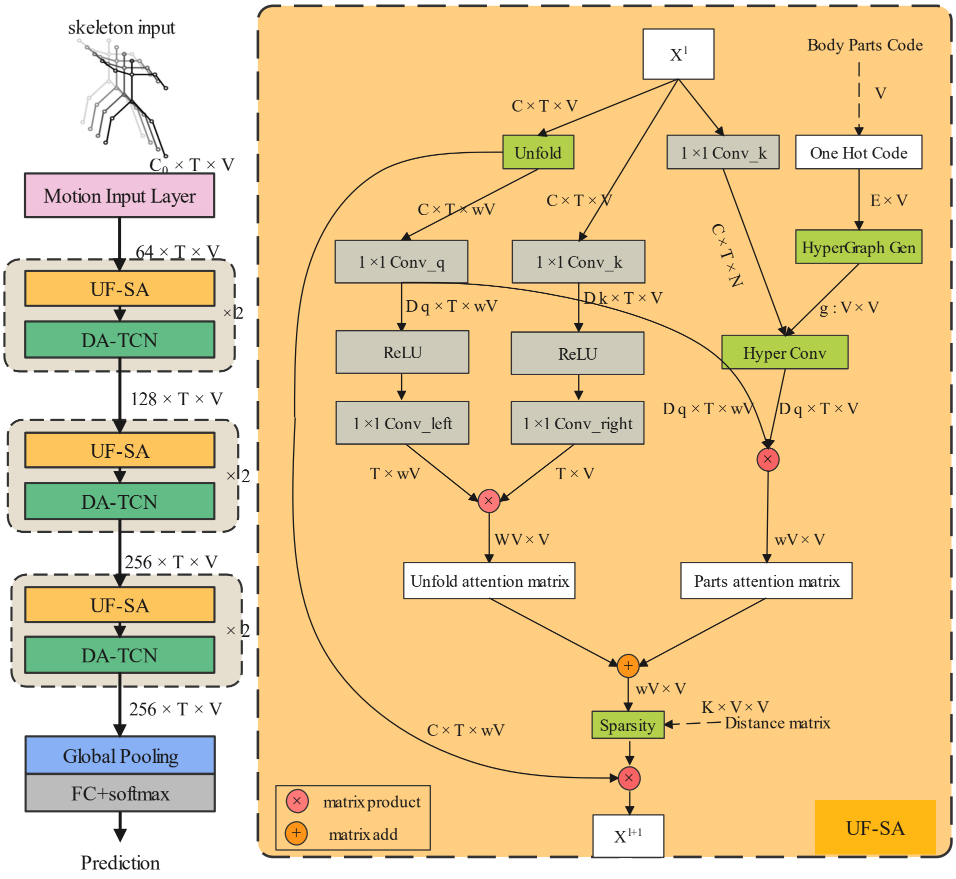


Figure 2: Architecture of Spatial Temporal Unfold Transformer.

graph convolution and temporal attention for the temporal stream. DSTA-Net [13] decoupled spatial-temporal attention into spatial attention and temporal attention for skeleton-based action recognition. The performance of these plain Transformer-based methods has shown to be far from the state-of-the-art GCN-based methods due to three issues. First, they strictly employ a paradigm that separates spatial and temporal dimensions, with each layer of the network consisting of a spatial module (SAT) and a temporal module (TAT or TCN). As shown in Fig. 1, SAT extracts intra-frame features from every single frame, and TAT/TCN aggregates the inter-frame features for different frames. This strategy of complete decoupling of the spatial and temporal dimensions has been proven to be suboptimal [14, 15]. Second, they ignored the structural prior of the skeleton data. Specifically, the permutation invariant attention operation is agnostic to the bone connectivity between human body joints [16, 17]. Third, they ignored the sparsity of the skeleton data. The available human body structure provides us with a rich source of information, and not all joint pairs interact with each other. Ignoring these redundant interactions is necessary for accurate recognition [18].

To tackle these issues, in this paper, we propose a novel variant of spatial-temporal self-attention (STSA) called Unfold Transformer (STUT) as shown in Figure 2. Compared with classic STSA (Figure 1), it extends intra-frame spatial attention to local spatio-temporal sparsity attention and introduces human body structure information to guide the information interaction between different parts of the joints. In our STUT module, each basic vit block consists of local spatial-temporal unfold sparse attention (UF-SA) and dynamic temporal attention convolution (DA-TCN) which is similar to classic architecture shown in Figure 1. Specifically, in UF-SA, firstly, We utilize a sliding window algorithm to divide a continuous frame into multiple consecutive segments which is different from STTFormer [15] division method. They block the information exchange between different segments because they use a non-overlapping division strategy. That has been proven to be not optimal [19, 20, 21]. Secondly, We compute the attention score matrix at each local segment which is possible to make different joints interact with each other between consecutive frames. To reduce the computational complexity, we use a fast query operation that allows the current joint to pay attention to all other joints in the local segment. Thirdly, inspired by SGN [22], we realize the essential role of human part semantics in action recognition, so we introduce a human partition information integration technique based on hypergraph theory [23, 4] to guide the bias of the attention operator. Finally, based on the sparse property of the skeleton data, we use relative distance masks to sparse the obtained attention score matrix to improve the model performance. In the DA-TCN module, we observe that the spatial features at different times in a sequence play different important roles in the whole action. A simple but effective SE attention layer [24] is designed for adaptively weighting frames. And then, following [25, 26, 27], we replace the vanilla TCN for temporal modeling with a multi-branch TCN.

In addition, in order to sufficiently consider the motion features of the action, we use a difference algorithm to extract the velocity variation of different joint points in the network input layer (Motion Input Layer). The velocity feature is then used as the motion representation of the action and fused with the joint coordinate feature to improve the recognition accuracy of the action.

The main contributions of this paper are as follows:

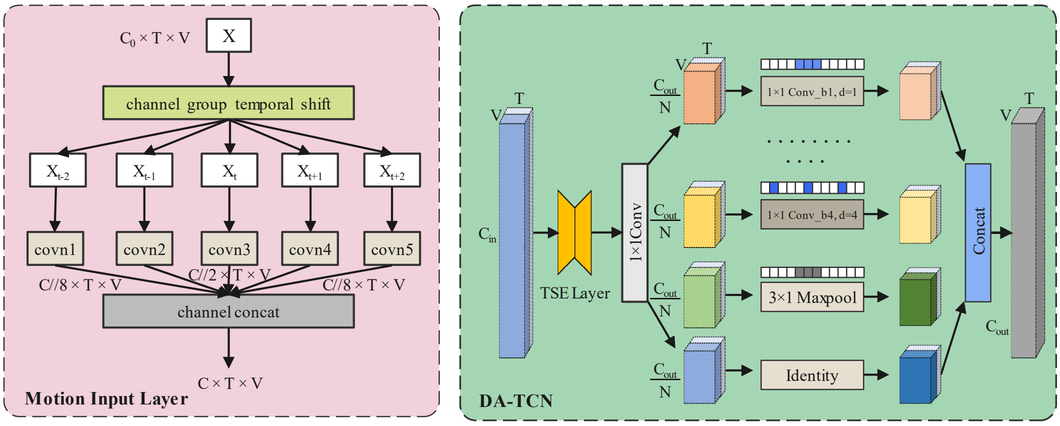
- We propose a novel spatial-temporal unfold transformer network, i.e., STUT, for skeleton-based action recognition. STUT is able to compute short-term spatial-temporal attention features in the local window through an efficient sparsity spatial-temporal attention, i.e., UF-SA. Moreover, it injects action-related human body structure priors into the model to improve performance. In the temporal module, DA-TCN is used for dynamic global temporal modeling.

- We have conducted extensive experiments on two challenging benchmarks. Ablation experiments verify our proposed method and the results achieve the existing state-of-the-art methods.

2 Related Work

Convolutional neural networks (CNNs) have achieved remarkable performance in processing Euclidean data, e.g. image classification, and object detection. However, GCNs and Transformers show better results when handling non-Euclidean data or graph-structured data, e.g. molecules, social networks, product recommendations, computer programs, etc. The skeleton-based action data consists of a number of skeleton joints that can be represented as a graph structure based on the human joint structure [5].

GCNs-based methods have shown powerful representation capabilities for skeleton-based action recognition [5, 28, 29, 27, 30]. ST-GCN [5] is an early application of GCN [2, 31] on skeleton-based action recognition which uses stacked GCN and TCN layers to process skeleton data. It treats skeleton joints as the vertices of the graph and bones as the



(a) Motion Input Layer

(b) Dynamic Temporal Attention Convolution

Figure 3: Illustration of Motion Input Layer and DA-TCN module

edges of the graph according to the human body structure, and fixes the graph topology during training. Due to the importance of topology in GCN, GCN-based methods focus on topology modeling. MS-G3D [27] introduce multi-scale graph topology into GCN for multi-scale joint relationship modeling. 2s-GCN [30] introduce an adaptive graph convolution for skeleton-based action recognition. Dynamic GCN [28] introduces context-enriched topology learning for skeleton-based action recognition. CTR-GCN [29] uses a topology refinement graph convolution to dynamically learn different topologies in different channels for skeleton-based action recognition. Moreover, DGNN [32] represents the skeleton data as a directed graph based on the kinematic dependency between the joints and bones in the natural human body, and uses message-passing neural networks (MPNN) to learn the representation for skeleton-based action recognition. InfoGCN [9] uses information bottleneck theory to guide learning the informative and compact latent representation for skeleton-base action recognition. It is worth noting that it has become the general trend among these GCN-based methods to use attention mechanisms to help learn graph topology.

Transformer-based methods, in contrast to the GCN-based methods, can non-locally learn the relationships between nodes in non-Euclidean data without any topological prior. ST-TR [12] and DSTA-Net [13] decouple the spatial and temporal attention and treat each joint as an independent token for spatial and temporal self-attention computation, which fall behind the state-of-the-art GCN-based methods due to the lack of topology prior. HGCT [33] and Hyperformer [17] try to provide prior to the model from the hypergraph and hierarchical graph perspectives, respectively. Inspired by these works, we propose a novel attention operation with body part information guidance, which is more generalized than the traditional transformer attention mechanism (DPGAT) [34, 3].

3 Method

Notations. Given the skeleton action in the form of $\mathbf{X} \in \mathbb{R}^{C\times T\times V}$, where C denotes the number of channels, T denotes the length of temporal dimension, V denotes the number of number of skeleton joints. $\mathbf{X}^{\ell}$ denotes the input of ℓ-th layer. $\mathbf{A} \in \mathbb{R}^{K\times V\times V}$ denotes the distance matrix where K is the distance, as shown in Figure 4. we use $\sigma(\cdot)$ denotes the activation, $LN(\cdot)$ denotes the normalization function. In this section, we first introduce the

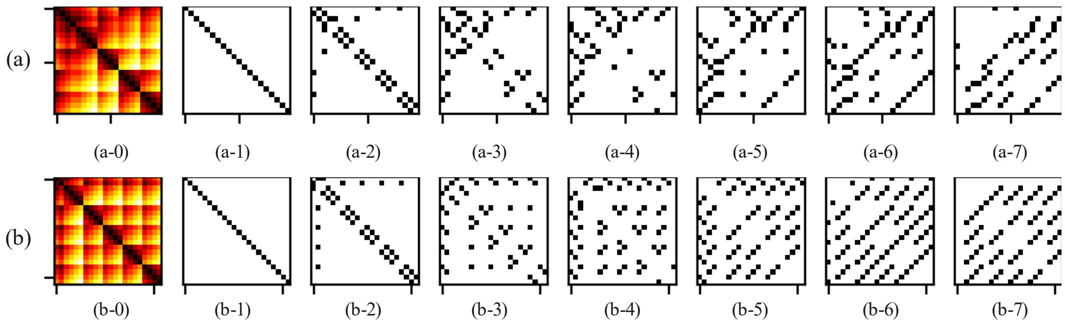
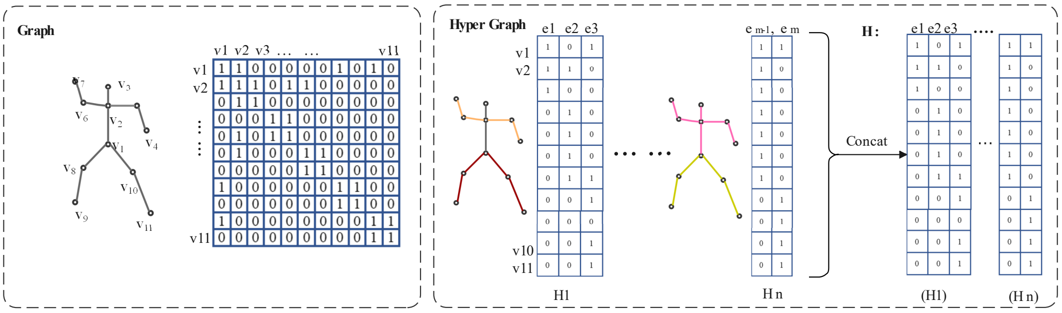


Figure 4: Visualization of the distance matrix. (a) The visualization of the relative distance matrix of the NW-UCLA dataset from 1 to 7. (b) The visualization of the relative distance matrix of the SHREC17 dataset from 1 to 7.



	v1	v2	v3	…	…						v11
v1	1	1	0	0	0	0	0	1	0	1	0
v2	1	1	1	0	1	1	0	0	0	0	0
	0	1	1	0	0	0	0	0	0	0	0
⋮	0	0	0	1	1	0	0	0	0	0	0
	0	1	0	1	1	0	0	0	0	0	0
⋮	0	1	0	0	0	1	1	0	0	0	0
	0	0	0	0	0	1	1	0	0	0	0
	1	0	0	0	0	0	0	1	1	0	0
	0	0	0	0	0	0	0	1	1	0	0
	1	0	0	0	0	0	0	0	0	1	1
v11	0	0	0	0	0	0	0	0	0	1	1

Figure 5: The comparison between Graph and Hyper Graph

overall architecture of the network, and then each module of the network is described in detail.

3.1 Model Architecture

The architecture of our proposed STUT is shown in Figure 2. Firstly, jnjecting motion features into independent appearance features has been proven to be effective in improving the accuracy of video-based action recognition [35]. Therefore, we use a difference algorithm to extract the velocity variation of different joints by a Motion Input Layer (MIL).

$$\mathbf{X}^1 = \mathcal{MIL}(\mathbf{X}) \tag{1}$$

In our STUT, the basic vit layer consists of UF-SA and DA-TCN in series. For the ℓ-th layer, it can be expressed as :

$$\hat{\mathbf{X}}^\ell = \mathcal{S}^\ell(\mathbf{X}^\ell) \tag{2}$$

$$\mathbf{X}^{\ell+1} = \mathcal{T}^\ell(\hat{\mathbf{X}}^\ell) \tag{3}$$

where $\mathcal{S}(\cdot)$ and $\mathcal{T}(\cdot)$ denotes the UF-SA module and DA-TCN module respectively. After the last vit layer, a global pooling on dimensions V and T and a fully connected layer are used to generate the final prediction logits.

Algorithm 1 Pytorch-like pseudo code of $\mathcal{MIL}$

```
def MIL(x, motion = True, g = 2):
    # x is a tensor with a shape of
    # [Batch, Channel, Time, Joints]
    B, C, T, V = x.shape
    if motion == True:
        dif1 = x[:, :, 1:] - x[:, :, 0:-1] # distance = 1
        dif2 = x[:, :, 2:] - x[:, :, 0:-2] # distance = 2
        dif3 = x[:, :, :-1] - x[:, :, 1:] # distance = -1
        dif4 = x[:, :, :-2] - x[:, :, 2:] # distance = -2
        out = torch.cat((input_map(x), diff_map1(dif1), diff_map2(dif2), diff_map3(
            dif3), diff_map4(dif4)), dim = 1) # diff_map_x is 1x1 conv output
            channels/8; input_map is 1x1 conv with output channels/8.
    else:
        out = input_map(x) # 1x1 conv output channels.
    return out
```

3.2 Motion Input Layer

As we mentioned, the utilization of motion features, e.g., optical flow, and velocity, can effectively improve the accuracy of action recognition [36, 35]. In our experiments, the feature differences between the two frames are treated as velocity features. As shown in Figure 3-a, in order to get more abundant features, we calculate the velocity features with different distances. The output of the Motion Input Layer (MIL) has half of the motion and coordinate features. The pseudo-code of MIL is presented in Algorithm 1.

3.3 Spatial Temporal Unfold Self-attention

The previous transformer-based methods strictly separate the spatial from the temporal representation process and ignore the ignored structural prior of the skeleton data (Sec.1). In contrast, as in Figure 2, our UF-SA module performs a human body partition-semantically guided local spatio-temporal representation (Eq. 2). The attention score matrix can be expressed as :

$$\mathcal{A} = \mathcal{A}^u + \mathcal{A}^h + \mathcal{A}^b \tag{4}$$

where $\mathcal{A}^b \in \mathbb{R}^{V\times V}$ is a learnable diagonal bias matrix initialized to **1**. It is essential because each physical joint plays a unique role and has different relationships with other joints. $\mathcal{A}^u \in \mathbb{R}^{wV\times V}$ and $\mathcal{A}^h \in \mathbb{R}^{wV\times V}$ denotes the local unfold spatio-temporal attention matrix and human body semantic attention matrix. we first extract sliding local blocks from input tensor $\mathbf{X}^{\ell} \in \mathbb{R}^{C\times T\times V}$ as :

$$\mathbf{u}_t = \mathbf{X}^{\ell}_{t-w/2:t+w/2} = \{x_{t-w/2},\ldots,x_t,\ldots,x_{t+w/2}\} \tag{5}$$

$$\mathbf{U} = \{\mathbf{u}_0,\mathbf{u}_1,\ldots,\mathbf{u}_t,\ldots\} \in \mathbb{R}^{C\times T\times wV}, 0 \le t \le T. \tag{6}$$

where w denotes the size of the sliding window along the temporal dimension. $\mathbf{u}_t \in \mathbb{R}^{C\times wV}$ denotes the local spatial-temporal joints patch, $x_t \in \mathbb{R}^{C\times V}$ denotes current frame, and $\mathcal{A}^u$ can be expressed as :

$$\mathcal{A}^u_{i,j} = \frac{1}{T}\sum_{t=0}^{T} \alpha(\mathbf{u}_{t,i}, x_{t,j}) \tag{7}$$

$$\alpha(\mathbf{u}_{t,i}, x_{t,j}) = \mathbf{a}^{\top} LeakyReLU(\mathbf{W}\cdot[\mathbf{u}_{t,i} \| x_{t,j}]) \tag{8}$$

where $\mathbf{W} = [\mathbf{W}_{right} \| \mathbf{W}_{left}]$ which is the parameters of $Conv_q$ and $Conv_k$ in Figure 2, $\mathbf{W}_{right} \in \mathbb{R}^{C\times d}$, $\mathbf{W}_{left} \in \mathbb{R}^{C\times d}$, $\mathbf{a} = [\mathbf{a}_{right} \| \mathbf{a}_{left}]$ which is the parameters of $Conv_{left}$ and $Conv_{right}$ in Figure 2, $\mathbf{a}_{left} \in \mathbb{R}^{d\times 1}$, $\mathbf{a}_{right} \in \mathbb{R}^{d\times 1}$.

In order injects action-related human body structure priors into the attention operation, inspired by HyperGraph[23, 4], as shown in Figure 5, we use hyperedges to represent the individual parts of the human body. Compared to the traditional graph structure, the hypergraph can be represented as $\mathcal{H} = (\mathcal{V}, \mathcal{E})$, which includes a vertex set $\mathcal{V}$ and hyperedge set $\mathcal{E}$, $V = |\mathcal{V}|$ denotes the number of joints and $E = |\mathcal{E}|$ denotes the number of human body partitions. The hyperedge is beyond pairwise connections among data which can encode high-order data correlation of the joints in every partition. The matrix $\mathbf{H}$ of hypergraph $\mathcal{H}$ with shape $V \times E$ can be defined as:

$$h(v,e) = \begin{cases} 1, & if \quad v \in e \\ 0, & if \quad v \notin e \end{cases} \tag{9}$$

For every joint $v \in \mathcal{V}$, its degree can be defined as $d(v) = \sum_{e \in \mathcal{E}} h(v,e)$, for every hyperedge $e \in \mathcal{E}$, its degree can be defined as $\delta(e) = \sum_{v \in \mathcal{V}} h(v,e)$. Further, we use $\mathbf{D}_v$ and $\mathbf{D}_e$ to denote the diagonal matrices of the vertex degrees and hyperedge degrees, respectively. Then we define our hyperedge convolution according HyperGNN[23] as :

$$\overline{\mathbf{X}}^{\ell} = \mathbf{D}_v^{-1/2} \mathbf{H} \mathbf{D}_e^{-1} \mathbf{H}^{\top} \mathbf{D}_v^{1/2} \mathbf{X}^{\ell} \mathbf{W}_h \tag{10}$$

where $\mathbf{W}_h$ is the weight matrix of HyperConv in Figure 2, $\overline{\mathbf{X}}^{\ell} \in \mathbb{R}^{C \times T \times V}$ is the semantic features with human body partitions which are generated based on the specified input action. Then the semantic attention matrix can be expressed as :

$$\mathcal{A}^h = \mathbf{U}^{\top} \star \overline{\mathbf{X}}^{\ell} \tag{11}$$

where $\star$ denotes matrix product as shown in Figure 2. And then the $\mathcal{S}^{\ell}(\cdot)$ in Equation 2 can expressed :

$$\mathcal{S}^{\ell}(\mathbf{X}^{\ell}) = \mathbf{U} \star \mathcal{SP}(\mathcal{A}) \tag{12}$$

$$\mathcal{SP}(\mathcal{A}_{i,j}) = \begin{cases} softmax(\mathcal{A})_{i,j}, & if \quad M_k[i,j] > 0, \quad 0 < k < K \\ \omega, & if \quad M_k[i,j] = 0 \end{cases} \tag{13}$$

where $\mathcal{SP}(\cdot)$ denotes the sparse functions which use the distance matrix $\mathbf{M}$ shown in Figure 4 to sparse attention matrix, K denotes the relative distance of joints, $\omega = -9e15$ in our experiments.

3.4 Dynamic Temporal Attention Convolution

Squeeze-and-Excitation (SE) [24] was originally used for modeling interdependencies between channels of image features. It has been proven to be a simple and effective mechanism for channeling attention. In order to adaptively distinguish the importance of different frames in the action, we adapt it to the temporal dimension of the action feature. As shown in Figure 3-b, a multi-group TCN consists of multiple dilated convolution branches with different dilation d. Each branch has a different temporal receptive field, and the temporal modeling capability is significantly improved compared to the vanilla TCN in ST-GCN [5] with fixed temporal kernel size.

4 Experiments

4.1 Datasets and Setings

Northwestern-UCLA (NW-UCLA) dataset [37] contains 1494 action samples with 10 categories which are captured by three Kinect cameras from multiple viewpoints. We follow the same evaluation method in [37]. **SHREC'17 Track** (SHREC17) dataset [38] contains 2800 gesture samples performed by one finger or the whole hand performed by 28 persons. We follow the same evaluation protocol in [10].

All experiments are conducted on one NVIDIA RTX A6000 GPU and Pytorch-1.12 platform. We use SDG to train our STUT model and we use a five epochs warm-up to make the training more stable. The base learning rate is 0.1, momentum is 0.9, weight decay is 0.0004, batch size is 32, and training epoch is 130, let the learning rate decay with a factor of 0.1 at epochs 50 and 80. All random seeds were set to 1 in all experiments, without using any data augmentation technique.

4.2 Ablation Study

To analyze the effects of the components of our STUT, we examined the accuracy of the different configurations of our model. All ablation studies are performed on the NW-UCLA dataset (bone modality).

4.2.1 Study on model structure

We test the effectiveness of the motion input layer (MIL), unfold operation (UF), and sparsity operation (SP) in our model. As shown in Tabel 1, the baseline model is a classic spatial-temporal attention model as shown in Figure 1. We equip it with our modules step by step to verify their effectiveness. Firstly, when equipped with MIL, the baseline model obtains 8.19% performance boost with a slight number of parameters. Then, when continuing to equip the UF and SP modules, the model is boosted by 3.27% and 0.21% respectively. It demonstrates the non-negligible role of local motion features in action modeling.

In Tabel 2, we verify the components of spatial-temporal unfold self-attention (section 3.3). Firstly, we confirm the validity of positional encoding (PE) [34] in our model. It improves the performance by 3.04%. Secondly, we can find our body partition attention (PA) generated by our HyperConv in equation 3.3 can continue to enhance the result to 94.40% (1.09%). This means that PA and PE are completely different. PA provides body semantic information related to specific actions, while PE provides the model with positional information, they are complementary. Then, we replace the DPGAT with our unfold self-attention in equation 3.3, the performance is improved by 0.2% which denotes DPGAT is weaker than ours, that is consistent with GATv2 [3].

The temporal module (DA-TCN) consists of TSE Layer and multi-group TCN. As shown in Tabel 3, compared to TCN [5], DA-TCN improves the performance to 90.95% (1.72%). It suggests that adaptive receptive fields are necessary for different actions. We also observe that TSE improves the TCN and DA-TCN to 92.89% and 94.40%, respectively. This suggests that some frames in the action are non-critical and some are critical, and it is necessary to distinguish between them. In Tabel 4, We further test the effect of different squeeze rates in the TSE layer.

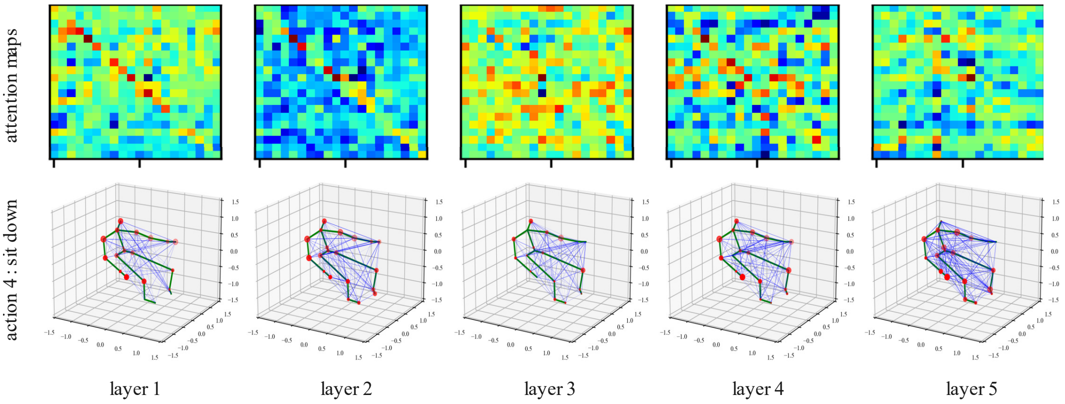


Figure 6: Visualization of attention maps and topologies.

Table 1: The effectiveness of Motion Input Layer (MIP), Temporal Unfold Operation(UF) and Sparsity for Attention Martix (SP).

Model	Parameters	Acc(%)
baseline	1.55M	81.03
MIL	1.73M	89.22
MIL + UF	1.73M	92.46
MIL + UF + SP	1.73M	92.67

Table 2: The effectiveness of Positional Encoding and Parts Attention for traditional transformer attention (DPGAT) and our unfold self-attention (UF-SA).

Model	Parameters	Acc(%)
W/o PE	1.73M	89.65
PE	1.73M	92.67
PA	1.75M	93.31
PE + PA	1.75M	94.40
PE + PA + UF-SA	1.76M	94.60

Table 3: Ablations on NW-UCLA for DA-TCN module.

Model	Acc(%)
TCN	89.23
DA-TCN	90.95
TCN + TSE	92.89
DA-TCN + TSE	94.40

Table 4: Ablations on NW-UCLA for different squeeze rates of temporal squeeze layer (TSE Layer).

squeeze rate	Acc(%)
2	93.31
4	94.61
6	92.02
8	93.97
10	92.24

Table 5: Ablations on NW-UCLA for attention sparsity with different distance r.

r	Acc(%)
2	92.03
4	94.61
8	92.67
10*	91.16

Table 6: Ablations on NW-UCLA for different modalities

Modality	Acc(%)
Joint	93.75
Joint Motion	90.09
Bone	94.61
Bone Motion	88.58
J+B+M+BM	95.48
J+B+M	96.42

As we described in Section 1, we use relative distance masks to sparse the obtained attention score matrix in order to improve the model. The distance matrix is shown in Figure 4. As shown in Table 5, it gets the best accuracy when we use the first 4 relative distance

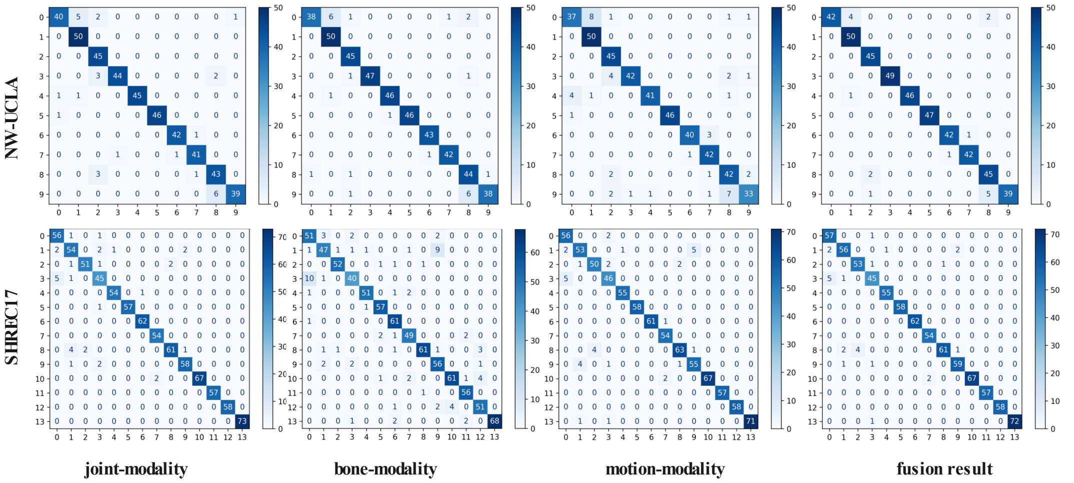


Figure 7: Visualization of confusion matrices for different modalities.

matrices, while the worst results are achieved when no sparse operation is performed. This suggests sparsity of the skeleton data cannot be ignored.

Following [30, 9], we train our model on joints, bones, joint motion, and bone motion modalities, then fusion the four streams. As shown in Tabel 6, we find that our model only needs to fuse joint, bone, and motion modes for the best results. The visualization of confusion matrices is shown in Figure 7.

4.2.2 *Visualization of attention matrix*

As shown in Figure 6, we visualize the first five attention maps of action *sit down*. In the first two layers, the joints pay more attention to their own local windows. However, in the last three layers, the attention of the nodes is gradually discrete. This indicates that the hidden topological information related to the action has been learned. It is more clearly visible from the 3D visualization diagram, green lines indicate natural connections, blue lines indicate learned hidden interactions, and blue lines are denser in the latter three layers.

4.3 Comparisons with the State-of-the-Art

We compare our results with the previous state-of-the-art method on NW-UCLA and SHREC17 in Table 7. On NW-UCLA, our results are very comparable to Info-GCN [9] (96.4% vs 96.6%), while it is worth noting that Info-GCN relies on two additional MMD losses. On the SHREC17 14-Gestures dataset, our approach is comparable to DSTA-Net [13] (96.91% vs. 97.0%) which fusion four streams, namely, spatial-temporal stream (original data), spatial stream, fast-temporal stream and slow-temporal stream. In 28-Gestures, our model also achieves state-of-the-art performance.

5 Conclusion

In this paper, we propose a novel spatial-temporal unfold attention network (STUT) for skeleton-based action recognition. In order to adequately extract the local motion information, it expands classic intra-frame spatial attention to local spatial-temporal sparsity attention (UF-SA) and applies hypergraph convolution to extract body semantic features to

Table 7: Action classification performance on the NW-UCLA and SHREC17.

Method	NW-UCLA(%)	SHREC17(%)	
		14Gesture	28Gesture
ST-GCN[5]	-	92.7	87.7
ST-TS-HGR-Net[39]	-	94.3	89.4
Shift-GCN[6]	94.6	-	-
FGCN[7]	95.5	-	-
HG-GCN[8]	-	92.8	88.3
InfoGCN[9]	**96.6**	-	-
MS-ISTGCN[10]	-	96.6	**94.8**
Ta-CNN[40]	96.1	-	-
DD-Net[41]	-	94.6	91.9
DSTA-Net[13]	-	**97.0**	93.9
DG-STA[11]	-	94.4	90.7
STUT(joint)	93.8	96.07	93.69
STUT(j+b+jm)	**96.4**	**96.91**	**94.88**

guide the information interaction. In addition, we use the difference algorithm to extract the velocity variation as the motion input to the network for a more comprehensive understanding of human actions and gestures. To verify the effectiveness of STUT, extensive experiments are conducted and our model achieves state-of-the-art performance.

References

[1] Y. Chen, Y. Tian, and M. He, “Monocular human pose estimation: A survey of deep learning-based methods,” *Computer vision and image understanding*, vol. 192, p. 102897, 2020.

[2] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” *arXiv preprint arXiv:1609.02907*, 2016.

[3] S. Brody, U. Alon, and E. Yahav, “How attentive are graph attention networks?” *arXiv preprint arXiv:2105.14491*, 2021.

[4] S. Bai, F. Zhang, and P. H. Torr, “Hypergraph convolution and hypergraph attention,” *Pattern Recognition*, vol. 110, p. 107637, 2021.

[5] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[6] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, “Skeleton-based action recognition with shift graph convolutional network,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 183–192.

[7] H. Yang, D. Yan, L. Zhang, Y. Sun, D. Li, and S. J. Maybank, “Feedback graph convolutional network for skeleton-based action recognition,” *IEEE Transactions on Image Processing*, vol. 31, pp. 164–175, 2021.

[8] Y. Li, Z. He, X. Ye, Z. He, and K. Han, “Spatial temporal graph convolutional networks for skeleton-based dynamic hand gesture recognition,” *EURASIP Journal on Image and Video Processing*, vol. 2019, no. 1, pp. 1–7, 2019.

[9] H.-g. Chi, M. H. Ha, S. Chi, S. W. Lee, Q. Huang, and K. Ramani, “Infogcn: Representation learning for human skeleton-based action recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 186–20 196.

[10] J.-H. Song, K. Kong, and S.-J. Kang, “Dynamic hand gesture recognition using improved spatio-temporal graph convolutional network,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 6227–6239, 2022.

[11] Y. Chen, L. Zhao, X. Peng, J. Yuan, and D. N. Metaxas, “Construct dynamic graphs for hand gesture recognition via spatial-temporal attention,” in *Proc. Brit. Mach. Vis. Conf*, 2019, pp. 1–13.

[12] C. Plizzari, M. Cannici, and M. Matteucci, “Spatial temporal transformer network for skeleton-based action recognition,” in *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part III.* Springer, 2021, pp. 694–701.

[13] L. Shi, Y. Zhang, J. Cheng, and H. Lu, “Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition,” in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[14] X. Qin, R. Cai, J. Yu, C. He, and X. Zhang, “An efficient self-attention network for skeleton-based action recognition,” *Scientific Reports*, vol. 12, no. 1, p. 4111, 2022.

[15] H. Qiu, B. Hou, B. Ren, and X. Zhang, “Spatio-temporal tuples transformer for skeleton-based action recognition,” *arXiv preprint arXiv:2201.02849*, 2022.

[16] L. Rampášek, M. Galkin, V. P. Dwivedi, A. T. Luu, G. Wolf, and D. Beaini, “Recipe for a general, powerful, scalable graph transformer,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 14 501–14 515, 2022.

[17] Y. Zhou, C. Li, Z.-Q. Cheng, Y. Geng, X. Xie, and M. Keuper, “Hypergraph transformer for skeleton-based action recognition,” *arXiv preprint arXiv:2211.09590*, 2022.

[18] V. P. Dwivedi and X. Bresson, “A generalization of transformer networks to graphs,” *arXiv preprint arXiv:2012.09699*, 2020.

[19] A. Hatamizadeh, G. Heinrich, H. Yin, A. Tao, J. M. Alvarez, J. Kautz, and P. Molchanov, “Fastervit: Fast vision transformers with hierarchical attention,” *arXiv preprint arXiv:2306.06189*, 2023.

[20] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[21] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 009–12 019.

[22] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng, "Semantics-guided neural networks for efficient skeleton-based human action recognition," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1112–1121.

[23] Y. Feng, H. You, Z. Zhang, R. Ji, and Y. Gao, "Hypergraph neural networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 3558–3565.

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[25] Z. Chen, S. Li, B. Yang, Q. Li, and H. Liu, "Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 2, 2021, pp. 1113–1122.

[26] H. Duan, J. Wang, K. Chen, and D. Lin, "Dg-stgcn: dynamic spatial-temporal modeling for skeleton-based action recognition," *arXiv preprint arXiv:2210.05895*, 2022.

[27] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 143–152.

[28] F. Ye, S. Pu, Q. Zhong, C. Li, D. Xie, and H. Tang, "Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 55–63.

[29] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 13 359–13 368.

[30] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 026–12 035.

[31] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[32] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7912–7921.

[33] R. Bai, M. Li, B. Meng, F. Li, M. Jiang, J. Ren, and D. Sun, "Hierarchical graph convolutional skeleton transformer for action recognition," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 2022, pp. 01–06.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[35] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.

[36] L. Sevilla-Lara, Y. Liao, F. Güney, V. Jampani, A. Geiger, and M. J. Black, "On the integration of optical flow and action recognition," in *Pattern Recognition: 40th German Conference, GCPR 2018, Stuttgart, Germany, October 9-12, 2018, Proceedings 40.* Springer, 2019, pp. 281–297.

[37] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning and recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2649–2656.

[38] Q. D. Smedt, H. Wannous, J.-P. Vandeborre, J. Guerry, B. L. Saux, and D. Filliat, "3D Hand Gesture Recognition Using a Depth and Skeletal Dataset," in *Eurographics Workshop on 3D Object Retrieval*, I. Pratikakis, F. Dupont, and M. Ovsjanikov, Eds. The Eurographics Association, 2017.

[39] X. S. Nguyen, L. Brun, O. Lézoray, and S. Bougleux, "A neural network based on spd manifold learning for skeleton-based hand gesture recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 036–12 045.

[40] K. Xu, F. Ye, Q. Zhong, and D. Xie, "Topology-aware convolutional neural network for efficient skeleton-based action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2866–2874.

[41] F. Yang, Y. Wu, S. Sakti, and S. Nakamura, "Make skeleton-based action recognition model smaller, faster and better," in *Proceedings of the ACM multimedia asia*, 2019, pp. 1–6.