Local Shuffled Skeleton Position Embedding Vision Transformer for Human Activity Recognition

Zihui Yan Z.YAN@LBORO.AC.UK
Xiyu Shi X.SHI@LBORO.AC.UK
Varuna De Silva V.D.DE-SILVA@LBORO.AC.UK
Institute for Digital Technologies, Loughborough University London, 3 Lesney Avenue, The Broadcast Centre, Here East, Queen Elizabeth Olympic Park, London, E20 3BS

Editors: Hung-yi Lee and Tongliang Liu

Abstract

Vision Transformers (ViTs) in human activity recognition tasks suffer from inadequate spatial modeling through conventional position embeddings, leading to over-reliance on fixed positional information. This paper proposes Shuffled Positional Embedding (SPE), a mechanism that randomly disrupts the order of positional encoding during each forward propagation, reducing model dependence on position embedding and encouraging exploration of intrinsic spatial relationships. While SPE enhances general spatial awareness, it lacks targeted guidance for human-centric modeling. To address this limitation, Local Shuffled Skeleton Position Embedding (LSSPE) is developed, which leverages 2D skeleton data to provide human body structure-aware spatial representation. LSSPE computes attention weights based on spatial distances between image patches and skeleton keypoints, incorporating joint motion amplitudes for enhanced modeling. To further utilize skeleton data, a dual-stream architecture is designed combining TimeSFormer with LSSPE (LSSPE-TimeSFormer) for RGB processing and SkateFormer for skeleton processing. The proposed dual-stream model achieves outstanding performance of 95.8% and 98.7% accuracy on NTU RGB+D cross-subject and cross-view settings, establishing the effectiveness of skeleton-aware position embedding for human activity recognition.

Keywords: Shuffled position embedding, Vision transformer, Human activity recognition.

1. Introduction

Human Activity Recognition (HAR) has emerged as a fundamental task in computer vision, with applications spanning security surveillance, human-computer interaction, and medical monitoring (Karim et al., 2024). Recent advances in Vision Transformers (ViTs) have demonstrated superior performance in capturing long-range dependencies through self-attention mechanisms (Dosovitskiy et al., 2020). However, their application to HAR reveals fundamental limitations in spatial modeling capabilities.

A critical bottleneck lies in the positional embedding layer, which compensates for the permutation invariance of self-attention operations. Conventional approaches employ absolute or relative position coding, assuming fixed grid-like arrangements—an assumption that proves suboptimal for dynamic human activities. These general-purpose position embeddings demonstrate three key limitations in HAR: (1) lack of explicit human body structure modeling, (2) requirement for substantial datasets due to their generic nature, and (3) fail-

ure to capture hierarchical articulated relationships in human body parts (Souza Leite et al., 2024).

Through systematic experimentation, we discovered a counterintuitive phenomenon: Shuffled Position Embedding (SPE) during training consistently outperforms standard position embeddings on activity recognition benchmarks. This performance enhancement becomes more pronounced with smaller training datasets, suggesting that conventional embeddings may create harmful inductive biases that limit generalization across diverse human poses.

Motivated by these findings, we propose Local Shuffled Skeleton Position Embedding (LSSPE), which leverages two-dimensional skeleton information to compute attention weights for image patches based on their proximity to skeletal joints. Unlike previous approaches that treat skeleton data as auxiliary information, LSSPE fundamentally reimagines position embedding as an explicit data fusion mechanism, directly incorporating human body structure into spatial representation learning.

To fully exploit the complementary nature of RGB and skeleton data, we design a dual-stream architecture combining TimeSFormer with LSSPE for RGB processing and SkateFormer for skeleton processing.

The main contributions of this work are as follows:

1. Discovery and analysis of the counterintuitive benefits of shuffled position embeddings in transformer-based activity recognition.

2. Introduction of LSSPE, a novel position embedding mechanism that integrates skeletal structure and motion information, serving as an explicit data fusion tool.

3. Design of a dual-stream architecture achieving state-of-the-art performance on multiple HAR benchmarks.

2. Related work

This section reviews the relevant literature across two key areas that inform the proposed approach. The review first examines the evolution of human activity recognition methods, covering RGB-based approaches, skeleton-based methods, and multimodal fusion strategies that combine both modalities. The analysis then focuses on position embedding techniques in Vision Transformers, with particular attention to their limitations in human-centric tasks and recent advances in HAR-specific position embeddings.

2.1. Human Activity Recognition

2.1.1. RGB-based Methods

Deep learning approaches for RGB-based HAR have evolved from 2D Convolutional Neural Networks (CNNs) (Karpathy et al., 2014) to 3D CNNs (Tran et al., 2015), which incorporate temporal dimensions but significantly increase computational costs (Carreira and Zisserman, 2017). Two-stream networks (Simonyan and Zisserman, 2014) processed spatial and temporal features separately but relied on computationally intensive optical flow (Feichtenhofer et al., 2016). Recurrent Neural Networks (RNNs) and Long Short-term Memory

Networks (LSTMs) improved temporal modeling but struggled with vanishing gradients in long sequences (Donahue et al., 2015; Yue-Hei Ng et al., 2015).

Recently, transformer architectures have shown remarkable promise in HAR through self-attention mechanisms. TimeSFormer (Bertasius et al., 2021) introduced divided space-time attention for video understanding, while Video Swin Transformer (Liu et al., 2022) achieved superior performance through hierarchical feature learning. However, these approaches primarily focus on temporal modeling while treating spatial relationships through conventional position embeddings, limiting their effectiveness for human-centric tasks.

2.1.2. Skeleton-based Methods

Skeleton data provides a structured representation of human pose that directly reflects activity patterns and can be obtained from RGB videos, depth maps, or motion capture devices (Sarker et al., 2021). Traditional approaches utilized hand-crafted features and conventional machine learning techniques, but recent transformer-based methods have demonstrated significant improvements. Notable approaches include attention-enhanced methods using MMPose-extracted skeletons (Qin et al., 2024), DualMotion combining skeletons from RGB and motion capture devices (Xu et al., 2021), and graph transformer architectures for enhanced spatial-temporal modeling (Zhang et al., 2023).

2.1.3. Multimodal Fusion Methods

Fusion methods combine RGB and skeleton modalities to leverage their complementary strengths. Early approaches employed feature-level or late fusion strategies due to significant distribution differences between modalities (Zhao et al., 2017; Zolfaghari et al., 2017). More sophisticated methods include Video Pose Networks (Das et al., 2020, 2021) and multimodal networks (Bruce et al., 2022) that perform feature-level fusion. Recent transformer-based fusion approaches include 3D deformable transformers (Kim et al., 2023), dual-stream transformers (Shi et al., 2023), and STAR-Transformer (Ahn et al., 2023). However, these methods typically treat fusion as a post-processing step rather than integrating multimodal information at the fundamental representation level.

2.2. Position Embedding in Vision Transformers

Position embedding serves as a crucial component in transformer architectures, compensating for the permutation invariance of self-attention mechanisms. Early approaches in NLP employed sinusoidal encodings (Vaswani et al., 2017), while vision transformers typically utilize learnable absolute position embeddings (Dosovitskiy et al., 2020). Relative position encoding methods (Shaw et al., 2018; Dai et al., 2019) have shown advantages in certain scenarios by modeling pairwise relationships between positions.

For video understanding, specialized position embedding strategies have been developed. TimeSFormer (Bertasius et al., 2021) employs separate spatial and temporal position embeddings with divided attention mechanisms. Video Swin Transformer (Liu et al., 2022) extends relative position bias to three dimensions, incorporating both spatial and temporal relationships. However, these general-purpose approaches do not account for the specific structural properties of human activities.

Recent work has begun exploring domain-specific position embeddings for human activity recognition. Souza Leite et al. (2024) demonstrated that relative position embeddings do not always outperform absolute encodings in HAR tasks, highlighting the need for task-specific designs. Zhou et al. (2024) explored concatenation-based position embedding methods as alternatives to additive approaches. However, these works primarily focus on general improvements rather than incorporating human body structure information.

The integration of pose information into video transformers has gained attention recently. Duan et al. (2022) proposed pose-guided attention mechanisms for action recognition, while other works have explored skeleton-aware feature learning (Qin et al., 2024). These approaches typically treat pose information as auxiliary features rather than fundamental components of spatial representation.

Unlike existing approaches that treat position embedding and multimodal fusion as separate problems, our LSSPE method uniquely combines skeletal information with local shuffling strategies to create a unified solution. LSSPE serves as both an improved position embedding mechanism and an explicit data fusion tool, directly addressing the limitations of conventional spatial encoding in human activity recognition while providing a principled approach to RGB-skeleton integration at the representation level.

3. Proposed methods

This section presents the proposed methods for improving human activity recognition through novel position embedding strategies. First, SPE is introduced to address the over-reliance on fixed positional information in standard Vision Transformers. Building upon the insights from SPE, LSSPE is developed, which incorporates human body structure information to provide more targeted spatial representations. Finally, a dual-stream architecture is presented that combines the strengths of both RGB and skeleton modalities to achieve comprehensive human activity recognition. The progression from general position embedding improvements to human-specific enhancements and finally to multimodal fusion demonstrates a systematic approach to addressing the limitations of current methods.

3.1. Shuffled position embedding

Position embedding is a key component used to obtain position information in ViT. The mechanisms of position embedding and the comparative effectiveness of different embedding types remain active research topics. Relative position embedding is, to some extent, a human-assisted inductive bias on position information, which should be able to help the model converge faster. However, this is not always the case (Souza Leite et al., 2024). Some work has pointed out that relative position embedding is sometimes not more effective than absolute position coding. Therefore, the potential of position embedding warrants further exploration.

A Shuffled Positional Embedding (SPE) mechanism is introduced to improve the standard Vision Transformer (ViT) model. Unlike the fixed-order learnable positional encoding in ViT, to enhance the robustness of the model to spatial position variations, SPE employs a learnable positional embedding but randomly permutes its order during each forward propagation as shown in (1):

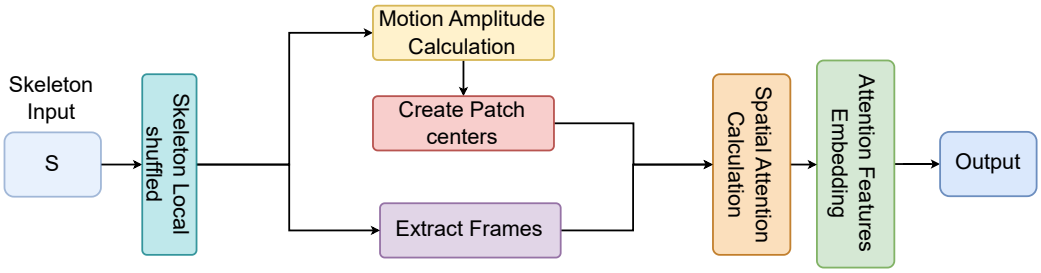

Figure 1: Proposed distance-based skeleton attention

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \cdots; \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{\text{pos}}^{\pi}, \tag{1}$$

where $\mathbf{z}_0$ represents the input representation after adding positional encoding, $\mathbf{x}_{\text{class}}$ is the classification token , $\mathbf{x}_p^i$ is the i-th image patch , $\mathbf{E}$ is the patch embedding matrix , $\mathbf{E}_{\text{pos}}^{\pi}$ is the position encoding matrix after random permutation π , $\pi \in S_N$ is a random permutation of $\{1, 2, \ldots, N\}$.

We suggest that artificially shuffling the position embedding prevents the model from over-relying on the position information provided by the position embedding. This can motivate the model to explore the positional connections that exist within each patch. Wang et al. (2024) proposed a similar shuffled position embedding, but primarily for knee osteoarthritis.

3.2. Local Shuffled Skeleton Position Embedding

The local shuffle skeleton position embedding (LSSPE) was developed from the findings in Section 3.1. Position embedding has more potential than previously recognized, motivating the fusion of additional information for ViT embedding. For human activity recognition, the skeleton is a very intuitive and important modality. It is typically represented as coordinate data (e.g., joint positions), which is significantly more lightweight than RGB frames and well-suited for computational embedding.

To enhance robustness while preserving structural information, a local shuffling mechanism that randomly permutes joint positions within each body part is introduced. Unlike global shuffling, which would destroy anatomical structure, LSSPE maintains the overall body configuration while creating variations that help the model generalize better. Specifically, the skeleton is divided into anatomically significant parts (e.g. arms, legs, torso, head) and performs random permutations only within these local regions. This preserves the relative positioning between body parts while introducing beneficial noise at the joint level. Experiments show that this controlled shuffling significantly improves model performance compared to both standard position embeddings and completely random shuffling approaches.

The LSSPE mechanism consists of two main components: motion amplitude calculation and spatial attention calculation, as illustrated in Figure 1.

Motion Amplitude Calculation: The motion amplitude for each skeleton joint is first computed by measuring the displacement between consecutive frames. This captures the temporal dynamics of human movement and helps the model focus on more active body parts. The motion amplitude for joint j, m_j, is calculated as (2):

$$m_j = \sum_{t=1}^{T-1} |p_{j,t+1} - p_{j,t}|, \tag{2}$$

where $p_{j,t}$ represents the position of joint j at time t, and T is the sequence length. The motion amplitudes are then normalized to ensure balanced contribution across different joints.

Spatial Attention Calculation: A distance-based skeleton attention mechanism is proposed that allocates attention weights by computing spatial distances between image patches (geometric centre) and skeleton keypoints. The core idea is to use a Gaussian kernel function to assign higher attention weights to image patches that are closer to skeleton joints. The attention weight between patch i and joint j, w_{ij}, and the normalized attention weights $\hat{w}_{ij}$ are calculated as (3):

$$w_{ij} = \exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right), \hat{w}_{ij} = \frac{w_{ij} \cdot (\hat{m}_j + 0.2) - \min_j(w'_{ij})}{\max_j(w'_{ij}) - \min_j(w'_{ij}) + \epsilon}, \tag{3}$$

where d_{ij} is the Euclidean distance from image patch i to skeleton joint j, and σ is an adaptive parameter that adjusts based on both skeleton size and input image dimensions to ensure appropriate attention spread. The attention weights are then refined by incorporating motion amplitude information, giving higher priority to more active joints. After normalization to prevent dominance by any single joint, the skeleton embedding features are processed through a Fourier feature mapping function and combined with positional information to generate the final skeleton embedding.

3.3. Dual stream HAR model based on RGB frames and skeleton data

To fully leverage the advantages of skeleton data, a dual-stream architecture is designed that combines SkateFormer (Do and Kim, 2024), which processes 3D skeleton data with LSSPE-TimeSFormer (Bertasius et al., 2021). This model has the ability to capture the complementary strengths of both modalities, improving the accuracy and robustness of human activity recognition. The architecture can be seen in Figure 2.

In the current stage of human activity recognition research, skeleton data remains an extremely effective modality for capturing relative relationships between major body parts (Yue et al., 2022). However, for fine-grained recognition scenarios such as facial expression changes, finger movements, or interactions with small objects, skeleton data alone may not provide sufficient detail (Song et al., 2023). RGB data offers crucial complementary information, including texture, colour, and subtle changes that skeleton data cannot capture. Therefore, it is believed that fusing these complementary modalities is necessary to simultaneously leverage the structured representation of skeleton data and the rich details of RGB data, achieving more comprehensive and accurate human activity recognition.

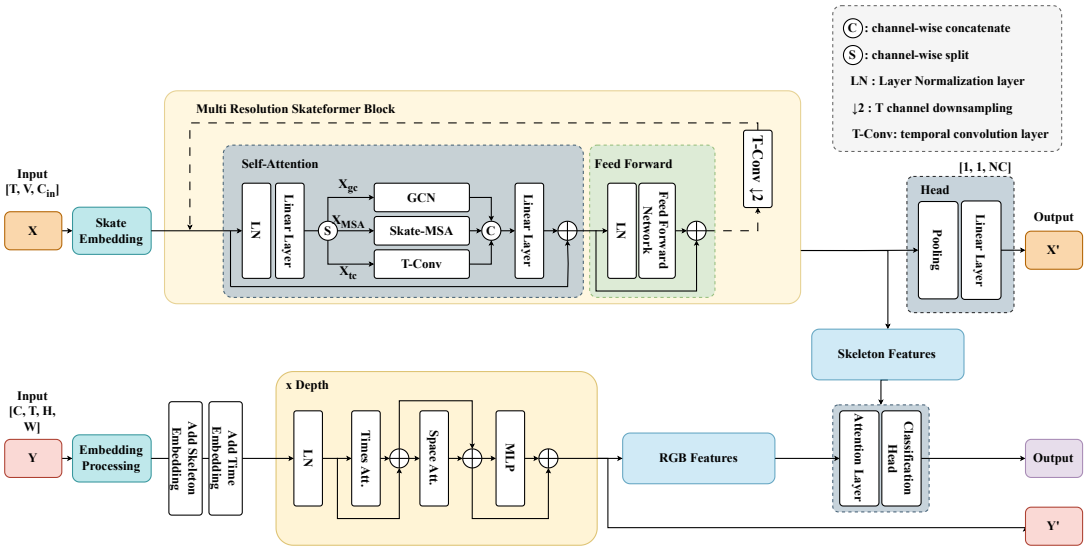

Figure 2: Proposed dual-stream architecture for human activity recognition, combining RGB frames with LSSPE and 3D skeleton data.

4. Experiments

In this section, comprehensive experiments are presented to evaluate the effectiveness of the proposed methods. The experimental setup, including datasets, implementation details, and evaluation metrics will be introduced. Then, proposed methods are compared with state-of-the-art approaches and conduct ablation studies to analyse the contribution of each component.

4.1. Experiments Setup

4.1.1. Datasets

Experiments are conducted on three widely used human activity recognition datasets:

- **UCF101** (Soomro et al., 2012): A dataset containing 13,320 videos across 101 action categories, primarily focused on sports and daily activities.
- **HMDB51** (Kuehne et al., 2011): A smaller dataset with 6,766 video clips from 51 action categories, featuring greater diversity in camera viewpoints and backgrounds.
- **NTU RGB+D** (Shahroudy et al., 2016): A large-scale dataset containing 56,880 action samples in 60 action classes, captured using Kinect v2 sensors. This dataset provides synchronised RGB videos and 3D skeleton data, making it ideal for evaluating our dual-stream architecture.

For NTU RGB+D, the experiments follow the standard evaluation protocols: Cross-Subject (CS) setting, where subjects in training and testing sets are different, and Cross-View (CV) setting, where camera views in training and testing sets are different.

4.1.2. Implementation Details

All experiments are conducted on an Intel Core i9-10980XE processor with an NVIDIA A6000 GPU. For the Shuffled Position Embedding ViT, it is validated on the UCF101 and HMDB51 datasets. For the dual-stream model and its component branches, training and evaluation are performed on the NTU RGB+D dataset.

For the LSSPE-TimeSFormer, the model is initialised with weights pre-trained on Something-Something V2 (Goyal et al., 2017). The patch size is setted to 16×16 pixels and hidden dimension is 768. The model was trained for 30 epochs using the AdamW optimiser with a learning rate of 10^{-4} and weight decay of 0.05. For the SkateFormer branch, all the settings are based on config files offered on their official website.

The dual-stream model combined both branches with a feature fusion module followed by a classification head. During training, a two-stage strategy is employed: first, pre-training each branch independently, then fine-tuning the entire model end-to-end.

4.1.3. Evaluation Metrics

For the shuffled position embedding ViT on the UCF101 and HMDB51 datasets, proposed methods are evaluated through standard classification metrics: accuracy (Acc), precision, recall, and F1 score. For the LSSPE-TimeSFormer and the dual stream model on the NTU RGB+D dataset, the results are reported under official Cross-Subject (CS) and Cross-View (CV) settings to assess the generalisation capability of our models.

4.2. Comparison with State-of-the-Art Methods

Models with similar or larger parameters are selected to compare with proposed SPE-ViT. The comparison results can be seen in the Table 1. All the reported performance of other models are gotten from their papers. As reported, on the UCF101 dataset, proposed SPE-ViT performs second only to VideoMAE v2 (Wang et al., 2023a), and is comparable to BIKE ViT-L/14 (Wu et al., 2023), both achieving 98.8%, but with significantly fewer parameters. On the HMDB51 dataset, SPE-ViT outperforms all other comparison models, achieving an accuracy of 90.3%.

Table 2 presents the comparison of our methods with state-of-the-art approaches on the NTU RGB+D dataset under both CS and CV settings. Proposed dual-stream model with LSSPE achieves superior performance compared to existing methods, demonstrating the effectiveness of our approach. It is worth mentioning that pi-vit (Reilly and Das, 2024) using RGB and Skeleton modalities slightly outperforms the dual-stream model in cross-subject and cross-view, and these two models use the same baseline TimeSFormer. This is because pi-vit uses 16 frames of tracks of human-centred crops (Das et al., 2020) as input and achieved 93% and 97.2% performance in training the baseline model, which is better than 87.2% and 91.5% using 8 frames of random cropping. And pi-vit also uses the Hyperformer (Zhou et al., 2022) to preprocess the skeleton data. While this paper focuses on exploring the potential of position embedding, mainly concentrated on the performance gains brought about by LSSPE; therefore, we did not train the model with same frames crop and Frame extraction strategy settings to compare performance.

Table 1: Comparison with reported state-of-the-art methods on HMDB51&UCF101 dataset.

Method	Params	UCF101(%)	HMDB51(%)
MVD-L (Wang et al., 2023b)	87M	97.5	79.7
VideoMAE (Tong et al., 2022)	87M	96.1	73.3
VideoMAE V2 (Wang et al., 2023a)	1050M	**99.6**	**88.1**
UniFormerV2-B/16 (Li et al., 2022)	115M	96.8	80.0
FTP-UniFormerV2-B/16 (Lu et al., 2024)	136M	98.7	83.9
BIKE ViT-L/14 (Wu et al., 2023)	230M	**98.8**	82.2
MMVFAC (Alayrac et al., 2020)	94M	95.2	75.0
SPE-ViT (Ours)	88M	**98.8**	**90.3**

The best scores for each dataset are bolded, and the second-best scores are highlighted in blue.

Table 2: Comparison with reported state-of-the-art methods on NTU RGB+D dataset.

Method	Architecture	Modalities	CS (%)	CV (%)
π-ViT (Reilly and Das, 2024)	ViT	RGB+Skeleton	**96.3**	**99.0**
π-ViT (Reilly and Das, 2024)	ViT	RGB only	94.0	97.9
STAR-Transformer (Ahn et al., 2023)	ViT	RGB+Skeleton	92.0	96.5
3DA (Kim et al., 2023)	ViT	RGB+Skeleton	94.3	97.2
UMDR (Zhou et al., 2023)	Transformer	RGB-D	**96.2**	98.0
MaskCLR (Abdelfattah et al., 2024)	Transformer	Skeleton	93.9	97.3
JPFormer (Cui and Hayama, 2024)	Transformer	Skeleton	93.2	96.9
ProtoGCN (Liu et al., 2025)	GCN	Skeleton	93.8	97.8
DSTSA-GCN (Cui et al., 2025)	GCN	Skeleton	92.8	97.0
IPP-Net (Ding et al., 2023)	\	Parsing+Skeleton	93.8	97.1
PoseC3D[3D Heatmap] (Duan et al., 2022)	\	Skeleton	94.1	97.1
VPN (Das et al., 2020)	\	RGB+Skeleton	95.5	98.0
DVANet (Siddiqui et al., 2024)	\	RGB	93.4	**98.1**
Dual-Stream Model (Ours)	ViT	RGB+Skeleton	**95.8**	**98.7**

The best scores for each dataset are bolded, the second-best scores are highlighted in blue and the third-best scores are highlighted in red.

4.3. Ablation Studies

4.3.1. Effect of Different Position Embedding Methods

To validate the effectiveness of proposed LSSPE, experiments with different position embedding methods are conducted on the NTU RGB+D 60 datasets. Table 3 shows the results, demonstrating that LSSPE consistently outperforms other position embedding methods.

Table 3: Comparison of different position embedding methods.

Position Embedding Type	Cross-sub(%)	Cross-view (%)
Learnable Absolute	87.2	91.5
Relative (Ma and Wang, 2024)	87.8	92.8
SPE (Ours)	87.5	93.6
Skeleton embedding	87.3	93.2
LSSPE (Ours)	**89.8**	**93.7**

The best scores are bolded.

4.3.2. Contribution of Skeleton Embedding and Dual-Stream Architecture

To further analysis the contribution of proposed LSSPE and the dual-stream architecture on the NTU RGB+D dataset. Ablation study of different components are necessary. Table 4 presents the results, showing that both components significantly improve the performance, and their combination yields the best results.

Table 4: Ablation study on the contribution of different components.

Method	CS (%)	CV (%)
TimeSFormer (Baseline)	87.2	91.5
+ LSSPE	89.8	93.7
+ Skeleton Branch	94.9	98.0
+ LSSPE + Skeleton Branch	**95.8**	**98.7**

The best scores are bolded.

4.4. Generalisation Performance on Small Datasets

To evaluate the generalisation capability of proposed SPE-ViT, experiments with limited training data are conducted on the HMDB51 & UCF101 dataset. Table 5 and Table 6 illustrate the performance of SPE-ViT with varying percentages of training data. From these two tables, it can be noticed that shuffled position embedding seems to be viewed as a kind of data augmentation, with shuffled position embedding delivering a somewhat larger performance gain as the dataset decreases. However, when the proportion of data used is lower than that, the model performance improvement again decreases instead.

4.5. Visualisation Analysis of Skeleton Embedding

To provide insights into how LSSPE method improves the model's attention mechanism, the attention maps of different models are visualised. Figure 3 shows the comparison of attention maps between the baseline TimeSFormer and LSSPE-TimeSFormer. The visualisation demonstrates that LSSPE enables the model to focus more accurately on human motion-related regions, leading to better recognition performance.

Table 5: SPE ablation experiments on different training set sizes of UCF101

Training Data	Embedding Type	Acc.	Precision	Recall	F1
80%	SPE	0.9883	0.9866	0.9858	0.9861
	Standard	0.9876	0.9865	0.9845	0.9854
70%	SPE	0.9865	0.9853	0.9844	0.9847
	Standard	0.9843	0.9820	0.9805	0.9810
60%	SPE	0.9855	0.9829	0.9812	0.9818
	Standard	0.9808	0.9771	0.9760	0.9764
50%	SPE	0.9816	0.9793	0.9767	0.9777
	Standard	0.9786	0.9747	0.9737	0.9739

Table 6: SPE ablation experiments on different training set sizes of HMDB51

Training Data	Embedding Type	Acc.	Precision	Recall	F1
80%	SPE	0.9032	0.9086	0.8849	0.8939
	Standard	0.8997	0.8950	0.8884	0.8893
70%	SPE	0.8890	0.8834	0.8740	0.8755
	Standard	0.8838	0.8823	0.8699	0.8734
60%	SPE	0.8762	0.8703	0.8591	0.8617
	Standard	0.8712	0.8684	0.8547	0.8559
50%	SPE	0.8407	0.8451	0.8158	0.8225
	Standard	0.8381	0.8418	0.8160	0.8179

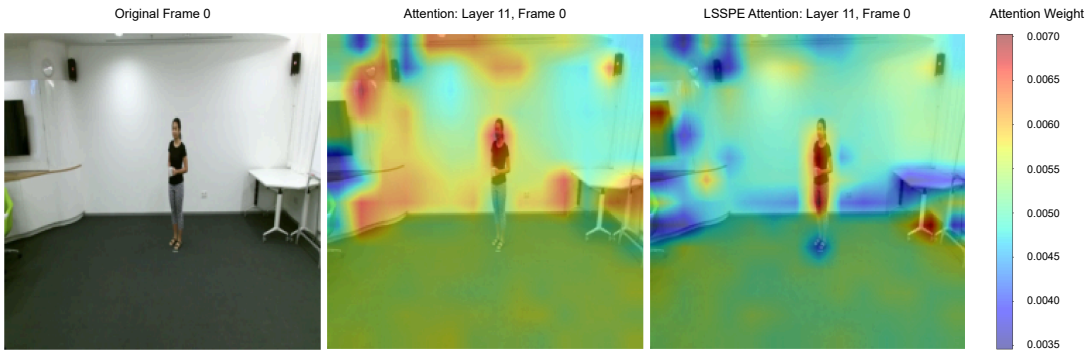

Figure 3: Visualisation of attention maps

5. Conclusion

In this work, we propose Shuffled Position Embedding (SPE) and its enhanced variant, Local Shuffled Skeleton Position Embedding (LSSPE), as novel position embedding methods for human activity recognition. These approaches innovatively explore shuffling mechanisms while incorporating skeletal structural information. Extensive experiments on UCF101,

HMDB51, and NTU RGB+D datasets demonstrate that both SPE and LSSPE significantly outperform conventional position embedding techniques. Furthermore, our results establish position embedding as an effective modality fusion strategy. The proposed dual-stream architecture successfully integrates RGB and skeleton modalities, achieving outstanding performance on NTU RGB+D dataset with 95.8% (cross-subject) and 98.7% (cross-view) accuracy. Future work will investigate the applicability of shuffled position embedding to broader video understanding tasks and explore its generalization capabilities.

6. Acknowledgement

The authors appreciate financial support from the China Scholarship Council.

References

Mohamed Abdelfattah, Mariam Hassan, and Alexandre Alahi. Maskclr: Attention-guided contrastive learning for robust action representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18678–18687, 2024.

Dasom Ahn, Sangwon Kim, Hyunsu Hong, and Byoung Chul Ko. Star-transformer: a spatio-temporal cross attention transformer for human action recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3330–3339, 2023.

Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *Advances in neural information processing systems*, 33:25–37, 2020.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Icml*, volume 2, page 4, 2021.

XB Bruce, Yan Liu, Xiang Zhang, Sheng-hua Zhong, and Keith CC Chan. Mmnet: A model-based multimodal network for human action recognition in rgb-d videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3522–3538, 2022.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

Hu Cui and Tessai Hayama. Joint-partition group attention for skeleton-based action recognition. *Signal Processing*, 224:109592, 2024.

Hu Cui, Renjing Huang, Ruoyu Zhang, and Tessai Hayama. Dstsa-gcn: Advancing skeleton-based gesture recognition with semantic-aware spatio-temporal topology modeling. *Neurocomputing*, 637:130066, 2025.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

Srijan Das, Saurav Sharma, Rui Dai, Francois Bremond, and Monique Thonnat. Vpn: Learning video-pose embedding for activities of daily living. In *European conference on computer vision*, pages 72–90. Springer, 2020.

Srijan Das, Rui Dai, Di Yang, and Francois Bremond. Vpn++: Rethinking video-pose embeddings for understanding activities of daily living. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9703–9717, 2021.

Runwei Ding, Yuhang Wen, Jinfu Liu, Nan Dai, Fanyang Meng, and Mengyuan Liu. Integrating human parsing and pose network for human action recognition. In *CAAI International Conference on Artificial Intelligence*, pages 182–194. Springer, 2023.

Jeonghyeok Do and Munchurl Kim. Skateformer: skeletal-temporal transformer for human action recognition. In *European Conference on Computer Vision*, pages 401–420. Springer, 2024.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2969–2978, 2022.

Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

Misha Karim, Shah Khalid, Aliya Aleryani, Jawad Khan, Irfan Ullah, and Zafar Ali. Human action recognition systems: A review of the trends and state-of-the-art. *IEEE Access*, 2024.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

Sangwon Kim, Dasom Ahn, and Byoung Chul Ko. Cross-modal learning with 3d deformable attention for action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10265–10275, 2023.

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *arXiv preprint arXiv:2211.09552*, 2022.

Hongda Liu, Yunfan Liu, Min Ren, Hao Wang, Yunlong Wang, and Zhenan Sun. Revealing key details to see differences: A novel prototypical perspective for skeleton-based action recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29248–29257, 2025.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.

Hui Lu, Hu Jian, Ronald Poppe, and Albert Ali Salah. Enhancing video transformers for action understanding with vlm-aided training. *arXiv preprint arXiv:2403.16128*, 2024.

Yujun Ma and Ruili Wang. Relative-position embedding based spatially and temporally decoupled transformer for action recognition. *Pattern Recognition*, 145:109905, 2024.

Jing Qin, Shugang Zhang, Yiguo Wang, Fei Yang, Xin Zhong, and Weigang Lu. Improved skeleton-based activity recognition using convolutional block attention module. *Computers and Electrical Engineering*, 116:109231, 2024.

Dominick Reilly and Srijan Das. Just add?! pose induced video transformers for understanding activities of daily living. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18340–18350, 2024.

Sujan Sarker, Sejuti Rahman, Tonmoy Hossain, Syeda Faiza Ahmed, Lafifa Jamal, and Md Atiqur Rahman Ahad. Skeleton-based activity recognition: Preprocessing and approaches. In *Contactless Human Activity Analysis*, pages 43–81. Springer, 2021.

Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

Jing Shi, Yuanyuan Zhang, Weihang Wang, Bin Xing, Dasha Hu, and Liangyin Chen. A novel two-stream transformer-based framework for multi-modality human action recognition. *Applied Sciences*, 13(4):2058, 2023.

Nyle Siddiqui, Praveen Tirupattur, and Mubarak Shah. Dvanet: Disentangling view and action features for multi-view action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4873–4881, 2024.

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, pages 568–576, 2014.

Wenwei Song, Wenxiong Kang, and Liang Lin. Hand gesture authentication by discovering fine-grained spatiotemporal identity characteristics. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(1):461–474, 2023.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Clayton Frederick Souza Leite, Henry Mauranen, Aziza Zhanabatyrova, and Yu Xiao. Transformer-based approaches for sensor-based human activity recognition: Opportunities and challenges. *Available at SSRN 5131703*, 2024.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14549–14560, 2023a.

Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Lu Yuan, and Yu-Gang Jiang. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6312–6322, 2023b.

Zhe Wang, Aladine Chetouani, Mohamed Jarraya, Didier Hans, and Rachid Jennane. Transformer with selective shuffled position embedding and key-patch exchange strategy for early detection of knee osteoarthritis. *Expert Systems with Applications*, 255:124614, 2024.

Wenhao Wu, Xiaohan Wang, Haipeng Luo, Jingdong Wang, Yi Yang, and Wanli Ouyang. Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6620–6630, 2023.

Chunyan Xu, Rong Liu, Tong Zhang, Zhen Cui, Jian Yang, and Chunlong Hu. Dual-stream structured graph convolution network for skeleton-based action recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17 (4):1–22, 2021.

Rujing Yue, Zhiqiang Tian, and Shaoyi Du. Action recognition based on rgb and skeleton data sets: A survey. *Neurocomputing*, 512:287–306, 2022.

Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015.

Jiaxu Zhang, Wei Xie, Chao Wang, Ruide Tu, and Zhigang Tu. Graph-aware transformer for skeleton-based action recognition. *The Visual Computer*, 39(10):4501–4512, 2023.

Rui Zhao, Haider Ali, and Patrick Van der Smagt. Two-stream rnn/cnn for action recognition in 3d videos. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4260–4267. IEEE, 2017.

Benjia Zhou, Pichao Wang, Jun Wan, Yanyan Liang, and Fan Wang. A unified multimodal de-and re-coupling framework for rgb-d motion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11428–11442, 2023.

Xin Zhou, Zhaohui Ren, Shihua Zhou, Zeyu Jiang, TianZhuang Yu, and Hengfa Luo. Rethinking position embedding methods in the transformer architecture. *Neural Processing Letters*, 56(2):41, 2024.

Yuxuan Zhou, Zhi-Qi Cheng, Chao Li, Yanwen Fang, Yifeng Geng, Xuansong Xie, and Margret Keuper. Hypergraph transformer for skeleton-based action recognition. *arXiv preprint arXiv:2211.09590*, 2022.

Mohammadreza Zolfaghari, Gabriel L Oliveira, Nima Sedaghat, and Thomas Brox. Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2904–2913, 2017.