TemPose

a new skeleton-based transformer model designed for fine-grained motion recognition in badminton

Ibh, Magnus; Grasshof, Stella; Witzner, Dan; Madeleine, Pascal

Published in:
Proceedings - 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2023

DOI (link to publication from Publisher):
10.1109/CVPRW59228.2023.00548

Publication date:
2023

Document Version
Accepted author manuscript, peer reviewed version

TemPose: a new skeleton-based transformer model designed for fine-grained motion recognition in badminton

Magnus Ibh Stella Grasshof Dan Witzner Pascal Madeleine
Machine learning group, IT University of Copenhagen Aalborg University
{ibhq, stgr, witzner}@itu.dk pm@hst.aau.dk

Abstract

This paper presents TemPose, a novel skeleton-based transformer model designed for fine-grained motion recognition to improve understanding of the detailed player actions in badminton. The model utilizes multiple temporal and interaction layers to capture variable-length multi-person human actions while minimizing reliance on non-human visual context. TemPose is evaluated on two fine-grained badminton datasets, where it significantly outperforms other baseline models by incorporating additional input streams, such as the shuttlecock position, into the temporal transformer layers of the model. Additionally, TemPose demonstrates great versatility by achieving competitive results compared to other state-of-the-art skeleton-based models on the large-scale action recognition benchmark NTU RGB+D. Experiments are conducted to explore how different model parameter configurations affect TemPose's performance. Additionally, a qualitative analysis of the temporal attention maps suggests that the model learns to prioritize frames of specific poses relevant to different actions while formulating an intuition of each individual's importance in the sequences. Overall, TemPose is an intuitive and versatile architecture that has the potential to be further developed and incorporated into other methods for managing human motion in sports with state-of-the-art results.

1. Introduction

Badminton is a fast-paced racket sport that requires a high level of skill, athleticism, and tactical awareness. As the sport's popularity grows, the need for objective and data-driven methods for evaluating player performance has become increasingly important. One area of particular interest uses automatic analysis, specifically human action recognition (HAR), to provide insights into a player's performance [13] and inform decision-making in the sport. Fine-grained action recognition deals with action classes closely related in both type (e.g. badminton strokes) and motion (i.e., strokes may look similar) and is appropriate for highly technical sports disciplines that require high precision and accuracy in movement execution. The required attention to detail in badminton results in small and subtle differences in how players execute specific movements, which are difficult to capture using RGB-based methods [42]. Skeleton motion as a primary feature in fine-grained action recognition has been effective in various sports disciplines [8, 17], including badminton [19, 21]. Skeleton-data provides a detailed representation of the body movement through spatiotemporal sequences of joint and bone positions, which enables extracting features crucial for recognizing specific actions and movements, even those that may be subtle or difficult to detect with traditional imaging techniques. While existing methods for skeleton-based action recognition have achieved good results on controlled action benchmark datasets [23, 39], many tend to lack robustness and scalability for real-world applications. In an approach to address this issue, recent research has explored the use of transformer models, which have shown excellent capabilities in natural language processing (NLP) [7] and image segmentation [9, 14], to model sequential data for video action recognition [1, 18, 22].

This paper presents *TemPose*, a new skeleton-based transformer model designed for fine-grained motion recognition in badminton. The model offers several significant contributions, including a novel factorized transformer model that combines temporal and interaction layers, multi-person interaction modeling, and improved recognition rates using fewer parameters. The proposed action recognition model, *TemPose*, is outlined in Figure 1. The model takes processed skeleton data as input and passes it through a sequence of transformer layers in the TemPose encoder. This process creates tokens of the temporal data and captures the temporal body dynamics and sequential interactions between an arbitrary number of people involved in the action. The MLP head at the top predicts the action based on the information embedded in a class token. The primary and badminton-specific version of the model includes the fu-

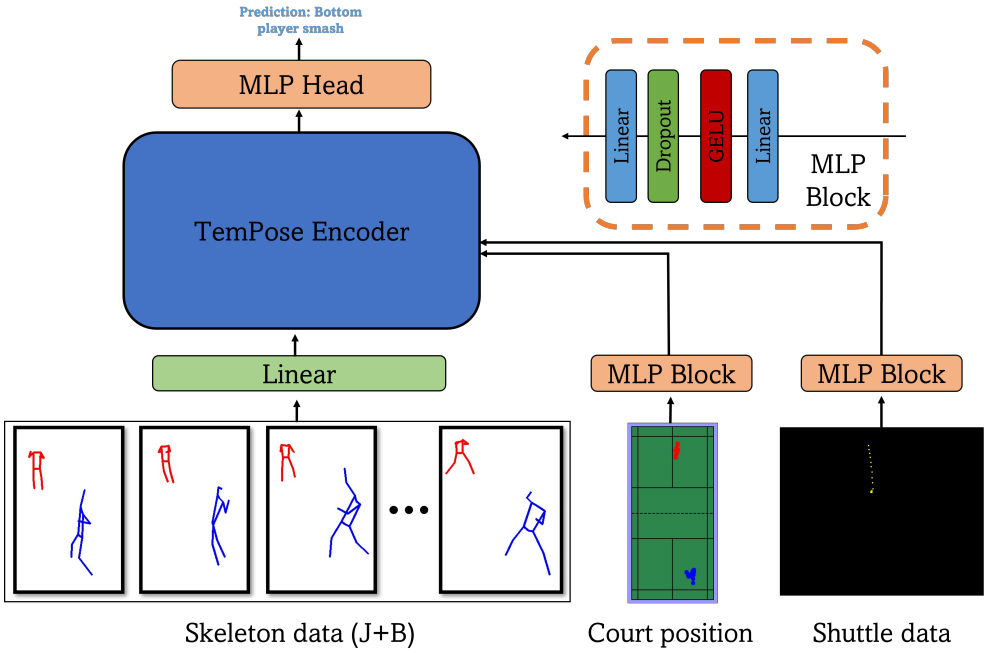

Figure 1. Illustration of our proposed action recognition framework, TemPose. The framework uses human skeleton data, consisting of joint and bone information, and incorporates additional features such as player court position and shuttlecock position from a badminton action (e.g., smash). The TemPose encoder, composed of multiple transformer layers, processes the input to embed relevant features into a class token. Finally, an MLP head utilizes these features to predict the action class. The composition of the MLP block is shown in the upper right corner.

sion of skeleton-data with player court positions (CP) and shuttlecock position (SP). We exhaustively test two different versions of TemPose, where the additional modalities are integrated at different stages of the TemPose encoder. In one version (TemPose-NF), the CP and SP sequences are tokenized and appended to the embedded skeleton-data before the interaction transformer layers. Figure 2 depicts TemPose-NF. The other version (TemPose-TF) prioritizes an early fusion of the skeleton, SP, and CP modalities.

An overview of related work is provided in Section 2, followed by a description of pose and shuttle estimation, pre-processing, and the model architecture in Section 3. The experimental results on fine-grained badminton datasets are presented in Section 4, along with testing on a standard benchmark action recognition datasets NTU RGB+D [20, 32]. The paper concludes with a qualitative analysis of the information stored in the different transformer layers and future works in 5.

2. Related Work

Action recognition in sports Most work on action recognition in badminton uses convolutional neural network (CNN) architectures for feature extraction on RGB images [29–31]. Decision-making algorithms like Support Vector Machines then use the extracted features to make predictions. Other approaches involve using handcrafted features such as Histogram of Oriented Gradients, along with temporal convolutional networks (TCN) to process the action's spatial and temporal aspects [5, 13]. Instead of using image data, skeleton data has been successfully used for the analysis and recognition tasks of other sports, such as Tai Chi [8, 10, 36] and fencing [26, 42]. But despite its potential, skeleton poses have yet to be thoroughly tested for badminton tasks. In one recent example [21], skeleton data is used in a gated recurrent unit (GRU) model to perform binary hit detection. However, like other recurrent models, GRUs can struggle with training issues. This paper proposes an architecture more suited for utilizing skeleton data for badminton recognition tasks.

Human-action recognition using skeleton data. Graph convolutional networks (GCN) are a popular method for skeleton-based action recognition [35, 40]. GCNs uses nodes to represent every human joint at every time. Connecting nodes, both spatially and temporally, to the other nodes with edges allows GCNs to capture both spatial and temporal aspects of human motion. Spatio-temporal GCNs have demonstrated promising results for skeleton-based ac-

tion recognition [23, 39, 40], but they also possess some limitations. One limitation is their limited ability to model long-term dependencies in complex actions, as they typically use a fixed-length temporal window. Moreover, they are sensitive to missing data [41] and require a carefully designed graph structure based on the recognized actions' characteristics, which can be challenging. CNNs are also commonly used for analyzing skeleton data. One approach is to stack heatmaps along the temporal dimension into a 3D input and use 3D-CNNs to extract information [3, 11]. Other studies, such as [17, 42], generate a temporal sequence of joint coordinates and use TCNs to encode the information.

Transformers for human action recognition. The emergence of ViT [9] has led to many applications of vision transformer backbones [2, 22, 37, 38]. Including vision transformers [1, 12, 18, 22] used for HAR. These works are typically trained on Kinetics-400 [16] to mitigate the issue involved with over-fitting. But only a few works have considered transformers on other modalities than RGB data, in our case, skeleton data. Utilizing transformer based models on skeleton data has, however, been attempted in other recent work. [28] combines self-attention with a GCN and TCN to model spatial and temporal attention. Similarly, [27] performs temporal encoding of the skeleton poses with a sequence of temporal transformer layers.
Unlike previous work, we present a factorized transformer encoder. Embedding individual skeleton data into temporal and interaction tokens allows the method to encode information about the motion of multiple people across the entire sequence length.

3. Model

This section first outlines the extraction process for the skeleton, CP, and SP data. We introduce the temporal transformer layer module for a single individual and subsequent action prediction. Subsequently, the model is extended to a factorized temporal and interaction encoder, which embeds information about the interaction between individuals in the class token. Last, we describe two methods of incorporating CP and SP data into the model.

3.1. Extraction of the skeleton, player position, and shuttlecock data

A visual representation of the skeleton-data retrieval is shown in Figure 3. In a video sequence with T frames, the poses of a person are given by the sequence $P = [P_1, \ldots, P_T]^{\mathrm{T}} \in \mathbb{R}^{T\times 2J}$. A pose $P_t \in \mathbb{R}^{J\times 2}$ in frame t is represented by J keypoints $(x_i^{(t)}, y_i^{(t)})$, where $x_i^{(t)}$ and $y_i^{(t)}$ are 2D joint coordinates for the joint i. The bones $B_t \in \mathbb{R}^{B\times 2}$ in frame t are represented by the keypoint differences $(x_i^{(t)} - x_j^{(t)}, y_i^{(t)} - y_j^{(t)})$, where i and j are specific

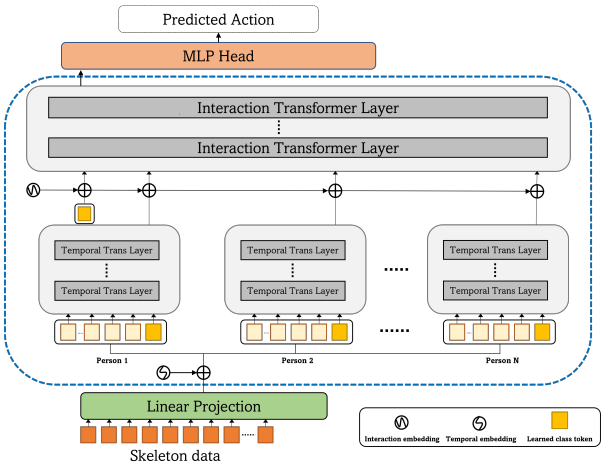

Figure 2. Illustration of *TemPose encoder* shows the factorized transformer structure. First, the temporal token for each person is encoded by the temporal transformer layer. Second, The interaction between actors is modeled based on the temporal context of each person.

joint pairs that make up the human bones. The final skeleton data sequence, S, of an individual, is defined to be

$$S = [[P_1, B_1], \ldots, [P_T, B_T]]^{\mathrm{T}} \in \mathbb{R}^{T\times 2(J+B)} \quad (1)$$

The pose extraction pipeline consists of two stages using tools from previous studies, including [4, 6] to detect humans and perform pose estimation. We employ HRnet [33], a pre-trained framework, to estimate the 2D poses. However, irrelevant individuals, such as spectators in the crowd, can limit the quality of the skeleton data. To address this issue in badminton, we calculate a homography using the court's known dimensions and map the feet of the detected individuals to the ground plane. By doing so, we only consider skeletons within the court and can identify each sequence's top and bottom player. In cases where a whole skeleton is missing, we replace it with the pose from the previous frame. Finally, we normalize the poses by centering them and scaling them to have a bounding box diagonal of 1. Additionally, we sample the players' 2D ground plane feet position (i.e., CP) for each time frame as an additional input feature. The sequence PC is represented as $\in \mathbb{R}^{T\times 2}$. The shuttlecock's position holds valuable information for categorizing the different strokes in badminton. To extract the shuttlecock's position, we use a pre-trained model from [34] to obtain its image coordinates in each frame of the video, represented as (u, v, c), where u and v are the image coordinates of the shuttlecock, and c is the confidence of the prediction. We only consider predictions with a confidence score above 0.75; we pad failed predictions with zeros.

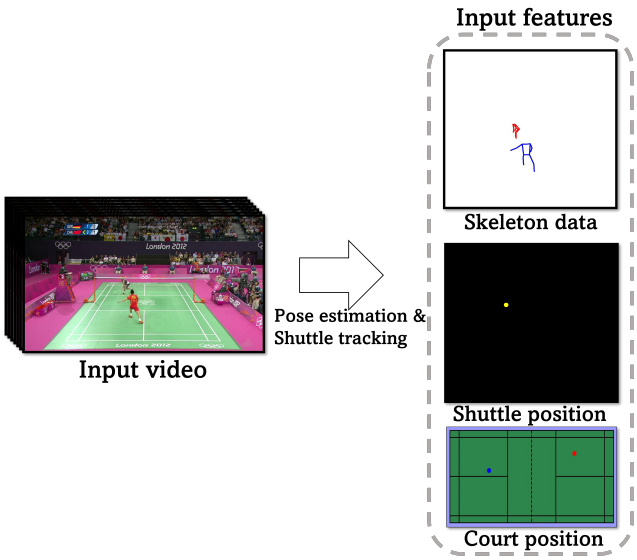

Figure 3. The figure illustrates the input data utilized by our proposed action recognition framework, *TemPose*. The framework takes in centered and normalized skeleton data of the badminton players, along with their court position and the scaled position of the shuttlecock, all of which are extracted from RGB video input. Specifically, HRNet [33] estimates the poses of the badminton players, while TrackNet [34] estimates the position of the shuttlecock.

3.2. Skeleton-based temporal self-attention for action prediction

As a first step, we consider a single-person sequence without the interaction layers from Figure 2. The skeleton data is mapped through a linear projection to a sequence of temporal tokens $[x_1, \ldots, x_T]^{\mathrm{T}} \in \mathbb{R}^{T \times D_L}$, where each token $x_t \in \mathbb{R}^{D_L}$ is the vector representation of the skeleton at that particular time frame. A learnable temporal embedding $E_T \in \mathbb{R}^{T+1 \times D_L}$ is added to the tokens to capture the underlying temporal structure better. The sum of the embedding and projection yields x, the input of the transformer layers. x is formally defined as

$$x = [x_{cls}, \mathrm{Linear}(S)]^{\mathrm{T}} + E_T \tag{2}$$

$$= [x_{cls}, x_1, \ldots, x_T]^{\mathrm{T}} + E_T, \in \mathbb{R}^{T \times D_L} \tag{3}$$

where $x_{cls} \in \mathbb{R}^{D_L}$ is a learned class token, D_L is the dimension of the embedded feature space, and Linear is a learned linear projection. The representation of x_{cls} at the final transformer layer is used by the MLP head to make predictions. The tokens defined in (3) are then passed through transformer layers, where L is the transformer depth. To distinguish between the tokens at different layers, we define them as $x^{(l)}$ after having passed through layer l. Each layer is composed of a multi-head self-attention (MHSA), layer normalization (LN), and a multi-layer perceptron (MLP), which consists of two linear projections only separated by a GELU activation [15] and dropout, see Figure 1. The design of a single transformer layer is illustrated in Figure 4 on the left. The transformer block is described by Equation 4 and Equation 5

$$\tilde{x}^{(l+1)} = x^{(l)} + \mathrm{MHSA}(\mathrm{LN}(x^{(l)})), \tag{4}$$

$$x^{(l+1)} = \tilde{x}^{(l+1)} + \mathrm{MLP}(\mathrm{LN}(\tilde{x}^{(l+1)})), \tag{5}$$

where $\tilde{x}^{(l+1)}$ is the in-between embedded obtained after the self-attention module. The following equation describes a single head of self-attention

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{D_K}}\right) V, \tag{6}$$

where $Q = W_Q x^{(l)} \in \mathbb{R}^{T \times D_A}$, $K = W_K x^{(l)} \in \mathbb{R}^{(T+1) \times D_A}$, and $V = W_V x^{(l)} \in \mathbb{R}^{(T+1) \times D_A}$ are learned linear projections of the input sequence that respectively represent the current token, the other tokens, and their associated values, used to calculate attention scores and output. After scaling and softmax activation of the input variables, (6) serves as an attention map that provides temporal context to the value-array V, where D_A is the attention head latent dimension. The output of the MHSA yields N_{heads} value vectors V weighted by the temporal attention maps. As illustrated in Figure 4, these weighted value vectors are concatenated and mapped to the updated representation of the temporal tokens $\tilde{x}^{(l+1)}$ with another learned linear projection. Note that the number of transformer layers L is an adjustable hyperparameter.
Finally, the class token x_{cls}^{L} is fed to an MLP block to predict the action category of the samples

$$x_{act} = \mathrm{MLP}(x_{cls}), \tag{7}$$

where $x_{act} \in \mathbb{R}^{D_{cls}}$ is the model prediction, and D_{cls} the number of different action categories.

3.3. Factorized temporal and interaction structure

Actions in video sequences often contain multiple people (e.g., two people in badminton singles matches) making different subactions and movements in parallel and often reacting to the other people involved in the action. Hence, in the TemPose encoder, we want to account for multiple people and their actions. *TemPose* utilizes a factorized encoder structure of transformer layers, inspired by ViViT [1], to model the interaction between people. Figure 2 depicts the TemPose encoder. The model consists of 1.) a temporal transformer layer that captures the temporal interaction between poses extracted from the same person, identical to the temporal structure outlined in section 3.2. However, the temporal encoding is now performed in parallel

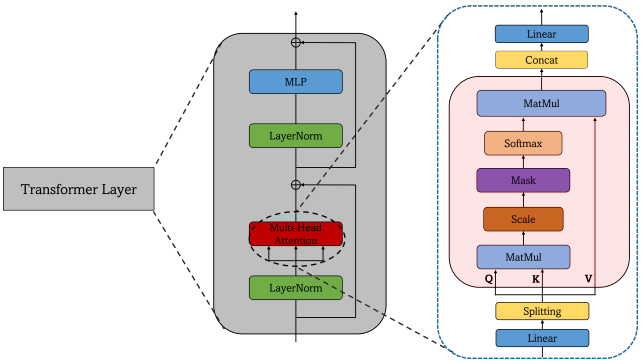

Figure 4. Illustration of transformer layers to the left. The structure of the transformer layers is identical for the temporal and interaction transformer layers. The right block shows the composition of a single-head self-attention module. The Mask leaves out the (zero) padded temporal and interaction tokens in the attention map, which allows the model to handle action sequences of varying lengths proficiently.

for up to N people involved in the action sequence, see Figure 2. As a result, the class token notation $x_{cls}^{(l)}$ is updated to $\tau_{cls,n}^{(l_T)}$. $\tau_{cls,n}^{(l_T)}$ is the temporal class token for person $n \in \{1, \ldots, N\}$, at temporal layer l_T 2.) After being processed by all L_T temporal layers, the N temporal class tokens are concatenated to $[\tau_{cls,1}^{(L_T)}, \ldots, \tau_{cls,N}^{(L_T)}]$, prepended with a interaction class token η_{cls}, and assigned an interaction embedding E_I identical to the step of Equation 3. Summing up the input for the interaction encoder becomes

$$z = [\eta_{cls}, \tau_{cls,1}, \ldots, \tau_{cls,N}]^{\mathrm{T}} + E_I, \in \mathbb{R}^{T \times D_L}. \quad (8)$$

Subsequently, the input tokens, z, are passed through L_N transformer layers to capture interactions between the embedded temporal class tokens of people participating in the action. An MLP head uses the final representation of the interaction classification token $\eta_{cls}^{(L_N)} \in \mathbb{R}^{D_L}$, see Equation 7, to predict the action class.

3.4. Player and shuttlecock position infusion

This section discusses two configurations for integrating badminton-specific CP and SP data into the encoder.

Temporal fusion: In the first fusion configuration, the SP and CP input data passes through separate TCN blocks consisting of two 1D-convolutional layers separated only by a GELU activation and dropout. The two layers have dilation 1 and 3, respectively, with a kernel size of 5 and stride 1. Through the TCN block, the channels (i.e., dimensionality) of SP and CP data are increased to the embedded dimension of the transformer layers D_L. The two new input streams are then appended to the transformer input equivalent to adding additional people (see Figure 2). The remaining architecture is identical to factorized encoder described above.

Interaction fusion In a different approach, the CP and SP data are first incorporated into the *TemPose* encoder after the temporal transformer layers. Here the SP and CP data are flattened along the temporal and coordinate dimensions and then separately passed through an MLP block. The resulting representations are appended to z in Equation 8, which is then processed identically to the original *TemPose* architecture but with two additional interaction tokens. In the experiments section, *TemPose* without fusion is referred to as *TemPose-V*. In contrast, *TemPose-TF* and *TemPose-NF* refer to temporal fusion and interaction fusion, respectively.

Temporal & multi-person padding A property of transformer architectures is the ability to handle sequences of different lengths. We implement this for TemPose by always creating a set number of temporal tokens T, corresponding to the maximal clip length for a video. Shorter videos are padded with zeros and assigned pad tokens, so they are not considered when calculating self-attention. Specifically, the MASK step in the MHSA reduces attention scores on the padded tokens to zero, see Figure 4. TemPose extends this process by choosing an upper limit N of people to model interactions from. Clips with fewer people are padded with zeros and assigned pad tokens, so they are not considered in the interaction attention.

4. Experiments

This section presents the results of our experiments to assess the performance of the factorized transformer layers compared to other state-of-the-art skeleton-based human activity recognition models. Specifically, we evaluated *TemPose* on two fine-grained badminton datasets [13] and demonstrated the versatility of the architecture by including experiments on the large-scale human motion dataset NTU RGB+D [20, 32].

Badminton Olympics (Bad OL). A fine-grained badminton dataset from 10 videos containing 15300 samples from 13 classes of badminton strokes with the following classes: top/bottom player forehand, backhand, smash, lob, react strokes, and a none class. The train/test split is match/video splitting, where all clips from one video (match) are kept as the test set.

Dataset on Badminton Stroke Placement (Bad PL). The dataset is confidential but contains 5500 samples of backcourt badminton strokes categorized into either attack or transport strokes. Additionally, the dataset includes information on the approximate location of the shuttlecock

Table 1. Hyperparameters for the TemPose training procedure. The right part of the table includes regularization and data augmentation choices.

Training	
Batch size	64
Optimizer	AdamW
Warm-up	25%
Learning rate	1e-04
LR scheduler	cosine decay

Regularization	
Label smoothing	0.1
Flipping	30%
Random shifting	30%
Dropout	0.3
Weight decay	0.01

placement for each stroke grouped into three different areas, such as left backcourt or middle midcourt, resulting in 12 different classes based on stroke type and placement. The train/test splitting is done cross-matches.

NTURGB+D. Is a large-scale human action dataset collected in a controlled setting. The dataset contains two versions, NTU-60 and NTU-120. NTU-60 has 57000 videos categorized into 60 different actions. NTU-120 has 114000 videos belonging to 120 different categories. Test and training data can be split in three different ways: cross-setup (**XSet**), cross-view (**XView**), and cross-subject (**X-sub**).

4.1. Implementation details

Table 1 list all settings and hyperparameters used for the training procedure. The choice is made based on a randomized search across the datasets. The AdamW optimization algorithm [24] is used for all training runs along with cosine-annealing [25]. Each training run is initialized with a sequence of warm-up epochs, slowly increasing the learning rate linearly from 0 to prevent overfitting. Unless specified otherwise, joint and bone data, J and B, respectively, are used together as skeleton data input.

4.2. Component studies

We analyze the individual components and different model configurations of TemPose. Unless stated otherwise, the performance is reported as classification accuracy on the Bad OL dataset. The default configuration uses the depth $L_T = L_N = 2$, $N_{heads} = 6$, embedded dimensions of $D_L = 100$ and $D_A = 128$, and lastly, an MLP scale factor of 4 between the input and hidden layers in MLP blocks.

Model configurations. To validate the multi-modal fusion approaches of the CP and SP data, we examine the performance of TemPose-V, TemPose-TF, and TemPose-NF for many different model settings shown in Table 2. AcT [27], a purely temporal skeleton-based model, is shown as the baseline model. Among the TemPose versions, TemPose-TF with $D_L = 100$ and $D_A = 128$ has the highest accuracy of 90.7% while only having 1.7 million parameters. The results suggest that temporal fusion of SP and CP is the best approach, as it achieves the highest accuracy using the fewest parameters.

Table 2. Accuracy and model size for different settings of the 3 TemPose versions. The number of attention heads $N_{heads} = 6$ and depth $L_T = L_N = 2$ are shared for all model configurations.

Model configuration	Params	Acc
Baseline (AcT [27])	2.1M	83.7%
TemPose-V		
with ($D_L = 75$, $D_A = 100$)	0.9M	85.6%
with ($D_L = 200$, $D_A = 200$)	5.2M	83.6%
TemPose-TF		
with ($D_L = 50$, $D_A = 75$)	0.5M	88.6%
with ($D_L = 100$, $D_A = 128$)	1.7M	90.7%
with ($D_L = 200$, $D_A = 256$)	6.7M	88.0%
TemPose-NF		
with ($D_L = 50$, $D_A = 75$)	2.5M	88.1%
with ($D_L = 100$, $D_A = 128$)	3.8M	89.3%
with ($D_L = 200$, $D_A = 256$)	9.0M	86.2%

Exploring joint-bone skeleton data. We investigated the impact of incorporating bone data into the joint data of the skeleton on the Bad OL and NTU RGB+D datasets for TemPose without CP and SP input. The results are presented in Table 3 and Table 6. Our findings are consistent with previous studies [11, 23]. The performance of TemPose significantly improves by utilizing both bone and joint data as input.

Importance of transformer depth. The effect of varying transformer depth is a crucial aspect of transformer models. Table 4 shows the results of a depth study on the TemPose-TF model. The model settings are kept constant throughout the study, except for the number of transformer layers, and report the model's performance for different combinations of L_T and L_N. The results show that increasing the transformer depth beyond a certain point leads to a drop in performance. The best accuracy is achieved for the combination of $L_T = 2$ and $L_N = 2$, with a top-1 accuracy of 90.7%. Increasing the depth further to $L_T = 3$ and $L_N = 3$ leads to a significant drop in accuracy to 88.3%. The performance continues to degrade as the depth is further increased. The continued drop could suggest that the performance continually worsens due to overfitting as the depth is increased. Thus, the study exemplifies the importance of finding the right balance between model complexity and data size. The overfitting can possibly be attributed to issues such as vanishing gradients or the relatively limited number of training samples in the Bad OL dataset.

Table 3. Joint + Bone architecture study

Models	Acc
Baseline (AcT [27])	
with (J)	81.8%
with (J+B)	83.7%
TemPose-V	
with (J)	81.4%
with (J+B)	85.6%

Table 4. Transformer depth study of the TemPose-TF. The remaining model settings are constant for the study, where $D_L = 100$, $D_A = 128$, and $N_{heads} = 6$. The performance of TemPose drops when the transformer depth increases.

Model	Acc
TemPose-TF	
with ($L_T = 1,\ L_N = 1$)	89.7%
with ($L_T = 1,\ L_N = 2$)	89.9%
with ($L_T = 2,\ L_N = 1$)	90.0%
with ($L_T = 2,\ L_N = 2$)	90.7%
with ($L_T = 3,\ L_N = 3$)	88.3%
with ($L_T = 4,\ L_N = 4$)	86.6%
with ($L_T = 6,\ L_N = 2$)	85.5%
with ($L_T = 2,\ L_N = 6$)	86.1%
with ($L_T = 6,\ L_N = 6$)	85.4%
with ($L_T = 8,\ L_N = 8$)	85.2%

Table 5. Top-1 accuracy results for *TemPose* with temporal (TF) and interaction (NF) fusion, to state-of-the-art (HAR) models on Badminton placement (Bad PL) and Badminton Olympics (Bad OL).

Model	Bad PL	Bad OL	Params
Bidirectional TCN [42]	80.4%	86.1%	4.1M
TCN Hog [13]	66.6%	77.0%	1.1M
ST-GCN [40]	72.3%	82.0%	3.4M
AcT-M [27]	77.9%	83.7%	2.1M
MS-G3D [23]	78.0%	83.2%	3.2M
TemPose-TF	83.9%	90.7%	1.7M
TemPose-NF	84.3%	89.3%	3.8M

4.3. Evaluation

Fine-grained sports action recognition - badminton. Table 5 shows the Top-1 accuracy results for TemPose-TF and TemPose-NF on two different Badminton datasets - Bad PL and Bad OL. The table also includes results for other state-of-the-art models on the same datasets. Overall, the results show that TemPose outperforms all other models on both datasets, with TemPose-TF achieving the highest accuracy on Bad OL and TemPose-NF achieving the highest accuracy on Bad PL. As observed in the configuration study, both fusion approaches achieve strong results, and based on our studies, no method is superior by a significant margin. However, we conclude that TemPose can accurately be used to classify the different types of movements in badminton.

Large-scale human action recognition. We showcase the versatility of TemPose, by testing TemPose on other more generic large-scale HAR benchmarks and comparing TemPose-V to other top-performing skeleton-based actions recognition models. Table 6 shows the results of *TemPose-V* on the NTU datasets. Despite being slightly worse than MS-G3D [23], and PoseConv3D [11] overall, TemPose achieves competitive results to other state-of-the-art models on all splittings of NTU RGB+D.

4.4. Qualitative analysis of temporal and interaction attention

We examine the attention maps of the transformer layers. To inspect what information is captured by the encoder. The temporal attention maps of two forehand strokes shown in Figure 6 reveal that similar patterns emerge between actions of the same class. The similar attention maps suggest that the model has learned to focus on specific temporal aspects of the actions to predict the entire sequence.
The attention maps are used to determine a temporal and interaction attention score for all actions. We define the attention score as the self-attention of the x_{cls}-token in the last transformer layer, aggregated and normalized across all attention heads. The temporal attention score is averaged over all individuals but weighted according to their interaction attention score. For a badminton smash, the attention score is depicted in Figure 5. TemPose identifies the frames around contact with the shuttlecock as the most significant section. The red and purple text represent the target and prediction class of the action, respectively. The model accurately predicts the action as a smash from the bottom player. Additionally, more attention is given to the smashing individual. The logical distribution of attention suggests that the model has developed the ability to gauge the relevance of each individual for the action based on their skeleton movement.

5. Future prospects

TemPose demonstrates top results on badminton action recognition tasks. However, in the experiments, the larger configurations of TemPose show clear signs of overfitting. The result indicates that the performance of TemPose may be further improved if additional steps to combat overfitting are taken. One prospect involves generating synthetic data to increase the amount of training data.

Table 6. Top-1 accuracy on the NTU RGB+D for state-of-art skeleton-based action recognition models.

	NTU RGB+D 120 (XSet)	NTU RGB+D 120 (XSub)	NTU RGB+D 60 (XSub)	NTU RGB+D 60 (XView)
ST-GCN [39]	73.2%	70.7%	81.5%	88.3%
ST-TR-agcn [28]	87.1%	85.1%	90.3%	96.3%
PoseConv3D [11]	89.6%	86.9%	93.7%	96.6%
MS-G3D [23]	88.4%	86.9%	91.5%	96.2%
TemPose-V (B)	85.1%	84.1%	91.0%	93.1%
TemPose-V (B+J)	88.5%	87.0%	92.7%	95.2%

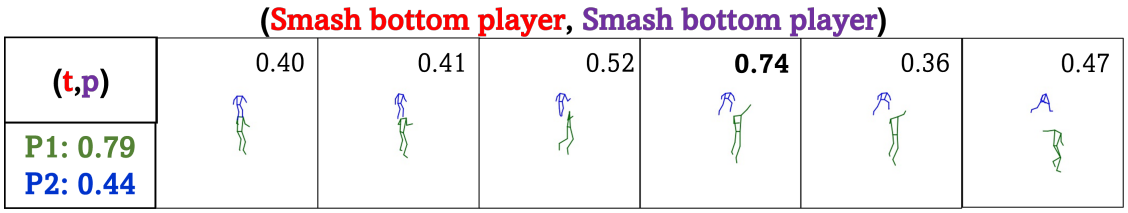

Figure 5. Prediction and attention score produced by TemPose-V for a skeleton sequence from Badminton Olympics. (t,p) refers to t as the target and p as the prediction. The interaction attention score is shown at the left, with the color corresponding to the person in the action sequence. The weighted temporal attention score is shown atop each frame in the sequence. For visual clarity, the frames are grouped by three, showing only the middle one, and the listed attention score is the average between them.

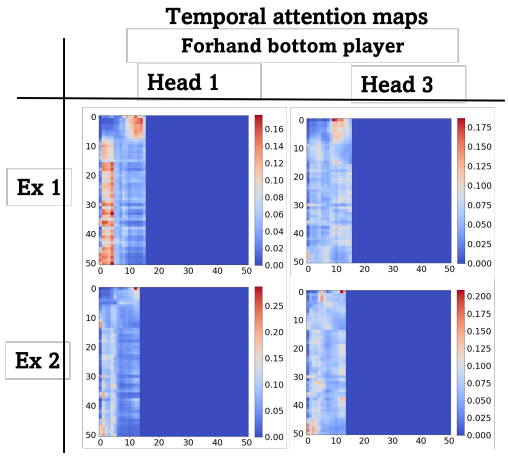

Figure 6. Temporal attention maps for a forehand by the bottom player (from Bad OL). The distribution of attention shows that TemPose prioritizes similar information when the actions are of the same class. Additionally, the attention maps also show the effect of the padding mask. The padding tokens are given no attention.

6. Conclusion

TemPose is a new skeleton-based action recognition model that uses temporal transformer layers to capture human motion dynamics and factorized interaction transformer layers to model the interaction between humans. The model outperforms existing methods in recognizing fine-grained badminton actions by fusing shuttlecock data, player court positions, and skeleton movements. It also achieves state-of-the-art performance on a large-scale action recognition dataset. Further studies will reveal how well the general TemPose architecture applies to other action recognition tasks.

Acknowledgements. We acknowledge the financial support from the Novo Nordisk Fonden as a part of Team-SPORTek, which enabled us to conduct this research. We would also like to thank Badminton Danmark and Team Danmark for their contributions to this study, including data collection and expert insights into badminton.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6816–6826, 2021. 1, 3, 4

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 813–824. PMLR, 2021. 3

[3] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. 3

[4] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 3

[5] Wei-Ta Chu and Samuel Situmeang. Badminton video analysis based on spatiotemporal and stroke features. In *Other Conferences*, pages 448–451, 06 2017. 2

[6] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 3

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. 1

[8] Lingxiao Dong, Dongmei Li, Shaobin Li, Shanzhen Lan, and Pengcheng Wang. Tai chi action recognition based on structural lstm with attention module. In *Other Conferences*, 2019. 1, 2

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 1, 3

[10] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1118, 2015. 2

[11] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2959–2968, 2022. 3, 6, 7, 8

[12] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021. 3

[13] Anurag Ghosh, Suriya Singh, and C. V. Jawahar. Towards structured analysis of broadcast badminton videos. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 296–304, 2018. 1, 2, 5, 7

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CoRR*, abs/2111.06377, 2021. 1

[15] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 4

[16] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. 3

[17] Kaustubh Milind Kulkarni and Sucheth Shenoy. Table tennis stroke recognition using two-dimensional human pose estimation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4571–4579, 2021. 1, 3

[18] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, 2022. 1, 3

[19] Jiatong Liu and Bo Liang. An Action Recognition Technology for Badminton Players Using Deep Learning. *Mobile Information Systems*, 2022:1–10, May 2022. 1

[20] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2684–2701, 2020. 2, 5

[21] Paul Liu and Jui-Hsien Wang. MonoTrack: Shuttle Trajectory Reconstruction From Monocular Badminton Video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, page 10, 2022. 1, 2

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 3

[23] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 140–149, Seattle, WA, USA, June 2020. IEEE. 1, 3, 6, 7, 8

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 6

[25] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. *In ICLR*, 2017. 6

[26] Filip Malawski and Bogdan Kwolek. Improving multimodal action representation with joint motion history context. *Journal of Visual Communication and Image Representation*, 61:198–208, 2019. 2

[27] Vittorio Mazzia, Simone Angarano, Francesco Salvetti, Federico Angelini, and Marcello Chiaberge. Action transformer: A self-attention model for short-time pose-based human action recognition. *Pattern Recognition*, 124:108487, 2022. 3, 6, 7

[28] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vision and Image Understanding*, 208-209:103219, 2021. 3, 8

[29] Nur Azmina Rahmad and Muhammad Amir As'ari. The new convolutional neural network (cnn) local feature extractor for automated badminton action recognition on vision based data. *Journal of Physics Conference Series*, 1529, 2020. 2

[30] Nur Azmina Rahmad, Muhammad Amir As'ari, Mohamad Fauzi Ibrahim, Nur Anis Jasmin Sufri, and Keerthana Rangasamy. Vision based automated badminton action recognition using the new local convolutional neural network extractor. In Mohd Hasnun Arif Hassan, Ahmad Munir Che Muhamed, Nur Fahriza Mohd Ali, Denise Koh Choon Lian, Kok Lian Yee, Nik Shanita Safii, Sarina Md Yusof, and Nor Farah Mohamad Fauzi, editors, *Enhancing Health and Sports Performance by Design*, pages 290–298, Singapore, 2020. Springer Singapore. 2

[31] N A Rahmad and M A As'ari. The new convolutional neural network (cnn) local feature extractor for automated badminton action recognition on vision based data. *Journal of Physics: Conference Series*, 1529(2):022021, apr 2020. 2

[32] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1010–1019, 2016. 2, 5

[33] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep High-Resolution Representation Learning for Human Pose Estimation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5686–5696, Long Beach, CA, USA, June 2019. IEEE. 3, 4

[34] Nien-En Sun, Yu-Ching Lin, Shao-Ping Chuang, Tzu-Han Hsu, Dung-Ru Yu, Ho-Yi Chung, and Tsì-Uí İk. Tracknetv2: Efficient shuttlecock tracking network. In *2020 International Conference on Pervasive Artificial Intelligence (ICPAI)*, pages 86–91, 2020. 3, 4

[35] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 2

[36] Pengcheng Wang and Shaobin Li. Structural-attentioned lstm for action recognition based on skeleton. In *Other Conferences*, 2018. 2

[37] Wenxiao Wang, Lu Yao, Long Chen, Binbin Lin, Deng Cai, Xiaofei He, and Wei Liu. Crossformer: A versatile vision transformer hinging on cross-scale attention. In *International Conference on Learning Representations, ICLR*, 2022. 3

[38] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *CoRR*, abs/2106.03348, 2021. 3

[39] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. 1, 3, 8

[40] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *AAAI*, 32, 2018. 2, 3, 7

[41] Dingyuan Zhu, Ziwei Zhang, Peng Cui, and Wenwu Zhu. Robust graph convolutional networks against adversarial attacks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 1399–1407, New York, NY, USA, 2019. Association for Computing Machinery. 3

[42] Kevin Zhu, Alexander Wong, and John McPhee. Fencenet: Fine-grained footwork recognition in fencing. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3588–3597, 2022. 1, 2, 3, 7