

STAR-Transformer: A Spatio-temporal Cross Attention Transformer for Human Action Recognition

Dasom Ahn¹, Sangwon Kim¹, Hyunsu Hong², Byoung Chul Ko^{1*}

¹Dept. of Computer Engineering, Keimyung University, Daegu, South Korea

²Difine, Seongnam, South Korea

{ektha772, eddiesangwonkim}@gmail.com, phil@difine.co.kr, niceko@kmu.ac.kr

Abstract

In action recognition, although the combination of spatio-temporal videos and skeleton features can improve the recognition performance, a separate model and balancing feature representation for cross-modal data are required. To solve these problems, we propose Spatio-Temporal cRoss (STAR)-transformer, which can effectively represent two cross-modal features as a recognizable vector. First, from the input video and skeleton sequence, video frames are output as global grid tokens and skeletons are output as joint map tokens, respectively. These tokens are then aggregated into multi-class tokens and input into STAR-transformer. The STAR-transformer encoder consists of a full spatio-temporal attention (FAttn) module and a proposed zigzag spatio-temporal attention (ZAttn) module. Similarly, the continuous decoder consists of a FAttn module and a proposed binary spatio-temporal attention (BAttn) module. STAR-transformer learns an efficient multi-feature representation of the spatio-temporal features by properly arranging pairings of the FAttn, ZAttn, and BAttn modules. Experimental results on the Penn-Action, NTU-RGB+D 60, and 120 datasets show that the proposed method achieves a promising improvement in performance in comparison to previous state-of-the-art methods.

1. Introduction

Action recognition is a traditional research topic that classifies human actions using video frames and has been applied in various applications, including human-robot interaction [3], healthcare [36], and video surveillance [18]. With the recent development of deep learning, action recognition research trends have been divided into three approaches. First, in a video-based approach [56, 4, 46, 30, 22, 33, 21], deep learning models use only video frames to

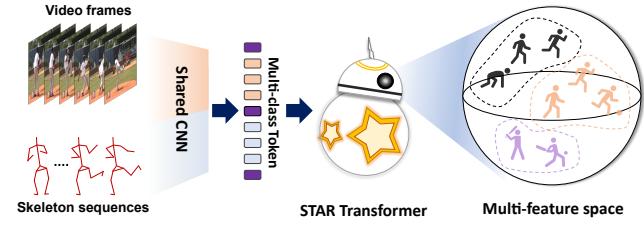


Figure 1. The proposed STAR-transformer takes a multi-class token as input and transforms it into a class-separable multi-feature representation. The multi-class token is an aggregation of global grid and joint map tokens obtained by feeding video and pose sequences into a shared CNN.

recognize the action. This approach results in a significant degradation in performance owing to various noises from the wild, such as differences in the camera angles and sizes of the human targets and complex backgrounds. The second is a skeleton-based approach [41, 6, 12, 17, 43, 19]. Actions are recognized using human skeletons and joint trajectories in different time zones as inputs into the deep learning model. However, an additional deep learning model is required to extract the human skeleton from an image. In addition, the action recognition is largely dependent on the accuracy of the skeleton extractor and the degree of overlap of the skeleton. The third approach is the use of cross-modal data, video, and skeletons together [15, 14, 39]. A deep learning model learns the RGB of the video frames and human skeletal features together; thus, it generally shows a high recognition performance. However, combining video and skeleton data is an ambiguous process and requires a separate submodel for cross-modal learning.

As a new learning paradigm in the deep learning field, Vision Transformer (ViT) [16] has recently attracted attention owing to its excellent performance in various computer vision fields such as image classification [26], image segmentation [48], object tracking [52], and action recognition [9]. The self-attention mechanism, which is a key element of ViT, is specialized for determining the spatial relation-

*Corresponding author

ship of each image and is effectively applied to image classification. However, in action recognition, the features of long-range frames and multiple features that change over time must both be considered; therefore, ViT based on the existing multi-head attention mechanism has limitations in terms of a high computational cost [27].

In this study, we propose a multi-feature representation method based on cross-modal learning and a Spatio-Temporal cRoss transformer (STAR-transformer) attention mechanism. For cross-modal learning, we propose a method of aggregating the cross-modal data of spatio-temporal video and a skeleton into a multi-class token to solve the problem of combining cross-modal action data. STAR-transformer consists of a new cross-attention module that replaces the multi-head attention of a simple ViT. The proposed STAR-transformer has demonstrated an excellent performance through various experiments.

Figure 1 shows the overall operational structure of STAR-transformer. Two cross-modal features are fed into the shared convolutional neural network (CNN) model and separated into multi-class tokens. STAR-transformer consists of an L -layer encoder-decoder output separable multi-class feature, which is used as input for the downstream action recognition network.

The contributions of this paper can be summarized as follows.

- Cross-modal learning: It is possible to flexibly aggregate spatio-temporal skeleton features as well as video frames and effectively learn cross-modal data to create multi-class tokens.
- STAR-transformer: The existing self-attention mechanism is limited to the application of action recognition because it focuses on the relation between spatial features. We therefore propose a STAR attention mechanism that can learn cross-modal features. The encoder and decoder of STAR-transformer are composed of zigzag and binary skip STAR attention.
- Various performance evaluation experiments: A performance evaluation was conducted based on several benchmark datasets, and the proposed model showed a better performance than existing state-of-the-art (SoTA) models.

2. Related Work

Video and image-based action recognition : It aims to recognize actions using only sequential [56, 4, 46, 30, 43, 19] or still images [22, 21]. The general process of video-based action recognition involves breaking the action into smaller semantic components and understanding the importance of each component in action recognition [21]. Because this method uses video frames, it can be processed using a simple single model. However, if the video is long, the recognition speed will be slow and the performance will be significantly affected by the various noises from the wild.

Skeleton-based action recognition: It aims to recognize actions by applying a list of spatio-temporal joint coordinates of video frames extracted from a pose estimator to a graph convolutional network (GCN) [41, 12, 7], 3D-CNN [17], and CNN [6]. A skeleton sequence has an advantage of being unaffected by contextual disturbances such as changes in background and lighting [17] but has a disadvantage in that the recognition performance is largely dependent on the pose extractor and requires an extra classifier for recognition.

Video and skeleton-based action recognition: It aims to achieve a high action recognition performance by fusing multi-modal (cross-modal) information into an integrated set of discriminative features [15, 14, 39, 35]. The video-pose network (VPN) action recognition mechanism [14], which uses cross-modal features and knowledge distillation to infuse poses into RGB streams, has proven that cross-modal features can achieve a better performance than unimodal features. Despite its relatively high recognition performance, this method still has problems in the design of a subnetwork for cross-modal learning and methods for combining cross-modal data.

Transformer-based action recognition: Because a transformer is a powerful tool in terms of long-range temporal modeling when using a self-attention module [2], an increasing number of studies in this area, particularly action recognition, have been conducted [2, 5, 42]. Most action recognition approaches using a transformer apply video frames as input tokens [45, 20, 47, 53, 50, 33], and relatively few methods use the skeleton [38, 34, 33] of the transformer. However, transformer-based action recognition often suffers from high computational costs owing to self-attention given to the large number of 3D tokens in a video [27]. Moreover, an approach to coupling cross-modal information using a transformer has yet to be developed. Therefore, this study is the first attempt at using spatio-temporal cross-modal data as input tokens for ViTs without applying separate sub-models.

3. Approach

Figure 2 shows the overall structure of the proposed action-recognition model based on the STAR-transformer module. Sixteen video frames and the corresponding skeleton sequences were received as input. Each frame goes through the pre-trained ResNet mixed convolution 18 (MC18) [43] to extract local and global feature maps. ResNet MC18 models are unsuitable for the proposed zigzag and binary operations because they reduce the video frame size after an operation. The global feature map, which is the output of the last layer of ResNet MC18, is transformed into a global grid token (GG-token) that represents the visual features of the images (Fig. 2 (a)). The local feature map, the output of the middle layer of ResNet

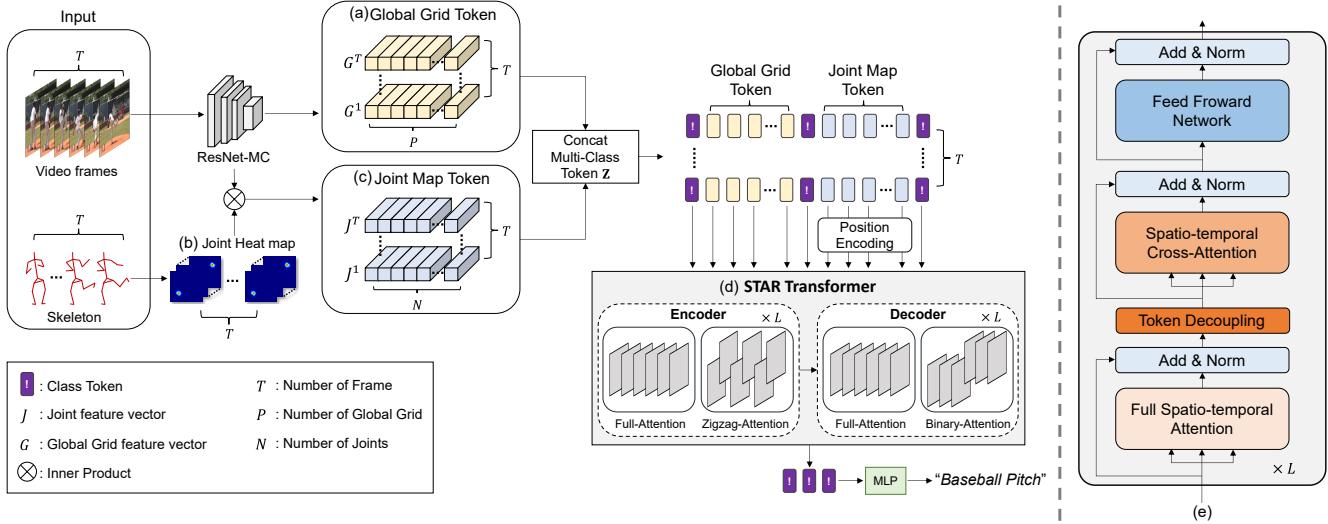


Figure 2. Overall architecture of the proposed action recognition model. (a) global grid token, (b) joint heat map, (c) joint map token, (d) STAR-transformer module, and (e) encoder and decoder structure of STAR transformer. An encoder and a decoder have the same structure.

MC18, is combined with the joint heat map (Fig. 2 (b)) and then transformed into a joint map token (JM-token), as shown in Fig. 2 (c). The JM tokens represent the local features of each skeleton joint. The two tokens are aggregated into a multi-class token and then fed into STAR-transformer, as shown in Fig. 2 (d), to infer the final action label.

3.1. Cross-Modal Learning

We first propose a cross-modal learning method that can combine video frames and skeleton features. The video frames are fed to ResNet MC18, and two feature maps are extracted from the middle and last layers. Because the feature map of the middle layer contains more detailed local features than the last layer, it is used for JM-token extraction, and the last layer is applied for GG-token extraction.

Global grid token (GG-token): Let a GG-token \mathbb{T}_g consisting of P tokens be $\mathbb{T}_g^t = \{g_1^t, \dots, g_P^t\}$ in video frame t . To extract the element of GG-token g_p^t from the t -th frame, the input frame is adjusted to a size of 224×224 , and the global feature map generated through ResNet MC18 has a size of $h \times w$. The global feature map is again flattened into a vector of size hw (P), which becomes the number of elements of \mathbb{T}_g^t . Because a global feature map consists of C channels, the number of dimensions of each element of \mathbb{T}_g^t is $g_P^t \in \mathbb{R}^C$. This process continues for every video frame, and thus we can obtain T temporal GG-tokens, as shown in Fig. 2 (a).

Joint map token (JM-token): In each t -th pose corresponding to the t -th video frame, we obtain N joint heat maps emphasizing joints for each frame and JM-token set $\mathbb{T}_j^t = \{j_1^t, \dots, j_N^t\}$ based on such maps. First, local feature

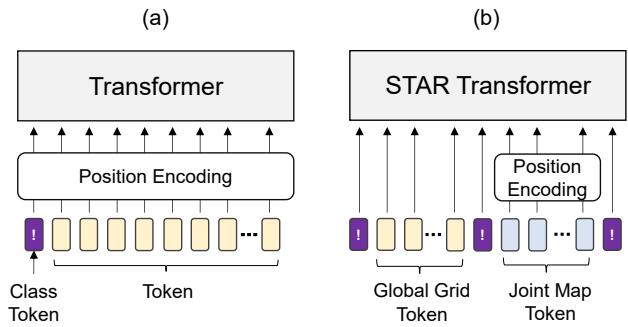


Figure 3. Multi-class token aggregation. (a) single class token generation of pure ViT and (b) proposed multi-token aggregation based on cross-modal learning.

maps $\mathbf{F} \in \mathbb{R}^{C' \times h' \times w'}$ of ResNet MC18 are obtained. The n -th joint heat map $h_n \in \mathbb{R}^{h' \times w'}$ is the result of projecting the n -th joint onto a temporary map with a size of $h' \times w'$ and applying Gaussian blurring at a scale of σ . Because a local feature map consists of C' channels, the number of dimensions of each joint element of \mathbb{T}_j^t is $j_N^t \in \mathbb{R}^{C'}$. The joint element j_n^t on the t -th pose is obtained through the concatenation (\oplus) of the local feature map \mathbf{F} and the n -th joint heat map h_n^t , as shown in the following equation:

$$\mathbb{T}_{j,n}^t = \oplus_{c'=1}^{C'} \left(\sum_i^{h'} \sum_j^{w'} \mathbf{F}_{c'}(i, j) \times h_n^t(i, j) \right). \quad (1)$$

This process continues for every pose sequence, and thus we can obtain T temporal JM-tokens, as shown in Fig. 2 (c).

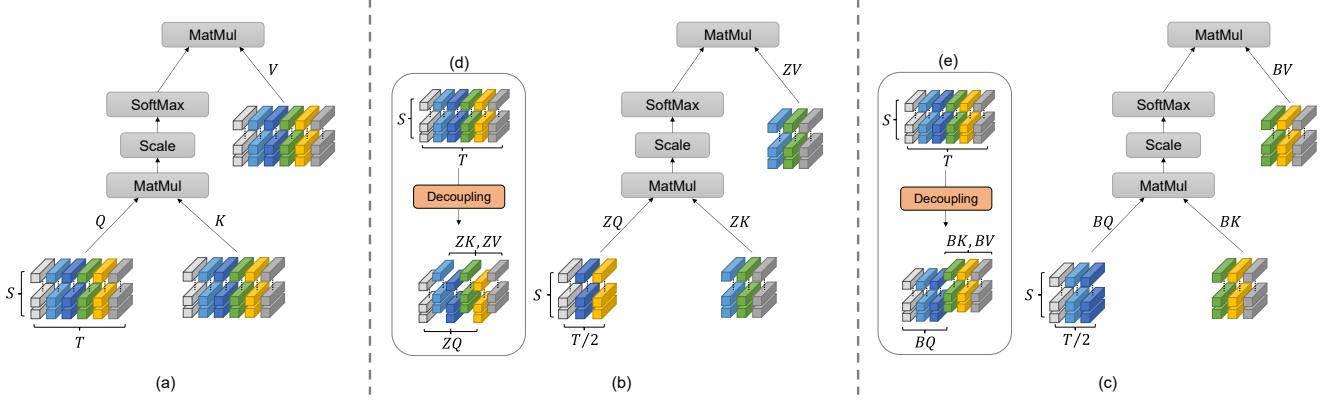


Figure 4. Proposed spatio-temporal cross attention modules. (a) full spatio-temporal attention (FAttn), (b) zigzag spatio-temporal attention (ZAttn), and (c) binary spatio-temporal attention (BAttn) modules.

Multi-class token: To aggregate the GG- and JM-tokens generated using cross-modal data, we propose a multi-class token aggregation, as shown in Fig. 3. Pure ViT [16], shown in Fig. 3 (a), focuses on learning the global relationships between the input tokens using single-class tokens. However, the proposed action recognition model must cooperatively learn the multi-class tokens generated from the cross-domain data. The proposed aggregation method of multi-class tokens therefore effectively learns the characteristics of different feature representations, as shown in Fig. 3 (b).

Multi-class token \mathbf{Z} is created by concatenating (\oplus) the class tokens for a GG-token (CLS_{glob}) and JM-token (CLS_{joint}) as follows:

$$\mathbb{T}_g = \text{CLS}_{glob} \oplus \mathbb{T}_g, \quad (2)$$

$$\mathbb{T}_j = \text{CLS}_{joint} \oplus (\mathbb{T}_j + \mathbf{pos}), \quad (3)$$

$$\mathbf{Z} = \mathbb{T}_g \oplus \mathbb{T}_j \oplus \text{CLS}_{total} \quad (4)$$

where CLS_{total} is the class token for all tokens. Unlike a GG-token, with a JM-token, the joint position information \mathbf{pos} is important, and thus \mathbf{pos} is only added to the JM-token.

3.2. Spatio-temporal cross attention

Inspired by [2], we first propose full spatio-temporal attention (FAttn), as shown in Fig. 4 (a), which applies the attention mechanism for all tokens within the spatio-temporal dimension. When FAttn is applied to all tokens of time dimension T and spatial dimension S , the complexity increases to $O(T^2S^2)$. However, because FAttn alone is insufficient to handle spatio-temporal features, we propose two additional cross-attention mechanisms, *i.e.*, zigzag spatio-temporal attention (ZAttn), as shown in Fig. 4 (b), and binary spatio-temporal attention (BAttn), as shown in

Fig. 4 (c). There is no need to pay attention to all tokens of time dimension T . Instead, all tokens are divided into two token groups for ZAttn and BAttn. When ZAttn and BAttn are applied to all tokens of the time dimension T and spatial dimension S , the computational complexity is reduced by 0.25-fold in comparison to FAttn with $O(\frac{1}{4}T^2S^2)$ because the tokens in the time dimension are divided into two groups.

We first obtain the same-sized query (Q), key (K), and value (V) $\in \mathbb{R}^{S \times T}$ matrices from the multi-class token \mathbf{Z} and compute the FAttn outputs as follows:

$$\hat{\mathbf{z}} = \text{FAttn}(Q, K, V) \quad (5)$$

$$\text{FAttn}(Q, K, V) = \sum_t^T \sum_s^S \text{Softmax} \left\{ \frac{Q_{s,t} \cdot K_{s,t}}{\sqrt{d_h}} \right\} V_{s,t} \quad (6)$$

ZAttn learns the detailed process of changing actions. To calculate ZAttn, the odd-numbered vectors in \mathbf{Z} are divided into $ZQ' \in \mathbb{R}^{S \times T/2}$, and the even-numbered vectors in \mathbf{Z} are divided into ZK' and $ZV' \in \mathbb{R}^{S \times T/2}$ in a zigzag manner, as shown in Fig. 4 (b). By contrast, the odd-numbered tokens in \mathbf{Z} are divided into ZK'' and $ZV'' \in \mathbb{R}^{S \times T/2}$, and the even-numbered vectors in \mathbf{Z} are divided into $ZQ'' \in \mathbb{R}^{S \times T/2}$.

We calculate $a' \in \mathbb{R}^{S \times T/2}$ and $a'' \in \mathbb{R}^{S \times T/2}$ individually using the two types of matrices extracted in a zigzag manner using the following formulas, and then concatenate the outputs a' and a'' as the result of ZAttn.

$$a' = \sum_t^{T/2} \sum_s^S \text{Softmax} \left\{ \frac{ZQ'_{s,t} \cdot ZK'_{s,t}}{\sqrt{d_h}} \right\} ZV'_{s,t} \quad (7)$$

$$a'' = \sum_t^{T/2} \sum_s^S \text{Softmax} \left\{ \frac{ZQ''_{s,t} \cdot ZK''_{s,t}}{\sqrt{d_h}} \right\} ZV''_{s,t} \quad (8)$$

$$\text{ZAttn}(Q, K, V) = a' \oplus a'' \quad (9)$$

BAttn is also generated into two groups by dividing the time-dimensional tokens back and forth, as shown in Fig. 4 (c). Through this process, it is possible to learn the change at the beginning and end of the action. In the case of BAttn, after dividing \mathbf{Z} into two groups in a binary manner, the front and rear vectors, $BQ' \in \mathbb{R}^{S \times T/2}$ and BK' , respectively, and $BV' \in \mathbb{R}^{S \times T/2}$ matrices are calculated. By contrast, the front vectors in \mathbf{Z} are divided into BK'' and $BV'' \in \mathbb{R}^{S \times T/2}$, and the rear vectors in \mathbf{Z} are divided into $BQ'' \in \mathbb{R}^{S \times T/2}$. We calculate the individual $b' \in \mathbb{R}^{S \times T/2}$ and $b'' \in \mathbb{R}^{S \times T/2}$ using the two types of matrices with the same formula of ZAttn, and concatenate the output b' and b'' as the result of BAttn.

$$\text{BAttn}(Q, K, V) = b' \oplus b'' \quad (10)$$

3.3. STAR-transformer encoder and decoder

The proposed STAR-transformer follows a encoder-decoder structure of pure transformer [44] than pure ViT [16], as shown in Fig. 2 (e). However, the encoder is composed of a series of FAttn (self-attention) and ZAttn L layers, and the decoder is composed of a series of FAttn and BAttn layers. The encoder uses ZAttn to focus on the learning relationships for detailed changes in action, and the decoders use BAttn to learn the relationships for large changes in action.

The structure of the STAR-transformer layer is as follows.

$$\bar{z}_l = \text{LN}\{\text{FSTA}(z_{l-1}) + z_{l-1}\}, l \in \{1, 2, \dots, L\} \quad (11)$$

$$z'_l, z''_l = \text{Decoupling}(\bar{z}_l) \quad (12)$$

$$\tilde{z}_l = \text{LN}\{(\text{STA}(z'_l) + z'_l) \oplus (\text{STA}(z''_l) + z''_l)\} \quad (13)$$

$$z_l = \text{LN}\{\text{MLP}(\tilde{z}_l) + \tilde{z}_l\} \quad (14)$$

Here, l is the number of transformer layers, LN is the layer normalization, and FSTA is the multi-head self-attention for FAttn. Decoupling refers to zigzag or binary grouping. STA represents spatio-temporal attention for ZAttn and BAttn, and MLP is a multi-layer perceptron.

The multi-class tokens output by STAR-transformer are combined into a single class token by averaging and feeding into the MLP to infer the final action label.

Method	Pre-train	Feature		Acc. (%)
		RGB	Annot Pose	
3D Deep [8]	✗	✓	✓	98.1
PoseMap [29]	✗	✓	✓	98.2
Multitask CNN [32]	✗	✓	✓	98.6
HDM-BG [55]	✗		✓	93.4
Pr-VIPE [40]	✓		✓	97.5
UNIK [49]	✓		✓	97.9
STAR-Transformer	✗	✓	✓	98.7

Table 1. Performance comparison with other state-of-the-art methods on the Penn-Action dataset (Annot, annotated pose(skeleton); Acc, accuracy).

4. Experimental Results

In this section, we describe the implementation details, including the dataset and training hyperparameters applied. After conducting a quantitative analysis based on SoTA approaches, ablation studies and a qualitative analysis were applied on the effectiveness of multi-expression learning, the number of transformer layers, and spatio-temporal cross attention.

4.1. Experiment Setup

Dataset Description: The experiment was conducted using the representative action recognition datasets, Penn-Action [54], NTU-RGB+D 60 [37], and 120 [28]. The Penn-Action dataset includes 15 different action classes, such as *baseball swings*, *jumping jacks*, and *pushups*, for a total of 2,326 RGB video sequences. The NTU-RGB+D 60 dataset is a large dataset used for human action recognition containing 56,880 samples of 60 action classes collected from 40 subjects. Actions are divided into three categories having 40 daily actions (e.g., *drinking*, *eating*, and *reading*), 9 health-related actions (e.g., *sneezing*, *staggering*, and *falling*), and 11 mutual actions (e.g., *punching*, *kicking*, and *hugging*), respectively, based on multi-modal information of the action characterization, including depth maps, 3D skeletal joint positions, RGB frames, and infrared sequences. NTU-RGB+D 60 has two evaluation protocols, cross-subject (XSub) and cross-view (XView). NTU-RGB+D 120 extends this version of NTU-RGB+D 60 by adding another 60 classes and containing 114,480 samples in total. NTU-RGB+D 120 has two evaluation protocols, XSub and cross-setup (XSet).

Implementation details: The proposed STAR-transformer was implemented using PyTorch, and ResNet MC18 pre-trained with Kinetics-400 was applied as the backbone network. When training the model, the Penn-Action and NTU-RGB+D datasets used 16 fixed frames. For all datasets, we utilized a batch size of 4, 300 epochs, an stochastic gradient descent (SGD) optimizer, a learning

Method	Pre-training	Feature			NTU60		NTU120	
		RGB	Est. Pose	Annot. Pose	XSub	XView	XSub	XSet
PoseMap [29]	✗	✓	✓		91.7	95.2	-	-
MMTM [35]	✗	✓		✓	91.9	-	-	-
VPN [14]	✗	✓		✓	95.5	98.0	86.3	87.8
DualHead-Net [10]	✗			✓	92.0	96.6	88.2	89.3
Skeletal GNN [51]	✗			✓	91.6	96.7	87.5	89.2
CTR-GCN [11]	✗			✓	92.4	96.8	88.9	90.6
InfoGCN [13]	✗			✓	93.0	97.1	89.8	91.2
3s-AimCLR [23]	✓			✓	86.9	92.8	80.1	80.9
PoseC3D [17]	✓	✓	✓		97.0	99.6	95.3	96.4
KA-AGTN [31]	✗			✓	90.4	96.1	86.1	88.0
STAR-Transformer	✗	✓		✓	92.0	96.5	90.3	92.7

Table 2. Performance comparison with other state-of-the-art methods on the NTU-RGB+D Action dataset (Est, estimated pose(skeleton); XSub,cross-subject; XView, cross-view; XSet, cross-setup).

rate of 2e-4, and a momentum of 0.9. The experiments were conducted in an environment configured with four NVIDIA Tesla V100 GPUs.

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

Penn-Action Dataset: Table 1 shows the results of the comparison experiments with other SoTA action recognition technologies for the Penn-Action dataset: 1) body joint guided 3D deep convolutional descriptors (3D Deep) [8], 2) evolution of pose estimation maps (EV-Pose) [29], 3) multitask CNN [32], 4) Bayesian hierarchical dynamic model (HDM-BG) [55], 5) view-invariant probabilistic embedding (Pr-VIPE) [40], and 6) a unified framework for skeleton-based action recognition (UNIK) [49]. UNIK [49] was pre-trained using the Posetics dataset reconstructed from the Kinect-400 [25] dataset, and Pr-VIPE [40] was pretrained using the Human3.6M dataset [24]. STAR-transformer and the other methods were trained and tested only on the given data, without any pre-training.

The pre-trained UNIK [49] model showed a 0.8% lower accuracy than the proposed model at 97.9%, and the Pr-VIPE [40] model showed 97.5% accuracy, which is 1.2% lower than that of the proposed model.

During the experiment, STAR-transformer and the three methods [8, 29, 32] using the RGB of the video frames and the pose (skeleton) feature together showed a high overall performance of 98% or higher. However, the three methods [55, 40, 49] using only the pose feature showed a relatively low performance of 93% to 97%. As the results in Table 1 indicate, we can confirm that the action recognition performance can be improved when the RGB of the video frames and pose features are used together. Although STAR-transformer did not use any pre-training, the highest accuracy was derived through the proposed cross-attention using the cross-modal features together.

NTU-RGB+D Dataset: Table 2 shows the results of the comparison experiments with SoTA action recognition technologies when applying the NTU-RGB+D dataset: 1) long-term localization using 3D LiDARs (PoseMap) [29], 2) multimodal transfer module (MMTM) [35], 3) video-pose embedding (VPN) [14], 4) multi-granular spatio-temporal graph network (DualHead-Net) [10], 5) skeletal graph neural networks (Skeletal GNN) [51], 6) channel-wise topology refinement GCN (CTR-GCN) [11], 7) information bottleneck-based GCN (InfoGCN) [13], 8) contrastive learning (3s-AimCLR) [23], 9) 3D skeleton and heatmap stack (PoseC3D) [17], and 10) kernel attention adaptive graph transformer network (KA-AGTN) [31]. Because a transformer-based action recognition method that uses RGB of the video frames and cross-modal features of skeleton together has not yet been published, we compare the performance with KA-AGTN, a SoTA for skeleton and transformer-based action recognition.

During this experiment, the accuracy was measured separately for the NTU-RGB+D 60 and NTU-RGB+D 120 datasets, and the cross-subject (XSub), cross-view (XView), and cross-setup (XSet) were measured separately for each dataset. The performances of four methods [17, 14, 35, 29] using the RGB of the video frames and pose together, and six methods using only the pose [31, 10, 51, 11, 13, 23], were compared with STAR-transformer. Pre-training was conducted using only 3s-AimCLR [23]. As shown in Table 2, the accuracy was higher for NTU60 and NTU120 when the cross-modal features of the RGB and pose were used together than when a unimodal feature was applied. PoseC3D [17] performed 5% better on NTU60 XSub and 3.7% better on NTU120 XSet than the proposed STAR-transformer because PoseC3D did not use annotated poses but applied a separate pre-trained poseConv3D model for 3D pose estimation to achieve better action recognition. As

Multi-Class token	Accuracy(%)
✗	97.3
✓	98.7

Table 3. Effectiveness of multi-class token.

the results indicate, PoseC3D achieved a relatively high accuracy because it extracted the optimal pose features suitable for its own model and used them for learning. However, this method still has certain disadvantages in that it requires a pre-trained model for additional pose detection, and the pose detection and action recognition models cannot be trained end-to-end as a single model. KA-AGTN [31] used a transformer structure as in our method. However, because it uses only skeleton information and the transformer is used only for spatial information processing between joints, the performance is inferior to the proposed method by up to 1.6% for NTU60 and up to 4.7% for NTU120.

Although transformers need to be pre-trained using a large dataset, the proposed STAR-transformer combines the RGB and annotated poses without any pre-training, achieving a promising accuracy even on a larger class dataset NTU120. In particular, NTU120 XSub and XSet showed the second highest performance with accuracy rates of 90.3% and 92.7%, respectively. This indicates that STAR-transformer is capable of an excellent action recognition, although the action class is increased or the cross view is changed.

4.3. Ablation Study

In this section, the detailed performance of the modules constituting the proposed STAR transformer model is verified based on several experiments. All experiments were conducted using the Penn-Action dataset.

Effectiveness of multi-expression learning: To confirm the effect of the proposed multi-expression learning, Table 3 presents a comparison experiment conducted with a single-class token used in pure ViT [16] and the multi-class token proposed in this study. The proposed multi-class token performed 1.4% higher than a single-class token. Although the existing single-class tokens did not effectively conduct learning between the cross-modal tokens, it was confirmed that the proposed multi-class token can effectively increase the performance of the model under the same cross-modal condition.

Effectiveness of the number of transformer layers: Figure 5 shows the difference in performance according to the number of transformer layers for the proposed spatio-temporal cross-attention module structure. As shown in Fig. 5, the overall performance improves as the number of layers increases; however, when there are more than four

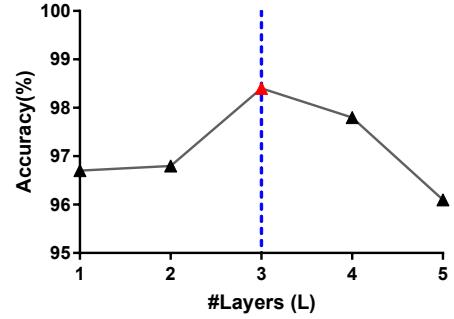


Figure 5. Variation in accuracy according to the number of spatio-temporal cross attention layers.

Attention Struture		Accuracy (%)
Encoder	Decoder	
F-F	F-F	96.1
F-Z	F-Z	97.3
F-B	F-B	97.8
F-B	F-Z	97.3
F-Z		98.7

Table 4. Difference in accuracy between spatio-temporal cross-attention mechanisms. F represents FAttn, Z indicates ZAttn, and B represents the BAttn.

layers, the model is easily overfitted. Therefore, based on the experimental results, we set the number of transformer layers to three.

Effectiveness of spatio-temporal cross attention: Figure 6 shows the relative frame importance score for the spatio-temporal cross-attention mechanism proposed in this study. The scores were calculated using an attention rollout [1] to calculate the relative concentration for each frame. The attention rollout recursively receives the embedding attention as the input for each layer of the transformer model and computes the token attention.

In Fig. 6 (a), FAttn, which has a structure similar to that of ViTs, shows a high score in the last frame, indicating that the front frames are barely considered in action recognition. Here, only the final top-three frames of *after throwing the ball* significantly contributed to the performance, and thus we can see that FAttn does not consider the overall temporal characteristics.

In the case of ZAttn and BAttn, which are spatio-temporal cross-attention mechanisms, the importance scores are equally high in all frames, as shown in Fig. 6 (b) and (c). When checking the top-three frames of ZAttn and BAttn, sequentially varied frames such as *before throwing the ball*, *while throwing the ball*, and *after throwing the ball*

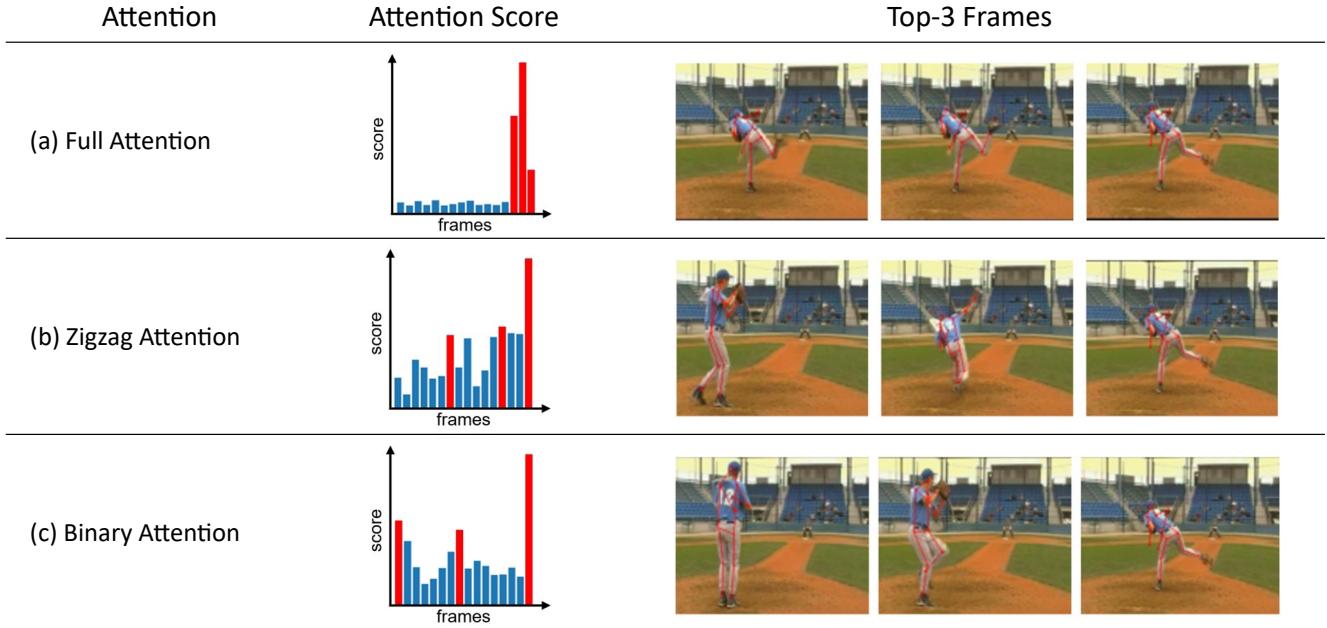


Figure 6. The relative importance score of 16 input frames of a validation video. The bar graph shows the attention score for each frame. When only full spatio-temporal attention is used, the attention score appears high at the end of the action. In the case of zigzag spatio-temporal attention, high attention scores were obtained in the middle and last frames when the action was large. In the case of binary spatio-temporal attention, a high attention score appears in the entire frame of an action.

ball are considered for a performance improvement.

Difference in accuracy of cross attention modules: Table 4 shows the differences in accuracy according to the structure of the spatio-temporal cross-attention module. Based on the experimental results, we can see that when ZAttn (Z) and BAttn (B) are used together, the performance is higher than when FAttn is used alone. When FAttn and BAttn were used equally for the encoder and decoder (F-B, F-B), the second highest accuracy was achieved at 97.8%; however, the importance of the entire frame was still not accurately reflected, and thus the performance was slightly lower than that of the F-Z and F-B combinations. The combination of F-B and F-Z, in which BAttn is applied to the encoder and ZAttn is applied to the decoder, showed the second-lowest performance at 97.3%. Based on the experimental results, we used F-Z as the encoder and F-B as the decoder.

Through these experimental results, we can see that for an accurate action recognition, it is necessary to learn the frame characteristics evenly in all frames through the proposed spatio-temporal cross attention mechanism.

5. Conclusion

In this paper, we proposed STAR-transformer, an algorithm based on a spatial-temporal cross-attention module that simultaneously uses video frames and skeleton-based

features for action recognition. In addition, the proposed multi-feature representation learning approach was able to flexibly combine the RGB of the video frames, skeleton, and joint trajectories using multi-class tokens. As a result of testing the proposed algorithm using the Penn-Action and NTU-RGB+D action datasets, it was confirmed that the proposed STAR-transformer model achieved substantial improvements in comparison to previous SoTA methods. In a future study, we plan to develop an algorithm that can efficiently learn a model without an overfitting, even with a small number of data. In addition, by extending the proposed STAR-transformer to a model that combines a pose estimation rather than annotated poses, we will modify the STAR-transformer into an end-to-end model that can simultaneously apply pose feature estimation and action recognition optimized for action recognition.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2022R1I1A3058128), and partly supported by the Scholar Research Grant of Keimyung University in 2022.

References

[1] S. Abnar and W. Zuidema. Quantifying attention flow in transformers. In *The Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online, July 2020. Association for Computational Linguistics.

[2] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid. Vivit: A video vision transformer. In *The International Conference on Computer Vision (ICCV)*, pages 6836–6846, 2021.

[3] C. Bandi and U. Thomas. Skeleton-based action recognition for human-robot interaction using self-attention mechanism. In *The International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2021.

[4] F. Baradel, C. Wolf, J. Mille, and G. W Taylor. Glimpse clouds: Human activity recognition from unstructured feature points. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 469–478, 2018.

[5] G. Bertasius, H. Wang, and L. Torresani. Is space-time attention all you need for video understanding? In *The International Conference on Machine Learning (ICML)*, volume 2, page 4, 2021.

[6] C. Caetano, J. Sena, F. Brémont, J. A Dos S., and W. R. Schwartz. Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition. In *The Conference on Advanced Video and Signal based Surveillance (AVSS)*, pages 1–8. IEEE, 2019.

[7] J. Cai, N. Jiang, X. Han, K. Jia, and J. Lu. Jolo-gcn: mining joint-centered light-weight information for skeleton-based action recognition. In *The Winter Conference on Applications of Computer Vision (WACV)*, pages 2735–2744, 2021.

[8] C. Cao, Y. Zhang, C. Zhang, and H. Lu. Body joint guided 3-d deep convolutional descriptors for action recognition. *IEEE Transactions on Cybernetics*, 48(3):1095–1108, 2017.

[9] J. Chen and C. M. Ho. Mm-vit: Multi-modal video transformer for compressed video action recognition. In *The Winter Conference on Applications of Computer Vision (WACV)*, pages 1910–1921, 2022.

[10] T. Chen, D. Zhou, J. Wang, S. Wang, Y. Guan, X. He, and E. Ding. Learning multi-granular spatio-temporal graph network for skeleton-based action recognition. In *The ACM International Conference on Multimedia*, pages 4334–4342, 2021.

[11] 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 *The International Conference on Computer Vision (ICCV)*, pages 13359–13368, 2021.

[12] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu. Skeleton-based action recognition with shift graph convolutional network. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 183–192, 2020.

[13] H.-g. Chi, M. H. Ha, S. Chi, S. W. Lee, Q. Huang, and K. Ramani. Infogen: Representation learning for human skeleton-based action recognition. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20186–20196, 2022.

[14] S. Das, S. Sharma, R. Dai, F. Bremond, and M. Thonnat. Vpn: Learning video-pose embedding for activities of daily living. In *The European Conference on Computer Vision (ECCV)*, pages 72–90. Springer, 2020.

[15] M. Davoodikakhki and K. Yin. Hierarchical action classification with network pruning. In *International Symposium on Visual Computing*, pages 291–305. Springer, 2020.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *The International Conference on Learning Representations (ICLR)*, pages 72–90, 2020.

[17] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai. Revisiting skeleton-based action recognition. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2969–2978, 2022.

[18] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, A. Bouridane, and A. Beghdadi. A combined multiple action recognition and summarization for surveillance video sequences. *Applied Intelligence*, 51(2):690–712, 2021.

[19] C. Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 203–213, 2020.

[20] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman. Video action transformer network. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 244–253, 2019.

[21] D. Girish, V. Singh, and A. Ralescu. Understanding action recognition in still images. In *The Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 370–371, 2020.

[22] G. Guo and A. Lai. A survey on still image based human action recognition. *Pattern Recognition*, 47(10):3343–3361, 2014.

[23] T. Guo, H. Liu, Z. Chen, M. Liu, T. Wang, and R. Ding. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In *The AAAI Conference on Artificial Intelligence*, volume 36, pages 762–770, 2022.

[24] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(7):1325–1339, 2013.

[25] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[26] S. Kim, J. Nam, and B. C. Ko. Vit-net: Interpretable vision transformers with neural tree decoder. In *The International Conference on Machine Learning (ICML)*, pages 1–13, 2022.

[27] Y. Liang, P. Zhou, R. Zimmermann, and S. Yan. Dualformer: Local-global stratified transformer for efficient video recognition. *arXiv preprint arXiv:2112.04674*, 2021.

[28] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d

human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(10):2684–2701, 2019.

[29] M. Liu and J. Yuan. Recognizing human actions as the evolution of pose estimation maps. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1159–1168, 2018.

[30] X. Liu, S. L. Pintea, F. K. Nejadasl, O. Booij, and J. C. van G. No frame left behind: Full video action recognition. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14892–14901, 2021.

[31] Y. Liu, H. Zhang, D. Xu, and K. He. Graph transformer network with temporal kernel attention for skeleton-based action recognition. *Knowledge-Based Systems*, 240:108146, 2022.

[32] D. C. Luvizon, D. Picard, and H. Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5137–5146, 2018.

[33] V. Mazzia, S. Angarano, F. Salvetti, F. Angelini, and M. Chiarberge. Action transformer: A self-attention model for short-time pose-based human action recognition. *Pattern Recognition*, 124:108487, 2022.

[34] C. Plizzari, M. Cannici, and M. Matteucci. Skeleton-based action recognition via spatial and temporal transformer networks. *The Computer Vision and Image Understanding (CVIU)*, 208:103219, 2021.

[35] V. Reza, H. Jozé, A. Shaban, M. L. Iuzzolino, and K. Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13289–13299, 2020.

[36] F. Serpush, M. B. Menhaj, B. Masoumi, and B. Karasfi. Wearable sensor-based human activity recognition in the smart healthcare system. *Computational Intelligence and Neuroscience*, 2022, 2022.

[37] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1010–1019, 2016.

[38] L. Shi, Y. Zhang, J. Cheng, and H. Lu. Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In *The Asian Conference on Computer Vision (ACCV)*, 2020.

[39] L. Su, C. Hu, G. Li, and D. Cao. Msaf: Multimodal split attention fusion. *arXiv preprint arXiv:2012.07175*, 2020.

[40] J. Sun, J. Zhao, L.-C. Chen, F. Schroff, H. Adam, and T. Liu. View-invariant probabilistic embedding for human pose. In *The European Conference on Computer Vision (ECCV)*, pages 53–70. Springer, 2020.

[41] K. Thakkar and PJ Narayanan. Part-based graph convolutional network for action recognition. In *The British Machine Vision Conference (BMVC)*, pages 1–13, 2018.

[42] Z. Tong, Y. Song, J. Wang, and L. Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022.

[43] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6450–6459, 2018.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *The Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

[45] X. Wang, S. Zhang, Z. Qing, Y. Shao, Z. Zuo, C. Gao, and N. Sang. Oadtr: Online action detection with transformers. In *The International Conference on Computer Vision (ICCV)*, pages 7565–7575, 2021.

[46] Z. Wang, Q. She, and A. Smolic. Action-net: Multipath excitation for action recognition. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13214–13223, 2021.

[47] M. Xu, Y. Xiong, H. Chen, X. Li, W. Xia, Z. Tu, and S. Soatto. Long short-term transformer for online action detection. *The Advances in Neural Information Processing Systems (NeurIPS)*, 34:1086–1099, 2021.

[48] X. Yan, H. Tang, S. Sun, H. Ma, D. Kong, and X. Xie. After-unet: Axial fusion transformer unet for medical image segmentation. In *The Winter Conference on Applications of Computer Vision (WACV)*, pages 3971–3981, 2022.

[49] D. Yang, Y. Wang, A. Dantcheva, L. Garattoni, G. Francesca, and F. Bremond. Unik: A unified framework for real-world skeleton-based action recognition. In *The British Machine Vision Conference (BMVC)*, pages 1–13, 2021.

[50] F. Yi, H. Wen, and T. Jiang. Asformer: Transformer for action segmentation. In *The British Machine Vision Conference (BMVC)*, pages 1–15, 2021.

[51] A. Zeng, X. Sun, L. Yang, N. Zhao, M. Liu, and Q. Xu. Learning skeletal graph neural networks for hard 3d pose estimation. In *The International Conference on Computer Vision (ICCV)*, pages 11436–11445, 2021.

[52] F. Zeng, B. Dong, T. Wang, X. Zhang, and Y. Wei. Motr: End-to-end multiple-object tracking with transformer. *arXiv preprint arXiv:2105.03247*, 2021.

[53] B. Zhang, J. Yu, C. Fifty, W. Han, A. M. Dai, R. Pang, and F. Sha. Co-training transformer with videos and images improves action recognition. *arXiv preprint arXiv:2112.07175*, 2021.

[54] W. Zhang, M. Zhu, and K. G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *The International Conference on Computer Vision (ICCV)*, pages 2248–2255, 2013.

[55] R. Zhao, W. Xu, H. Su, and Q. Ji. Bayesian hierarchical dynamic model for human action recognition. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7733–7742, 2019.

[56] J. Zhu, W. Zou, L. Xu, Y. Hu, Z. Zhu, M. Chang, J. Huang, G. Huang, and D. Du. Action machine: Rethinking action recognition in trimmed videos. *arXiv preprint arXiv:1812.05770*, 2018.