Cross-Modal Learning with 3D Deformable Attention for Action Recognition

Sangwon Kim Dasom Ahn Byoung Chul Ko*
Keimyung University
{eddiesangwonkim, tommydasomahn}@gmail.com, niceko@kmu.ac.kr

Abstract

An important challenge in vision-based action recognition is the embedding of spatiotemporal features with two or more heterogeneous modalities into a single feature. In this study, we propose a new 3D deformable transformer for action recognition with adaptive spatiotemporal receptive fields and a cross-modal learning scheme. The 3D deformable transformer consists of three attention modules: 3D deformability, local joint stride, and temporal stride attention. The two cross-modal tokens are input into the 3D deformable attention module to create a cross-attention token with a reflected spatiotemporal correlation. Local joint stride attention is applied to spatially combine attention and pose tokens. Temporal stride attention temporally reduces the number of input tokens in the attention module and supports temporal expression learning without the simultaneous use of all tokens. The deformable transformer iterates L-times and combines the last cross-modal token for classification. The proposed 3D deformable transformer was tested on the NTU60, NTU120, FineGYM, and PennAction datasets, and showed results better than or similar to pre-trained state-of-the-art methods even without a pre-training process. In addition, by visualizing important joints and correlations during action recognition through spatial joint and temporal stride attention, the possibility of achieving an explainable potential for action recognition is presented.

1. Introduction

Spatiotemporal feature learning is a crucial part of action recognition, which aims to fuse not only the spatial features of each frame but also the temporal correlation between input sequences. Previous studies on action recognition [19, 6, 5, 42, 9, 48] investigated the application of 3D convolutional kernels with an additional temporal space beyond the 2D spatial feature space. Since then, 3D convolutional neural networks (CNN) have achieved a promising performance and have eventually become the *de facto*

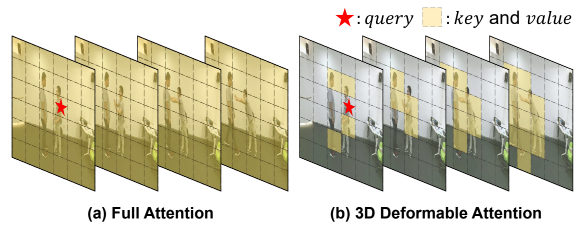


Figure 1: **Comparison between (a) Full attention and (b) the proposed 3D deformable attention.** Full attention considers all tokens against a specific *query* in a complete sequence. By contrast, 3D deformable attention considers only intense tokens with adaptive receptive fields.

standard for various action recognition tasks using sequential data. Vision transformers (ViTs) for action recognition, which have peaked in popularity, have recently been used to explore a 3D token embedding to fuse the temporal space within a single token. However, ViTs-based action recognition methods [1, 34] are limited in that they can only conduct spatiotemporal feature learning within restricted receptive fields.

To avoid this problem, several studies [15, 57, 47] have been conducted to allow more flexible receptive fields for deep learning models. Deformable CNN leverages dynamic kernels to capture the intense object regions. First, they determine the deformable coordinates using embedded features. The kernel is then applied to the features extracted from the deformable coordinates. Deformable ViTs [47, 57] encourage the use of an existing attention module to learn deformable features. The *query* tokens are projected onto the coordinates to obtain deformable regions from the *key* and *value* tokens. The deformed value tokens are then applied to the attention map, which is generated through a scaled dot product of the input *query* and deformed *key* tokens. These methods suggest a new approach that can overcome the limitations of existing standardized feature learning. However, despite some impressive results, these studies are still limited in that they are only compatible with the spatial dimensions. Therefore, as a primary challenge, there is a need for the development of novel and deformable

*Corresponding author

ViTs that can learn spatiotemporal features from image sequences.

Another challenge is the efficient application of multimodal input features to an action recognition model. Action recognition is classified into three categories based on the feature type. The first is a video-based approach [56, 4, 46, 29, 20, 43, 33], which has traditionally been used for action recognition. This approach is limited by a degraded performance caused by noise, such as varying object sizes, occlusions, or different camera angles. The second is a skeleton-based approach [51, 25, 12, 13, 11], which mainly converts poses into graphs for recognizing actions through a graph neural network (GNN). Although this approach is robust against noise, its performance is highly dependent on the pose extraction method. To overcome the shortcomings of the previous two approaches, the third method aims to simultaneously fuse heterogeneous domain features using multimodal or cross-modal learning. With this approach, video and skeleton features are jointly trained simultaneously. However, because most related studies use a separate model composed of a GNN + CNN or CNN + CNN for each modality, there is a limit in constructing an effective single model.

To alleviate the drawbacks stated above, we propose the use of transformer with 3D deformable attention for dynamically utilizing the spatiotemporal features for action recognition. In this way, the proposed model applies flexible cross-modal learning, which handles the skeletons and video frames in a single transformer model. The skeletons are projected onto sequential joint tokens, and each joint token contains an activation at every joint coordinate. To provide effective cross-modal learning between each modality, the proposed method adopts a cross-modal token that takes the role of mutually exchanging contextual information. Therefore, the proposed model is capable of achieving a boosted performance without an auxiliary submodel for the cross-modalities. Figure 1 shows a comparison between the previous full attention and the proposed 3D deformable attention. In the case of the full attention shown in Fig. 1 (a), all tokens within a spatiotemporal space are covered against a specific *query* token. By contrast, our proposed 3D deformable attention scheme, shown in Fig. 1 (b), considers only tokens with high relevance within the entire spatiotemporal space. The main contributions of this study are as follows:

- We propose the first 3D deformable attention that adaptively considers the spatiotemporal correlation within a transformer as shown in Fig. 1 (b), breaking away from previous studies that consider all tokens against a specific *query* in a complete sequence.
- We propose a cross-modal learning scheme based on complementary cross-modal tokens. Each cross-modal token delivers contextual information between the different modalities. This approach can support a simple yet effective cross-modal learning within a single-transformer model structure.
- We present qualitative evidence for 3D deformable attention with visual explanations and prove that the proposed model outperforms several previous state-of-the-art (SoTA) methods.

2. Related Works

Spatiotemporal learning for action recognition. Early studies in this area focused primarily on employing a 3D CNN, which is an extension of a 2D CNN. This has become a central remedy in vision-based action recognition in recent years. PoseC3D [19] combines 3D volumetric heatmaps from the skeletons and frames of the input video. SlowFast [21] makes a significant contribution to the field by providing a frame-fusion scheme between different frame rates. There are also related methods [22, 42, 43, 44, 9, 48, 4, 20, 24] that explore the use of a 3D CNN architecture for action recognition. STDA [24] applies a 3D deformable CNN that captures substantial intense regions for spatiotemporal learning. Over the last few years, focus has shifted toward skeleton-based action recognition with respect to the emergence of a GNN. ST-GCN [51] has become a baseline adopting separate spatial and temporal representation modules for spatiotemporal modeling. In addition, ViTs have attracted considerable attention owing to their superior performance in sequential tasks. STAR [1] applies cross-attention for the fusing of temporal correlations between spatial representations. ViViT [2] embeds an input video with a 3D tokenizer to compose the spatiotemporal features in a single token. Other studies [7, 1, 31] have adopted a temporal stride to capture the diversity between different time steps. However, the concept of a 3D deformation, despite its excellent performance, cannot be applied to the attention of ViTs owing to various structural constraints.

Cross-modal learning for action recognition. Most current action recognition methods use various modalities with video frames and skeletons. Several methods [17, 6, 5, 16] employ a graph convolutional network (GCN) to handle a raw skeleton input and a CNN for the video frames. VPN [17] applies GCN subnetworks to support the CNN. The footages of GCN networks are linearly combined with the CNN feature maps. MMNet [6] introduced a multimodal network with two GCN subnetworks and CNNs. Each subnetwork embeds the features separately, and these features are then summed at the end of the network. Other studies [14, 49, 3, 1, 19, 39] transformed graphical skeletons into the heatmaps. PoseC3D [19] uses dual 3D CNN branches for video frames and 3D volumetric heatmaps. It does not explicitly consider the spatial relationships between joints

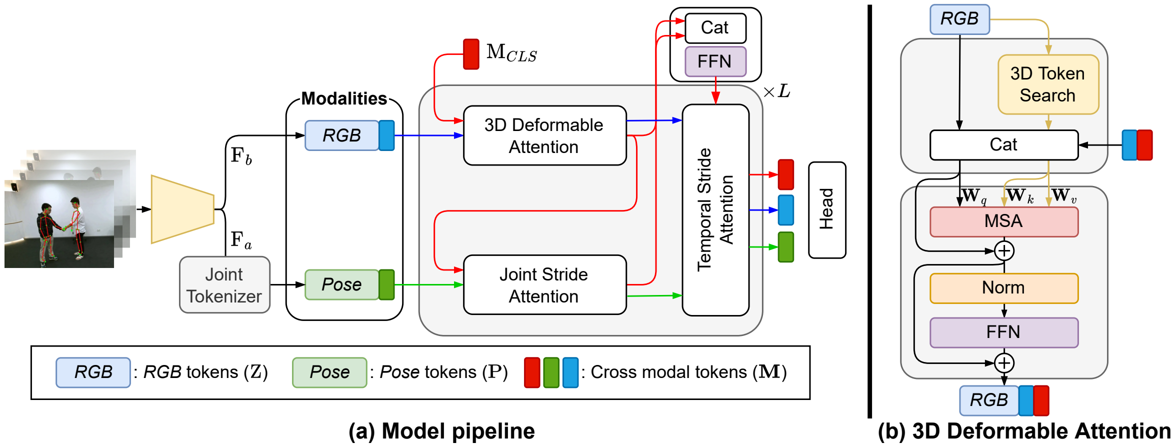


Figure 2: **Overview of our 3D deformable transformer.** (a) Our model consists of a backbone and a series of transformer blocks. Each transformer block uses different modality tokens to process intrinsic properties from various domains and fuse the modalities with cross-modal tokens. The proposed model includes joint stride and temporal stride attention to reduce computational costs. (b) The proposed 3D deformable attention includes the 3D token search (3DTS) and the attention block. The input *RGB* token Z is embedded as *query* token after concatenating with modal tokens. The deformable token from the 3DTS is also concatenated with modal tokens, then it is multiplied with *key* ($\mathbf{W}_k$) and *value* ($\mathbf{W}_v$) weights. These are then fed to the multi-head self-attention (MSA) to interact with the *query* token.

in the skeleton. This may limit its ability to capture complex and subtle movements, or to distinguish between similar actions that involve different joint configurations. STAR [1] proposed joint tokens generated by combining CNN feature maps with 2D joint heatmaps. To fuse the two modalities, they concatenated multiclass tokens by combining different modal tokens. Despite the improved performance of cross-modal learning, video frames and skeleton modalities are merely integrated, thus neglecting a careful design. We propose an effective feature fusion method called a cross-modal token. To exchange contextual information, each token is dispatched to another modality.

Transformer with deformable attention. The idea of a 2D deformable CNN for the learning of deformable features has been applied to the attention module of a ViT, achieving an excellent performance in various applications, including image classification. Deformable DETR [57] was applied to object detection and demonstrated its ability to accurately detect objects of various sizes. A deformable attention transformer (DAT) [47] with an improved numerical stability and a robust performance was recently proposed. In terms of action recognition, 3D deformable CNNs [24, 26] for spatiotemporal learning showed a better performance than a 2D deformable CNN but were not applied to a transformer owing to the structural constraints of an attention optimized for spatial feature embedding. Therefore, in this study, we propose a new 3D deformable transformer capable of fusing cross-modal features using cross-modal tokens. The proposed deformable method enables 3D deformable feature embedding based on local joint stride and temporal stride attention. The remainder of this paper is organized as follows. Section 3 provides a detailed explanation of the proposed approach. Section 4 provides an experimental analysis of several benchmarks as well as visual descriptions. Finally, Section 5 provides some concluding remarks regarding this research.

3. Approach

We propose a 3D deformable transformer for action recognition with adaptive spatiotemporal receptive fields and a cross-modal learning scheme. The overall architecture of the proposed model is shown in Fig. 2 and is described in detail in the following sections.

3.1. Cross-modal learning

In action recognition, cross-modal learning has been the mainstream, leveraging various modalities such as video frames and skeletons. Several successful studies [19, 6, 5, 17, 23, 16] have employed subnetworks that handle different domain features. However, these designs eventually increase the redundancy and complexity owing to the domain-specific subnetworks. We propose simple yet effective cross-modal learning for mutually exchanging contextual information. Our cross-modal learning method consists of a backbone [45], which provides intermediate feature maps and sequential tasks. When the image has a height H, width W, temporal dimension T, and feature dimension C, the backbone network feeds visual feature maps $\mathrm{F}_a \in \mathbb{R}^{C\times T\times \frac{H}{2}\times\frac{W}{2}}$ and $\mathrm{F}_b \in \mathbb{R}^{4C\times T\times \frac{H}{8}\times\frac{W}{8}}$, extracted

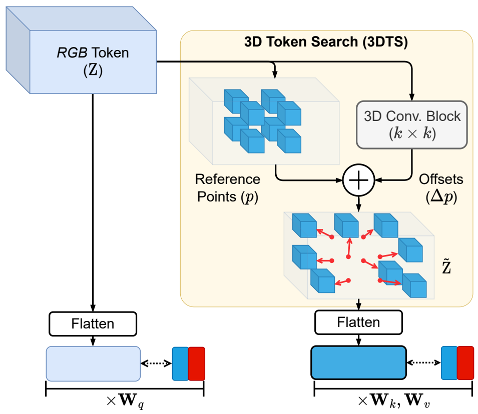


Figure 3: **Illustration of proposed 3D deformable attention for adaptive spatiotemporal learning.** Using 3DTS, an offset vector (Δp) from a 3D conv block finds the deformed tokens by moving the reference points (p) scattered on the input *RGB* tokens Z. The 3D deformable tokens $\tilde{\mathrm{Z}}$ have the same number of tokens as the reference points (p).

from intermediate layers. In the case of F_b, we consider it as the *RGB* modality input for visual representation learning, whereas the local level feature map, F_a, is regarded as the *pose* modality input by combining with skeletons. To fuse both modalities, we apply the following concepts:

***Pose* modality.** To design a cross-modal learning scheme with an alleviated redundancy, we propose visual feature-oriented *pose* tokens combined with joint heatmaps, such as [19, 1]. First, the sequential skeletons are decomposed into single-joint units. Each joint is then recomposed into a joint heatmap $\mathcal{H} \in \mathbb{R}^{T\times R\times \frac{H}{2}\times\frac{W}{2}}$ by projecting the joints toward an empty voxel at the corresponding coordinates $(x_{t,r}, y_{t,r})$. Here, R is the number of joints, and spatial dimension follows feature map size of F_a. Finally, *pose* tokens P using the joint tokenizer shown in Fig. 2 (a) using the multiplication of F_a and the Gaussian blur output at time t as follows:

$$\mathrm{P}_t = ||_r \sum_{j}^{\frac{H}{2}} \sum_{i}^{\frac{W}{2}} \mathrm{F}_{a,t}(i,j)\mathcal{H}_{t,r}(i,j) \tag{1}$$

$$\mathcal{H}_{t,r}(i,j) = e^{-\frac{(i-x_{t,r})^2+(j-y_{t,r})^2}{2\sigma^2}} \tag{2}$$

where $\mathrm{P} \in \mathbb{R}^{C\times T\times R}$ consists of R *pose* tokens for every skeleton sequence with C feature dimensions. $||$ indicates concatenation. To meet the feature dimensions with *RGB* modality, F_b, linear projection is applied to *pose* tokens, resulting in $\mathrm{P} \in \mathbb{R}^{4C\times T\times R}$.

***RGB* modality.** With *RGB* modality, the extracted visual feature map F_b is regarded as *RGB* tokens $\mathrm{Z} \in \mathbb{R}^{4C\times T\times\frac{H}{8}\times\frac{W}{8}}$ and is fused with the *position embedding*.

3.2. 3D deformable transformer

Cross-modal tokens. An intuitive method is to concatenate all tokens from both modalities, considering the characteristics of each token, and then combining information through the transformer stacks. However, to combine different modalities in a single transformer, a deliberate design is required, and the modalities must be cooperative and complementary. Similarly, in STAR [1], the authors employ multi-class tokens for cross-modal learning. Despite a simple yet effective approach, on par with other transformers, it is aimed only at an information fusion for all tokens without considering the intrinsic properties and complementarities of various modalities. Therefore, we propose a cross-modal token that effectively combines the different modalities within the transformer. The cross-modal token $\mathbf{M} \in \mathbb{R}^{4C\times T\times 3}$ is a set of three trainable tokens: *CLS*, *RGB* and *pose* modal tokens. In previous studies [18, 41], *CLS* tokens were used as the final embeddings fusing information by interacting with other tokens. We consider the *CLS* token $\mathrm{M}_{CLS} \in \mathbb{R}^{4C\times T\times 1}$ as a 'modality-mixer' compiling the remaining two modal tokens, which are dispatched to mutual modalities to trade their domain knowledge. The first M_{RGB} and M_{CLS} tokens are fed to the 3D deformable attention. Then, the output *RGB* and *CLS* modal tokens, M_{RGB} and M_{CLS} of a 3D deformable attention, reflect information from their own domains through separated transformer blocks cooperating with the dispatched *CLS* tokens. Hereafter, we introduce the 3D deformable attention shown in Fig. 2 (b), which is the core of the proposed transformer.

3D deformable attention. Although transformers have recently become a new standard in vision tasks, relatively few studies have been conducted on action recognition tasks. Because the nature of a transformer considers long-term relations between the input tokens, it may lead to an exponentially increasing computational complexity with the time steps. In addition, to solve the problem of static transformers, DAT [47] that flexibly selects the *key* and *value* positions in a self-attention has been proposed; however, it is unsuitable for an action recognition that has to deal with cross-modalities and spatiotemporal features. To alleviate the complexity while maintaining the nature of the transformer, inspired by [47], we propose the use of 3D deformable attention for action recognition, as shown in Fig. 2 (b). 3D deformable attention can adaptively capture spatiotemporal features on the *RGB* modality.

The 3D deformable attention module consists of a 3D token search (3DTS) and multi-head self-attention (MSA) with a feed-forward network (FFN), as shown in Fig. 2 (b). First, the input of the module, *RGB* token Z, is fed to the 3DTS, which contains a two-layered Conv3D with kernel k. After the first Conv3D, layer normalization (LN) and GELU non-linearity are applied. The last Conv3D generates offsets (Δp) that contain flow fields against the reference points (p). The reference points are defined as being regularly scattered within a 3D space. The offsets guide the reference points to find discriminative token coordinates in the spatiotemporal tokens Z, as shown in Fig. 3. 3D deformable tokens $\tilde{\mathrm{Z}}$ are configured by selecting the tokens from the adjusted coordinates taken from the offsets.

$$\tilde{\mathrm{Z}} = \mathrm{3DTS}(\mathrm{Z};\omega) \tag{3}$$

where $\mathrm{Z} \in \mathbb{R}^{4C\times T\times \frac{H}{8}\times \frac{W}{8}}$ and $\tilde{\mathrm{Z}} \in \mathbb{R}^{4C\times \tilde{T}\times \tilde{H}\times \tilde{W}}$ are the input and selected *RGB* tokens respectively. The size of $\tilde{T}$, $\tilde{H}$ and $\tilde{W}$ are determined based on the kernel size k. In our case, we set the k as 7 without padding for sparsely extracting deformable tokens and increasing the efficiency. In addition, $\mathbf{W}_q \in \mathbb{R}^{4C\times 4C}$ and ω are trainable weight and model parameters for the MSA and 3D conv block in the 3DTS, respectively. It should be noted that while *query* tokens are composed in the same manner as the transformer, the *key* and *value* tokens are composed of selected tokens from the 3DTS. Further details on our implementation are in Appendix B.

These tokens are then embedded into the *key* and *value* tokens using $\mathbf{W}_k$ and $\mathbf{W}_v$, respectively. Herein, we aim to make the M_{RGB} token faithfully learn the *RGB* modality features, and M_{CLS} trades domain knowledge between the *RGB* and *pose* modalities. To fuse cross-modal tokens with the *RGB* modality, three tokens, M_{RGB}, M_{CLS} and spatiotemporal feature tokens Z, are concatenated to token $\mathbf{X}$.

$$\mathbf{X} = [\mathrm{Z}||\mathrm{M}_{RGB}||\mathrm{M}_{CLS}] \tag{4}$$

where M_{RGB} and M_{CLS} are obtained from a portion of the proposed cross-modal tokens representing the *RGB* modality and modality head, respectively.

Similarly, the selected deformable tokens $\tilde{\mathrm{Z}}$ are coupled with two cross-modal tokens to produce $\tilde{\mathbf{X}}$.

$$\tilde{\mathbf{X}} = [\tilde{\mathrm{Z}}||\mathrm{M}_{RGB}||\mathrm{M}_{CLS}] \tag{5}$$

Then, $\mathbf{X}$ is multiplied with *query* weight $\mathbf{W}_q$ and, $\tilde{\mathbf{X}}$ is multiplied with *key* and *value* weights, $\mathbf{W}_k$ and $\mathbf{W}_v$, respectively. Those recomposed tokens are fed into multi-head self-attention as a *query*, *key* and *value* for each.

$$\mathbf{X} = \mathbf{X} + \mathrm{MSA}(\mathbf{X}\mathbf{W}_q, \tilde{\mathbf{X}}\mathbf{W}_k, \tilde{\mathbf{X}}\mathbf{W}_v) \tag{6}$$

The output $\mathbf{X}$ of 3D deformable attention is finally obtained by applying a LN and FFN.

$$[\mathrm{Z}, \mathrm{M}_{RGB}, \mathrm{M}'_{CLS}] = \mathbf{X} + \mathrm{FFN}(\mathrm{LN}(\mathbf{X})) \tag{7}$$

We visualized the attention scores of the selected tokens from the proposed 3D deformable attention shown in Fig. 5.

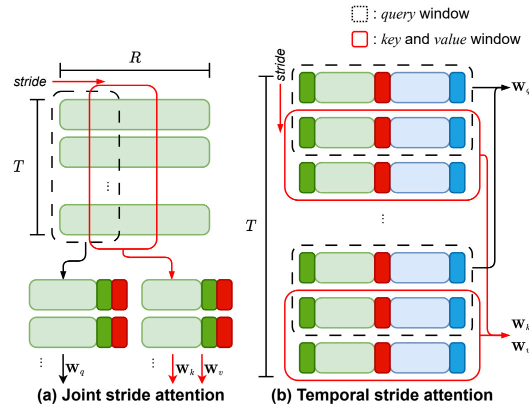


Figure 4: **Proposed stride attention modules.** (a) Joint stride attention, where a series of joint tokens are grouped into *query*, *key*, and *value* including all temporal dimensions with a stride window. (b) Temporal stride attention where tokens are bundled with temporal stride windows to fuse the changes with the time steps.

As indicated in Fig. 5, our proposed 3DTS faithfully identifies the fundamental intense regions with adaptive receptive fields against entire sequences.

Local joint stride attention. In action recognition, there are often multiple people appearing in a scene; therefore, the number of joint tokens increases with the number of people. To reduce the computational complexity, we concatenate the joints of multiple people into a series of joint tokens. Although this approach is an efficient way to process multiple people in the same scene simultaneously without significantly increasing the complexity, it still results in a problem in that the size of the joint token increases exponentially as the number of people increases. To avoid this problem, we configure the *query*, *key*, and *value* tokens using a sliding window on the joint tokens, as shown in Fig. 4 (a). All tokens in each sliding window are flattened and then concatenated with M_{pose} and M'_{CLS} dispatched from a 3D deformable attention to apply a scaled-dot product. This is more efficient than calculating all tokens at once and maintaining the relations with each other. The output of the joint stride attention is the *pose* token P and modal tokens M_{pose} and M^*_{CLS}.

The calculated *RGB* tokens Z and *pose* tokens P are fed to the temporal stride attention module. Before this step, to fuse contextual information from each modality, M'_{CLS} memorized from the 3D deformable attention and M^*_{CLS} calculated from the joint stride attention are projected together into a new single M_{CLS} as shown in Fig. 2 (a). Subsequently, the temporal stride attention module learns the correlations between temporal changes against tokens concatenated with cross-modal tokens.

Temporal stride attention. Several limitations of an atten-

Table 1: Accuracy comparisons with SoTA approaches on NTU60 and NTU120. P and R denote *pose* and *RGB* modalities, respectively. † indicates estimated *pose*.

Method	**Modality**	**Pre-trained**	**NTU60**		**NTU120**	
			XSub (%)	XView(%)	XSub (%)	XSet(%)
3s-AimCLR [40]	P	✓	86.9	92.8	80.1	80.9
PoseC3D [19]	P†	✓	93.7	96.6	86.0	89.6
VPN [17]	R+P	✓	95.5	98.0	86.3	87.8
VPN++ (3D *pose*) [16]	R+P	✓	96.6	99.1	90.7	92.5
MMNet [6]	R+P	✓	96.0	98.8	92.9	94.4
PoseC3D [19]	R+P†	✓	97.0	99.6	95.3	96.4
ST-GCN [51]	P		81.5	88.3	-	-
DualHead-Net [10]	P		92.0	96.6	88.8	89.3
Skeletal GNN [52]	P		91.6	96.7	87.5	89.2
CTR-GCN [11]	P		92.4	96.8	88.8	90.6
InfoGCN [13]	P		93.0	97.1	89.8	91.2
KA-AGTN [30]	P		90.4	96.1	86.1	88.0
PoseMap [28]	R+P†		91.7	95.2	-	-
MMTM [23]	R+P		91.9	-	-	-
STAR [1]	R+P		92.0	96.5	90.3	92.7
TSMF [5]	R+P		92.5	97.4	87.0	89.1
Ours	R+P		**94.3**	**97.9**	**90.5**	**91.4**

tion module exist when the transformer handles the input tokens. In general, the attention module covers all input tokens with scaled-dot products. Thus, the complexity of the attention module is highly dependent on the number of input tokens. In the case of sequential data, this problem is more serious because the input tokens grow with the sizes of the temporal dimensions. Ahn *et al*. [1] divided temporal dimensions into two groups that contain regularly interleaved tokens. Despite the halved temporal dimensions, the complexity was only slightly reduced, and the temporal correlations of the neighborhood were decoupled. Unlike Ahn *et al*. [1], we propose a temporal stride with mitigated complexity and an enhanced temporal correlation for cross-attention. When building input *query*, *key*, and *value* tokens, the temporal dimensions are split into regularly increasing strides to couple various sequential relationships with reduced complexity. As shown in Fig. 4 (b), we first set a local time window for a given stride. This window traverses all tokens and specifies the *query*, *key*, and *value* tokens. It not only reduces the number of input tokens of the attention module but also supports temporal representation learning without using all tokens at once.

All of the deformable transformers stated above are repeated L-times, as shown in Fig. 2 (a). To make the final logits, we concatenate the cross-modal tokens only along with the channel dimension and then feed them to the classification head.

4. Experiments

Datasets. We conducted experiments using several representative benchmark datasets: FineGYM [36], NTU60 [35], NTU120 [27], and PennAction [54]. FineGYM contains 29K videos with 99 fine-grained action labels collected from gymnastic video clips. NTU60 and NTU120 are representative multimodal datasets that are used for human action recognition. NTU60 consists of 57K videos of 60 action labels collected in a controlled environment. NTU120, which is a superset of NTU60, contains 114K videos of 120 action labels. The NTU datasets use three types of validation protocols following the action subjects and camera settings, *i.e*., cross-subject (XSub), cross-view (XView), and cross-setup (XSet). In addition, we validated our proposed model with a smaller dataset, PennAction, which contains 2K videos for 15 action labels.

Settings. We adopted an AdamW optimizer for 90 epochs with a cosine scheduler with a 5-epoch warm-up. A batch consisted of randomly cropped videos with a pixel resolution of 224×224 for training and center-cropped videos for testing. Training and testing were conducted using four NVIDIA Tesla V100 32GB GPUs with APEX.

4.1. Comparison with state-of-the-art approaches

NTU60 & 120. In action recognition, pre-training for feature extraction and pose information for recognition have a great impact on performance. To obtain objective results in

Table 2: Accuracy comparisons with other SoTA approaches on FineGYM.

Method	Modality	Pre-trained	Mean (%)
ST-GCN [51]	P†		25.2
TQN [53]	R	✓	90.6
PoseC3D [19]	R+P†	✓	95.6
Ours	R+P†		**90.3**

Table 3: Accuracy comparisons with other SoTA approaches on PennAction.

Method	Modality	Pre-trained	Top-1 (%)
Pr-VIPE [37]	P	✓	97.5
UNIK [50]	P	✓	97.9
HDM-BG [55]	P		93.4
3D Deep [8]	R+P		98.1
PoseMap [28]	R+P		98.2
Multitask CNN [32]	R+P		98.6
STAR [1]	R+P		98.7
Ours	R+P		**99.7**

action recognition, it is desirable to use pose information given in the dataset without pre-training. As shown in Table 1, PoseC3D [19] shows the best performance at NTU60 and 120, but unlike other methods, it used an estimated pose optimized for recognition models. Therefore, in the experiment, we exclude PoseC3D from quantitative comparison and analysis with SoTA methods.

In Table 1, we first compare the top-1 accuracy with various SoTA action recognition methods for the NTU60 dataset. When compared with the TSMF and STAR models trained under fair conditions, our model shows a 0.5%-2.5% higher performance, respectively. Among the GNN-based methods, InfoGCN ranks remarkably even when it is trained with a single modality; nonetheless, our model shows higher performance of 94.3% and 97.9%, respectively.

We also provide benchmark results against the NTU120 dataset, which has double the number of videos and action labels compared to NTU60. With this result in Table 1, the GNN-based SoTA method, InfoGCN, achieved accuracy of 89.8% and 91.2% for both protocols. In the case of the multimodal trained method, PoseC3D, it achieved accuracy of 95.3% and 96.4% with pre-training and estimated poses. Unlike PoseC3D, the scratch-trained multimodal method, STAR, showed accuracy of 90.3% and 92.7%. Finally, our proposed method showed a promising performance under similar conditions, *i.e.*, multimodal and scratch training, with accuracy of 90.5% and 91.4%, respectively.

FineGYM. In the case of the FineGYM dataset, which is harsher than other datasets because clips have dynamic motions and camera movements from sports games. This dataset has rarely been used for multimodal action recognition, because it does not contain ground-truth skeletons. We used the estimated skeletons from HRNet [38] to apply it to cross-modal learning. Table 2 shows the results of the comparison experiments with other SoTA methods for the FineGYM. In the case of a single modality, ST-GCN showed relatively low accuracy because it used the estimated skeletons as our approach. The TQN achieved 90.6% with the *RGB* single modality. Cross-modality-based methods, including PoseC3D and our approach, showed a relatively high performance compared to other methods. Pre-trained PoseC3D demonstrated a SoTA performance of 95.6%; however, our method showed promising accuracy despite not applying a pre-training step.

PennAction. We validated our method by applying smaller datasets with clearly recorded videos as shown in Table 3. In the GNN regime using the single-pose modality, Pr-VIPE and UNIK achieved high performance of 97.5% and 97.9%, respectively. Among cross-modality-based methods, Multitask CNN and STAR show the best performances among the CNN and transformer approaches at 98.6% and 98.7%, respectively. Our proposed method outperforms above SoTAs by a large margin of 1.0%-6.3%. This result indicates that the proposed method can achieve a good performance regardless of whether small or large datasets are applied.

4.2. Qualitative analysis

Visualization of 3D deformable attention. We propose a 3D deformable attention mechanism supporting an adaptive receptive field for action recognition. To demonstrate its effectiveness, we provide qualitative evaluations that show the visualized attention values of selected tokens against different video sequences. As shown in Fig. 5, the proposed method finds intensive regions through a 3D token search of the entire sequence. The activation is relatively low in inactive scenes, but, on the contrary, high activation occurs at a large transition among sequences. In particular, activations appear finely in joints used for actual action and do not appear coarsely. We can confirm that the proposed method not only properly finds tokens that are practically required for recognition in the entire sequence but also shows that strong activation appears in those tokens.

Visualization of joint stride attention. To make the proposed model independent of the number of input joints, we propose a local joint stride attention by grouping joint tokens with a sliding window. This approach provides improved efficiency in attention, but we need to verify whether the configured tokens with overlapping contribute to the at-

Figure 5: **Visualized 3D deformable attentions.** The proposed 3D deformable attention found discriminative tokens with strong attention among entire frame sequences. In particular, strong attentions are discovered at only noticeable changes in action.

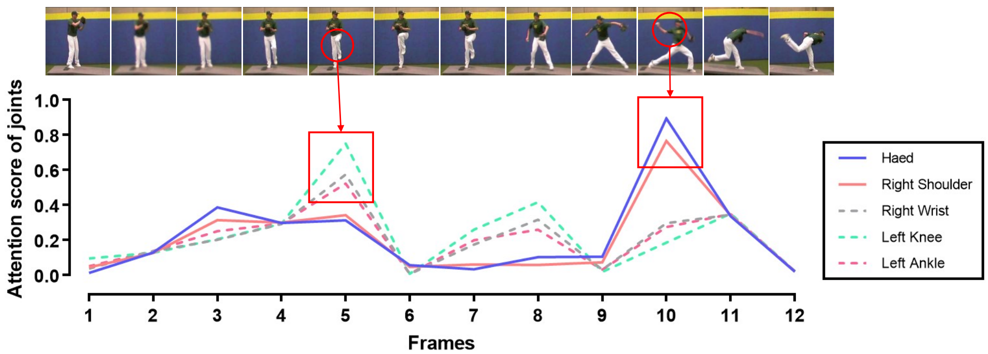


Figure 6: **Visualized joint stride attentions.** The proposed 3D deformable attention activates the attention score of each joint differently depending on the size of the '*pitching*' action for each frame (shown in color).

Table 4: Ablation study with attention blocks on PennAction.

	3D def.	Joint	Temp.	Top-1 (%)
①		✓	✓	94.8
②	✓			97.9
③	✓		✓	98.2
④	✓	✓		98.5
Ours	✓	✓	✓	**99.7**

tention module without full attention. In this experiment, we charted the attention levels of each joint according to sequence frames, as shown in Fig. 6. In practical terms, the joints that move the most when the ball is pitched in the sample video are the head, right hand, and right leg. The chart shows that the attention was largely activated according to this action flow. From this experiment, it is clear that the proposed local joint attention maintains the correlations between the entire joint token and the efficiency.

4.3. Ablation studies

Attention modules. We provide ablation studies on three types of attention modules with PennAction dataset. According to the Table 4, when ① 3D deformable attention is

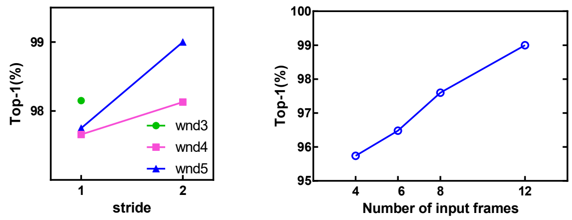
Table 5: Comparison of the Top-1 accuracies for different modal token configurations on PennAction.

Modal Tokens	Top-1 (%)
No tokens	92.2
Single token	97.0
Cross-modal tokens	**99.7**

eliminated from the model, it is confirmed that the accuracy is abruptly degraded (4.9%). Conversely, ②-④ other attention modules are found to have relatively less impact. The proposed 3D deformable attention is pivotal for achieving high accuracy in action recognition tasks.

Cross-modal Tokens. To analyze the impact of the cross-modal tokens, we conducted ablation studies as shown in Table 5. In the 'No tokens' scenario, averaging the last RGB and *pose* tokens led to a performance drop to 92.2%. Using a single token for all modalities resulted in a slight decrease to 97.0%. However, our approach achieved the best performance, highlighting the effectiveness of learning multiple modalities in a single transformer design.

Temporal stride. To solve the limitations of complexity dependent on the number of tokens in an attention module, we propose a local window cross attention using a temporal



(a) (b)

Figure 7: (a) Ablation study with different stride on various window (wnd) sizes on PennAction. (b) Ablation study with different numbers of input frames during the training phase on PennAction.

stride. In this experiment, to evaluate the validity of the proposed approach, we observe the changes in performance by applying various strides to a fixed-sized window using the PennAction dataset. As shown in Fig. 7 (a), the best performance is obtained when the stride is about half the size of the window, regardless of the window size. This is because more overlap between temporal tokens fuels the temporal correlations with enhanced efficiency. Therefore, the proposed method faithfully maintains the correlations, even when tokens are divided into local windows.

Number of input frames. One of the important factors to determine in the learning of sequential data is the number of input frames. If the number of frames is sufficiently large, better feature representation learning is possible, although the computational cost significantly increases. We therefore empirically determined the optimal number of input frames using the PennAction dataset, as illustrated in Fig. 7 (b). In our model, the best performance was achieved using 12 input frames. When the number of frames is smaller than 12, the performance drops slightly, and there is a limit to learning the continuity of the actions.

5. Conclusion

ViTs have become the mainstream in various vision tasks, achieving an overwhelming performance; however, it has been used relatively little in action recognition tasks. Therefore, we first proposed a 3D deformable attention consisting of stride window cross attention for better spatiotemporal feature learning, as well as a cross-modal framework for action recognition. The proposed method achieved a newly demonstrated SoTA performance on representative action recognition datasets. Based on the results of qualitative experiments, we can confirm that our proposed method has a strong spatiotemporal feature learning capability 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).

References

[1] Dasom Ahn, Sangwon Kim, Hyunsu Hong, and Byoung Chul Ko. Star-transformer: a spatio-temporal cross attention transformer for human action recognition. In *WACV*, pages 3330–3339, 2023. 1, 2, 3, 4, 6, 7

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, pages 6836–6846, 2021. 2

[3] Sadjad Asghari-Esfeden, Mario Sznaier, and Octavia Camps. Dynamic motion representation for human action recognition. In *WACV*, pages 557–566, 2020. 2

[4] Fabien Baradel, Christian Wolf, Julien Mille, and Graham W Taylor. Glimpse clouds: Human activity recognition from unstructured feature points. In *CVPR*, pages 469–478, 2018. 2

[5] XB Bruce, Yan Liu, and Keith CC Chan. Multimodal fusion via teacher-student network for indoor action recognition. In *AAAI*, pages 3199–3207, 2021. 1, 2, 3, 6

[6] 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 TPAMI*, pages 01–18, 2022. 1, 2, 3, 6

[7] Adrian Bulat, Juan Manuel Perez Rua, Swathikiran Sudhakaran, Brais Martinez, and Georgios Tzimiropoulos. Space-time mixing attention for video transformer. In *NeurIPS*, pages 1–14, 2021. 2

[8] Congqi Cao, Yifan Zhang, Chunjie Zhang, and Hanqing Lu. Body joint guided 3-d deep convolutional descriptors for action recognition. *IEEE Trans. Cybernetics*, 48(3):1095–1108, 2017. 7

[9] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 1, 2

[10] Tailin Chen, Desen Zhou, Jian Wang, Shidong Wang, Yu Guan, Xuming He, and Errui Ding. Learning multi-granular spatio-temporal graph network for skeleton-based action recognition. In *ACM MM*, pages 4334–4342, 2021. 6

[11] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *ICCV*, pages 13359–13368, 2021. 2, 6

[12] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *CVPR*, pages 183–192, 2020. 2

[13] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *CVPR*, pages 20186–20196, 2022. 2, 6

[14] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid. Potion: Pose motion representation for action recognition. In *CVPR*, pages 7024–7033, 2018. 2

[15] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *CVPR*, pages 764–773, 2017. 1

[16] Srijan Das, Rui Dai, Di Yang, and Francois Bremond. Vpn++: Rethinking video-pose embeddings for understanding activities of daily living. *IEEE TPAMI*, pages 1–15, 2021. 2, 3, 6

[17] Srijan Das, Saurav Sharma, Rui Dai, Francois Bremond, and Monique Thonnat. Vpn: Learning video-pose embedding for activities of daily living. In *ECCV*, pages 72–90, 2020. 2, 3, 6

[18] 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. In *ICLR*, pages 1–21, 2020. 4

[19] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *CVPR*, pages 2969–2978, 2022. 1, 2, 3, 4, 6, 7

[20] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *CVPR*, pages 203–213, 2020. 2

[21] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019. 2

[22] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE TPAMI*, 35(1):221–231, 2012. 2

[23] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *CVPR*, pages 13289–13299, 2020. 3, 6

[24] Jun Li, Xianglong Liu, Mingyuan Zhang, and Deqing Wang. Spatio-temporal deformable 3d convnets with attention for action recognition. *PR*, 98:1–9, 2020. 2, 3

[25] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *CVPR*, pages 3595–3603, 2019. 2

[26] Wei Li, Zehuan Yuan, Dashan Guo, Lei Huang, Xiangzhong Fang, and Changhu Wang. Deformable tube network for action detection in videos. *arXiv preprint arXiv:1907.01847*, 2019. 3

[27] 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 TPAMI*, 42(10):2684–2701, 2019. 6

[28] Mengyuan Liu and Junsong Yuan. Recognizing human actions as the evolution of pose estimation maps. In *CVPR*, pages 1159–1168, 2018. 6, 7

[29] Xin Liu, Silvia L Pintea, Fatemeh Karimi Nejadasl, Olaf Booij, and Jan C van Gemert. No frame left behind: Full video action recognition. In *CVPR*, pages 14892–14901, 2021. 2

[30] Yanan Liu, Hao Zhang, Dan Xu, and Kangjian He. Graph transformer network with temporal kernel attention for skeleton-based action recognition. *Knowledge-Based Systems*, 240:1–16, 2022. 6

[31] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, pages 3202–3211, 2022. 2

[32] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *CVPR*, pages 5137–5146, 2018. 7

[33] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *ICCV*, pages 3163–3172, 2021. 2

[34] Kunyu Peng, Alina Roitberg, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhagen. Delving deep into one-shot skeleton-based action recognition with diverse occlusions. *IEEE TMM*, pages 1–16, 2023. 1

[35] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *CVPR*, pages 1010–1019, 2016. 6

[36] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *CVPR*, pages 2616–2625, 2020. 6

[37] Jennifer J Sun, Jiaping Zhao, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, and Ting Liu. View-invariant probabilistic embedding for human pose. In *ECCV*, pages 53–70, 2020. 7

[38] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019. 7

[39] Ruiqi Sun, Qin Zhang, Chuang Luo, Jiamin Guo, and Hui Chai. Human action recognition using a convolutional neural network based on skeleton heatmaps from two-stage pose estimation. *Biomimetic Intelligence and Robotics*, 2(3):1–12, 2022. 2

[40] Guo Tianyu, Liu Hong, Chen Zhan, Liu Mengyuan, Wang Tao, and Ding Runwei. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In *AAAI*, pages 762–770, 2022. 6

[41] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *ICML*, pages 10347–10357, 2021. 4

[42] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015. 1, 2

[43] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *ICCV*, pages 5552–5561, 2019. 2

[44] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, pages 6450–6459, 2018. 2

[45] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, pages 6450–6459, 2018. 3

[46] Zhengwei Wang, Qi She, and Aljosa Smolic. Action-net: Multipath excitation for action recognition. In *CVPR*, pages 13214–13223, 2021. 2

[47] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *CVPR*, pages 4794–4803, 2022. 1, 3, 4

[48] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, pages 305–321, 2018. 1, 2

[49] An Yan, Yali Wang, Zhifeng Li, and Yu Qiao. Pa3d: Pose-action 3d machine for video recognition. In *CVPR*, pages 7922–7931, 2019. 2

[50] Di Yang, Yaohui Wang, Antitza Dantcheva, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. Unik: A unified framework for real-world skeleton-based action recognition. In *BMVC*, pages 1–13, 2021. 7

[51] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In *IJCAI*, pages 3634–3640, 2018. 2, 6, 7

[52] Ailing Zeng, Xiao Sun, Lei Yang, Nanxuan Zhao, Minhao Liu, and Qiang Xu. Learning skeletal graph neural networks for hard 3d pose estimation. In *ICCV*, pages 11436–11445, 2021. 6

[53] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Temporal query networks for fine-grained video understanding. In *CVPR*, pages 4486–4496, 2021. 7

[54] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *ICCV*, pages 2248–2255, 2013. 6

[55] Rui Zhao, Wanru Xu, Hui Su, and Qiang Ji. Bayesian hierarchical dynamic model for human action recognition. In *CVPR*, pages 7733–7742, 2019. 7

[56] Jiagang Zhu, Wei Zou, Liang Xu, Yiming Hu, Zheng Zhu, Manyu Chang, Junjie Huang, Guan Huang, and Dalong Du. Action machine: Rethinking action recognition in trimmed videos. *arXiv preprint arXiv:1812.05770*, pages 1–13, 2018. 2

[57] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, pages 1–16, 2021. 1, 3