Just Add π! Pose Induced Video Transformers for Understanding Activities of Daily Living

Dominick Reilly
UNC Charlotte
dreilly1@charlotte.edu

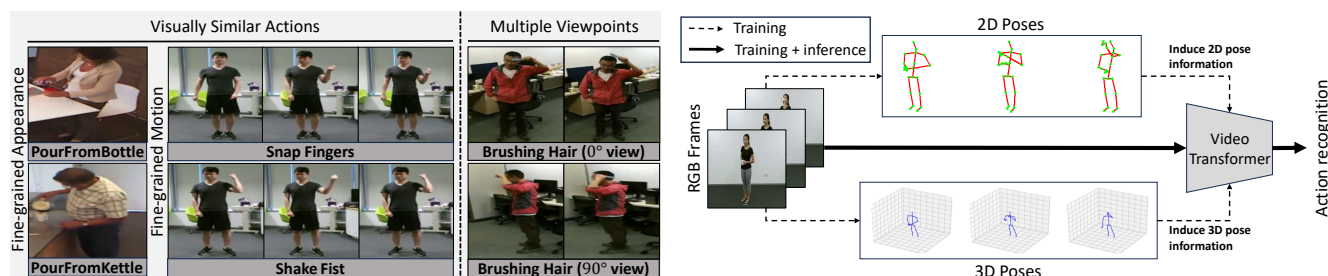Srijan Das
UNC Charlotte
sdas24@charlotte.edu

Figure 1. **Left:** we illustrate the challenges of Activities of Daily Living. Notice the visual similarity in the action pairs (*PourFromBottle*, *PourFromKettle*) and (*Snap Fingers*, *Shake Fist*). Also notice the significant change in appearance when viewing *Brushing Hair* from different angles. **Right:** we present an overview of our proposed approach to address these challenges. We induce human pose information into the representations learned by video transformers. This induction of information is only required during training.

Abstract

*Video transformers have become the de facto standard for human action recognition, yet their exclusive reliance on the RGB modality still limits their adoption in certain domains. One such domain is Activities of Daily Living (ADL), where RGB alone is not sufficient to distinguish between visually similar actions, or actions observed from multiple viewpoints. To facilitate the adoption of video transformers for ADL, we hypothesize that the augmentation of RGB with human pose information, known for its sensitivity to fine-grained motion and multiple viewpoints, is essential. Consequently, we introduce the first Pose Induced Video Transformer: **PI-ViT** (or π-ViT), a novel approach that augments the RGB representations learned by video transformers with 2D and 3D pose information. The key elements of π-ViT are two plug-in modules, 2D Skeleton Induction Module and 3D Skeleton Induction Module, that are responsible for inducing 2D and 3D pose information into the RGB representations. These modules operate by performing pose-aware auxiliary tasks, a design choice that allows π-ViT to discard the modules during inference. Notably, π-ViT achieves the state-of-the-art performance on three prominent ADL datasets, encompassing both real-world and large-scale RGB-D datasets, without requiring poses or additional computational overhead at inference. We release code and models at https://github.com/dominickrei/pi-vit/.*

1. Introduction

Recently, the task of monitoring Activities of Daily Living (ADL) has gained prominence as it enables many applications, such as advanced security systems or assisting the elderly. Historically, the learning of ADL representations has borrowed heavily from the action recognition literature, where advanced vision models are trained primarily on internet-sourced videos [8, 26, 28, 46] such as sports, YouTube, and movie clips. These models, however, are predominantly appearance-based, aligning actions strongly with their scenes [24], and do not adequately capture the intrinsic challenges of ADL. In Figure 1, we present the challenges in ADL which broadly includes the presence of visually similar actions and actions captured from different camera views. Firstly, ADL involves visually similar but *fine-grained* actions, distinguishable through subtle appearance or motion cues. Secondly, the actors of ADL can be observed from multiple viewpoints, thus requiring the learning of view-agnostic representations that can recognize actions regardless of camera viewpoint.

Video representation learning has undergone a paradigm shift with the emergence of video transformers [3, 6, 34]. However, despite their effectiveness in learning action representations, they predominantly rely only on the RGB modality. These unimodal RGB representations are insuf-
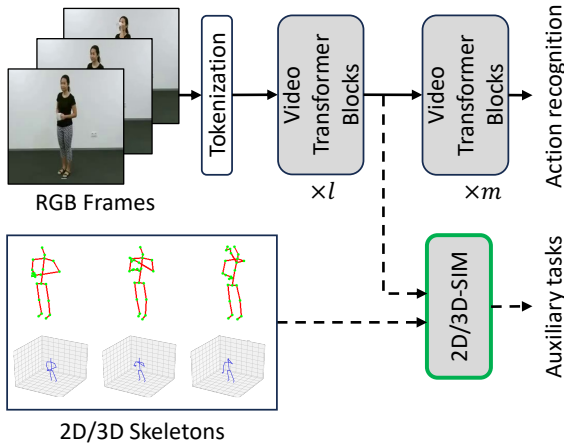
Figure 2. **An overview of our Pose Induced Video Transformer (π-ViT).** During training (indicated by dashed lines), the video transformer incorporates 2D-SIM and 3D-SIM, which process skeletons and the intermediate visual representations. During inference the video transformer is used independently.

ficient to capture the fine-grained details in the videos and are sensitive to changes in viewpoint. Thus, these models are suboptimal for ADL and warrant the need for research directed towards multi-modal representation learning.

Furthermore, huge advancements have been made in the field of skeleton based action recognition [29, 55, 56, 58] using 3D human poses, which are inherently viewpoint-agnostic and offer key positional information for modeling human motion. They are effective for some ADL challenges but they cannot encode appearance information. Therefore, a natural idea is to combine RGB and poses [2, 14, 17, 27, 41]. However, these multi-modal methods require depth sensors for obtaining 3D poses or they incur high latency due to the computational demands of estimating 3D poses from RGB [36, 38]. This brings us to the main question: *What is the best strategy to combine RGB and Poses without compromising model latency?*

We observe that crucial appearance cues for distinguishing visually similar actions are localized in the RGB regions corresponding to the salient human skeleton joints. Additionally, the temporal evolution of 3D skeletons effectively captures fine-grained motion and is viewpoint-agnostic. We hypothesize that infusing these inherent properties of 3D skeletons into video transformers will optimize the RGB representations to address the challenges of ADL.

To this end, we introduce the **Pose Induced Video Transformer**, dubbed as PI-ViT or π-ViT. It utilizes both 2D and 3D skeletons to infuse complementary information into the visual token representations learned by video transformers (see Figure 1). π-ViT is composed of two novel plug-in modules: 2D Skeleton Induction Module (2D-SIM) and 3D Skeleton Induction Module (3D-SIM). 2D-SIM leverages 2D skeletons to perform an auxiliary task

of mapping the skeleton joints and visual tokens to provide extra supervision to the RGB regions containing the relevant skeleton joints involved in an action. This task refines the RGB representations and enforces the video transformer to discriminate actions with fine-grained appearance. Conversely, 3D-SIM utilizes 3D skeletons to address the challenges of fine-grained motion and multiple viewpoints. To realize this objective, 3D-SIM performs an auxiliary feature alignment task, refining the RGB representations by integrating an optimized 3D skeleton representation for action classification. These modules, 2D-SIM and 3D-SIM, can be inserted after any layer of the existing video transformer. What sets π-ViT apart is its pose induction through auxiliary tasks performed by the modules. This not only enforces that the video transformer learn a pose-augmented RGB representation but also allows for the removal of these modules during inference, eliminating the need for poses at inference time.

We summarize the key contributions of our work as:

- The introduction of the first Pose Induced Video Transformer, π-ViT, that leverages both 2D and 3D Poses for enhancing video representation learning on ADL.
- π-ViT incorporates two novel plug-in modules, 2D-SIM and 3D-SIM, that are designed to address the challenges of ADL. These modules perform distinct pose-aware auxiliary tasks that enable video transformers to learn fine-grained and view-invariant representations.
- π-ViT's efficacy is demonstrated through superior performance on the real-world ADL dataset Toyota-Smarthome, and the largest RGB-D human action recognition datasets: NTU120 and NTU60. π-ViT achieves state-of-the-art results without requiring poses, or additional computational overhead, during inference.

2. Preliminaries: Video Transformers

In this section, we provide a brief description of the working principle of existing video transformers [3, 6, 18, 34]. Given a video V with the shape $T \times H \times W \times 3$, video transformers decompose V into disjoint spatio-temporal patches, each of size $\tau \times p \times p$, where $\tau = 1$ is similar to image patches [6] and $\tau > 1$ is similar to the tublets employed in [3, 18, 34]. Then, these patches are tokenized via a linear projection that projects the patches to the shape $T_v \times S_v \times d_v$, where the spatial dimension $S_v = \left\lceil \frac{H}{p} \right\rceil \cdot \left\lceil \frac{W}{p} \right\rceil$ and d_v is the embedding dimension of the video transformer. Spatio-temporal positional embeddings are then added to each of the tokens, enabling them to encode their location in the video. Additionally, a classification token is appended, resulting in a total of $T_v \cdot S_v + 1$ tokens. The resulting sequence of tokens $\mathbf{z}_0$ can then be processed by a series of video transformer blocks. Consequently, the output of the lth video
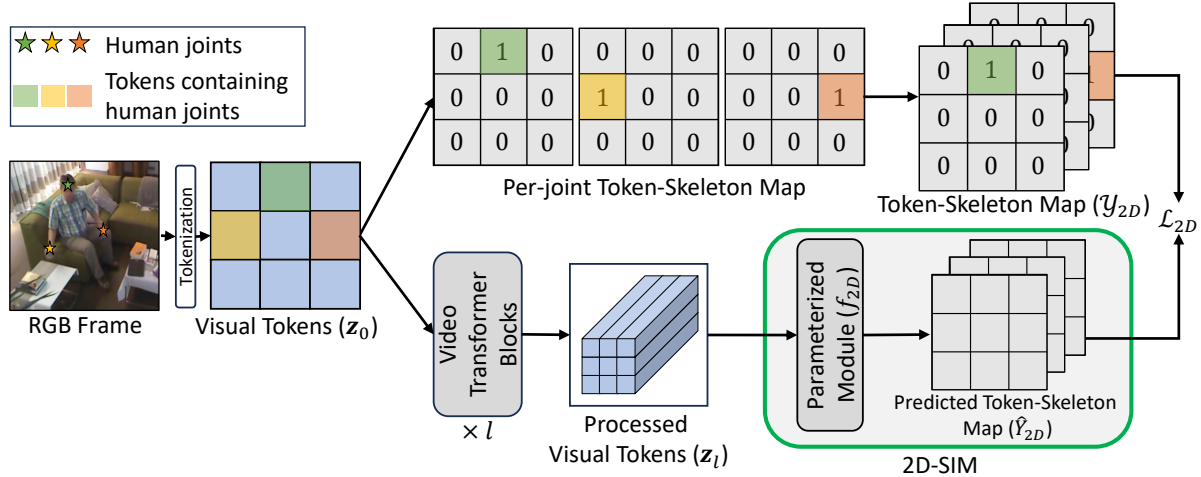
Figure 3. **An overview of the 2D Skeleton Induction Module (2D-SIM).** We only visualize a single frame from the video for clarity. A token-skeleton map is constructed that indicates the presence of human skeleton joints in the RGB regions corresponding to the visual token. 2D-SIM refines the visual tokens from the video transformer, and uses them to predict the token-skeleton map.

transformer block, $\mathbf{z}_l$, can be obtained as

$$\mathbf{z}_l = \mathbf{z}_{l-1} + \text{ST-MHSA}(\text{LN}(\mathbf{z}_{l-1})) \qquad (1)$$

$$\mathbf{z}_l = \mathbf{z}_l + \text{MLP}(\text{LN}(\mathbf{z}_l)) \qquad (2)$$

where ST-MHSA denotes spatio-temporal multihead self-attention [6], LN denotes layer normalization [4], and MLP denotes multi-layer perceptron. In a video transformer composed of L layers, class predictions are computed from the classification token in $\mathbf{z}_L$ via a fully connected layer. The video transformer is then trained with an entropy loss ($\mathcal{L}_v^{cls}$) computed using the class predictions and ground-truth.

3. Proposed Video Transformer

In this section, we present our Pose Induced Video Transformer (π-ViT) which implicitly learns discriminative representations for understanding ADL videos. This is accomplished by introducing 2D and 3D human skeleton information into a vanilla video transformer through the addition of two modules: (1) the 2D Skeleton Induction Module (2D-SIM) and (2) the 3D Skeleton Induction Module (3D-SIM). Throughout this paper, we use the term "pose" to refer to the abstract configuration of the human joints, and the term "skeleton" to refer the low-level positional coordinates of the human joints. A high-level overview of π-ViT is shown in Figure 2. Both modules are implemented to perform auxiliary tasks that are only required during training, and are removed during inference, thus requiring no extra computation during inference.

3.1. 2D Skeleton Induction Module (2D-SIM)

We first describe 2D-SIM, a plug-in module that can be inserted after any layer in an existing video transformer architecture, and that is designed to address the challenge of

fine-grained appearance in ADL. During the training phase, the role of 2D-SIM is to refine the video representations learned by the transformer with detailed human anatomy information obtained from 2D skeletons. Specifically, 2D-SIM provides extra supervision to the RGB regions containing the specific human skeleton joints involved in an action. This extra supervision is achieved through an auxiliary task that learns the mapping between the visual tokens and the skeleton joints. At inference time, 2D-SIM is removed.

Token-Skeleton Map. The first step in 2D-SIM is the construction of a token-skeleton map that defines the correspondence between the RGB regions and the 2D skeleton joints. A visual illustration of this mapping can be found in Figure 3. The tokenization performed by the video transformer provides an elegant way to obtain distinct and discrete RGB regions. Hence, we construct a mapping from visual tokens to 2D skeleton joints. We denote the set of 2D skeleton joints for a given video V as

$$\mathcal{S}_{2D} = \{(t, j, x, y)\} : 1 \le t \le T, 1 \le j \le J \qquad (3)$$

where J is the number of human joints contained in the 2D skeleton and x, y indicates the spatial location of jth joint in the tth frame. Then, a binary map $\mathcal{M}$ with the shape $T \times H \times W \times J$ is computed through

$$\mathcal{M}_{thwj} = \begin{cases} 1 & \text{if } (t, j, h, w) \in \mathcal{S}_{2D} \\ 0 & \text{otherwise} \end{cases} \qquad (4)$$

where each value in $\mathcal{M}$ indicates the presence or absence of a particular joint *at the pixel level* of the input video. Since the objective is to obtain the mapping at the token level, we perform max pooling on the binary map $\mathcal{M}$ across the spatio-temporal resolution. Consequently, the token-skeleton map $\mathcal{Y}_{2D}$ is computed with a pooling kernel of size
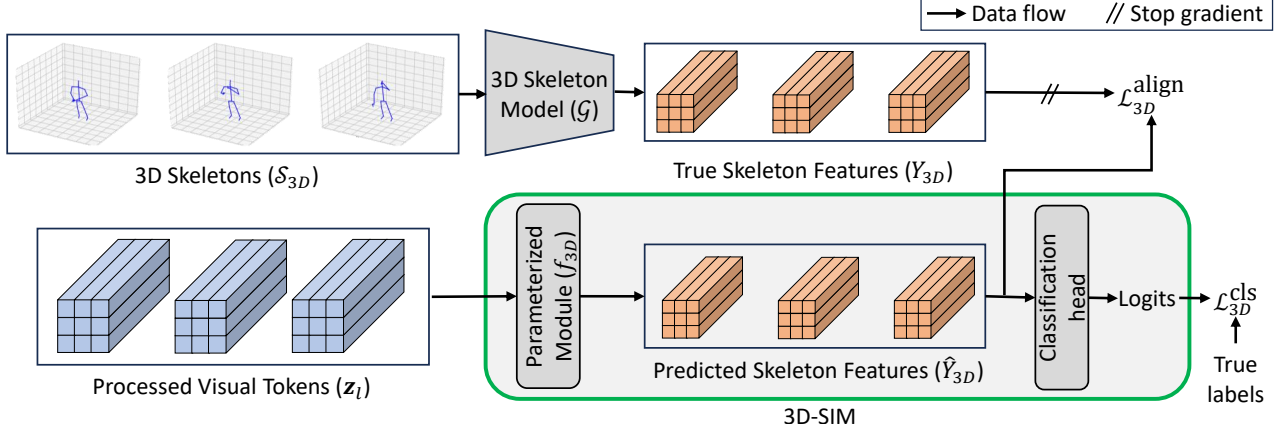
Figure 4. **An overview of the 3D Skeleton Induction Module (3D-SIM).** 3D-SIM processes visual tokens from the video transformer layer preceding it. A parameterized module transforms these tokens to obtain predicted skeleton features, which are then aligned with corresponding features generated by a pre-trained 3D skeleton model. During training, the weights of the skeleton model are not updated.

equal to $\tau \times p \times p$ as

$$\mathcal{Y}_{2D} = \text{MaxPool}_{\tau \times p \times p}(\mathcal{M}) \tag{5}$$

Now, $\mathcal{M}$ takes the shape of $T_v \times S_v \times J$ and indicates the presence or absence of a particular joint *at the token level*.

Auxiliary Task of 2D-SIM. Since the goal of 2D-SIM is to learn representations for fine-grained appearance actions, it enforces extra supervision to the RGB regions containing human joints. This is achieved through the use of a multi-class skeleton joint classification in the video transformer. Given the output tokens of the video transformer at the layer preceding 2D-SIM, $\mathbf{z}_l$, 2D-SIM directly predicts the presence or absence of every joint for each token in $\mathbf{z}_l$. Thus, it is a parameterized module $f_{2D}()$ with a skeleton joint classification head. The token-skeleton map predictions of 2D-SIM are computed as

$$\hat{Y}_{2D} = W f_{2D}(\mathbf{z}_l) + b \tag{6}$$

where W and b are the parameters of the skeleton joint classification head. In practice, we find that $f_{2D}()$ implemented using a simple fully-connected layer that projects z_l from $d_v \rightarrow d_b$ embedding space is sufficient to learn the token-skeleton mapping. After performing the prediction, the output shape of 2D-SIM aligns with that of the token-skeleton map, i.e., $T_v \times S_v \times J$. The loss for 2D-SIM is computed as the binary cross-entropy between $\mathcal{Y}_{2D}$ and $\hat{Y}_{2D}$

$$\mathcal{L}_{2D} = -\frac{1}{J} \sum_{i=1}^{J} [\mathcal{Y}_{2D}^i \log(\hat{Y}_{2D}^i) + (1 - \mathcal{Y}_{2D}^i) \log(1 - \hat{Y}_{2D}^i)] \tag{7}$$

3.2. 3D Skeleton Induction Module (3D-SIM)

Next we describe 3D-SIM, which shares the same modular design as 2D-SIM, allowing it to be inserted after any layer within the video transformer architecture. Unlike its 2D counterpart, 3D-SIM leverages 3D skeletons, which are inherently viewpoint-agnostic and effective at modeling fine-

grained motion. Specifically, 3D-SIM performs the auxiliary tasks of feature alignment and feature classification to enrich the visual representations with information from 3D skeletons. Like 2D-SIM, 3D-SIM is only applied during the training phase and removed during inference. An illustration of 3D-SIM is provided in Figure 4.

3D Skeleton Representation. The first step of 3D-SIM is to compute the 3D skeleton representation for a given video. Given a trained 3D skeleton action recognition model, $\mathcal{G}$, 3D-SIM enforces the RGB representations learned by the video transformer to align with the skeleton representations from $\mathcal{G}$. In practice, any 3D skeleton-based model can be chosen for $\mathcal{G}$, as long as the temporal and joint dimensions are preserved throughout the model. Given the visual tokens, $\mathbf{z}_l$, of the video transformer layer preceding 3D-SIM, and the corresponding set of 3D skeleton joints:

$$\mathcal{S}_{3D} = \{(t, j, x, y, x)\} : 1 \leq t \leq T, 1 \leq j \leq J, \tag{8}$$

we first obtain 3D skeleton representations as $Y_{3D} = \mathcal{G}(\mathcal{S}_{3D})$, where the shape of Y_{3D} is $T_s \times J \times d_s$.

Auxiliary Tasks of 3D-SIM. Since the goal of 3D-SIM is to learn skeleton representations from visual representations, it employs two hallucination based auxiliary tasks. The first is RGB-skeleton feature alignment, and the second is classifying the hallucinated skeleton features into action classes in the training distribution. For skeleton feature alignment, we propose two levels of alignment: (1) global, and (2) local. Depending on the level of alignment, we update the skeleton representation Y_{3D} as follows

$$Y_{3D} = \begin{cases} \frac{1}{T_s} \frac{1}{J} \sum_{t=1}^{T_s} \sum_{j=1}^{J} Y_{3D}^{t,j} & \text{if global} \\ \frac{1}{J} \sum_{j=1}^{J} Y_{3D}^{j} & \text{if local} \end{cases} \tag{9}$$

In 3D-SIM, an intermediate visual representation $\tilde{\mathbf{z}}_l$ is

computed for the desired alignment level as

$$\tilde{\mathbf{z}}_l = \begin{cases} \frac{1}{T_v} \frac{1}{S} \sum_{t=1}^{T_v} \sum_{h=1}^{S} \mathbf{z}_{t,s} & \text{if global} \\ \frac{1}{S} \sum_{h=1}^{S} \mathbf{z}_s & \text{if local} \end{cases} \quad (10)$$

For brevity, we ignore the class token in the above equation. For clarity, we present the shapes of Y_{3D} and $\tilde{\mathbf{z}}_l$ obtained at different alignment levels in Table 1. Due to differences in temporal sampling between the visual and skeleton models, T_s and T_v may not be equal. In the case where $T_s > T_v$, we uniformly sample T_v frames from Y_{3D}. In the case where $T_s < T_v$, we repeat the last frame.

The final step is to project the visual representations $\tilde{\mathbf{z}}_l$ to the embedding space of the 3D skeleton representation. A parameterized module $f_{3D}()$ can be used to perform the projection:

$$\hat{Y}_{3D} = f_{3D}(\tilde{\mathbf{z}}_l) \quad (11)$$

Similar to 2D-SIM, this parameterized module $f_{3D}()$ can be implemented using a fully-connected layer that projects $\tilde{\mathbf{z}}_l$ from $d_v \rightarrow d_s$ embedding space. Thus, the shape of $\hat{Y}_{3D}$ is same as Y_{3D} based on the alignment level (see Table 1). Then, the alignment loss of 3D-SIM is computed using Mean Squared Error (MSE) as

$$\mathcal{L}_{3D}^{\text{align}} = \frac{1}{\lambda} \sum_{i=1}^{\lambda} (Y_{3D} - \hat{Y}_{3D})^2 \quad (12)$$

where $\lambda = 1$ for global alignment and $\lambda = T$ for local alignment in the common visual skeleton semantic space. Finally, 3D-SIM performs another auxiliary task of classifying the predicted skeleton features into actions. $\hat{Y}_{3D}$ is processed by a classification head implemented using a FC layer to obtain logits. Then, the cross-entropy loss $\mathcal{L}_{3D}^{\text{cls}}$ between the logits and ground truth action label of V is computed. This classification auxiliary task improves the discriminability of the predicted skeleton features, thus inducing discriminative 3D skeleton features into the RGB cue. Thus, the total 3D-SIM loss is obtained as

$$\mathcal{L}_{3D} = \mathcal{L}_{3D}^{\text{align}} + \mathcal{L}_{3D}^{\text{cls}} \quad (13)$$

Table 1. Shape of Y_{3D} and $\tilde{\mathbf{z}}_l$ based on alignment level.

Alignment Level	Shape of Y_{3D}	Shape of $\tilde{\mathbf{z}}_l$
Global	d_s	d_v
Local	$T_s \times d_s$	$T_v \times d_v$

3.3. Pose Induced Video Transformer (π-ViT)

Finally, we integrate 2D-SIM and 3D-SIM together into the video transformer architecture to obtain π-ViT. π-ViT learns unified RGB and pose representations that are view-invariant and can discriminate fine-grained and similar actions, facilitated by the complementary nature of 2D-SIM and 3D-SIM. During the training phase, the modules can be inserted after any layer of the video transformer independently of one another. The total loss that is optimized

during training is:

$$\mathcal{L}_{total} = \mathcal{L}_v^{cls} + \mathcal{L}_{2D} + \mathcal{L}_{3D} \quad (14)$$

During the inference phase, both modules are removed and only the backbone video transformer architecture is used.

4. Experimental Results

Datasets. We assess our methods on three popular ADL datasets: Toyota-Smarthome [13] (Smarthome, SH), NTU120 [32], and NTU60 [42]. Smarthome comprises 16K videos across 31 actions, using cross-subject (CS) and two cross-view protocols (CV1, CV2), measured by mean class accuracy (mCA). NTU120, with 114K videos and 120 actions, follows CS and cross-setup (CSet) protocols. NTU60, a subset of NTU120, includes 57K videos of 60 actions, using CS and cross-view (CV) protocols. NTU60's challenging cross-view-subject [50] (CVS) protocols are used for ablation studies. Further information on the dataset details can be found in the supplementary material.

Implementation details. In all experiments, we use a pre-trained TimeSformer [6] as the backbone video transformer architecture into which we insert our 2D-SIM and 3D-SIM. We use Hyperformer [58] as our backbone 3D skeleton model, $\mathcal{G}$, for 3D-SIM. Default hyperparameter settings for both modules are shown in Table 4. Additional implementation details can be found in the supplementary material.

4.1. Comparison with State-of-the-art

We compare our π-ViT with state-of-the-art (SoTA) approaches under three categories: (1) Pose Only, (2) RGB + Pose, and (3) RGB Only (at inference). For our RGB + Pose approach, we perform a late fusion between the logits from π-ViT and Hyperformer. This shows that if 3D Poses are available at inference, our method performs competitively against other RGB + Pose approaches, including PoseC3D [17] which holds SoTA on NTU120 and NTU60.

Toyota-Smarthome. In Table 2, we present the comparison of π-ViT with the SoTA on Smarthome. π-ViT using only RGB at inference achieves SoTA over approaches in all three categories, notably on approaches using both RGB and 3D Poses at inference. π-ViT also achieves a considerable improvement over the TimeSformer backbone, with relative improvements of +6.6%, +10.4%, and +6.9% on CS, CV1, and CV2 protocols respectively. We also find that 2D-SIM and 3D-SIM learn complementary representations, improving the performance of π-ViT over 2D-SIM and 3D-SIM by +0.6% and +2.0% respectively on CS protocol.

NTU120 and NTU60. We present the comparison of π-ViT with the SoTA on NTU120 and NTU60 in Table 3 and Table 5 respectively. On NTU120, π-ViT achieves SoTA compared to other RGB only approaches on the CS and CSet protocols. When compared to the baseline video transformer, π-ViT achieves notable performance boosts by up

Table 2. **Comparison with SoTA on Toyota-Smarthome dataset.** We report the mean class accuracy on cross-subject (CS) and cross-view (CV$_1$, CV$_2$) protocols. ∘ indicates that the modality has been used only in training. Bold text indicates best performance, underline indicates second best performance. † indicates results produced by the authors.

Methods	Modality		CS	CV$_1$	CV$_2$
	Pose	RGB			
Pose Only					
2s-AGCN [29]	✓	✗	60.9	21.6	32.3
PoseC3D† [17]	✓	✗	50.6	20.0	28.2
Hyperformer† [58]	✓	✗	57.5	31.6	35.2
RGB + Pose					
P-I3D [12]	✓	✓	54.2	35.1	50.3
Separable STA [13]	✓	✓	54.2	35.2	50.3
VPN [14]	✓	✓	65.2	43.8	54.1
VPN++ + 3D Poses [15]	✓	✓	71.0	-	58.1
PoseC3D [17]	✓	✓	53.8	21.5	33.4
π-ViT + 3D Poses (Ours)	✓	✓	73.1	55.6	65.0
RGB Only (at inference)					
AssembleNet++ [39]	✗	✓	63.6	-	-
LTN [57]	✗	✓	65.9	-	54.6
VPN++ [15]	∘	✓	69.0	-	54.9
Video Swin† [34]	✗	✓	69.8	36.6	48.6
MotionFormer† [35]	✗	✓	65.8	45.2	51.0
TimeSformer† [6]	✗	✓	68.4	50.0	60.6
+ 2D-SIM (Ours)	∘	✓	72.5	54.8	62.9
+ 3D-SIM (Ours)	∘	✓	71.4	51.2	62.3
π-ViT (Ours)	∘	✓	72.9	55.2	64.8

Table 3. **Comparison with SoTA on NTU120 dataset.** We report the top-1 accuracy on cross-subject (CS) and cross-setup (CS) protocols. ∘ indicates that the modality has been used only in training. Bold text indicates best performance, underline indicates second best performance. † indicates results produced by the authors.

Methods	Modality		CS	CSet
	Pose	RGB		
Pose Only				
InfoGCN [10]	✓	✗	89.8	91.2
PoseC3D [17]	✓	✗	86.0	89.6
Hyperformer [58]	✓	✗	86.6	88.0
3Mformer [55]	✓	✗	92.0	93.8
RGB + Pose				
VPN [14]	✓	✓	86.3	87.8
VPN++ + 3D Poses [15]	✓	✓	90.7	92.5
PoseC3D [17]	✓	✓	95.3	96.4
STAR-Transformer [2]	✓	✓	90.3	92.7
3D-Def-Transformer [27]	✓	✓	90.5	91.4
π-ViT + 3D Poses (Ours)	✓	✓	95.1	96.1
RGB Only (at inference)				
VPN++ [15]	∘	✓	86.7	89.3
ViewCLR [11]	✗	✓	86.2	84.5
ViewCon [41]	✗	✓	85.6	87.5
Video Swin† [34]	✗	✓	91.4	92.1
MotionFormer† [35]	✗	✓	87.0	87.9
TimeSformer† [6]	✗	✓	90.6	91.6
+ 2D-SIM (Ours)	∘	✓	90.5	91.6
+ 3D-SIM (Ours)	∘	✓	91.8	92.7
π-ViT (Ours)	∘	✓	91.9	92.9

to +1.4% on CSet protocol. On NTU60, π-ViT achieves SoTA when compared to other RGB only approaches on the CS protocol, and competitive performance on the CV protocol (97.9% vs ViewCon's 98.0%). Consistent with other datasets, we also observe a performance boost of π-ViT over the TimeSformer backbone (up to +1.1% on CS).

In contrast to Smarthome, we find that 2D-SIM does not improve the action classification performance on NTU datasets. We attribute this to the lack of fine-grained appearance actions in NTU, as these videos are predominantly captured in controlled laboratory settings. Nonetheless, the integration of both 2D-SIM and 3D-SIM within the unified architecture of π-ViT facilitates the model's ability to still acquire complementary representations from 3D-SIM. We also note that PoseC3D [17] outperforms π-ViT on NTU. However, it exhibits significantly lower accuracy in real-world scenarios, particularly in Smarthome environments. This suggests that PoseC3D's effectiveness is heavily dependent on the high quality of pose data.

Other video transformers. We argue that the action recognition performance of the video transformer architectures are already competitive to the previous SoTA [15]. Across all datasets, π-ViT consistently outperforms all the representative video transformers [6, 34, 35] substantiating the effectiveness of π-ViT for understanding ADL.

Runtime vs performance. In Fig. 5, we compare runtime and accuracy of π-ViT with the prior SoTA (VPN++) and pose-only methods on the Toyota-Smarthome dataset. Runtimes include forward pass and modality extraction (e.g., 2D/3D pose) times. We find that π-ViT surpasses VPN++ in both runtime and performance, and outperforms pose-only approaches with significantly shorter runtimes. These achievements are due to π-ViT's use of poses at training time only. If 3D poses are available to π-ViT during inference, it can match the runtimes of pose only approaches while delivering superior performance.

4.2. Ablation Study

In this section, we ablate the design choices of 2D-SIM and 3D-SIM. Ablations are performed on the Smarthome cross-subject (SH CS) and NTU60 CVS1 (NTU CVS1) protocols. **What parameterized module should be used?** We ablate the choice of parameterized modules, f_{2D} and f_{3D}, in Table 4a. We evaluate a single FC layer, a 4 layer multi-layer perceptron (MLP), and a 2 layer transformer encoder [51] (Transformer). We observe that the performance often improves when a heavier MLP/transformer module is used (2D-SIM on NTU60 CVS1). On the other hand, these modules can also degrade the performance (2D-SIM and 3D-SIM on SH CS). We find that a single fully-connected layer yields strong and consistent performance.
Where should 2D-SIM and 3D-SIM be placed? In Ta-

Table 4. **Ablation studies.** We ablate the design choices of 2D-SIM and 3D-SIM on the Smarthome cross-subject (SH CS) and NTU60 CVS1 (NTU CVS1) protocols. For each experiment, we highlight the default design choice in gray.

(a) **Choice of parameterized module.** A simple fully connected layer is sufficient.

Dataset	Module	Choice of f_{2D} and f_{3D}		
		FC	MLP	Transformer
SH CS	2D-SIM	72.5	69.9	68.4
	3D-SIM	71.4	70.7	69.9
NTU CVS1	2D-SIM	88.0	88.74	89.1
	3D-SIM	90.61	90.53	90.32

(b) **Comparison of 3D-SIM with traditional distillation.** 3D-SIM's auxiliary alignment is best.

Approach	SH CS	NTU CVS1
Baseline (TimeSformer)	68.4	86.5
+ FD with class token	67.1	85.1
+ FD with distillation token	67.6	84.4
+ LD with class token	68.3	87.9
+ LD with distillation token	69.1	86.7
+ 3D-SIM (Ours)	71.4	90.6

(c) **Alignment level and position of 3D-SIM.** Global alignment at position 12 performs best.

Dataset	Baseline	Alignment level	3D-SIM Position			
			1	6	12	1,6,12
SH CS	68.44	Global	70.3	68.9	71.4	67.7
		Local	68.8	68.3	70.6	69.3
		Global+Local	68.0	68.3	71.3	70.5
NTU CVS1	86.52	Global	89.0	89.0	90.6	89.7
		Local	88.8	89.1	90.3	89.3
		Global+Local	88.5	89.3	90.1	90.6

(d) **Token-skeleton map variants.** The proposed joint-specific token-skeleton map performs best.

Variant	SH CS	NTU CVS1
Token-Skeleton Map	72.5	88.0
Flat Variant	71.1	87.8
Depth Variant	71.3	87.8

(e) **Classification task of 3D-SIM.** The classification task yields consistently better performance.

Dataset	Alignment level	Classification task	
		✓	✗
SH CS	Global	71.4	70.6
	Local	70.6	70.4
NTU CVS1	Global	90.3	88.6
	Local	90.6	89.1

(f) **Position of 2D-SIM.** 2D-SIM performs best when placed near the beginning of the model.

Dataset	2D-SIM Position				
	1	6	12	1,6	1,12
NTU CVS1	88.0	87.9	87.5	87.8	87.7
SH CS	72.5	70.9	69.9	70.7	70.2

Table 5. **Comparison with SoTA on NTU60 dataset.** We report the top-1 accuracy on cross-subject (CS) and cross-view (CV) protocols. ○ indicates that the modality has been used only in training. Bold text indicates best performance, underline indicates second best performance. † indicates results produced by the authors.

Methods	Modality		CS	CV
	Pose	RGB		
Pose Only				
InfoGCN [10]	✓	✗	93.0	97.1
PoseC3D [17]	✓	✗	93.7	96.6
Hyperformer [58]	✓	✗	90.7	95.1
3Mformer [55]	✓	✗	94.8	98.7
RGB + Pose				
Separable STA [13]	✓	✓	92.2	94.6
VPN [14]	✓	✓	95.5	98.0
VPN++ + 3D Poses [15]	✓	✓	96.6	99.1
PoseC3D [17]	✓	✓	**97.0**	**99.6**
STAR-Transformer [2]	✓	✓	92.0	96.5
ViewCon [41]	✓	✓	93.7	98.9
3D-Def-Transformer [27]	✓	✓	94.3	97.9
π-ViT + 3D Poses (Ours)	✓	✓	96.3	99.0
RGB Only (at inference)				
Glimpse Clouds [5]	○	✓	86.6	93.0
VPN++ [15]	○	✓	93.5	96.1
Vyas *et al.* [52]	✗	✓	82.3	86.3
ViewCLR [11]	✗	✓	89.7	94.1
Piergiovanni *et al.* [37]	✗	✓	-	93.7
ViewCon [41]	✗	✓	91.4	**98.0**
Video Swin† [34]	✗	✓	93.4	96.6
MotionFormer† [35]	✗	✓	85.7	91.6
TimeSformer† [6]	✗	✓	93.0	97.2
+ 2D-SIM (Ours)	○	✓	93.0	97.0
+ 3D-SIM (Ours)	○	✓	94.0	97.8
π-ViT (Ours)	○	✓	**94.0**	97.9

ble 4f and Table 4c, we ablate the insertion positions of 2D-SIM and 3D-SIM. 2D-SIM consistently performs best when placed near the initial layers of the transformer, where it has access to low-level, non-contextualized [1] tokens. On the contrary, 3D-SIM performs best when placed near the deeper layers of the transformer, where it has access to more high-level tokens with abstract representations.

How should we construct the token-skeleton map? We explore two alternative variants of the token-skeleton map and present the results in Table 4d. Recall from Section 3.1 that for a single token, the goal of 2D-SIM is to predict the presence or absence of specific joints within the corresponding RGB patch. In the flat variant of the token-skeleton map, 2D-SIM's task is to predict if the corresponding patch contains *any* joint. In the depth variant, we provide an additional dimension corresponding to depth, and thus 2D-SIM's task is to predict the specific joints contained in each patch, as well as their depth. This demonstrates that the auxiliary task of token-joint mapping provides relevant human anatomy-based supervision to the RGB cue, while depth does not offer the same level of relevant supervision.

Which alignment level is best for 3D-SIM? In Table 4c, we explore different alignment levels of 3D-SIM. We observe that global alignment consistently yields performance boosts at all positions. In global alignment, the pooling over the temporal dimension mitigates noise and reduces variability, facilitating the auxiliary feature alignment task.

Is 3D-SIM's classification task necessary? In Table 4e, we present the results of 3D-SIM with and without the additional classification task. We find that the inclusion of this task consistently improves the performance. This additional task enforces 3D-SIM to learn more discriminative skeleton representation through feature alignment.

3D-SIM or traditional distillation? In Table 4b, we ablate the choice of alignment between RGB and 3D skeleton representations. A naive approach is to perform traditional knowledge distillation [25, 48] (KD) from a 3D skeleton model to the RGB video transformer. These traditional approaches can be summarized by distilling logits or features
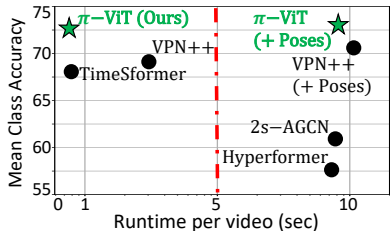
Figure 5. **Runtime vs accuracy plot** on Toyota-Smarthome. Red dashed line indicates use of 3D poses at inference.
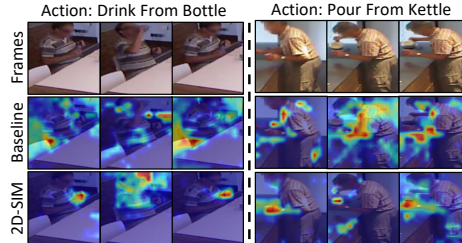

Figure 6. **GradCAM visualization** of the baseline TimeSformer and our 2D-SIM on two fine-grained appearance actions.
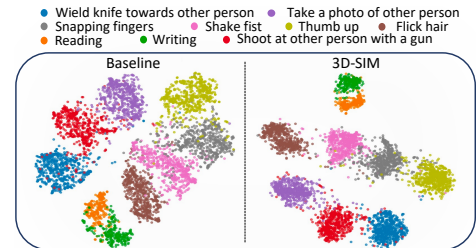

Figure 7. **TSNE Visualization** of the baseline TimeSformer and our 3D-SIM embeddings on nine visually similar actions.

into a class token or distillation token. We perform all four combinations of traditional KD and find them ineffective for feature distillation. However, 3D-SIM's feature alignment task in a different embedding space overcomes the limitations of traditional KD limitations and effectively induces pose information into video transformers.

4.3. Does π-ViT Address ADL Challenges?

In this section, we analyze 2D-SIM and 3D-SIM on the challenges of ADL. In Figure 6, we compare the Gradient Class Activation Maps (GradCAM) [40] of the baseline TimeSformer against 2D-SIM on two fine-grained action videos from Smarthome. We observe that 2D-SIM prioritizes RGB regions containing the relevant human joints that characterize the actions. In Figure 7, we use T-SNE [49] to visualize embeddings of the baseline TimeSformer and 3D-SIM on nine visually similar actions from NTU120. We find that 3D-SIM learns more densely grouped clusters than TimeSformer, highlighting its discriminative power and ability to disambiguate visually similar actions.

5. Related Works

Recent advancements in vision transformers [16, 18, 33, 48] have surpassed CNNs [9, 23, 47] on image-based tasks. Similarly, video transformers [3, 6, 18, 30, 34] have excelled over 3DCNNs [7, 19, 31] and two-stream CNNs [20, 21, 45] in video tasks. These video transformers, optimized for web videos, struggle on ADL videos [13, 32, 42, 44, 53], which pose unique challenges. Approaches using human poses [10, 22, 43, 56] are effective in laboratory settings [32, 42, 54] but limited in real-world videos [13, 44].

Therefore, several approaches combine the RGB and pose modalities [2, 14, 27] to address the challenges of ADL. Recently, STAR-Transformer [2] and 3D deformable transformer [27] have effectively utilized these modalities within video transformers. Both STAR-Transformer and 3D deformable transformer introduce specialized spatio-temporal attention mechanisms to enable cross-modal learning in video transformers. These methods capitalize on the efficacy of video transformers discussed above.

While the above approaches combining RGB and pose are effective, the burden of collecting poses at inference time is high, as specialized sensors or expensive pose estimation is required. Thus, approaches [5, 15, 37] using only the RGB are desirable. Our proposed π-ViT is one such approach that does not require any pose information at inference. Closest to our work is VPN++ [15], which integrates 3D pose information into a CNN-based RGB backbone through feature and attention level distillations, and does not require poses at inference. However, our experiments show that these distillations are sub-optimal in video transformers. Moreover, the previous RGB + pose approaches overlook the correspondence between 2D skeleton joints and spatial regions of interest. In contrast, π-ViT induces both 2D and 3D pose information within the RGB cue in video transformers through pose-aware auxiliary tasks.

6. Conclusion

In conclusion, we propose π-ViT, the first video transformer model leveraging both 2D and 3D human poses for understanding ADL videos. π-ViT consists of two novel plug-in modules, 2D-SIM and 3D-SIM, which are inserted into a video transformer model. Each module performs a distinct auxiliary task that induces human pose knowledge into the RGB representation space of the model, enabling π-ViT to address the specific challenges of ADL. Notably, the modules are only required during training and thus there is no requirement of poses at inference, drastically reducing the computational cost of π-ViT. We show that π-ViT effectively addresses the challenges of ADL and achieves state-of-the-art performance on popular ADL datasets.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online, 2020. Association for Computational Linguistics. 7

[2] Dasom Ahn, Sangwon Kim, Hyun Wook Hong, and ByoungChul Ko. Star-transformer: A spatio-temporal cross attention transformer for human action recognition. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3319–3328, 2023. 2, 6, 7, 8

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6836–6846, 2021. 1, 2, 8

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. 3

[5] Fabien Baradel, Christian Wolf, Julien Mille, and Graham W. Taylor. Glimpse clouds: Human activity recognition from unstructured feature points. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 7, 8

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. 1, 2, 3, 5, 6, 7, 8

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4724–4733. IEEE, 2017. 8

[8] João Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *CoRR*, abs/1808.01340, 2018. 1

[9] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *Brit. Mach. Vis. Conf.*, 2014. 8

[10] 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20154–20164, 2022. 6, 7, 8

[11] Srijan Das and Michael S. Ryoo. Viewclr: Learning self-supervised video representation for unseen viewpoints. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, pages 5562–5572. IEEE, 2023. 6, 7

[12] Srijan Das, Arpit Chaudhary, Francois Bremond, and Monique Thonnat. Where to focus on for human action recognition? In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 71–80, 2019. 6

[13] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome: Real-world activities of daily living. In *Int. Conf. Comput. Vis.*, 2019. 5, 6, 7, 8

[14] Srijan Das, Saurav Sharma, Rui Dai, Francois Bremond, and Monique Thonnat. Vpn: Learning video-pose embedding for activities of daily living. In *European Conference on Computer Vision*, pages 72–90. Springer, 2020. 2, 6, 7, 8

[15] Srijan Das, Rui Dai, Di Yang, and Francois Bremond. Vpn++: Rethinking video-pose embeddings for understanding activities of daily living. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 6, 7, 8

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *Int. Conf. Learn. Represent.*, 2021. 8

[17] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2959–2968, 2022. 2, 5, 6, 7

[18] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, 2021. 2, 8

[19] Christoph Feichtenhofer. X3D: expanding architectures for efficient video recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 200–210. Computer Vision Foundation / IEEE, 2020. 8

[20] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 1933–1941. IEEE, 2016. 8

[21] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6201–6210. IEEE, 2019. 8

[22] Ryo Hachiuma, Fumiaki Sato, and Taiki Sekii. Unified keypoint-based action recognition framework via structured keypoint pooling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22962–22971, 2023. 8

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 8

[24] Yun He, Soma Shirakabe, Yutaka Satoh, and Hirokatsu Kataoka. Human action recognition without human. In *ECCV Workshops*, 2016. 1

[25] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. 7

[26] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[27] Sangwon Kim, Dasom Ahn, and Byoungchul Ko. Cross-modal learning with 3d deformable attention for action

recognition. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 6, 7, 8

[28] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011. 1

[29] Shi Lei, Zhang Yifan, Cheng Jian, and Lu Hanqing. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, 2019. 2, 6

[30] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, 2022. 8

[31] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 8

[32] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 5, 8

[33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Int. Conf. Comput. Vis.*, 2021. 8

[34] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3192–3201, 2021. 1, 2, 6, 7, 8

[35] Mandela Patrick, Dylan Campbell, Yuki M. Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F. Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 12493–12506, 2021. 6, 7

[36] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[37] AJ Piergiovanni and Michael S. Ryoo. Recognizing actions in videos from unseen viewpoints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4124–4132, 2021. 7, 8

[38] Grégory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2

[39] Michael S. Ryoo, AJ Piergiovanni, Juhana Kangaspunta, and Anelia Angelova. Assemblenet++: Assembling modality representations via attention connections. In *Eur. Conf. Comput. Vis.*, 2020. 6

[40] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 8

[41] Ketul Shah, Anshul Shah, Chun Pong Lau, Celso M. de Melo, and Rama Chellapp. Multi-view action recognition using contrastive learning. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3370–3380, 2023. 2, 6, 7

[42] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 5, 8

[43] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions on Image Processing*, 29:9532–9545, 2020. 8

[44] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In *European Conference on Computer Vision(ECCV)*, 2016. 8

[45] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014. 8

[46] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 1

[47] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. 8

[48] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers amp; distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 7, 8

[49] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 8

[50] Gül Varol, Ivan Laptev, Cordelia Schmid, and Andrew Zisserman. Synthetic humans for action recognition from unseen viewpoints. *International Journal of Computer Vision*, 129:2264 – 2287, 2019. 5

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 6

[52] Shruti Vyas, Yogesh Singh Rawat, and Mubarak Shah. Multi-view action recognition using cross-view video prediction. In *European Conference on Computer Vision*, 2020. 7

[53] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining Actionlet Ensemble for Action Recognition with Depth Cameras. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 8

[54] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning, and recognition. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2649–2656, 2014. 8

[55] Lei Wang and Piotr Koniusz. 3mformer: Multi-order multi-mode transformer for skeletal action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5620–5631, 2023. 2, 6, 7

[56] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018. 2, 8

[57] Di Yang, Yaohui Wang, Quan Kong, Antitza Dantcheva, Lorenzo Garattoni, Gianpiero Francesca, and François Brémond. Self-supervised video representation learning via latent time navigation. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2023. 6

[58] Yuxuan Zhou, Zhi-Qi Cheng, Chao Li, Yifeng Geng, Xuansong Xie, and Margret Keuper. Hypergraph transformer for skeleton-based action recognition. *arXiv preprint arXiv:2211.09590*, 2022. 2, 5, 6, 7