PGVT: Pose-Guided Video Transformer for Fine-Grained Action Recognition

Haosong Zhang[1,2], Mei Chee Leong[1], Liyuan Li[1], Weisi Lin[2]
Institute for Infocomm Research (I^{2}R), A*STAR, Singapore[1]
Nanyang Technological University[2]
haosong001@e.ntu.edu.sg

Abstract

Based on recent advancements in transformer-based video models and multi-modal joint learning, we propose a novel model, named Pose-Guided Video Transformer (PGVT), to incorporate sparse high-level body joints locations and dense low-level visual pixels for effective learning and accurate recognition of human actions. PGVT leverages the pre-trained image models by freezing their parameters and introducing trainable adapters to effectively integrate two input modalities, i.e., human poses and video frames, to learn a pose-focused spatiotemporal representation of human actions. We design two novel core modules, i.e., Pose Temporal Attention and Pose-Video Spatial Attention, to facilitate interaction between body joint locations and uniform video tokens, enriching each modality with contextualized information from the other. We evaluate PGVT model on four action recognition datasets: Diving48, Gym99, and Gym288 for fine-grained action recognition, and Kinetics400 for coarse-grained action recognition. Our model achieves new SOTA performance on the three fine-grained human action recognition datasets and comparable performance on Kinetics400 with a small number of tunable parameters compared with SOTA methods. Various ablation studies are performed which verify the benefits of our new designs.

1. Introduction

Recognizing human actions involves capturing both spatial and temporal dynamics as well as semantic representation of human movements. Existing methods that learn spatiotemporal visual features and dependencies from video clips tend to be computationally expensive and less effective in learning human movements. In [21], a network model fusing human 2D pose and video clip for human action recognition is proposed. The inputs are images, i.e., the heatmaps of joints and video frames, which are fed to two CNN channels for visual feature extraction. The outputs from the two channels are concatenated for final classification. It also shows that introducing fusion in both early and later stages in the model is more effective than simply applying fusion in the later stage for joint learning on pose and visual features.

Parameter-efficient transfer learning, where large pre-trained models are frozen while a few additional parameters are fine-tuned, has gained pace in computer vision. However, the direct utilization of pre-trained image models for video and pose tasks has received less attention. Furthermore, given the remarkable performance of 2D pose detectors, it is intuitive to leverage pose-guided representations for video understanding tasks, as explored in previous studies [7, 65] and more recently by Duan et al. [21]. Nevertheless, the effective way to handle pose information is still up for discussion. Relying solely on the pose modality limits the learning of non-body parts in the video, leading to a weakness in robust semantic understanding. Since video modality provides richer visual information compared to the pose modality alone, it is helpful to learn diverse and discriminate visual features. On human action recognition, the pose represented by joint points provides a high-level knowledge of articulated human body structure and body movement, and the video frames show the pixel-level dense visual information of human movement in a real-world environment. Intuitively, joint learning on these two modalities would achieve accurate and effective knowledge for fine-grained human action recognition where the difference between some classes is small. The self-attention and cross-attention techniques of recent transformer architectures should give rise to a suitable way to achieve the goal.

Motivated by the aforementioned observations and ideas, our primary target of model design is to explicitly integrate pose-guided spatiotemporal representations within vision transformer architectures [18], extending this fusion across different layers of the model. Such architecture presents several challenges, including effectively focusing on the appearance of human bodies as they move, capturing the interaction among body joints, and incorporating the dynamics of the pose with the diversity of visual appearance. In addition, it is also helpful to involve contextual information from the video content outside the human body as background information.

We design a novel *Pose-Guided Video Transformer* (*PGVT*) that effectively integrates pose tokens and video tokens, leading to effective spatiotemporal representations. Our key idea involves introducing pose (2D body joint coordinates) and video information in a manner similar to regular patches, while also naturally integrating temporal dynamics into this framework. By integrating pose information into the vision transformer architecture, we aim to improve the model's capacity to capture intricate spatial and temporal relationships for efficient action recognition. Through thorough experiments on various datasets, we illustrate the efficacy of our PGVT model, outperforming existing methods on three fine-grained action recognition datasets and highlighting the advantages of incorporating pose guidance within vision transformers. Additionally, we show that PGVT requires comparative or significantly lower computing costs by fixing the pre-trained image model and training several adapters [37,100]. The contributions can be summarized as:

1. A novel multi-modal approach PGVT (*Pose-Guided Video Transformer*) that integrates 2D poses and video frames to learn refined pose-guided spatiotemporal representations for fine-grained action recognition. We introduce two new modules: i) *Pose Temporal Attention* employs self-attention over pose representations to capture trajectory interactions; ii) *Pose-Video Spatial Attention* models appearance by applying co-attention over both video tokens and pose tokens to capture pose-weighted appearance features.

2. The PGVT leverages a pre-trained vision foundation model by freezing its parameters and introducing a lightweight *Refining Adapter* to effectively integrate the two input modalities, equipping them with pose-guided spatiotemporal reasoning capabilities.

3. We validate PGVT on four video action recognition datasets where the pose is given as part of the input. Our extensive empirical study shows improved results on Gym99, Gym288, Diving48, and comparative results on Kinetics400, compared with existing SOTA methods.

2. Related Work

Video action recognition. Action recognition has been a longstanding problem in computer vision. Previous methods focused on optical flow-based features [22], while recent advances have seen the emergence of transformer-based approaches [23]. The evolution of these approaches can be broadly categorized from temporal pooling for feature extraction [43] to recurrent networks [17,66], and eventually to 3D spatiotemporal kernels [11,34,38,60,84,86,88] and two-stream networks that capture complementary signals, such as motion and spatial cues [29, 30, 81]. The recent emergence of vision transformers has introduced powerful models for video understanding [1, 3, 23, 69], building upon the achievement of language transformers [16] and vision transformers [8, 19]. The use of transformers for video action recognition has gained significant attention, following the success of vision transformers(e.g.ViT [18]) in the image domain. Instead of traditional convolutional networks [9, 29, 60, 85, 99], recent approaches have focused on extending image pre-trained models to handle video data. This is achieved by introducing new temporal modules [1, 4, 98, 107] or by inflating image models to video models [62]. These models "patchify" each video frame and employ self-attention mechanisms to obtain contextualized representations for these patches. However, a notable limitation of this approach is the absence of an explicit representation of the human skeleton. In this paper, we highlight a finding that self-attention can be jointly adopted to pose features and video features, providing a straightforward and elegant mechanism to enhance spatiotemporal representations through the inclusion of pose information. Our work leverages two consecutive *Pose Temporal Attention* and *Pose-Video Spatial Attention* modules to exploit pose-centric information.

Pre-trained large vision models. Recent works [28,39,48, 71, 83, 92, 102, 110] have utilized large-scale multi-modal datasets, such as image/video-text pairs, to train models, resulting in robust visual representations. These pre-trained models demonstrated strong transferability and zero-shot learning capabilities, and have greatly facilitated transfer learning to various downstream tasks. However, full fine-tuning of video data requires high computational costs, which is impractical for many researchers and practitioners. Parameter-efficient fine-tuning techniques aim to reduce the number of trainable parameters, thereby lowering computational costs, while still achieving or surpassing the performance of full fine-tuning [2, 12, 31, 40, 41]. In our work, we efficiently adapt well-pre-trained image models (ViT pre-trained on CLIP [71]) to learn pose and video representations, by training only a few parameters and attaining comparable or even superior performance compared to previous SOTA. There are also several works that leverage pre-trained CLIP for video action recognition. ActionCLIP [90] and X-CLIP [67] require a text branch for full fine-tuning, while PromptCLIP [42] applies prompt tuning [52] and introduces blocks for temporal modeling. EVL [61] adds a decoder branch that can learn temporal information. While existing methods in computer vision focus on adapting models within the same modality, our method adapts image models for human pose, making it less computationally costly than previous methods.

Skeleton-based models. The progress of pose detection inspires skeleton-based action recognition. To address gradient explosion and vanishing gradient challenges in RNN [56, 79, 80, 105], some researchers [6, 32, 53, 106] propose neural networks (NN) to investigate skeleton sequences for spatial and temporal information through spatial-temporal convolutions, and GCN to treat the skeleton sequence as a spatial-temporal graph, leveraging the inherent graph structure of the skeleton [13, 14, 45, 49, 63, 77, 78, 96, 101]. JMRN [75] investigated to capture interdependencies between joints in heatmaps separately. Another line of work turns the skeleton sequence into image-like data using hand-crafted techniques [15, 97]. PoseConv3D [21] employs 3D CNN on a series of 2D skeleton heatmaps to extract features. 3D pose representation is also investigated, such as LART [72] which uses a 3D SMPL model, person-level tracking and tokenized human-centric vectors for action prediction. A multi-modal approach is also adopted [21, 58] to incorporate pose and video inputs for video understanding. However, our work differs conceptually as [21] feeds pose heatmap to a CNN, while [58] adopts a distillation framework using flow input. Our PGVT effectively integrates 2D body joint coordinates into the transformer while preserving the entire spatiotemporal representation.

3. Method

We introduce our *Pose-Guided Video Transformer* (PGVT) model, which explicitly incorporates pose trajectories and visual appearance into the transformer architecture. We provide a high-level overview of the framework in Section 3.1, and introduce how we embed pose and video clips in Section 3.2, followed by the more detailed architecture of a PGVT block in Section 3.3.

3.1. Overview of PGVT framework

The whole architecture is shown in Figure 1. It consists of an input layer, PGVT blocks and an output layer, where the PGVT block is the core novel part. The proposed model offers a powerful framework for efficiently leveraging pose information within vision transformers. We aim for the pose to exert influence on the scene's representation throughout a bottom-up process, allowing the calculation of attention at the pose level and other areas within the image. This approach ensures that the model can effectively attend to pose-related information and other significant image regions, ultimately enhancing its ability to understand and recognize actions in a more comprehensive manner. By incorporating both spatial and temporal attention, our model can capture the dynamics of pose and appearance-related semantics of pose-guided video. *Joint Prediction* layer, which takes the average of the CLS tokens over all frames for pose and video outputs separately and concatenates them for the final prediction, enhances the discriminative power of the network.

3.2. Input Layer: Embedding Pose and Video

In the *Pose Embedding* layer in Fig. 1, we first project the 2D pose sequence[1] of size $T_h \times N_P \times 2$ to a latent space of dimension d, which aligns with the latent space dimension of the pretrained image transformer, by a trainable Multilayer Perceptron (MLP). The MLP is trained end-to-end to learn the mapping of pose coordinate representation to high dimensional features, exploiting ViT tokens for effective representation of pose modality. We obtain pose tokens $P \in \mathbb{R}^{T_h N_P d}$, by adding a learnable position embedding and appending a classification token (CLS), where T_h denotes a high temporal resolution for the pose, $N_P = J + 1$ and J denotes the maximum number of joints. Similarly, we obtain video tokens through the *Video Embedding* layer. Building upon the architectural foundation established by ViT, we obtain video tokens $Z \in \mathbb{R}^{T_l N_Z d}$, where T_l denotes a low temporal resolution for the video, $N_Z = n_h n_w + 1$ and $n_h n_w$ denotes the number of spatial patches extracted in a frame. This representation allows us to apply transformer-based operations effectively for subsequent stages of action recognition. Self-attention mechanisms are applied iteratively to these tokens, producing the final contextualized CLS feature vectors for pose and video modalities.

3.3. The PGVT Block

The purpose of PGVT block is to learn the latent semantic dynamics of body movement using pose modality, and then move forward to joint learning with co-attention, i.e., learning pose-guided spatial dynamics by extracting information from sparse pose locations and utilising it to refine the video tokens, as well as to employ dense video pixels to refine the pose tokens. To be specific, each block takes these inputs and produces refined pose tokens and video tokens by leveraging information from the other modality. To achieve this, as shown in Fig. 1 (PGVT Block), we employ two essential modules within the block: the *Pose Temporal Attention* module (denoted as $\mathcal{T}$), which models pose trajectories, and the *Pose-Video Spatial Attention* module (denoted as $\mathcal{S}$), which captures appearance-related information guided by pose. In $\mathcal{T}$, we apply image pre-trained self-attention layers with Adapter [100] to the temporal dimension of pose input to learn the dependencies over frames. Adapter has a bottleneck structure as shown in Fig. 1 (Adapter). For $\mathcal{S}$, we introduce *Refining Adapter* after the image pre-trained self-attention layers. The PGVT block takes pose and video tokens from the previous blocks.

[1] We use an off-the-shelf pose extractor to extract 2D pose sequence from the raw video.

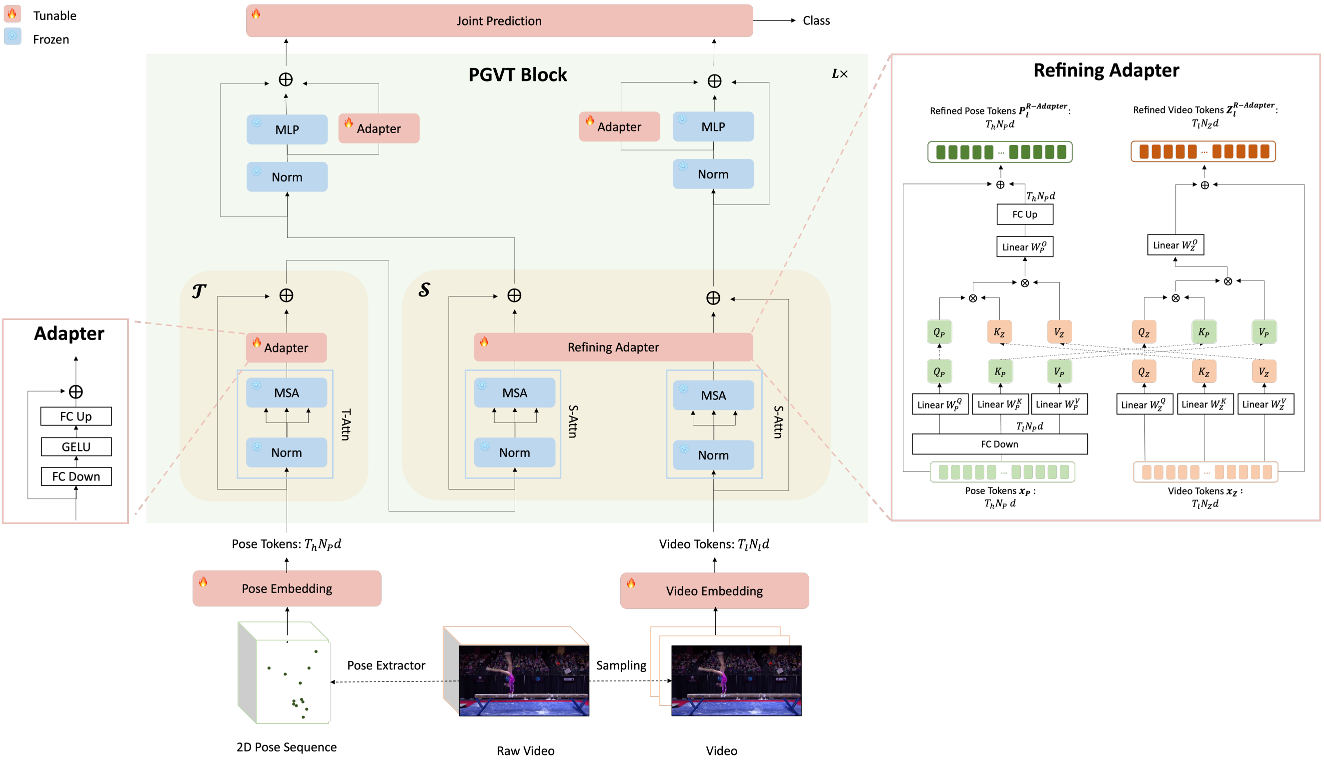

Figure 1. **PGVT Framework Architecture**. Our proposed framework consists of L stages of PGVT blocks. Each block is made up of two cascaded modules: a $\mathcal{T}$ module (*Pose Temporal Attention*) that models trajectories and a $\mathcal{S}$ module (*Pose-Video Spatial Attention*) that models appearance. The input pose tokens are fed into the $\mathcal{T}$ module and output pose tokens with learned temporal information. Afterwards, together with video tokens, temporal attended pose tokens are fed into the $\mathcal{S}$ module to output refined pose tensor and refined video tensor. The outputs of a PGVT block include the input residual connection for pose and video respectively. The final prediction is done by the *Joint Prediction* layer.

The output of the PGVT block consists of refined pose tokens and video tokens which are contextualized with information about each other.

Architecture. The architecture of our proposed method consists of multiple stages, denoted as L. Each stage contains two sets of tokens: pose tokens $\{\mathbf{p}_i\}_{i=1}^{N_P}$ and video tokens $\{\mathbf{z}_i\}_{i=1}^{N_Z}$. To simplify notation, we denote $\{\mathbf{p}_i^l\}_{i=1}^{N_P}$ as $\mathbf{P}_l$ and $\{\mathbf{z}_i^l\}_{i=1}^{N_Z}$ as $\mathbf{Z}_l$ for each stage l. Starting from l=1, in each stage, the pose tokens $\mathbf{P}_l$ are passed through $\mathcal{T}$, which facilitates the propagation of temporal information, resulting in temporal attended pose tokens as $\mathbf{P}_l^{\mathcal{T}} = \mathcal{T}(\mathbf{P}_l)$. Next, $\mathbf{P}_l^{\mathcal{T}}$ and the video tokens $\mathbf{Z}_l$ are passed through $\mathcal{S}$ to obtain the refined tokens:

$$\mathbf{P}_l^{\mathcal{S}}, \mathbf{Z}_l^{\mathcal{S}} = \mathcal{S}(\mathbf{P}_l^{\mathcal{T}}, \mathbf{Z}_l).$$

With residual connection, the above process is repeated for L stages.

Pose Temporal Attention. $\mathcal{T}$ (in Fig. 1) is in charge of capturing motion by functioning at a high refresh rate and with a high temporal resolution. It focuses solely on modeling pose dynamics independently of their appearance. Therefore, it only takes the pose P as input and produces the output $P^{\mathcal{T}}$. Considering its high temporal resolution, $\mathcal{T}$ is designed to have low computational costs by taking in dense high-level body structure. The focus of $\mathcal{T}$ is to model the geometry of motion using body joint coordinates and apply self-attention to them. Drawing inspiration from recent advancements in vision transformers, we propose a temporal transformer structure to thoroughly characterise the temporal correlations among human joints across frames. We denote the self-attention layer pre-trained in the image model as *Attn*, which is composed of a Layernorm and multiheaded self-attention (MSA) layer. Thereafter, we denote *T-Attn* for temporal modeling, and similarly, *S-Attn* for spatial modeling. Now given the pose tokens $P \in \mathbb{R}^{T_h \times N_P \times d}$, we first reshape it into $P \in \mathbb{R}^{N_P \times T_h \times d}$. Afterwards, we feed P into the *T-Attn* to capture the relationship among the T_h frames. Though *T-Attn* and *S-Attn* take in different input dimensions, they both share weights and are frozen. The computation of $\mathcal{T}$ can be written as

$$\begin{aligned}\mathbf{P}_l^{Adapter} &= \text{Adapter}(\text{T-Attn}(\mathbf{P}_{l-1})) \\ \mathbf{P}_l^{\mathcal{T}} &= \mathbf{P}_{l-1} + \mathbf{P}_l^{Adapter}\end{aligned} \tag{1}$$

where $\mathbf{P}_l^{\mathcal{T}}$ denotes the temporal attended output from $\mathcal{T}$ at stage l.

Pose-Video Spatial Attention. $\mathcal{S}$ (in Fig. 1) is in charge of capturing spatial semantics by functioning on frames of low temporal resolution. It models the attended appearance by considering both temporal attended pose representation $P^{\mathcal{T}}$ and video token representation Z and produces the output $\mathcal{S}(P^{\mathcal{T}}, Z)$. Inspired by multi-modal fusion methods [5, 20, 35, 95], *S-Attn* is applied independently to video and pose features, and followed by the *Refining Adapter* module (denoted as *R-Adapter*) for cross-modality learning. As shown in Fig. 1 (*Refining Adapter*), the left stream incorporates video tokens into pose tokens, while the right stream incorporates pose tokens into video patches, allowing cross-modal interaction in the embedding space. In the left stream, the pose tensor is first projected to a reduced number of frames (T_l) to match the temporal dimension of the video modality. This is done using the *FC Down* (denoted as FC_{down}), which is a linear layer for projecting the temporal dimension to a lower temporal resolution while learning the global temporal dynamics. Then the pose and video tokens are linear projected to obtain Q_i,K_i,V_i ($i \in P, Z$):

$$
\begin{aligned}
Q_P = W_P^Q \hat{x_P}, K_P = W_P^K \hat{x_P}, V_P = W_P^V \hat{x_P} \\
Q_Z = W_Z^Q x_Z, K_Z = W_Z^K x_Z, V_Z = W_Z^V x_Z
\end{aligned}
\tag{2}
$$

where $\hat{x_P}$ denotes $FC_{down}(x_P)$, x_P represents pose input to *R-Adapter* (i.e. S-Attn($\mathbf{P}_l^{\mathcal{T}}$)), x_Z represents video input to *R-Adapter* (i.e. S-Attn($\mathbf{Z}_{l-1}$)). Formally, attention between the pose tokens and video tokens is computed by the weighted sum of the tokens of the other modality. The fused features are projected by learned weights W_P^O and W_Z^O. Subsequently, the pose tensor is projected back to high temporal dimension T_h using the *FC Up* (denoted as FC_{up}), which is a linear layer to project pose to the same dimension as the input to our PGVT block. With a skip connection, we compute *Refining Adapter* (*R-Adapter*) as

$$
\begin{aligned}
\mathbf{P}_l^{R-Adapter} &= x_P + \mathrm{FC}_{\mathrm{up}}(W_P^O \cdot \mathrm{Softmax}\left(\frac{Q_P {K_Z}^T}{\sqrt{d_k}}\right) V_Z) \\
\mathbf{Z}_l^{R-Adapter} &= x_Z + W_Z^O \cdot \mathrm{Softmax}\left(\frac{Q_Z {K_P}^T}{\sqrt{d_k}}\right) V_P
\end{aligned}
\tag{3}
$$

where $\mathbf{P}_l^{R-Adapter}$ and $\mathbf{Z}_l^{R-Adapter}$ denote refined outputs from *R-Adapter*. This design allows for the repetitive application of the PGVT block, enabling the model to learn and refine representations iteratively. The *Refining Adapter* works similarly to the previously proposed semantic grouping methods [20, 64, 95]. While they learn instance-level grouping or cross attention of vision and language, our *Refining Adapter* refine pose and video tokens to adjust the weights using information from the other modality. The outputs of *Refining Adapter* are $T_h N_P$ pose tokens and $T_l N_Z$ video tokens, which can be perceived as refined token representations based on each source of information. The computation of $\mathcal{S}$ can be written as

$$
\begin{aligned}
\mathbf{P}_l^{\mathcal{S}} &= \mathbf{P}_l^{\mathcal{T}} + \mathbf{P}_l^{R-Adapter} \\
\mathbf{Z}_l^{\mathcal{S}} &= \mathbf{Z}_l + \mathbf{Z}_l^{R-Adapter}
\end{aligned}
\tag{4}
$$

Output of PGVT Block. The output of the l-th PGVT block is simply formed by an input residual connection for pose and video respectively. In addition to the MLP layer, we use an Adapter [100] to further tune the representations:

$$
\begin{aligned}
\mathbf{P}_l &= \mathbf{P}_l^{\mathcal{S}} + \mathrm{MLP}(\mathrm{LN}(\mathbf{P}_l^{\mathcal{S}}) + s_P \cdot \mathrm{Adapter}(\mathrm{LN}(\mathbf{P}_l^{\mathcal{S}})) \\
\mathbf{Z}_l &= \mathbf{Z}_l^{\mathcal{S}} + \mathrm{MLP}(\mathrm{LN}(\mathbf{Z}_l^{\mathcal{S}}) + s_Z \cdot \mathrm{Adapter}(\mathrm{LN}(\mathbf{Z}_l^{\mathcal{S}}))
\end{aligned}
\tag{5}
$$

where s_P and s_Z are scaling factors [100] to control weights from Adapter. PGVT block produces a refined version of standard spatiotemporal input tokens while maintaining the same dimension, functioning as a typical transformer layer. One can easily incorporate PGVT block into any transformer-based model for adaptation to cross-modality joint learning. This versatility allows the PGVT to easily leverage any vision pre-trained model, equipping with spatiotemporal reasoning capability and effective parameter fine-tuning.

4. Experiments

Datasets. We evaluate our PGVT on three fine-grained action recognition datasets, i.e., Diving48 [57], Gym99 and Gym288 [76], and one coarse-grained action recognition dataset Kinetics-400 (K400) [44]. Diving48 includes 15.9K training videos and 2K validation videos focusing on 48 fine-grained diving actions. Each diving class in Diving48 is distinguished by the sequence of takeoff, movements in flight, and entry, requiring models to differentiate fine-grained actions. The Gym99 dataset contains 20k training videos and 8.5k evaluation videos for 99 actions extracted from international competitions. Gym288 is an extended version of Gym99, which is a long-tailed setting with 23k training videos and 9.6k evaluation videos for 288 actions. The Diving and Gym datasets are created with neutrality towards static representations in mind, meaning that a model cannot solely rely on backgrounds to determine the action. K400 contains approximately 240K training videos and 20K validation videos across 400 human action classes. For evaluation, we employed standard classification accuracy as the performance metric.

Implementation Details. For all the experiments, we utilize ViT (ViT-B and ViT-L) models pre-trained by CLIP [71]. We largely follow the training settings from [100]. The implementation of PGVT was carried out in PyTorch. Our training procedures and code are based on the

Table 1. Comparison to SOTA on Diving48.

Model	Pretrain	Tunable Param (M)	Frames	Top-1
SlowFast [29]	IN-1K	54	128	77.6
TQN [103]	K400	-	all	81.8
TimeSformer-L [4]	IN-21K	121	96	81.0
VideoSwin-B [62]	IN-21K	88	-	81.9
BEVT [91]	IN-21K & K400	88	-	86.7
SIFAR-B-14 [25]	IN-21K	87	-	87.3
GC-TDN [33]	IN-21K	27.4	16	87.6
ORViT TimeSformer [36]	IN-21K	160	32	88.0
AIM ViT-B/16 [100]	CLIP	11	32x3	88.9
AIM ViT-L/14 [100]	CLIP	38	32x3	90.6
PGVT ViT-B/16	CLIP	75	48+16	89.5
PGVT ViT-L/14	CLIP	265	48+16	**91.3**

Table 2. Comparison to SOTA on Gym99 and Gym288. "M" stands for "Modality", "R", "T", "F" and "P" stand for "RGB", "Text", "Flow" and "Pose", respectively.

Model	Backbone	M	GFLOPs	Tunable Param (M)	Gym99		Gym288	
					Top-1	Mean	Top-1	Mean
TSN [89]	BNInception	R+F	33	-	86.0	76.4	79.9	37.6
TRNms [109]	BNInception	R+F	-	-	87.8	80.2	82.0	43.3
TSM [60]	ResNet-50	R+F	65	24.3	88.4	81.2	83.1	46.5
I3D [10]	3D ResNet-50	R	108	-	75.6	64.4	66.1	28.2
NL I3D [93]	3D ResNet-50	R	-	43.2	75.3	64.3	67.0	28.0
MTN [50]	SlowOnly	R	-	-	91.8	88.5	-	-
TQN [104]	S3D	R+T	-	-	93.8	90.6	89.6	61.9
SlowOnly [29]	ResNet101	R	-	-	93.9	90.6	86.8	51.2
3D VE [51]	SlowOnly	R	-	-	94.0	90.5	-	-
VT CE [51]	SlowOnly	R+T	-	-	94.6	91.4	90.1	62.6
PoseC3D [21]	SlowOnly-R50	P	-	-	93.2	-	-	-
RGBPose [21]	SlowOnly-R50	R+P	-	-	95.6	-	-	-
PGVT	ViT-B/16	R+P	495	75	96.1	91.4	90.7	63.4
PGVT	ViT-L/14	R+P	2227	265	**96.7**	**91.6**	**91.0**	**63.6**

AIM [100] and PoseConv3D [21]. The model is trained on one GPU for 50 epochs using AdamW [46] optimizer. The learning rate is 8e-6 and weight decay is 5e-2. The pose of each person in each frame is provided by the dataset provider, or otherwise pre-extracted for computational efficiency, using a pose extractor built on HRNet [82]. The number of joints J is the maximum number of joints in a frame and varies for different datasets (e.g. $J = 17$ for the Gym99 dataset). If there are multiple persons in a frame, J represents the number of body joints of all persons (i.e. $J = j \times n$, where j denotes the maximum number of joints per person and n denotes the number of persons). No weight sharing was performed between the pose extractor and our model. The computation of FLOPs and parameters for pose extraction is not included in the subsequent analysis. We provide an ablation study on the breakdown of the inference time in Supplementary Material Section C.2.

4.1. Comparisons to the State of the art

We examine four video action recognition datasets and compare our proposed approach with SOTA approaches. The results are presented in Tables 1, 2 and 3. For the "Frames" column, '$x + y$' denotes x pose frames and y video frames. We trained our models using the standard splits and followed the established evaluation procedure.

Table 3. Comparison to SOTA on Kinetics400.

Model	GFLOPs	Tunable Param (M)	Frames	Top-1	Top-5
TSM R50 [60]	330	24.3	8	74.1	91.2
CorrNet-101 [87]	-	-	32	79.2	-
SlowFast R101 [29]	7020	59.9	32	79.8	93.9
X3D-XXL [27]	4320	20.3	32	80.4	94.6
MoViNet-A6 [47]	386	31.4	120	81.5	95.3
MViT-B [24]	4095	37	64	81.2	95.1
UniFormer-B [55]	3108	50	32	83.0	95.4
TimeSformer-L [4]	7140	121	64	80.7	94.7
ViViT-L/16×2 FE [1]	3980	311	32	80.6	92.7
VideoSwin-L [62]	7248	197	32	83.1	95.9
MViTv2-L [59]	42420	218	32	86.1	97.0
TokenLearner-L/10 [74]	48912	450	64	85.4	96.3
PromptCLIP A7 [42]	-	-	16	76.8	93.5
ActionCLIP [90]	16890	142	32	83.8	97.1
X-CLIP-L/14 [67]	7890	420	8	87.1	97.6
EVL ViT-L/14 [61]	8088	59	32	87.3	-
MTV-L [98]	18050	876	32	84.3	96.3
Hiera-H [73]	1159x3x5	672	16	87.8	-
DualPath [68]	1868	27	32	87.7	97.8
EVA [26]	-	-	8	89.7	-
UMT-L [54]	1434x3x4	304	16	90.6	**98.7**
TubeViT [70]	17640	-	64	90.9	-
InternVideo [94]	-	1300	16	**91.1**	-
AIM ViT-B/16 [100]	2428	11	32x3	84.7	96.7
AIM ViT-L/14 [100]	11208	38	32x3	87.5	97.7
PGVT ViT-B/16	850	77	32+32	85.8	97.1
PGVT ViT-L/14	3832	268	32+32	89.2	98.0

Fine-grained Action Recognition. Table 1 and Table 2 illustrate that our PGVT model surpasses the SOTA methods on fine-grained action recognition datasets. On the Diving48 dataset (see Table 1), our PGVT achieves 89.5% and 91.3% using backbones ViT-B/16 and ViT-L/14 respectively, outperforming AIM [100] with the same backbones by 0.6% and 0.7%. When compared to ORViT [36] which leverages an object tracking model, our PGVT ViT-B/16 outperforms it by 1.5% with less than half of the tunable parameters. This demonstrates the effectiveness of incorporating pose representation and cross-learning of spatiotemporal dynamics. On the Gym99 and Gym288 datasets (see Table 2), our method outperforms all previous methods even when compared with RGBPose-Conv3D [21] which also take video and pose as inputs. This suggests that our pose representation can model motion trajectories more effectively, enhancing the learning of refined pose and video representations in fine-grained actions.

Coarse-grained Action Recognition. Table 3 presents the comparisons with SOTA video models on the K400 dataset. We observe that our PGVT consumes much fewer GFLOPs and tunable parameters than most of the previous methods. PGVT ViT-L/14 achieves 89.2% top-1 accuracy using 32 pose frames and 32 video frames, which is a comparable performance as the SOTA method InternVideo pre-trained on K400. We attribute the comparatively modest performance on the K400 dataset to the short videos which limit the effectiveness of temporal reasoning. We also investigate our model's performance on K400 subset with missing

Table 4. Effectiveness of proposed components.

Methods	GFLOPs	Tunable Param (M)	Frames	Top-1	Top-5
(1) Pose only	108	11	48	82.6	96.7
(2i) Video only	405	11	16	85.9	98.0
(2ii) Video only	850	11	32	87.0	98.2
(3) Late Fusion	513	23	48+16	88.1	98.8
(4) PGVT w/o $\mathcal{T}$	466	68	48+16	67.1	79.1
(5) R-Adapter → 2 Adapters	383	18	48+16	87.6	98.4
(6i) PGVT w/o refined pose	437	46	48+16	89.2	99.0
(6ii) PGVT w/o refined video	437	46	48+16	88.0	98.8
(7) **PGVT**	495	75	48+16	89.5	99.1

Table 5. Transformer for pose.

Pre-trained	Top-1	Top-5
Poseformer	86.4	97.5
CLIP Image Encoder	89.5	99.1

Table 6. Pose input forms.

Input Form	GFLOPs	Tunable Param (M)	Top-1	Top-5
Heatmaps	311	11	80.1	96.0
Coordinates	108	11	82.6	96.7

full body pose. The result and discussion are presented in Supplementary Material Section D.2.

4.2. Ablations

We conducted a comprehensive ablation study on the Diving48 dataset to analyze the contribution of different components in the PGVT and the performance of models with decreased computational cost. We evaluate the model on 48 frames of pose and 16 frames of video unless otherwise stated.

4.2.1 Effectiveness of Components

To illustrate the efficacy of the proposed components outlined in Section 3, we compare our method against the baselines including space-time pose-only and video-only transformer models. AIM [100] with ViT-B/16 backbone serves as the baseline architecture, where we freeze the image backbone. We examine the following versions of our model: (1) Pose only: single pose modality is fed into baseline architecture; (2) Video only: single video modality is fed into baseline architecture, with 16 frames for (2i) and 32 frames for (2ii); (3) Late fusion: a two-stage approach with pre-trained pose-only and video-only models, followed by concatenating the features for whole network fine-tuning. (4) PGVT w/o $\mathcal{T}$: $\mathcal{T}$ is removed; (5) *R-Adapter* → 2 Adapters: we replace *R-Adapter* with two parallel Adapters [100] for pose and video respectively without cross-modality attention; (6) Inside *R-Adapter*, we further investigate the effectiveness of the two streams separately: (6i) PGVT w/o refined pose; and (6ii) PGVT w/o refined video; (7) PGVT. The details of the architecture design are presented in the Supplementary Material Section B.

The results for the baselines are presented in the top part of Table 4. We observe that the pose-only model, despite 48 frames input, has significantly fewer GFLOPs than the video-only model (108 vs.405). The video-only model with 16 frames improves its spatial representations, resulting in a performance increase (82.6% to 85.9%). However, this approach substantially increases the GFLOPs. The video-only model with 32 frames further increases the performance but at the cost of further increasing the GLOPs to 850.

In the bottom part of Table 4, we intend to narrow the performance gap and even outperforming the video-only model by adding a small number of video frames to the pose-only model. The late fusion approach in Method (3) achieves better performance (88.1%) than baselines, but requires two-stage training. In Method (4), both the pose and video tokens are fed directly to $\mathcal{S}$, leading to a substantial performance drop from 89.5% to 67.1%. This affirms the critical role of temporal attention in PGVT. In Method (5), replacing *R-Adapter* with two original Adapters decreases the performance to 87.6%, yet surpassing video-only model (2ii) with its effective temporal modeling and joint pose-video learning. For Methods (6i) and (6ii), the addition of *R-Adapter* for cross-modality learning further improves the performance.Without refining pose features in Method (6i), the performance only drops by 0.3%, while Method (6ii) without refining video features suffers 1.5% performance drop. This shows the importance of pose-guided features in capturing strong spatiotemporal interaction between high-level temporal dynamics and video spatial representation. Finally, Method (7), which is our PGVT model, achieves not only better accuracy than the 32 frames video-only model (89.5% vs. 87.0%) but also with fewer GFLOPs (495M vs.850). It also outperforms the late fusion approach without the need for two-stage training. These results successfully validate the effectiveness of our proposed pose-focused strategies.

4.2.2 Ablations on Pose Learning

We evaluate different pre-trained models for pose stream and different pose input forms. First, we investigate replacing the pre-trained model for the pose in PGVT ViT-B/16 from CLIP image encoder to Poseformer [108]. As shown in Table 5, the CLIP image encoder achieves better performance. Secondly, we also investigate to use of pose heatmaps to replace the pose joint coordinate representation. We generate heatmaps as per [21], with a resolution of 112x112. The comparison is shown in Table 6. It can

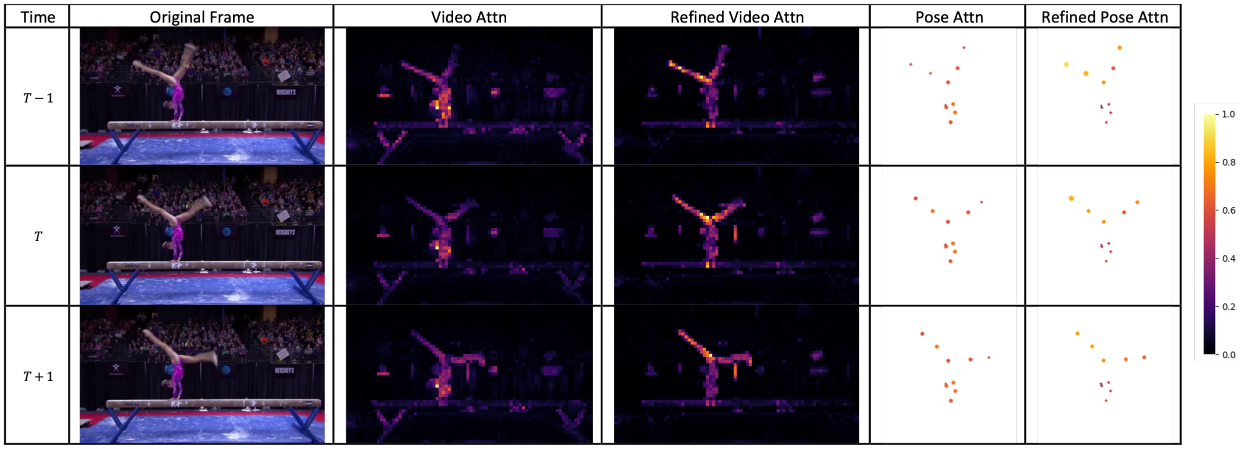

Figure 2. Visual examples show the improvements of visual attention by joint learning from pose tokens and video tokens in PGVT. A brighter colour and bigger dot in each pose frame means more attention.

be observed that the pose joint representation gives rise to better performance. The potential reasons are discussed in Supplementary Material Section B.2.

4.2.3 Decreasing Computational Cost

We demonstrate that the computational cost could be further reduced with only a minor decrease in performance. We experiment with how the embedding dimension in the pose branch and the number of video frames affect performance respectively. Details can be found in the Supplementary Material Section B.3.

4.3. Visualisation

A few examples for visual examination of the effectiveness of PGVT are shown in Fig. 2. “Video Attn” and “Pose Attn” are visualisation of video-only and pose-only baselines, and “Refined Video Attn” and “Refined Pose Attn” are visualisation of PGVT features. It can be observed that PGVT improves the spatial attention around corresponding body joints and temporal attention to fast-moving body joints, leading to a more effective representation of human action recognition. This feature enables our model to capture and represent diverse visual patterns effectively.

4.4. Discussions

In this study, we showcase the effectiveness of a pose-guided method that incorporates pose and video representations and propagates them into transformer layers. A limitation of our work is that we rely on externally provided pose rather than generating them within the model without supervision. Exploring the use of self-generated pose is an interesting area for future research.

Nevertheless, we believe that our work has a positive social impact, due to its potential to incorporate compositionality, achieving spatiotemporal reasoning capability with cost-effective training. Our method is straightforward and widely applicable, enabling the use of more effective foundation models. It is worth noting that our design of reusing image models for pose temporal modeling is sufficiently robust for datasets that place a stronger emphasis on the temporal dimension. Our approach is not limited to specific pre-trained models and can be applied to different architectures for future deployments. As pose modeling is a kind of sequence modeling, it is worthwhile to explore the reuse of pre-trained weights from more powerful sequential models instead of relying solely on image models, to enhance temporal modeling capabilities.

5. Conclusion

In this work, we have proposed *Pose-Guided Video Transformer* (PGVT), a novel approach for efficiently transferring pre-trained image models to video action recognition with pose and video inputs. We introduced *Pose Temporal Attention* module and *Pose-Video Spatial Attention* module to incorporate pose-guided spatiotemporal reasoning into an image model. Our model learns temporal dynamics semantics from pose and pose-guided spatial dynamics from pose and video to facilitate compositional understanding. With pose input, we can use fewer video frames than those models using dense video frames. By keeping the pre-trained image model frozen and updating the newly added adapters, our method reduces the training cost compared to most of the existing approaches, while achieving comparable or superior performance to prior SOTA methods on four benchmark datasets.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario LuÄiÄ, and Cordelia Schmid. Vivit: A video vision transformer, 2021. 2, 6

[2] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Visual prompting: Modifying pixel space to adapt pre-trained models. *arXiv preprint arXiv:2203.17274*, 2022. 2

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding?, 2021. 2

[4] 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)*, July 2021. 2, 6

[5] Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. Multimodal pretraining unmasked: Unifying the vision and language BERTs. *Transactions of the Association for Computational Linguistics (TACL)*, 2021. 5

[6] Congqi Cao, Cuiling Lan, Yifan Zhang, Wenjun Zeng, Hanqing Lu, and Yanning Zhang. Skeleton-based action recognition with gated convolutional neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(11):3247–3257, 2019. 3

[7] Congqi Cao, Yifan Zhang, Chunjie Zhang, and Hanqing Lu. Body joint guided 3d deep convolutional descriptors for action recognition, 2017. 1

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020. 2

[9] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2

[10] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 6

[11] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2018. 2

[12] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *arXiv preprint arXiv:2205.13535*, 2022. 2

[13] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition, 2021. 3

[14] Zhan Chen, Sicheng Li, Bing Yang, Qinghan Li, and Hong Liu. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition, 2022. 3

[15] Vasileios Choutas, Philippe Weinzaepfel, JÃ©rÃ´me Revaud, and Cordelia Schmid. Potion: Pose motion representation for action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7024–7033, 2018. 3

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 2

[17] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description, 2016. 2

[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 *International Conference on Learning Representations*, 2020. 1, 2

[19] 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, 2021. 2

[20] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers, 2022. 5

[21] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition, 2022. 1, 3, 6, 7

[22] Efros, Berg, Mori, and Malik. Recognizing action at a distance. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 726–733 vol.2, 2003. 2

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

[24] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021. 6

[25] Quanfu Fan, Chun-Fu Chen, and Rameswar Panda. Can an image classifier suffice for action recognition? In *International Conference on Learning Representations*, 2022. 6

[26] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale, 2022. 6

[27] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 203–213, 2020. 6

[28] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv preprint arXiv:2205.09113*, 2022. 2

[29] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition, 2019. 2, 6

[30] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition, 2016. 2

[31] Yunhe Gao, Xingjian Shi, Yi Zhu, Hao Wang, Zhiqiang Tang, Xiong Zhou, Mu Li, and Dimitris N. Metaxas. Visual Prompt Tuning for Test-time Domain Adaptation. *arXiv preprint arXiv:2210.04831*, 2022. 2

[32] Ryo Hachiuma, Fumiaki Sato, and Taiki Sekii. Unified keypoint-based action recognition framework via structured keypoint pooling, 2023. 3

[33] Yanbin Hao, Hao Zhang, Chong-Wah Ngo, and Xiangnan He. Group contextualization for video recognition, 2022. 6

[34] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition, 2017. 2

[35] Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. Decoupling the role of data, attention, and losses in multimodal transformers, 2021. 5

[36] Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Object-region video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3148–3159, 2022. 6

[37] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 2

[38] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013. 2

[39] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2

[40] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022. 2

[41] Shibo Jie and Zhi-Hong Deng. Convolutional bypasses are better vision transformer adapters. *arXiv preprint arXiv:2207.07039*, 2022. 2

[42] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. *arXiv preprint arXiv:2112.04478*, 2021. 2, 6

[43] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 2

[44] 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. 5

[45] Lipeng Ke, Kuan-Chuan Peng, and Siwei Lyu. Towards to-a-t spatio-temporal focus for skeleton-based action recognition, 2022. 3

[46] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 6

[47] Dan Kondratyuk, Liangzhe Yuan, Yandong Li, Li Zhang, Mingxing Tan, Matthew Brown, and Boqing Gong. Movinets: Mobile video networks for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16020–16030, 2021. 6

[48] Haofei Kuang, Yi Zhu, Zhi Zhang, Xinyu Li, Joseph Tighe, Sören Schwertfeger, Cyrill Stachniss, and Mu Li. Video contrastive learning with global context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3195–3204, 2021. 2

[49] Jungho Lee, Minhyeok Lee, Dogyoon Lee, and Sangyoun Lee. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition, 2022. 3

[50] Mei Chee Leong, Hui Li Tan, Haosong Zhang, Liyuan Li, Feng Lin, and Joo Hwee Lim. Joint learning on the hierarchy representation for fine-grained human action recognition. In *ICIP*, pages 1059–1063. IEEE, 2021. 6

[51] Mei Chee Leong, Haosong Zhang, Hui Li Tan, Liyuan Li, and Joo Hwee Lim. Combined cnn transformer encoder for enhanced fine-grained human action recognition, 2022. 6

[52] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 2

[53] Chuankun Li, Yonghong Hou, Pichao Wang, and Wanqing Li. Joint distance maps based action recognition with convolutional neural networks. *IEEE Signal Processing Letters*, 24(5):624–628, may 2017. 3

[54] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models, 2023. 6

[55] Kunchang Li, Yali Wang, Gao Peng, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatial-temporal representation learning. In *International Conference on Learning Representations*, 2021. 6

[56] Wenbo Li, Longyin Wen, Ming-Ching Chang, Ser Nam Lim, and Siwei Lyu. Adaptive rnn tree for large-scale human action recognition. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1453–1461, 2017. 3

[57] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528, 2018. 5

[58] Yinxiao Li, Zhichao Lu, Xuehan Xiong, and Jonathan Huang. Perf-net: Pose empowered rgb-flow net, 2021. 3

[59] 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4794–4804, 2022. 6

[60] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding, 2019. 2, 6

[61] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. *arXiv preprint arXiv:2208.03550*, 2022. 2, 6

[62] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 2, 6

[63] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition, 2020. 3

[64] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention, 2020. 5

[65] Diogo C. Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning, 2018. 1

[66] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification, 2015. 2

[67] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. *arXiv preprint arXiv:2208.02816*, 2022. 2, 6

[68] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Dual-path adaptation from image to video transformers, 2023. 6

[69] Mandela Patrick, Dylan Campbell, Yuki M. Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and JoÃ£o F. Henriques. Keeping your eye on the ball: Trajectory attention in video transformers, 2021. 2

[70] AJ Piergiovanni, Weicheng Kuo, and Anelia Angelova. Rethinking video vits: Sparse video tubes for joint image and video learning, 2022. 6

[71] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 5

[72] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, Christoph Feichtenhofer, and Jitendra Malik. On the benefits of 3d pose and tracking for human action recognition, 2023. 3

[73] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. Hiera: A hierarchical vision transformer without the bells-and-whistles, 2023. 6

[74] Michael S Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 6

[75] Anshul Shah, Shlok Mishra, Ankan Bansal, Jun-Cheng Chen, Rama Chellappa, and Abhinav Shrivastava. Pose and joint-aware action recognition, 2021. 3

[76] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[77] 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. 3

[78] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Adasgn: Adapting joint number and model size for efficient skeleton-based action recognition, 2021. 3

[79] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. An attention enhanced graph convolutional lstm network for skeleton-based action recognition, 2019. 3

[80] Chenyang Si, Ya Jing, Wei Wang, Liang Wang, and Tieniu Tan. Skeleton-based action recognition with spatial reasoning and temporal stack learning, 2018. 3

[81] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos, 2014. 2

[82] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation, 2019. 6

[83] Hao Tan, Jie Lei, Thomas Wolf, and Mohit Bansal. Vimpac: Video pre-training via masked token prediction and contrastive learning. *arXiv preprint arXiv:2106.11250*, 2021. 2

[84] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks, 2015. 2

[85] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 2

[86] GÃ¼l Varol, Ivan Laptev, and Cordelia Schmid. Long-term temporal convolutions for action recognition, 2017. 2

[87] Heng Wang, Du Tran, Lorenzo Torresani, and Matt Feiszli. Video modeling with correlation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 352–361, 2020. 6

[88] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos, 2017. 2

[89] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE TPAMI*, 41(11):2740–2755, 2018. 6

[90] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 2, 6

[91] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14733–14743, 2022. 6

[92] Wenhui Wang et al. Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks. *arXiv preprint arXiv:2208.10442*, 2022. 2

[93] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 6

[94] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning, 2022. 6

[95] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision, 2022. 5

[96] Kailin Xu, Fanfan Ye, Qiaoyong Zhong, and Di Xie. Topology-aware convolutional neural network for efficient skeleton-based action recognition, 2021. 3

[97] An Yan, Yali Wang, Zhifeng Li, and Yu Qiao. Pa3d: Pose-action 3d machine for video recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7914–7923, 2019. 3

[98] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3333–3343, 2022. 2, 6

[99] Taojiannan Yang, Sijie Zhu, Matias Mendieta, Pu Wang, Ravikumar Balakrishnan, Minwoo Lee, Tao Han, Mubarak Shah, and Chen Chen. Mutualnet: Adaptive convnet via mutual learning from different model configurations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):811–827, 2021. 2

[100] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. Aim: Adapting image models for efficient video action recognition, 2023. 2, 3, 5, 6, 7

[101] Fanfan Ye, Shiliang Pu, Qiaoyong Zhong, Chao Li, Di Xie, and Huiming Tang. Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition, 2020. 3

[102] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 2

[103] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Temporal query networks for fine-grained video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4486–4496, 2021. 6

[104] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Temporal query networks for fine-grained video understanding. In *CVPR*, pages 4486–4496, 2021. 6

[105] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive recurrent neural networks for high performance human action recognition from skeleton data, 2017. 3

[106] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive neural networks for high performance skeleton-based human action recognition, 2019. 3

[107] Yanyi Zhang, Xinyu Li, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe. Vidtr: Video transformer without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13577–13587, 2021. 2

[108] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers, 2021. 7

[109] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, pages 803–818, 2018. 6

[110] Mohammadreza Zolfaghari, Yi Zhu, Peter Gehler, and Thomas Brox. Crossclr: Cross-modal contrastive learning for multi-modal video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1450–1459, 2021. 2