Visual Knowledge Graph for Human Action Reasoning in Videos

Yue Ma*
y-ma21@mails.tsinghua.edu.cn
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences & Tsinghua Shenzhen International Graduate School, Tsinghua University

Yali Wang†
yl.wang@siat.ac.cn
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences & SIAT Branch, Shenzhen Institute of Artificial Intelligence and Robotics for Society

Yue Wu
yue.wu@siat.ac.cn
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences

Ziyu Lyu
zy.lv@siat.ac.cn
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences

Siran Chen
sr.chen@siat.ac.cn
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences & University of Chinese Academy of Science

Xiu Li
li.xiu@sz.tsinghua.edu.cn
Tsinghua Shenzhen International Graduate School, Tsinghua University

Yu Qiao‡
yu.qiao@siat.ac.cn
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences & Shanghai AI Laboratory

ABSTRACT

Action recognition has been traditionally treated as a high-level video classification problem. However, such a manner lacks the detailed and semantic understanding of body movement, which is the critical knowledge to explain and infer complex human actions. To fill this gap, we propose to summarize a novel visual knowledge graph from over 15M detailed human annotations, for describing action as the distinct composition of body parts, part movements and interactive objects in videos. Based on it, we design a generic multi-modal Action Knowledge Understanding (AKU) framework, which can progressively infer human actions from body part movements in the videos, with assistance of visual-driven semantic knowledge mining. Finally, we validate AKU on the recent Kinetics-TPS benchmark, which contains body part parsing annotations for detailed understanding of human action in videos. The results show that, our AKU significantly boosts various video backbones with explainable action knowledge in both supervised and few shot settings, and outperforms the recent knowledge-based action recognition framework, e.g., our AKU achieves 83.9% accuracy on Kinetics-TPS while PaStaNet achieves 63.8% accuracy under the same backbone. The codes and models will be released at https://github.com/mayuelala/AKU.

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**.

KEYWORDS

Action Recognition, Knowledge Graph, Video Understanding

ACM Reference Format:
Yue Ma, Yali Wang, Yue Wu, Ziyu Lyu, Siran Chen, Xiu Li, and Yu Qiao. 2022. Visual Knowledge Graph for Human Action Reasoning in Videos. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3548257

1 INTRODUCTION

Human action recognition is an important problem for video understanding. The great advancement of this research has been made with fast development of deep learning [26, 31, 34, 36, 54].

However, most existing approaches treat action recognition as a high-level video classification problem, and focus on designing backbones [16, 30, 53] for representation learning. Alternatively, human action is actually spatial-temporal evolution of body part movements and object interactions. Without such explicit understanding, these approaches often suffer from the performance bottleneck for recognizing confused actions with complex dynamics in the wild.

*This work was done during the internship of Yue Ma at Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences.
†Yue Ma and Yali Wang are equally-contributed first authors.
‡Yu Qiao is the corresponding author.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548257

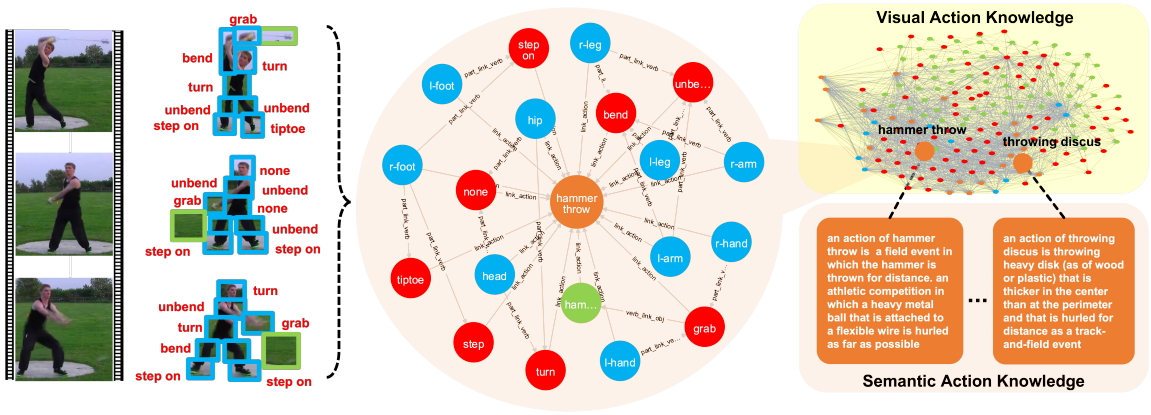

Figure 1: Visual Knowledge Graph for Video Action Recognition. For each video, we construct a knowledge sub-graph of the corresponding human action category (orange circle), based on visual connections of body parts (blue circle), their movements (also called as verbs, red circle) and interactive objects (green). Then, we summarize visual knowledge graph from all the video sub-graphs, and equip each action node with its semantic knowledge of Wikipedia descriptions. With such explainable guidance, we design a generic multi-modal framework for the detailed Action Knowledge Understanding (AKU) in the videos.

To alleviate such problem, several works have been proposed for more detailed action understanding [23, 38], by mimicking human beings with compositional knowledge. However, they mainly focus on hand (e.g., something-else [38]) or the whole human in the scene (e.g., action genome [23]), without deep studies on body part movements in the videos. Recently, a PastaNet [33] attempts to explore human activity knowledge by part-level state (i.e., movement/verb) annotations. But it works on human-object-interaction (HOI) in the image domain, without learning human dynamics in videos. More importantly, human actions refer to the more abstract concepts, compared to the HOI categories. Hence, PastaNet lacks the commonsense knowledge to describe human actions with discriminative semantics.

To tackle these difficulties, we propose to construct a novel visual knowledge graph for human action understanding in videos, based on the recent action parsing benchmark of Kinetics-TPS [25]. As shown in Fig. 1, we describe a human action by visual connections of body parts, part movements, and interactive objects. For each human-centric video, we first extract these connections from a frame and then integrate them over all the frames. This process generates a sub-graph of the corresponding action category. Subsequently, we summarize a distinct visual knowledge graph from sub-graphs of all the videos, which effectively constructs visual action knowledge about how the body part moves and interacts with objects, from over 15M part-level annotations of 24 human action categories in Kinetics-TPS. Moreover, we equip each action node with a description from Wikipedia, which provides semantic knowledge to understand human actions from discriminative texts. As a result, we build a robust commonsense knowledge base for human action reasoning, with complementary integration of visual and semantic knowledge.

Based on our distinct knowledge base, we develop a generic multi-modal Action Knowledge Understanding (AKU) framework, which can infer high-level human actions progressively, by mining semantic action knowledge with part-level visual guidance. Specifically, our AKU consists of three core modules in Fig. 2, i.e., visual learner, knowledge miner, and action recognizer. Visual learner can flexibly construct human-level representation at each frame, by adaptive aggregation of its part-level features. Knowledge miner can robustly leverage visual body movement as reasoning guidance, and exploit semantic action knowledge for each video. Action recognizer can effectively integrate visual learner and knowledge miner as multi-modal human feature, and learn spatial-temporal human relations among different frames. In this case, our AKU can take advantage of both visual and semantic knowledge for the detailed understanding of human action in videos.

Finally, we systematically evaluate our AKU on the action parsing benchmark, i.e., Kinetics-TPS [25]. In both fully-supervised and few-shot settings, our AKU can achieve over 3-5% accuracy improvement on various popular 2D, 3D and transformer backbones for human action recognition. Moreover, it significantly outperforms the recent knowledge-based action recognition framework under the same settings, i.e., our AKU achieves 83.9% accuracy on Kinetics-TPS while PaStaNet [33] achieves 63.8% accuracy. We will release our codes and models at https://github.com/mayuelala/AKU.

2 RELATED WORK

Action Understanding. Human action recognition is an important problem for video understanding. With fast development of deep learning models and large-scale datasets, action recognition has achieved remarkable successes. The previous approaches mainly work on spatial-temporal representation learning. In particular, 2D and/or 3D Convolution Neural Networks (CNNs) [7, 16, 17, 24, 29, 31, 34, 44–47, 52, 54, 57, 58, 60] have shown their superiority on this task. However, they often lack learning capacity to capture long-term dependencies in the video. Based on the recent trends of vision transformers [11–13, 32, 35, 51, 59, 62], researchers propose to leverage self-attention for spatial-temporal relation learning [2, 30, 41, 42, 65]. However, these works treat action recognition as a high-level video classification problem. Hence, they often ignore complex body movements and object interactions for detailed action understanding in videos. To tackle this problem, several

works have been recently introduced for action understanding, by exploiting human-object relations in the video [23, 38]. However, something-else [38] mainly works on human hands, while action genome [23] constructs scene graph about the whole human instances. They do not reflect the body part movements, which can be an important clue to recognize complex human actions in the video. Alternatively, Human-Object-Interaction (HOI) often takes such discriminative information to understand human poses and activities [8, 14, 28, 33, 66]. In particular, a PaStaNet [33] has been proposed with a human activity knowledge engine, which encodes semantic representation of body part states (i.e., verb or movement) to boost HOI. However, human action categories are often more abstract than HOI categories. Hence, PaStaNet may lack capacity of describing semantic knowledge of high-level actions. Moreover, this framework does not take video dynamics into account, due to its main application for image domain. Different from it, we propose a visual knowledge graph for human action in videos, with body part, part movement, interactive object and action nodes. Moreover, we equip each action node with semantic description. By complementary knowledge integration, our knowledge base is preferable for human action reasoning.

Knowledge Graph for Vision. Knowledge graph is a popular choice to represent semantic knowledge by concepts and relationships. A number of large-scale knowledge graphs are available commercially or in open source, which are generally constructed based on common sense [50], Wikipedia[3], English word [40]. Over the past years, knowledge graph has been widely used for Web search and social media, due to their superior reasoning abilities. Recently, researchers attempt to leverage it as an external knowledge to boost computer vision tasks [9, 15, 37, 55, 61]. In particular, as for video classification using knowledge graph [18, 19, 63], there are several limitations in these works. First, [18, 19] work on human actions, without detailed understanding of body movements. Hence, they are limited to distinguish confused actions with complex body movements and object interactions. Moreover, [63] works on the general video classification, instead of human action understanding. On the contrary, we exploit the detailed human actions with a distinct visual knowledge graph, which can effectively infer and explain human actions in the video. Second, they often work on limited data setting [18, 19], without investigating its validity on traditional settings. Alternatively, our AKU can largely boost both fully-supervised and few-shot settings to show its power in human action reasoning. Finally and most importantly, the paradigms of these previous works are not unified. Hence, it is hard to re-build their knowledge graph and re-implement these models for further studies in practice. In contrast, our visual knowledge graph is built upon an action parsing benchmark of Kinetics-TPS [25]. Hence, it can be systematically re-constructed without difficulty. Moreover, our model is based on a generic and flexible paradigm, which can be easily implemented with various video backbones.

3 ACTION KNOWLEDGE CONSTRUCTION

As mentioned before, it is critical to analysis visual composition of body movements and interactions for human action understanding in videos. Hence, we propose to construct a visual knowledge graph to represent such knowledge. Specifically, we use the recent

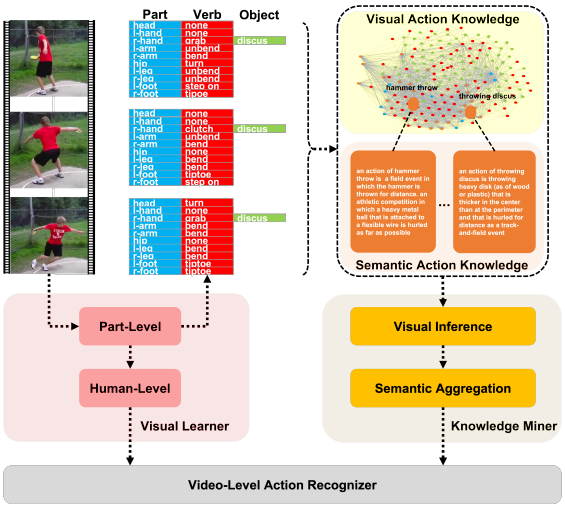

Figure 2: Overview of our AKU framework.

action parsing benchmark (Kinetics-TPS [25]) as video data base to achieve this goal. Different from the existing data sets in action recognition [26, 31, 34, 36, 54], Kinetics-TPS provides over 15M part-level annotations for detailed human action understanding, including 7.9M box annotations of body parts, 7.9M tags of part movements (i.e., verbs) , and 0.5M interactive objects. Note that, our goal is not to cover human action categories as many as possible. Alternatively, we aim at mastering the fundamental mechanism to construct visual commonsense knowledge of a human action. Hence, Kinetics-TPS is a preferable data set for this systematical investigation on 24 popular human action categories.

Visual Action Knowledge. Based on Kinetics-TPS, we build a visual knowledge graph with the following steps. **First**, we construct a sub-graph of *part-verb-object* from each annotated human instance, where we call the movement of body parts as verb for simplicity. In this sub-graph, the nodes refer to body parts, verbs and objects, and the edges refer to visual connections between them. For example, the actor uses his hands to grab a hammer in the first frame of Fig. 1. Hence, there are two paths of *left_hand-grab-hammer* and *right_hand-grab-hammer* in the sub-graph. Moreover, a body part may not interact with an object. In this case, we generate a path of *part-verb*. For example, the actor turns his head in the first frame of Fig. 1. Hence, there is a path of *head-turn*. **Second**, we integrate sub-graphs of all the annotated human instances for each video, by merging the repeated paths. As a result, we obtain a video-level sub-graph. To reflect the action category of this video, we add an extra action node which connects with all the nodes in the sub-graph. **Finally**, we combine sub-graphs of all the videos in Kinetics-TPS. This results in a visual knowledge graph with 181 nodes and 4,532 edges, including 10 part nodes, 84 verb nodes, 73 object nodes and 24 action nodes. Note that, we call it as visual knowledge graph, since it summarizes individual sub-graphs of all the human annotations to reflect visual commonsense knowledge of human actions in these videos.

Semantic Action Knowledge. Besides of visual knowledge graph, we propose to further extend our knowledge base with semantic descriptions of human actions, which often provide rich

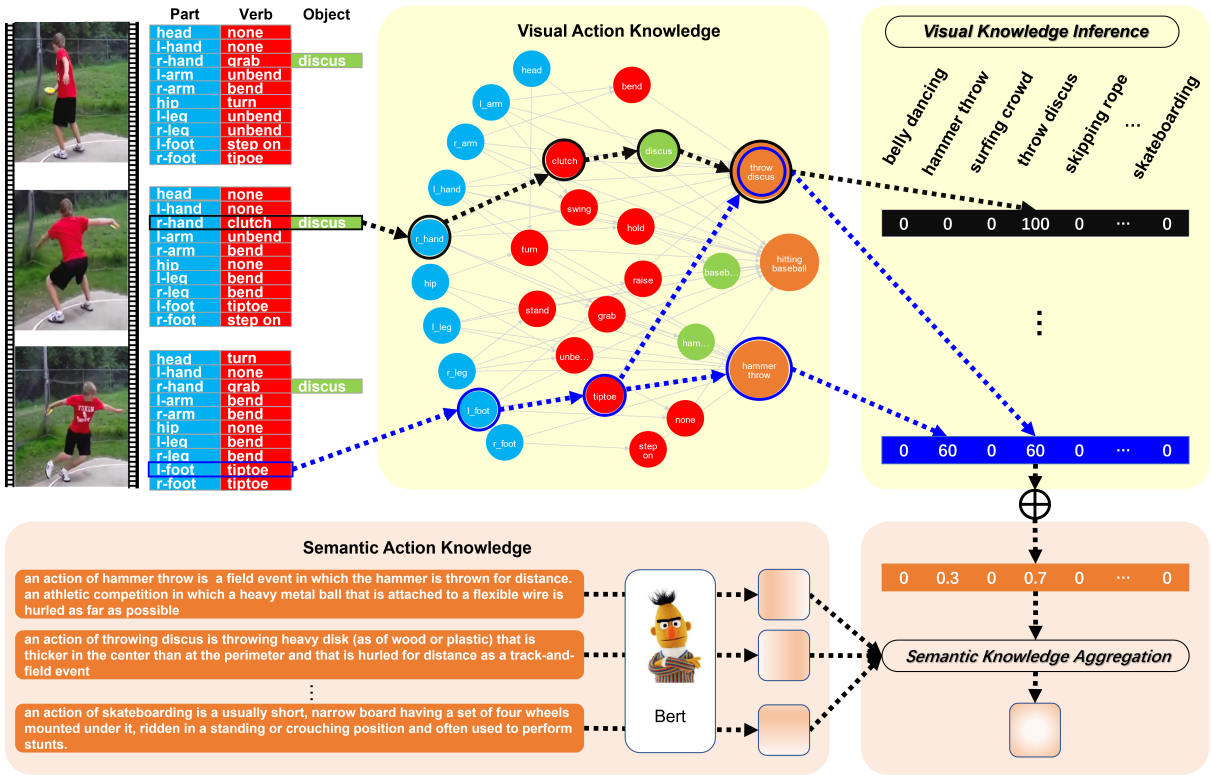

Figure 3: Knowledge Miner. It mainly consists of Visual Knowledge Inference (VKI) and Semantic Knowledge Aggregation (SKA). First, VKI can discover the weight of each action category for an input video, by part-level reasoning in our visual knowledge graph. Then, SKA leverages this weight vector as guidance, and attentively learn an action knowledge vector of this video, by weighted-sum over semantic knowledge of all the action categories.

and diversified contextual information. Specifically, we equip each action node with a textual description from Wikipedia, e.g., *an action of hammer throw is a field event in which the hammer is thrown for distance. An athletic competition in which a heavy metal ball that is attached to a flexible wire is hurled as far as possible.* Apparently, this text provides extra discriminative information such as which event this action belongs to, what a hammer looks like. Such knowledge has not been covered by visual information of the *hammer throw* video in Fig. 1. Hence, we integrate both visual and semantic action knowledge, and leverage their cooperative power for effective human action understanding in videos.

4 ACTION UNDERSTANDING FRAMEWORK

Based on our action knowledge, we propose a generic multi-modal AKU framework in this section. As shown in Fig. 2, it consists of three core modules for human action reasoning in videos. Knowledge Miner and Visual Learner are used to extract knowledge and visual representation of human action in the video. Subsequently, Action Recognizer integrates multi-modal features and perform spatial-temporal relation learning to identify the action category.

4.1 Knowledge Miner

To effectively infer and explain human action in a video, the first question is how to exploit discriminative knowledge of this video from our action knowledge base. We introduce a concise Knowledge Miner to address this problem. As shown in Fig. 3, it mainly consists of Visual Knowledge Inference (VKI) and Semantic Knowledge Aggregation (SKA). First, VKI can discover the uncertainty of each action category for an input video, by *part-verb-object* reasoning in our visual knowledge graph. Then, SKA leverages such category uncertainty as aggregation guidance, and attentively construct an action knowledge vector by weighted-sum over semantic knowledge of all the action categories.

Visual Knowledge Inference (VKI). Note that, our visual knowledge graph reflects part-level construction of human action. Hence, we first design VKI to infer action from each triple of *part-verb-object* in the input video. Specifically, we match such a triple with our visual knowledge graph to find the action node. Note that, each triple may identify many possible action nodes, since some body movements are actually shared among different action categories. For example, *left_foot-tiptoe* can match both *throw discus* and *hammer throw* in Fig. 3. Clearly, it brings the uncertainty for human action reasoning. To alleviate this problem, we introduce a simple but effective mechanism for uncertainty weighting. First, if a triple only matches one action category, we assign the weight of this action as 100 for this triple. Second, as the number of the matched actions increases, we reduce the weight of each action category, e.g., if a triple can match 2 (or 3, 4, 5) action categories,

we assign the weight of each category as 60 (or 30, 10, 3) to take uncertainty into account. Note that, the weight value is set to enlarge the confidence gap and prevent valid inference from being submerged. Third, if a triple matches more than five actions, we will not take the inference results of this triple into account, due to its large uncertainty. Based on this weighting mechanism $\mathcal{G}$, we can associate each part-level triple with an importance vector of all C action categories. Consequently, we sum the importance vectors over all the triples in the video,

$$\mathbf{W} = \sum\nolimits_p \mathcal{G}(< Part^{(p)}, Verb^{(p)}, Object^{(p)} >), \tag{1}$$

which generates an importance vector of action categories $\mathbf{W} \in \mathbb{R}^C$ for this video. We normalize this vector by softmax, and use it as an attention vector to describe visual action knowledge.

Semantic Knowledge Aggregation (SKA). After visual knowledge inference, we next extract semantic knowledge for aggregation. Specifically, we leverage Bert [10] to extract semantic vector of each action tag and its textual description,

$$\mathbf{S}^{(c)} = \mathcal{B}(< Action^{(c)}, Description^{(c)} >). \tag{2}$$

Subsequently, we use visual action knowledge (Eq. 1) of the input video as guidance, and summarize semantic representation of all the action categories as the knowledge vector of this video,

$$\mathbf{Z} = \sum\nolimits_c \mathbf{W}^{(c)} \cdot \mathbf{S}^{(c)}. \tag{3}$$

Additionally, we further learn semantic knowledge of human instances for multi-modal fusion afterwards. Hence, we use Bert [10] to extract semantic vector $\mathbf{S}^{(p)}$ of each part-level triplet,

$$\mathbf{S}^{(p)} = \mathcal{B}(< Part^{(p)}, Verb^{(p)}, Object^{(p)} >). \tag{4}$$

For a human instance, we construct a semantic vector by average pooling over all the corresponding part-level vectors, $\mathbf{S}^{(h)} = mean(\{\mathbf{S}^{(p)}\}_p)$. Finally, we concatenate $\mathbf{S}^{(h)}$ and $\mathbf{Z}$ to learn a knowledge vector for each human instance,

$$\mathbf{K}^{(h)} = \mathcal{H}(Concat(\mathbf{S}^{(h)}, \mathbf{Z})), \tag{5}$$

where $\mathcal{H}$ is a fully-connected layer for feature transformation.

4.2 Visual Learner

Visual Learner is another important module in our AKU framework. It mainly consists of two core submodules, i.e., part-level and human-level visual learners, which can progressively learn visual representation of human instances from an input video.

Part-Level Visual Learner. As mentioned before, the body part movement is critical to infer human actions from our knowledge base. Unfortunately, such verb annotations are often unavailable for a testing video. Hence, we design a part-level visual learner, which can learn robust features to predict the body movements from part and object annotations of each video. Note that, we do not further discuss how to obtain part and object annotations in this paper, since they are not the key designs in our AKU framework and easily obtained by remarkable pose estimators [8, 64] and object detectors [4, 5, 48]. Specifically, for each frame, we first extract visual features of body parts, human instances and objects (i.e., $\mathbf{F}^{(p)}$, $\mathbf{F}^{(h)}$ and $\mathbf{F}^{(o)}$), by ROI align [48] of their bounding boxes on an ImageNet-pretrained ResNet50 [22]. Second, we use a contextual

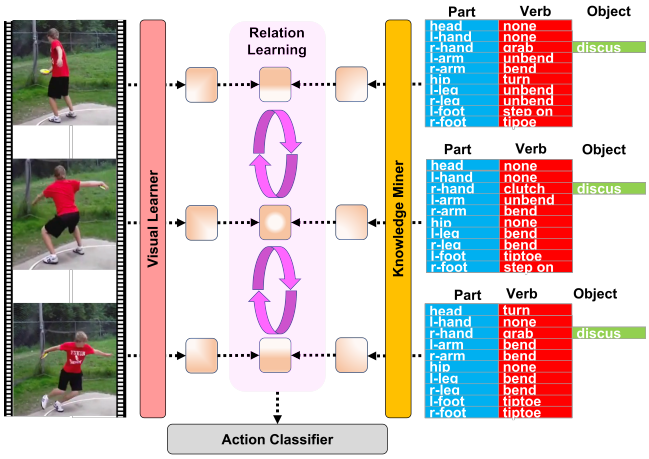

Figure 4: Action Recognizer with Knowledge and Visual Representations in the Video.

learning function $\mathcal{R}$ to enhance each part feature, by aggregating its corresponding human instance and object features,

$$\mathbf{V}^{(p)} = \mathcal{R}(\mathbf{F}^{(p)}, \mathbf{F}^{(h)}, \mathbf{F}^{(o)}). \tag{6}$$

One can simply use a FC layer as $\mathcal{R}$ for feature transformation. Finally, we use another FC layer to convert the enhanced part feature as the verb prediction vector $\mathbf{U}^{(p)} = \mathcal{F}(\mathbf{V}^{(p)})$. Then, we train part-level visual learner by cross-entropy loss between verb prediction $\mathbf{U}^{(p)}$ and ground truth verb vector $\mathbf{Y}^{(p)}$,

$$\mathcal{L}^{(p)} = \mathcal{CE}(\mathbf{U}^{(p)}, \mathbf{Y}^{(p)}). \tag{7}$$

Human-Level Visual Learner. After obtaining part-level features, we further summarize them as a visual feature of the corresponding human instance,

$$\mathbf{V}^{(h)} = \mathcal{A}(\{\mathbf{V}^{(p)}\}_p), \tag{8}$$

where $\mathcal{A}$ is a feature aggregation function. One can simply use average pooling over all part-level features of a human instance.

4.3 Action Recognizer

After generating knowledge vector $\mathbf{K}^{(h)}$ and visual vector $\mathbf{V}^{(h)}$ of each human instance in a video, we use action recognizer for multi-modal prediction of action label. Specifically, we first fuse knowledge and visual vectors per human instance,

$$\mathbf{P}^{(h)} = \mathcal{M}(\mathbf{K}^{(h)}, \mathbf{V}^{(h)}), \tag{9}$$

where $\mathcal{M}$ is a fusion function that can be the sum operation for example. Second, since human instances in different frames are actually with close relationships, we leverage such dependency to enhance multi-modal feature of all the human instances $\{\mathbf{P}^{(h)}\}_h$,

$$\{\mathbf{Q}^{(h)}\}_h = \mathcal{T}(\{\mathbf{P}^{(h)}\}_h), \tag{10}$$

where $\mathbf{Q}^{(h)}$ is the relation-enhanced feature for each human instance, and $\mathcal{T}$ is a relation function that can be a graph convolution network for example. Finally, we perform average pooling over human-level features, with regards to each frame. This generates a frame-level feature $\mathbf{Q}^{(f)} = mean(\{\mathbf{Q}^{(h)}\}_h)$. We concatenate $\mathbf{Q}^{(f)}$ and the frame-level feature from a video backbone $\mathbf{X}^{(f)}$, so that we

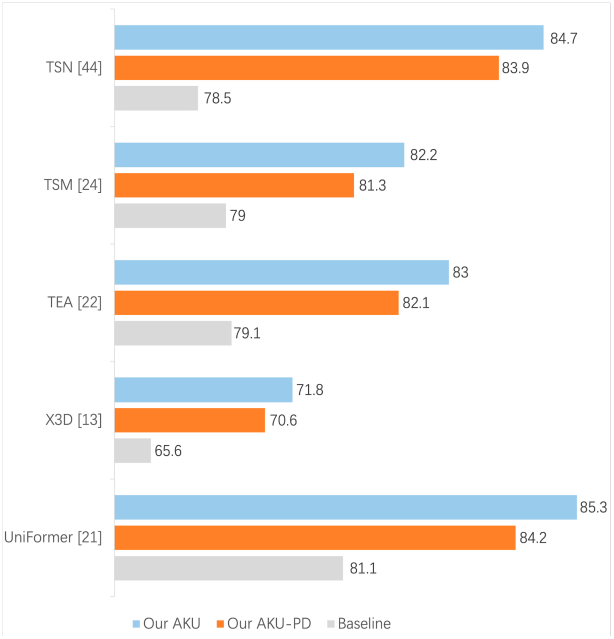

Figure 5: Recognition Accuracy on Different Video Backbones. PD means that we use the predicted verbs (part movements) in our approach. Our AKU framework can largely boost video baselines with human action knowledge.

can integrate both human-centric and scene-related representation together, in order to generate the action prediction vector $\mathbf{C}^{(f)}$ of this frame,

$$\mathbf{C}^{(f)} = \mathcal{J}(Concate(\mathbf{Q}^{(f)}, \mathbf{X}^{(f)})), \tag{11}$$

where $\mathcal{J}$ is a FC layer. Subsequently, we perform average pooling over all the frames to generate the action prediction vector of a video, i.e., $\mathbf{C}^{(v)} = mean(\{\mathbf{C}^{(f)}\}_f)$. We use cross entropy loss between $\mathbf{C}^{(v)}$ and ground truth action vector $\mathbf{Y}^{(v)}$,

$$\mathcal{L}^{(v)} = \mathcal{CE}(\mathbf{C}^{(v)}, \mathbf{Y}^{(v)}). \tag{12}$$

Then, we combine both verb and action prediction losses together for end-to-end training of our AKU framework,

$$\mathcal{L} = \mathcal{L}^{(p)} + \mathcal{L}^{(v)}. \tag{13}$$

5 EXPERIMENTS

Dataset. We evaluate our knowledge base and models in this section. Specifically, we investigate our paradigm on the recent Kinetics-TPS benchmark [25], since this dataset aims at parsing human actions by compositional learning of body part movements. It consists of 4,741 videos with a length of 30 seconds on average, which are selected from Kinetics-700 [6], and contain 24 action classes with complex human dynamics in the wild. Different from the existing datasets [1, 20, 49], it provides human part annotations, i.e., 7.9M annotations of 10 body parts, 7.9M part states, and 0.5M interactive objects, for detailed action understanding in videos.

Implementation Details. First, we use stochastic gradient descent (SGD) optimizer with a base learning rate of 0.01 and the momentum is 0.9. The model is trained with 100 epochs and the weight decay is 5×10^{-4}. The learning rate is warmed up for 10% of the total training epochs and decayed to zero following a cosine schedule for the rest of the training. We select batch size as 12. Moreover, we initialize visual backbone by leveraging the pre-trained weights from ResNet-50 [22] on ImageNet [27], and choose Xavier Initialization [21] for parameters in all the FC layers. The textual encoder uses the pre-trained Bert [10] to generate the 768-dim semantic embedding vector. Third, we uniformly sample 8 frames and set spatial resolution of the input frames to 224×224. Meanwhile, we follow the original train/test splits in Kinetics-TPS. The whole experiments are based on Pytorch with 8 NVIDIA RTX A5000 GPUs. For evaluation, we report the recognition accuracy on the test set to compare with other approaches in the literature.

Table 1: Recognition Accuracy on Each Category of Kinetics-TPS. We compare our AKU framework with the recent human activity knowledge approach PaStaNet [33].

Action Categories	Prediction		Ground Truth	
	PaStaNet [33]	Our AKU	PaStaNet [33]	Our AKU
belly dancing	76.2	92.9	81.0	95.2
push up	67.5	92.5	75.0	90.0
lunge	60.0	80.0	62.5	82.5
country line dancing	56.4	76.9	64.1	76.9
surfing crowd	73.0	97.3	73.0	100.0
skipping rope	50.0	75.0	55.0	75.0
hitting baseball	57.9	92.1	60.5	92.1
tobogganing	63.9	94.4	66.7	97.2
hammer throw	48.7	74.4	48.7	76.9
pull ups	85.0	85.0	85.0	85.0
salsa dancing	35.9	76.9	41.0	76.9
front raises	70.0	70.0	70.0	70.0
hurling (sport)	92.1	94.7	94.7	94.7
hopscotch	78.9	89.5	81.6	89.5
deadlifting	67.5	77.5	67.5	80.0
skateboarding	66.7	92.3	71.8	92.3
capoeira	59.0	82.1	59.0	82.1
throwing discus	61.5	87.2	66.7	87.2
clean and jerk	27.5	62.5	32.5	65.0
snatch weight lifting	42.5	57.5	50.0	57.5
punching bag	65.8	92.1	73.7	94.7
juggling balls	60.5	89.5	60.5	89.5
riding mechanical bull	73.7	89.5	78.9	89.5
crawling baby	92.3	94.9	97.4	97.4
Avg	63.8	**83.9**	67.3	**84.7**

5.1 Action Recognition

We first evaluate visual knowledge graph on action recognition of Kinetics-TPS. Specifically, we set the prediction and ground truth modes for our AKU paradigm. (1) Prediction Mode. As mentioned before, the body movement (i.e., verb) is not always available in practice. Hence, we propose to learn the verb label in the training, and predict the verb label in the testing for action reasoning in our visual knowledge graph. (2) Ground Truth Mode. We also use the ground truth verb annotations as the given input of video. It is an oracle method to compare and evaluate our paradigm, i.e., it means that we can recognize human part movements perfectly and reason out the uncertainty from the best starting point.

Results. First, we evaluate our AKU paradigm on various video backbones including 2D CNNs [31, 34, 58], 3D CNNs [16], and video transformer [30]. As shown in Fig. 5, our AKU paradigm

Table 2: Knowledge Miner.

Action Knowledge Base	ACC.
Without	78.5
<P,V,O>	82.3
<P,V,O> + Action Tag	84.1
<P,V,O> + Action Tag + Action Description	**84.7**

Table 3: Visual Learner.

Part-Level Learner	Human-Level Learner	ACC.
MLP	Tree	81.4
SCM	Tree	82.5
Graph	Tree	**84.7**
Graph	Mean	83.0
Graph	Linear	82.9
Graph	MLP	82.1
Graph	GCN	81.9

Table 4: Action Recognizer.

Modality Fusion	Relation Learning	ACC.
Max	GCN	83.3
Mean	GCN	83.8
Sum	GCN	**84.7**
Sum	Transformer	84.1
Sum	Mean	83.1

can significantly boost all these video backbones with large accuracy improvement. It illustrates that, our knowledge base provides discriminative semantic knowledge to enhance human action recognition. Additionally, the accuracy gap between prediction and ground truth modes of our AKU paradigm is small. It shows that, our knowledge base is actually robust for human action reasoning, even though the verb is predicted from video without ground truth annotations. Second, we compare our AKU paradigm with PaStaNet [33], a recent knowledge-based human action recognition approach. As shown in Table 1, our AKU significantly outperforms PaStaNet on most action categories, in terms of both prediction and ground truth modes. It shows that, our AKU paradigm is a preferable knowledge-based framework for human action recognition in videos. Finally, we also perform PaStaNet [33] with our visual knowledge graph. In ground truth mode, the action recognition accuracy of PaStaNet is improved from 67.3% to 73.3%, which clearly shows the effectiveness of our visual knowledge graph.

5.2 Ablation Studies

In this section, we do extensive experiments to verify that, each module plays an important role in contributing to the overall performance. Unless stated otherwise, we explore all the experiments on the ground truth mode for fair comparison.

Knowledge Miner. We first evaluate our knowledge miner with different types of action semantics. As shown in Table 2, the performance is the worst if we do not use any knowledge. When we gradually integrate the semantic knowledge of Part-Verb-Object (i.e., <P,V,O>), action tag and description, our knowledge miner can progressively boost accuracy. It clearly shows the effectiveness of this module, by integrating rich and commonsense knowledge from both visual and textual data.

Table 5: Different Knowledge Embedding.

Language Model	Input Type	ACC.
Bert[10]	word	83.2
Bert[10]	sentence	**84.7**
GloVe[43]	sentence	84.0
Word2Vec[39]	sentence	83.4

Table 6: Body Part Movement (Verb) Prediction.

Part-Level Learner	Verb ACC.	Action ACC.
MLP	68.4	82.5
Graph	**71.5**	**83.9**

Table 7: Our AKU Framework for Few-Shot Learning.

Method	Few@1	Few@5	Few@10	Few@50
TSN[58]	17.8	39.1	48.3	70.4
PaStaNet[33]	17.1	30.7	37.7	58.1
Our AKU (Action)	20.4	41.6	50.9	72.5
Our AKU (PD)	21.5	42.3	51.6	73.6
Our AKU	**22.8**	**43.5**	**52.7**	**74.6**

Part-Level Visual Learner. In the part-level visual learner, we introduce a contextual function to enhance human part features (Eq. 7). Table 3 shows different choices of contextual functions. (1) **Graph.** We use an attentive graph mechanism. First, we use a cross attention module, where we use the human feature as query, and use its part features as key and value. In this case, we can integrate the part features as a contextual feature, with guidance of human feature. Second, we concatenate each part feature, this contextual feature and object feature as a feature vector of the corresponding body part. Then we use a FC layer to map this vector as an enhanced part feature, which has the same dimension of the original part feature. Note that, if the part is not associated with an object, we use the frame feature to replace the object feature for concatenation. (2) **Spatial Configuration Map (SCM).** The spatial configuration map (SCM) contains the skeleton information of human body and the relative position relationship between human body and object. We use it to enhance body part feature, by following [56]. (3) **MLP.** We use two 1024-dim FC layers to enhance part-level representation. As shown in Table 3, the graph mechanism achieves the best. Hence, we set it as the default setting of part-level visual learner.

Human-Level Visual Learner. In the human-level visual learner, we introduce an aggregation function (Eq. 8) to summarize part-level features as a human-level feature. We design different settings of this function in Table 3. (1) **Mean.** We simply operate average

Figure 6: Visualization of Action Recognition.

pooling over all the part features to construct a human-level feature. (2) **Linear.** We concatenate all the part features as a vector, and use a FC layer to convert it as a human-level feature. (3) **MLP.** We feed the concatenated part feature vector into two 1024-dim FC layers (with relu and dropout) for learning human-level feature. (4) **Graph Convolution Network (GCN).** We use a 4-layer graph convolution network to learn relations between different part features. After that, we perform two 1024-dim FC layers to generate a human-level feature. (5) **Tree.** Human body has a natural hierarchy. Hence, we transform all the part features through this tree structure, as suggested in [33]. Specifically, 10 body parts are categorized into head, arm, hip, leg. For each category, we concatenate the corresponding part features and pass through a 512-dim FC layer to represent a middle-level vector. Then, these categories are further summarized as upper-body and lower-body components. For each component, we concatenate the middle-level vectors and pass through a 512-dim FC layer to represent a high-level vector. Finally, we concatenate these high-level vectors and the object feature into a 512-dim FC layer to construct a human-level feature. As shown in Table 3, the tree setting achieves the best performance. It indicates that, the physical connection in the body is effective to construct human-level feature from part-level features. Hence, we set it as the default setting of human-level visual learner.

Action Recognizer. We validate modality fusion on knowledge and visual features of a human instance. (1) **Mean**. It is the average pooling operation. (2) **Sum**. It is the sum pooling operation. (3) **Max**. It is the max pooling operation. Moreover, we investigate the relation learning function (Eq. 10) with the popular settings. (1) **GCN**. We feed all human features into 2-layer graph convolution network to learn the relation of different humans. Then the enhanced representations are feed into average pooling to generate a video-level feature. (2) **Transformer**. We feed all human features into ViT-B32 [12] to learn the relation of different human instances. After that, we perform average pooling to generate a video-level feature. (3) **Mean**. We operate average pooling over all the human features as a video-level feature. As shown in Table 4, we choose the sum operation of modality fusion and GCN-based relation learning in action recognizer for better performance.

Different Knowledge Embedding. We evaluate different textual encoders in Table 5. First, we investigate the types of input text. Clearly, the sentence type performs better due to more contextual information. Second, we replace Bert [10] with Word2Vec [39] and GloVe [43] to encode text. As expected, Bert performs the best with large-scale language pretraining.

Body Part Movement (Verb) Prediction. We finally evaluate the impact of verb prediction, based on different part-level visual learners. In Table 6, the final action accuracy depends on the accuracy of body part movement prediction. It indicates that, human action is actually the complex composition of body part movements.

5.3 Few-Shot Action Recognition

Settings. We investigate the potential of our knowledge graph reasoning for few-shot generalization. In this setting, we first train each model with only k examples for each category on Kinetics-TPS, where k = 1, 5, 10, 50 respectively. The iteration finishes on 100 epochs. Then we evaluate the trained models on all the testing videos in Kinetics-TPS.

Results. We report the few-shot performance in Table 7. "Few@i" indicates the recognition accuracy with i training videos per category. We compare our AKU paradigm with TSN [58] and PaStaNet [33]. For fairness, we use TSN as video backbone of our AKU paradigm. We can easily see that, our AKU paradigm significantly outperforms these approaches, with our distinct knowledge graph reasoning. Hence, it is a preferable framework for limited data scenarios in practice.

5.4 Visualization

To further verify the effectiveness of AUK, we conduct visualizations on action recognition. In Fig. 6, we show the comparison of PaStaNet[33] and our method. We can see that, it is difficult to distinguish between *hammer throw* and *hitting baseball* in PaStaNet. Alternatively, our AKU paradigm makes the prediction score as 93% for *hitting baseball*, with the proposed knowledge base. All the results indicate that, our AKU paradigm can effectively integrate visual and semantic knowledge to recognize complex actions.

6 CONCLUSION

In this paper, we propose a visual knowledge graph for human action reasoning in videos. Based on this distinct graph, we develop a generic multi-modal Action Knowledge Understanding (AKU) framework, which can progressively infer human actions by part-level parsing. Extensive experiments show its superior performance for boosting various video backbones. We expect that, it could provide a new perspective for detailed human action understanding, and potentially fill the knowledge gap between machine and human intelligence in this research and beyond.

ACKNOWLEDGEMENTS

This work is partially supported by National Natural Science Foundation of China (61876176), the Joint Lab of CAS-HK, the Shanghai Committee of Science and Technology (Grant No. 21DZ1100100).

REFERENCES

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675* (2016).

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

[3] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*. Springer, 722–735.

[4] A. Bochkovskiy, C. Y. Wang, and Hym Liao. 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. (2020).

[5] Zhaowei Cai and Nuno Vasconcelos. 2018. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6154–6162.

[6] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. 2019. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987* (2019).

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

[8] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. 2018. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7103–7112.

[9] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. 2014. Large-scale object classification using label relation graphs. In *European conference on computer vision*. Springer, 48–64.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[11] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. 2021. Cswin transformer: A general vision transformer backbone with cross-shaped windows. *arXiv preprint arXiv:2107.00652* (2021).

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[13] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. 2021. Multiscale Vision Transformers. *CoRR* abs/2104.11227 (2021). arXiv:2104.11227 https://arxiv.org/abs/2104.11227

[14] Hao-Shu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu. 2018. Pairwise body-part attention for recognizing human-object interactions. In *Proceedings of the European conference on computer vision (ECCV)*. 51–67.

[15] Yuan Fang, Kingsley Kuan, Jie Lin, Cheston Tan, and Vijay Chandrasekhar. 2017. Object detection meets knowledge graphs. International Joint Conferences on Artificial Intelligence.

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

[17] C. Feichtenhofer, H. Fan, J. Malik, and K. He. 2019. SlowFast Networks for Video Recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*.

[18] J. Gao, T. Zhang, and C. Xu. 2019. I Know the Relationships: Zero-Shot Action Recognition via Two-Stream Graph Convolutional Networks and Knowledge Graphs. *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (2019), 8303–8311.

[19] Pallabi Ghosh, Nirat Saini, Larry S. Davis, and Abhinav Shrivastava. 2020. All About Knowledge Graphs for Actions. https://doi.org/10.48550/ARXIV.2008.12432

[20] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6047–6056.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. 1026–1034.

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

[23] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. 2020. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10236–10247.

[24] Boyuan Jiang, Mengmeng Wang, Weihao Gan, Wei Wu, and Junjie Yan. 2019. STM: SpatioTemporal and Motion Encoding for Action Recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2000–2009. https://doi.org/10.1109/ICCV.2019.00209

[25] Kinetics-TPS. 2021. https://deeperaction.github.io/.

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

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012).

[28] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. 2021. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3383–3393.

[29] Kunchang Li, Xianhang Li, Yali Wang, Jun Wang, and Yu Qiao. 2021. CT-Net: Channel Tensorization Network for Video Classification. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=UoaQUQREMOs

[30] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. 2022. UniFormer: Unifying Convolution and Self-attention for Visual Recognition. *arXiv preprint arXiv:2201.09450* (2022).

[31] Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. 2020. Tea: Temporal excitation and aggregation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 909–918.

[32] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. 2021. Tokenpose: Learning keypoint tokens for human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11313–11322.

[33] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Shu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. 2020. Pastanet: Toward human activity knowledge engine. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 382–391.

[34] Ji Lin, Chuang Gan, and Song Han. 2019. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7083–7093.

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10012–10022.

[36] Zhaoyang Liu, Donghao Luo, Yabiao Wang, Limin Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Tong Lu. 2020. Teinet: Towards an efficient architecture for video recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11669–11676.

[37] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2016. Visual relationship detection with language priors. In *European conference on computer vision*. Springer, 852–869.

[38] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. 2020. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1049–1059.

[39] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[40] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[41] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. 2021. Video transformer network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3163–3172.

[42] Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F Henriques. 2021. Keeping your eye on the ball: Trajectory attention in video transformers. *Advances in Neural Information Processing Systems* 34 (2021).

[43] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[44] Zhaobo Qi, Shuhui Wang, Chi Su, Li Su, Qingming Huang, and Qi Tian. 2020. Towards more explainability: concept knowledge mining network for event recognition. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*. 3857–3865.

[45] Zhaobo Qi, Shuhui Wang, Chi Su, Li Su, Qingming Huang, and Qi Tian. 2021. Self-Regulated Learning for Egocentric Video Activity Anticipation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021). https://doi.org/10.1109/TPAMI.2021.3059923

[46] Zhaobo Qi, Shuhui Wang, Chi Su, Li Su, Weigang Zhang, and Qingming Huang. 2020. Modeling Temporal Concept Receptive Field Dynamically for Untrimmed Video Analysis. In *Proceedings of the ACM International Conference on Multimedia*

(ACM MM). 3798–3806.

[47] Zhaofan Qiu, Ting Yao, and Tao Mei. 2017. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 5534–5542. https://doi.org/10.1109/ICCV.2017.590

[48] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).

[49] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

[50] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

[51] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*. PMLR, 10347–10357.

[52] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[53] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. 2019. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5552–5561.

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

[55] Carl Vondrick, Deniz Oktay, Hamed Pirsiavash, and Antonio Torralba. 2016. Predicting motivations of actions by leveraging text. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2997–3005.

[56] B. Wan, D. Zhou, Y. Liu, R. Li, and X. He. 2019. Pose-Aware Multi-Level Feature Network for Human Object Interaction Detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*.

[57] Limin Wang, Wei Li, Wen Li, and Luc Van Gool. 2018. Appearance-and-relation networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1430–1439.

[58] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*. Springer, 20–36.

[59] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 2021. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. *CoRR* abs/2102.12122 (2021). arXiv:2102.12122 https://arxiv.org/abs/2102.12122

[60] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7794–7803.

[61] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4622–4630.

[62] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. 2021. Focal Self-attention for Local-Global Interactions in Vision Transformers. *CoRR* abs/2107.00641 (2021). arXiv:2107.00641 https://arxiv.org/abs/2107.00641

[63] Fang Yuan, Zhe Wang, Jie Lin, Luis Fernando D'Haro, Kim Jung Jae, Zeng Zeng, and Vijay Chandrasekhar. 2017. End-to-end video classification with knowledge graphs. *arXiv preprint arXiv:1711.01714* (2017).

[64] Hong Zhang, Hao Ouyang, Shu Liu, Xiaojuan Qi, Xiaoyong Shen, Ruigang Yang, and Jiaya Jia. 2019. Human pose estimation with spatial contextual information. *arXiv preprint arXiv:1901.01760* (2019).

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

[66] Penghao Zhou and Mingmin Chi. 2019. Relation parsing neural network for human-object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 843–851.