Research Statement – Yansong Tang

yansong.tang@eng.ox.ac.uk - https://andytang15.github.io

The world we live in is inherently dynamic. With the development of intelligent visual sensors and the Internet, there is an explosive growth of video data from the surveillance systems, smartphones, social media (*e.g.,* YouTube, TikTok), *etc*. As one of the most important and fundamental directions in computer vision, video understanding has wide practical applications, such as human-computer interaction, sport video analysis, video retrieval, and many others.

Recent decades have witnessed the success of deep learning techniques by designing various deep neural networks in the field of computer vision. However, there are still many challenges for video understanding, especially analysing human activity in different kinds of dynamic scenes shown in Figure 1. For example, (1) in the temporal domain, there is a lot of redundant information between adjacent frames of video data, and it is difficult for conventional models to capture the key information. (2) In the spatial domain, human activities (*e.g.,* skeleton-based action or group activity shown in Figure 2 (b)) usually lie in some non-Euclidean spaces, and the conventional deep learning models (*e.g.,* CNN, RNN) have limitations in modeling the high-level structures. (3) From the data perspective, although there have been various datasets presented for general action recognition, the most existing datasets in some emerging fields are still limited in the diversity and scale, *e.g.,* instructional activity (Figure 2 (c)) and RGBD egocentric action [11]. (4) From the knowledge perspective, there is inherent ambiguity in the score labels for action quality assessment, resulting in an unsatisfying performance for the most existing regression approaches. To address these issues, **my research focuses on enhancing the spatial-temporal modeling ability, building large-scale datasets, and effectively utilizing the knowledge to make intelligent systems perceive, understand, and interact with the surrounding dynamic environment.** Towards these goals, I have made research contributions in both academia and industry. Figure 1 presents a research map of my previous research and future plan, in which some representative works will be detailed as follows.

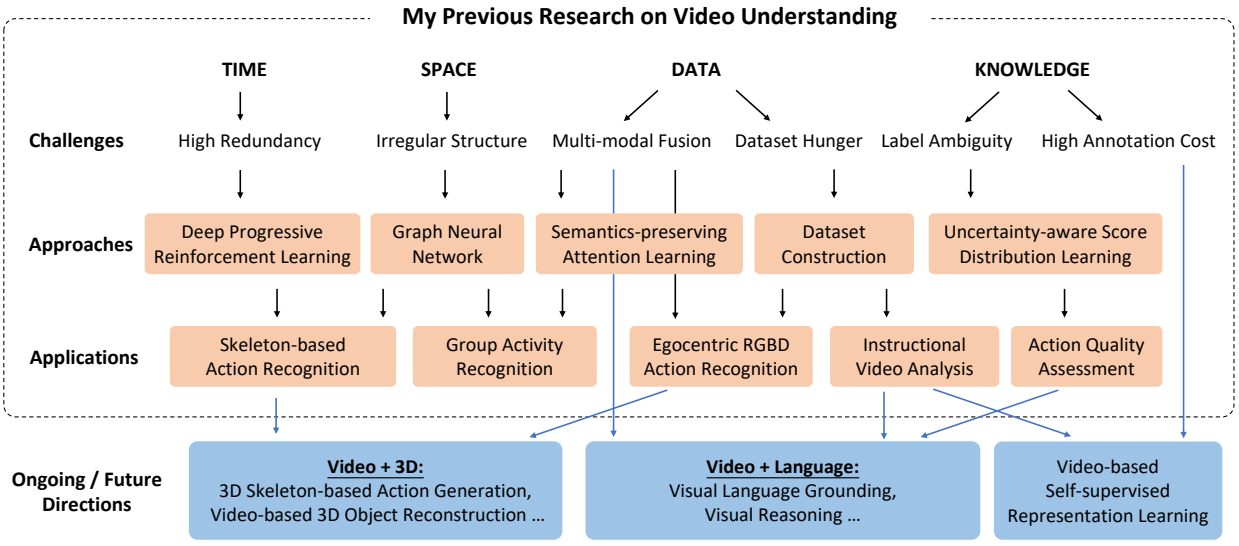
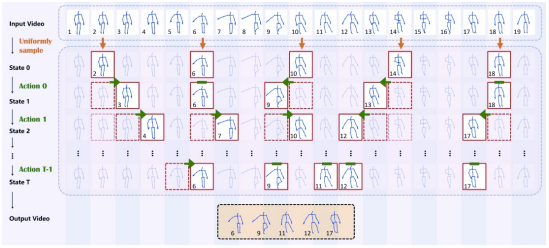
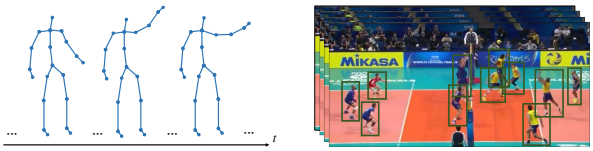


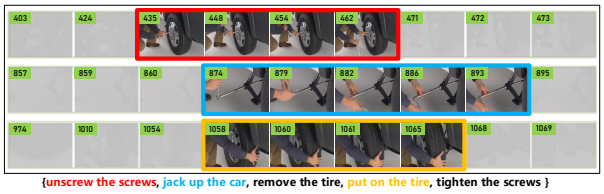
Figure 1: Overview of my previous research (orange), and future plan (blue) in video understanding.



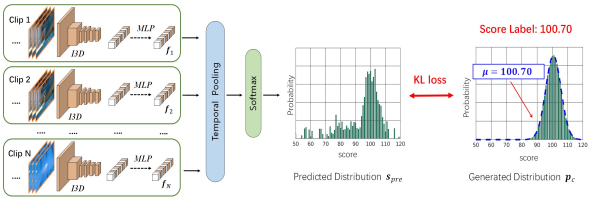
(a) Key Frames Selection with Progressive Reinforcement Learning



(b) Non-Euclidean Structure of Skeleton-based Action and Group Activity



(c) Different Steps in Instructional Video



(d) Uncertainty-aware Score Distribution Learning for Action Quality Assessment

Figure 2: (a) shows the key frames selection process with progressive reinforcement learning [9]. (b) illustrates the non-Euclidean structures lie in skeleton-based action represented by different joints of human body and group activity represented by different persons. (c) displays different steps of an instructional video in the COIN [7] dataset. (d) presents the pipeline of our proposed uncertainty-aware score distribution learning approach for action quality assessment [8].

1 My Previous Research on Video Understanding

Temporal Domain: Deep Progressive Reinforcement Learning for Key Frames Selection. Compared with the widely-studied image-based data, the videos with higher dimension has much more information, especially the high redundancy between adjacency frames may interfere with the recognition process of the model. While conventional deep learning approaches (*e.g.,* CNN or RNN based models) consider all of the frames in a sequence as equally important, they fail to focus on the most representative frames in an action sequence. To address this issue, we take skeleton-based action as an example, and propose a deep progressive reinforcement learning approach to select the most informative frames for action recognition as shown in Figure 2 (a). We formulate the selection procedure as a Markov decision process (MDP), and optimize a model to adjust selection results progressively by the deep reinforcement learning algorithm. Experiments on various datasets have shown the effectiveness of our proposed method. This work [9] has obtained *240+ citations* since it was published in CVPR 2018.

Spatial Domain: Semantic-preserving Attention and Graph Neural Network for Relation Modelling. As shown in Figure 2(b), human activities usually lie in some non-Euclidean spaces in the spatial domain. Different from conventional tensor-based methods, we propose two relation modelling approaches to capture the high-level dependency in the spatial domain.
(1) Semantics-preserving attention learning: I first consider group activity recognition, which aims to discern what a group of people are doing in a video. Compared with conventional action recognition based on a single person, it is a more challenging task as it requires further understanding of high-level relationships among different people. Different from the common attention mechanisms, we develop a Teacher Network to leverage the prior knowledge in the semantics domain, and explore the discriminative information of different people by transferring the attention learned by the Teacher Network in the semantics domain (*i.e.,* semantics-preserving attention) to the Student Network in the appearance domain (ACM MM 2018 ORAL [10]). *To my best knowledge, this is the original effort to leverage attention in both semantics and appearance clues to perform group activity recognition.*

(2) Graph Neural Network: Furthermore, since a group of people can be considered as a graph-based structure, where the node and edge represent each individual person and the relationship between two people respectively, we utilize the graph convolutional modules to reason about the relationship of different people and extend our method for group activity detection in untrimmed long videos (TIP 2019 [6]). Similarly, we could adopt the graph-based model for skeleton-based action recognition, where the vertices of the graph contain the 3D coordinates of human joints, and the edges reflect the relationships between joints [9] (CVPR 2018). Moreover, we discover that the relation between different graph models could be transferred by our proposed adversarial learning strategy to improve the performance of cross-dataset action recognition (TCSVT 2020 [12]).

Data Perspective: the COIN Dataset and Performance Evaluation. There are substantial instructional videos on the Internet, which enable novices to acquire knowledge for completing various tasks. Compared with short clips, instructional videos have longer durations, and an interesting task is to localize the different steps in the videos. To tackle this problem, we first collect a dataset called "COIN" for COmprehensive INstructional video analysis (CVPR 2019 [3]). It contains 11,827 videos of 180 tasks (*e.g.,* change car tire, replace the doorknob) from 12 domains (*e.g.,* vehicles, household item). Figure 2 (c) presents a video example of the COIN dataset. As a by-product, we contribute a new toolbox to effectively annotate a series of step descriptions and the corresponding temporal boundaries. Furthermore, we exploit two important characteristics (*i.e.,* task-consistency and ordering-dependency) for localizing important steps in instructional videos (TPAMI 2020 [7]). Accordingly, we propose two simple yet effective methods, which can be easily plugged into conventional proposal-based action detection models. *To my best knowledge, the COIN is the currently largest human-annotated instructional dataset.* Since the release of COIN dataset, it has been widely used in the community, including both academia (*e.g.,* University of Oxford [1], INRIA [15]) and industry (*e.g.,* DeepMind [1], Baidu [14], Google Research [2]). The dataset and source code are available at https://coin-dataset.github.io.

Knowledge Perspective: Label Distribution Learning for Action Quality Assessment (AQA). Assessing action quality from videos has attracted growing attention in recent years. Most existing approaches usually tackle this problem based on regression algorithms, which ignore the intrinsic ambiguity in the score labels caused by multiple judges or their subjective appraisals. To address this issue, we propose an uncertainty-aware score distribution learning (USDL) approach for action quality assessment (CVPR 2020 ORAL [8]). Specifically, we regard an action as an instance associated with a score distribution, which describes the probability of different evaluated scores. Moreover, under the circumstance where fine-grained score labels are available (*e.g.,* difficulty degree of an action or multiple scores from different judges), we further devise a multi-path uncertainty-aware score distributions learning (MUSDL) method to explore the disentangled components of a score. We conduct experiments on three AQA datasets containing various Olympic actions and surgical activities, where our approaches set state-of-the-arts under the Spearman's Rank Correlation.

2 Future Research Agenda

In the future, I will continue working in the field of video understanding. With the rapid growth of video data and accelerating advances in artificial intelligence, I foresee my work having impacts across science, society, and industry, including in robotics, drones, personal and wearable devices, AR/VR, and autonomous vehicles. Going beyond the video understanding, I am branching out to explore the field of 3D vision, multi-modal learning and self-supervised representation learning. Specifically, I outline three directions that I plan to pursue as follows.

Video + 3D. Humans live in a three-dimensional and dynamic world. To enhance the perception ability of 3D human action, my previous works introduced a graph neural network model to capture the

3D structure of skeleton-based action [9], and designed a multi-steam architecture to fuse the complementary information of the appearance, motion, and geometry from RGBD egocentric action [11]. Regarding my future research, I tend to agree with the view expressed by Richard Feynman, *"What I cannot create, I do not understand"*. As two possible goals I see in this direction, firstly I plan to investigate the problem of 3D skeleton-based action generation, which requires a better understanding of an action at the *atomic level* in both spatial and temporal domains. I believe this topic has great research potential and large application value for AR and 3D character animations. Secondly, I am now involving a 3D object reconstruction project within the University of Oxford and Apple Inc. jointly, and am considering to utilize the *temporal consistency* and *multi-view geometry* from video data to enhance the performance for 3D object reconstruction.

Video + Language. There has been growing attention on the interaction of CV and NLP. Based on my previous research, I plan to explore visual language grounding and visual reasoning in the future. For the first field, I am mentoring a DPhil student at the University of Oxford on video object segmentation from referring expressions. We proposed a hierarchical cross-attention model to effectively fuse the inputs of different modalities (RGB frames, optical flow, and text) and achieved state-of-the-art results on 3 public benchmarks [13]. As a promising direction, it is interesting to extend this work to the temporal domain for language-guided temporal localization. The second research direction is to investigate visual reasoning with the language. One interesting task is to re-establish the COIN dataset under the abductive visual reasoning paradigm, where the goal is to infer the most plausible sequence of steps between two observations. Besides, as an in-depth exploration of our AQA work [8], I expect to answer the question "why does the model output a certain score given an action?" To address this problem, one possible solution is to construct a knowledge graph describing the scoring rule of an action, and help the model to better understand how each body part or temporal segment contributes to the final action score. *This will be the pioneering work to explore the interpretability in AQA field.*

Video-based Self-supervised Representation Learning. Recent success of deep neural networks heavily relies on a large amount of human supervision, which is limited by the high cost of human labor. To address this problem, *an emerging direction is to leverage the structure and redundant information in data as self-supervisory signals to train deep networks.* Towards this goal, I have made some primary attempts by designing different proxy tasks for skeleton-based action recognition and dense label propagation. The first work [5] is to devise a temporal-spatial Cubism strategy, which guides the network to be aware of the permutation of the segments in the temporal domain and the body parts in the spatial domain separately, thus improves the generalization ability of the model. And the second work [4] is to introduce a fully convolutional model to take the cycle-consistency in time as free supervisory, and train an encoder tracking forward in time and backward to the initial position to form a cycle. As these two works focus on the high-level (instance level) and low-level (pixel level) representation learning, I will explore a *general* self-supervised representation learning approach to unify these tasks. Also, it is interesting to utilize the correspondence of multi-modal data (*e.g,* visual stream and caption in instructional video) for a better self-supervised representation learning.

References

[1] A. Miech, J. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, pages 9876–9886, 2020.

[2] P. H. Seo, A. Nagrani, and C. Schmid. Look before you speak: Visually contextualized utterances. In *CVPR*, pages 16877–16887, 2021.

[3] Y. Tang, D. Ding, Y. Rao, Y. Zheng, D. Zhang, L. Zhao, J. Lu, and J. Zhou. COIN: A large-scale dataset for comprehensive instructional video analysis. In *CVPR*, pages 1207–1216, 2019.

[4] Y. Tang*, Z. Jiang*, Z. Xie*, Y. Cao, Z. Zhang, P. H. S. Torr, and H. Hu. Breaking shortcut: Exploring fully convolutional cycle-consistency for video correspondence learning. In *submission*, 2021.

[5] Y. Tang*, X. Liu*, X. Yu, D. Zhang, J. Lu, and J. Zhou. Learning from temporal spatial cubism: A self-supervised domain adaptation approach for skeleton-based action recognition. *ACM TOMM*, 2021.

[6] Y. Tang, J. Lu, Z. Wang, M. Yang, and J. Zhou. Learning semantics-preserving attention and contextual interaction for group activity recognition. *TIP*, 28(10):4997–5012, 2019.

[7] Y. Tang, J. Lu, and J. Zhou. Comprehensive instructional video analysis: The COIN dataset and performance evaluation. *TPAMI*, 43(9):3138–3153, 2021.

[8] Y. Tang*, Z. Ni*, J. Zhou, D. Zhang, J. Lu, Y. Wu, and J. Zhou. Uncertainty-aware score distribution learning for action quality assessment. In *CVPR*, pages 9836–9845, 2020.

[9] Y. Tang*, Y. Tian*, J. Lu, P. Li, and J. Zhou. Deep progressive reinforcement learning for skeleton-based action recognition. In *CVPR*, pages 5323–5332, 2018.

[10] Y. Tang, Z. Wang, P. Li, J. Lu, M. Yang, and J. Zhou. Mining semantics-preserving attention for group activity recognition. In *ACM MM*, pages 1283–1291, 2018.

[11] Y. Tang, Z. Wang, J. Lu, J. Feng, and J. Zhou. Multi-stream deep neural networks for RGB-D egocentric action recognition. *TCSVT*, 29(10):3001–3015, 2019.

[12] Y. Tang, Y. Wei, X. Yu, J. Lu, and J. Zhou. Graph interaction networks for relation transfer in human activity videos. *TCSVT*, 30(9):2872–2886, 2020.

[13] Z. Yang*, Y. Tang*, L. Bertinetto, H. Zhao, and P. H. S. Torr. Hierarchical interaction network for video object segmentation from referring expressions. In *submission*, 2021.

[14] L. Zhu and Y. Yang. Actbert: Learning global-local video-text representations. In *CVPR*, pages 8743–8752, 2020.

[15] D. Zhukov, J. Alayrac, I. Laptev, and J. Sivic. Learning actionness via long-range temporal order verification. In *ECCV*, volume 12374, pages 470–487, 2020.