PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Video-based complex human event recognition with a probabilistic transformer

Hongji Guo, Alexander Aved, Collen Roller, Erika Ardiles-Cruz, Qiang Ji

Hongji Guo, Alexander Aved, Collen Roller, Erika Ardiles-Cruz, Qiang Ji, "Video-based complex human event recognition with a probabilistic transformer," Proc. SPIE 12525, Geospatial Informatics XIII , 125250J (15 June 2023); doi: 10.1117/12.2664106

SPIE.

Event: SPIE Defense + Commercial Sensing, 2023, Orlando, Florida, United States

Video-Based Complex Human Event Recognition With a Probabilistic Transformer

Hongji Guo[a], Alexander Aved[b], Collen Roller[b], Erika Ardiles-Cruz[b], and Qiang Ji[a]

[a]Rensselaer Polytechnic Institute, Troy, NY 12180, USA
[b]Air Force Research Laboratory, Rome, NY 13441, USA

ABSTRACT

Complex human events are high-level human activities that are composed of a set of interacting primitive human actions over time. Complex human event recognition is important for many applications, including security surveillance, healthcare, sports and games. Complex human event recognition requires recognizing not only the constituent primitive actions but also, more importantly, their long range spatiotemporal interactions. To meet this requirement, we propose to exploit the self-attention mechanism in the Transformer to model and capture the long-range interactions among primitive actions. We further extend the conventional Transformer to a probabilistic Transformer in order to quantify the event recognition confidence and to detect anomaly events. Specifically, given a sequence of human 3D skeletons, the proposed model first performs primitive action localization and recognition. The recognized primitive human actions and their features are then fed into the probabilistic Transformer for complex human event recognition. By using a probabilistic attention score, the probabilistic Transformer can not only recognize complex events but also quantify its prediction uncertainty. Using the prediction uncertainty, we further propose to detect anomaly events in an unsupervised manner. We evaluate the proposed probabilistic Transformer on FineDiving dataset and Olympics Sports dataset for both complex event recognition and abnormal event detection. The dataset consists of complex events composed of primitive diving actions. The experimental results demonstrate the effectiveness and superiority of our method against baseline methods.

Keywords: Complex human event recognition, probabilistic transformer, anomaly detection

1. INTRODUCTION

Complex human events are human activities that involve interactions between humans and their environments for a period of time. As shown in Figure 1, a complex human event such as "picking up a car" consists of a sequence of primitive actions, including "approaching the car", "getting into the car", and "driving away". It is these primitive actions and their spatio-temporal interactions over a period of time that produces a complex human event.

With the ubiquitous presence of surveillance cameras both on the ground and in the air, and an exponential increase in the volume of video data, it is becoming imperative to exploit the latest developments in the machine learning to automatically process the video data for accurate and fast complex human event analysis and recognition. Complex human event recognition has many practical applications such as visual surveillance,[1] human-machine interaction,[2] and anomaly detection.[3]

Despite the fast-growing progress in the field, video-based complex human event recognition remains challenging. Complex human events may last minutes long and it is difficult to capture the long-range dependencies among primitive actions. To address this issue, we adopt the Transformer[4] architecture to capture the long-range dependencies among sub-activities. The self-attention mechanism of Transformer models the pair-wise dependencies among all primitive actions through the scaled dot-product attention. In this way, the interactions

Further author information: (Send correspondence to Qiang Ji)
Qiang Ji: E-mail: jiq@rpi.edu

Geospatial Informatics XIII, edited by Kannappan Palaniappan, Gunasekaran Seetharaman, Joshua D. Harguess, Proc. of SPIE Vol. 12525, 125250J · © 2023 SPIE
0277-786X · doi: 10.1117/12.2664106

among the primitive actions can be modeled for complex human event recognition. However, existing deterministic Transformers cannot quantify their prediction uncertainty, which is crucial for understanding the model and enhancing the model performance. To address this limitation, we extend the deterministic Transformer to a probabilistic Transformer to quantify the prediction uncertainty. Specifically, we model the distribution of the attention scores. In the forward process, the attention scores are obtained by sampling from their distributions. Through the probabilistic modeling, we quantify the prediction uncertainty and use it to perform the abnormal event detection.

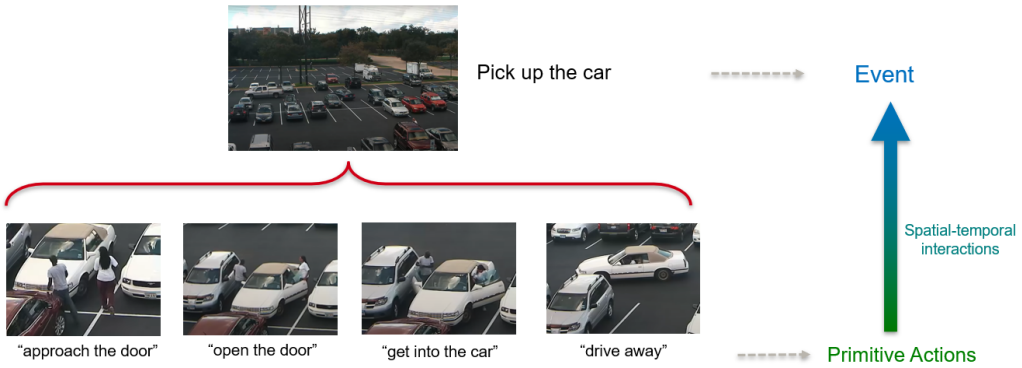

Figure 1. An illustration of a complex human event and primitive actions. It shows that a complex human event is composed of a the execution of a sequence of primitive actions.

In summary, the main contributions of this paper are:

- By modeling the distribution of the scaled dot-product attention, we introduce a probabilistic Transformer for video-based complex human event recognition.
- By probabilistic modeling, we quantify the predictive uncertainty and perform the abnormal event detection in an unsupervised manner.
- We evaluated our proposed method on FineDiving dataset for both complex human event recognition and abnormal human event detection. The experiment results demonstrate the effectiveness of our method.

2. RELATED WORK

2.1 Complex Human Event Recognition

Complex human event recognition, also know as complex human activity recognition or long-term action recognition, aims at identifying complex human actions in long-range videos. The task has been researched for long periods. Early work mainly handled the task with graphical models. Kuehne *et al.*[5] proposed to used an hidden Markov model (HMM) to recover the underlying dynamics of complex human events. Tang *et al.*[6] proposed a variable-duration HMM to tackle the videos with high variation. Wang *et al.*[7] proposed latent hierarchical model (LHM) to describe the decomposition of complex activity into sub-activities in a hierarchical way. The training was formulated in a latent kernelized SVM framework and an efficient cascade inference method was developed to speed up the classification. To capture the dependencies among time intervals, Zhang *et al.*[8] introduced the interval temporal Bayesian network (ITBN), based on combining the Bayesian network with the interval algebra to capture the temporal dependencies.

Recent work mostly focus on using the deep learning models to handle the task. Liu *et al.*[9] proposed latent task learning with privileged information (LTL-PI), a learning framework for complex action recognition by using a sequence of existing simple actions. A probability matrix of each action is constructed based on the training set and is used as the prior knowledge for the composition of complex actions. To make use of the power of graph

learning, Hussein *et al.*[10] proposed VideoGraph. The complex human actions are represented as undirected graphs so that the underlying temporal structure of the complex action can be captured through the graph learning process. To handle the variation of sub-activities, Hussein *et al.*[11] proposed a Timeception layer with multi-size window convolution kernels. The proposed Timeception layer can be embedded into the existing CNN-based models for performance improvement without dramatically increasing the number of parameters. Further, Zhou *et al.*[12] proposed GHRM, a graph-based method that models the high-order relation among sub-activities. By adopting the self-attention mechanism, Guo *et al.*[13] introduced uncertainty-guided probabilistic Transformer. By constructing a majority model for low-uncertainty input and a minority model for high-uncertainty input, the model can well utilize the prediction uncertainty to improve both training and inference.

2.2 Video Anomaly Detection

Video anomaly detection aims at finding the abnormal human event or out-of-distribution data from the video. Cheng *et al.*[14] introduced hierarchical feature representation and Gaussian process regression. To simultaneously detect local and global anomalies, this work formulates the extraction of normal interactions from training videos as the problem of efficiently finding the frequent geometric relations of the nearby sparse spatio-temporal interest points. A codebook of interaction templates is then constructed and modeled using Gaussian process regression. Liu *et al.*[15] introduced future frame prediction for anomaly detection. The anomaly detection is tackled within a prediction framework by leveraging the difference between a predicted future frame and its ground truth. Zhong *et al.*[16] introduced graph convolutional label noise cleaner for anomaly detection. Based on the feature similarity and temporal consistency, the network propagates supervisory signals from high-confidence snippets to low-confidence ones. So the network is capable of providing cleaned supervision for action classifiers. Morais *et al.*[17] performed the anomaly detection by learning regularity in skeleton. The skeleton movement is decomposed into two sub-components: global body movement and local body posture. It models the dynamics and interaction of the coupled features in a message-passing encoder-decoder recurrent network. Ionescu *et al.*[18] introduced object-centric auto-encoders and dummy anomalies for abnormal event detection. In the framework, the abnormal event detection is formulated as an one-verse-rest binary classification problem. An unsupervised feature learning framework is used to encode both motion and appearance information. A supervised classification approach based on clustering the training samples into normality clusters is proposed for the classification. In this paper, we aim to detect the abnormal event based on the prediction uncertainty.

3. METHOD

In this section, we first introduce the overall framework of our method. Then we introduce our proposed probabilistic Transformer. Finally, we show how we quantify the prediction uncertainty and perform the abnormal human event detection.

3.1 Overall Framework

Given the input video, we first performed the primitive action localization with an existing method.[19] The localized primitive actions are fed into our proposed probabilistic Transformer for recognizing the complex human event. At the same time, the probabilistic Transformer output the prediction uncertainty, which is used for abnormal human event detection. An overall framework is shown in Figure 2.

3.2 Probabilistic Transformer

To capture the long-range dependencies of primitive actions for complex human event detection, we adopt the Transformer[4] architecture. Given a sequence as the input tokens, the Transformer outputs a sequence that embeds the interactions among the tokens. The core of the Transformer is the self-attention mechanism, which models the pair-wise dependencies among the input tokens. The self-attention is achieved by the scaled dot-product attention. Here we briefly introduce the scaled dot-product attention.

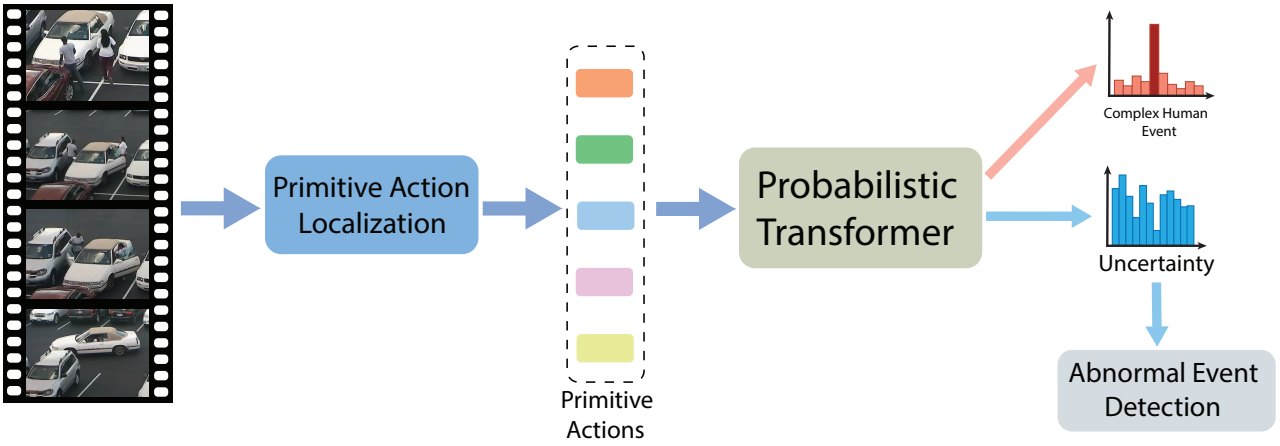

Figure 2. Overall framework of our proposed method. Given a video as input, we first perform the primitive action localization with an existing method.[19] Then the localized primitive actions are fed into the probabilistic Transformer to perform the complex human event recognition. At the same time, out probabilistic Transformer outputs the predictive uncertainty, which is used for abnormal human event detection.

Scaled dot-product attention. Given a sequence of input tokens, positional encoding is firstly added to each token so that the ordinal information of the input can be kept. Here, we use the positional encoding as:

$$\begin{aligned} PE_{(pos,2i)} &= sin(pos/10000^{2i/C}) \\ PE_{(pos,2i++1)} &= cos(pos/10000^{2i/C}) \end{aligned} \tag{1}$$

where pos is the input position and i is the dimension index.

Then, each token is projected to a query (q), a key (k), and a value (v) through linear transformations. The query and key are used to model the dependency and the value is used to store the information. The scaled dot-product attention between i-th and j-th tokens is defined as:

$$\alpha_{ij} = \frac{q_i \cdot k_j}{\sqrt{d}} \tag{2}$$

where d is the dimension of the query and the key.

After obtaining the attention, the Transformer generates the embedding sequence. The i-th embedding z_i is computed as:

$$z_i = \sum_{j=1}^{T} \frac{\alpha_{ij}}{\sum_{j'=1}^{T} \alpha_{ij'}} v_j \tag{3}$$

where T is the length of the sequence.

Probabilistic attention. To capture the prediction uncertainty, we extend the Transformer into a probabilistic one. We model the distributions of attention scores. Specifically, we assume these attention scores follows Gaussian distributions. For i-th input the j-th input token, the attention score α_{ij} follows a Gaussian distribution with mean μ_{ij} and variance σ_{ij}: $\alpha_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma_{ij}^2)$. The mean and variance are generated by a neural network: $\mu_{ij}, \sigma_{ij}^2 = MLP(q_i, k_j, \Phi)$, which takes the corresponding query and key as input. Thus, the attention score α_{ij} is stochastic during both traiing and inference. Besides Φ, we use Θ to denote the remaining parameters of the probabilistic Transformer. During training, we need sample $\mathcal{N}(\mu_{ij}, \sigma_{ij}^2)$ to obtain samples of the attention score α_{ij}. Given different samples of α_{ij}, the standard training procedure for the Transformer can then be used. After training, we obtain the trained probabilistic Transformer with parameters Θ^* and Φ^*.

Given a query input $\boldsymbol{X'}$, the event detection using an ensemble of Transformers can be written as:

$$P(y'|\boldsymbol{X'}, \Theta^*, \Phi^*) = \int_{\boldsymbol{\alpha}} P(y'|\boldsymbol{X'}, \Theta^*, \boldsymbol{\alpha}) p(\boldsymbol{\alpha}|\boldsymbol{X'}, \Phi^*) d\boldsymbol{\alpha} \tag{4}$$

where $\boldsymbol{\alpha} = \{\alpha_{ij} | i, j \in \{1, ..., T\}\}$ represents all the probabilistic attention of input $\boldsymbol{X}$. Thus, the complex human event recognition is formulated as:

$$y^* = \underset{y'}{\operatorname{argmax}} \int_{\boldsymbol{\alpha}} P(y'|\boldsymbol{X'}, \Theta^*, \boldsymbol{\alpha}) p(\boldsymbol{\alpha}|\boldsymbol{X'}, \Phi^*) d\boldsymbol{\alpha} \approx \underset{y'}{\operatorname{argmax}} \frac{1}{K} \sum_{k=1}^{K} P(y'|\boldsymbol{X'}, \Theta^*, \boldsymbol{\alpha_k}) \tag{5}$$

where K is the number of generated samples given one input and α_k represents k-th sample generated from $P(\boldsymbol{\alpha}|\boldsymbol{X'}, \Phi^*)$.

Uncertainty quantification. One advantage of our proposed probabilistic Transformer is the availability for uncertainty quantification. Given $P(y'|\boldsymbol{X'}, \Theta^*, \Phi^*)$, we can estimate the total predictive uncertainty in terms of aleatoric and epistemic uncertainty:

$$\underbrace{\mathcal{H}[P(y'|\boldsymbol{X'}, \Theta^*, \Phi^*)]}_{\text{Total Uncertainty}} = \underbrace{\mathcal{I}[y', \boldsymbol{\alpha}|\boldsymbol{X'}, \Theta^*]}_{\text{Epistemic Uncertainty}} + \underbrace{\mathbb{E}_{P(\boldsymbol{\alpha}|\boldsymbol{X'}, \Phi^*)}[\mathcal{H}[P(y'|\boldsymbol{X'}, \Theta^*, \boldsymbol{\alpha})]]}_{\text{Aleatoric Uncertainty}} \tag{6}$$

where $\mathcal{H}()$ and $\mathcal{I}$ represent the entropy and mutual information. The total uncertainty is the sum of the aleatoric uncertainty and the epistemic uncertainty. Aleatoric uncertainty captures the data uncertainty. it measures the amount of perturbation in the input. Epistemic uncertainty is also called model uncertainty. It measures the model's lack of knowledge about the input due to insufficient representation of the input in the training data. It is hence inversely proportional to the density of the input in the training data.

Training.

3.3 Uncertainty-Based Abnormal human activity detection

Anomaly activity detection aims at detecting the abnormal human activities from videos. The existing work for abnormal event detection is largely supervised, in that they require labelling for the abnormal events. Requiring annotations limits their methods to the known abnormal events. These methods fail to generalize to unknown abnormal events. To address this limitation, by leveraging the proposed probabilistic Transformer, we propose an uncertainty based method for abnormal event detection without using any anomaly annotations for detecting both known and unknown abnormal events. We define abnormal events as human events that rarely appear in the training data. As the epistemic uncertainty measures the density of the input data in the training data, it can be used as a metric for abnormal event detection. We hence use $\mathcal{I}[y', \boldsymbol{\alpha}|\boldsymbol{X'}, \Theta^*]$ in Eq. 6 to detect abnormal events. If the input significantly deviates from the events in the training data, its epistemic uncertainty should be large. We hence classify the input with sufficiently large epistemic uncertainty as abnormal event.

4. EXPERIMENTS

4.1 Datasets

FineDiving Dataset.[20] FineDiving is a dataset proposed for human action assessment. It contains different diving activities that are composed of various primitive actions. In this project, we use it for complex human event recognition and abnormal event detection. It contains 3000 videos in 52 action types with 29 primitive action types. We use 75% of the data for training and 25% of the data for testing. Some example data samples are shown in Figure 3

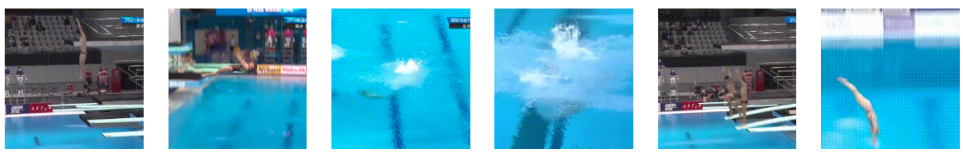

Figure 3. Data samples in FineDiving dataset.

Olympic Sports. This is a relative small dataset released in 2010, various methods have achieved near 100% accuracy on it. We just use this dataset to simply verify our methods. We get 98.9% accuracy on this dataset for complex action recognition. The decreasing of performance when we increase the complexity of 3D CNN backbone may due to the lack of training data, which causes the overfitting of our model. Some data samples are shown in Figure 4.

Figure 4. Data samples in Olympics Sports dataset.

4.2 Experiment Results

4.2.1 Complex event recognition

We evaluate our proposed probabilistic Transformer for complex human event recognition on FineDiving dataset. The experiment results are shown in Table 1. The experiments results show that our proposed probabilistic Transformer outperforms the deterministic Transformer, which demonstrates the effectiveness of our proposed method.

Table 1. Complex human event recognition results on FineDiving dataset

Method	Accuracy (%)
I3D	35.65
3D-ResNet-101	37.80
SlowFast	47.91
Timeception	51.62
Probabilistic Transformer (ours)	54.92

Table 2. Complex human event recognition results on Olympic Sports dataset

Method	Accuracy (%)
LHM[7]	83.2
Motion Atoms and Phrases[21]	84.9
MOF + MBH[22]	91.2
Probabilistic Transformer (ours)	98.6

4.2.2 Abnormal human event detection

For abnormal human event detection, we also evaluate our proposed uncertainty based method on FineDiving dataset. We use 45 action types as the training data and the remaining 7 classes of actions as abnormal events during the testing. We compare with the entropy-based deterministic Transformer. The experiment results are shown in Table 3. The experiment results show that by quantifying the epistemic uncertainty, our proposed probabilistic Transformer can better detect the abnormal human events than the entropy-based deterministic method.

Table 3. Abnormal human event detection results on FineDiving dataset

Method	Accuracy (%)
IT-AE	61.63
Late Fusion	71.14
AMDN	76.90
Stacked RNN	75.02
AbnormalGAN	83.98
Probabilistic Transformer - Uncertainty (ours)	82.15

Table 4. Abnormal human event detection results on Olympic Sports dataset

Method	Accuracy (%)
IT-AE	78.30
Late Fusion	80.67
AMDN	83.49
Stacked RNN	89.71
AbnormalGAN	95.07
Probabilistic Transformer - Uncertainty (ours)	93.15

4.3 Ablation Studies

Number of sampled models. To further study our proposed model, we evaluate our probabilistic Transformer with different sampling times. The experiment results are shown in Table 5. From the experiments results, the performance improves white we increase the number of model sampled. However, the computation cost also increase if we increase the number of sampling times.

Table 5. Experiment results for complex human event recognition with different sampling times. K denotes the number of models sampled. By increasing the number of sampled models, the performance can be improved.

Method	Accuracy (%)
Probabilistic Transformer ($K = 1$)	51.05
Probabilistic Transformer ($K = 2$)	52.79
Probabilistic Transformer ($K = 3$)	54.20
Probabilistic Transformer ($K = 4$)	54.35
Probabilistic Transformer ($K = 5$)	54.92

Number of heads. The multi-head attention mechanism is adopted for the probabilistic Transformer. To study the property of our proposed probabilistic Transformer, we vary the number of heads of the probabilistic Transformer. The comparison is shown in Table 6. From the results, using 5 attention heads gives the best performance.

Table 6. Comparison of performance using different number of heads. The results are reported on FineDiving dataset.

Number of heads	1	2	3	4	5	6	8	8
Accuracy (%)	42.19	47.04	50.20	52.75	54.92	54.80	54.38	54.49

5. CONCLUSION AND FUTURE WORK

In this paper, we proposed a probabilistic Transformer for complex human event recognition and abnormal human activity detection. We model the distribution of attention scores of the Transformer instead of using the scaled dot-product attention. We quantify the prediction uncertainty and use it as the metric for abnormal human activity detection. We evaluated the proposed method on FineDiving dataset. The experiment results demonstrate the effectiveness of our proposed method.

In the future, we may improve the proposed probabilistic Transformer in terms of accuracy and efficiency. In this work, we assume the attention scores follow Gaussian distribution, which is a reasonable but strong assumption. We may study and evaluate more probability distributions to better model the attention scores. And we may also evaluate the proposed method for large-scale datasets to further demonstrate its superiority over the state of the art methods.

ACKNOWLEDGMENTS

This project is supported, in part, by the U.S. Air Force Research Lab Summer Faculty Fellowship Program.

REFERENCES

[1] Poppe, R., "A survey on vision-based human action recognition," *Image and vision computing* **28**(6), 976–990 (2010).

[2] Yan, J., Yan, S., Zhao, L., Wang, Z., and Liang, Y., "Research on human-machine task collaboration based on action recognition," in [*2019 IEEE International Conference on Smart Manufacturing, Industrial & Logistics Engineering (SMILE)*], 117–121, IEEE (2019).

[3] Nguyen, T.-N. and Meunier, J., "Anomaly detection in video sequence with appearance-motion correspondence," in [*Proceedings of the IEEE/CVF international conference on computer vision*], 1273–1283 (2019).

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I., "Attention is all you need," *Advances in neural information processing systems* **30** (2017).

[5] Kuehne, H., Arslan, A., and Serre, T., "The language of actions: Recovering the syntax and semantics of goal-directed human activities," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 780–787 (2014).

[6] Tang, K., Fei-Fei, L., and Koller, D., "Learning latent temporal structure for complex event detection," in [*2012 IEEE Conference on Computer Vision and Pattern Recognition*], 1250–1257, IEEE (2012).

[7] Wang, L., Qiao, Y., and Tang, X., "Latent hierarchical model of temporal structure for complex activity classification," *IEEE Transactions on Image Processing* **23**(2), 810–822 (2013).

[8] Zhang, Y., Zhang, Y., Swears, E., Larios, N., Wang, Z., and Ji, Q., "Modeling temporal interactions with interval temporal bayesian networks for complex activity recognition," *IEEE transactions on pattern analysis and machine intelligence* **35**(10), 2468–2483 (2013).

[9] Liu, F., Xu, X., Zhang, T., Guo, K., and Wang, L., "Exploring privileged information from simple actions for complex action recognition," *Neurocomputing* **380**, 236–245 (2020).

[10] Hussein, N., Gavves, E., and Smeulders, A. W., "Videograph: Recognizing minutes-long human activities in videos," *arXiv preprint arXiv:1905.05143* (2019).

[11] Hussein, N., Gavves, E., and Smeulders, A. W., "Timeception for complex action recognition," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 254–263 (2019).

[12] Zhou, J., Lin, K.-Y., Li, H., and Zheng, W.-S., "Graph-based high-order relation modeling for long-term action recognition," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 8984–8993 (2021).

[13] Guo, H., Wang, H., and Ji, Q., "Uncertainty-guided probabilistic transformer for complex action recognition," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 20052–20061 (2022).

[14] Cheng, K.-W., Chen, Y.-T., and Fang, W.-H., "Video anomaly detection and localization using hierarchical feature representation and Gaussian process regression," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 2909–2917 (2015).

[15] Liu, W., Luo, W., Lian, D., and Gao, S., "Future frame prediction for anomaly detection–a new baseline," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 6536–6545 (2018).

[16] Zhong, J.-X., Li, N., Kong, W., Liu, S., Li, T. H., and Li, G., "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 1237–1246 (2019).

[17] Morais, R., Le, V., Tran, T., Saha, B., Mansour, M., and Venkatesh, S., "Learning regularity in skeleton trajectories for anomaly detection in videos," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 11996–12004 (2019).

[18] Ionescu, R. T., Khan, F. S., Georgescu, M.-I., and Shao, L., "Object-centric auto-encoders and dummy anomalies for abnormal event detection in video," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 7842–7851 (2019).

[19] Chen, M.-H., Li, B., Bao, Y., AlRegib, G., and Kira, Z., "Action segmentation with joint self-supervised temporal domain adaptation," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 9454–9463 (2020).

[20] Xu, J., Rao, Y., Yu, X., Chen, G., Zhou, J., and Lu, J., "Finediving: A fine-grained dataset for procedure-aware action quality assessment," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 2949–2958 (2022).

[21] Wang, L., Qiao, Y., and Tang, X., "Mining motion atoms and phrases for complex action recognition," in [*Proceedings of the IEEE international conference on computer vision*], 2680–2687 (2013).

[22] Wang, H. and Schmid, C., "Action recognition with improved trajectories," in [*Proceedings of the IEEE international conference on computer vision*], 3551–3558 (2013).