Probabilistic Temporal Modeling for Unintentional Action Localization

Jinglin Xu, Guangyi Chen, Nuoxing Zhou, Wei-Shi Zheng, and Jiwen Lu, *Senior Member, IEEE*

***Abstract*— Humans have the inherent advantage of understanding action intention, while it is an enormous challenge to train the machine to localize unintentional action in videos due to the lack of reliable annotations for stable training. The annotations of unintentional action are unreliable since different annotators are affected by their subjective appraisals and intrinsic ambiguity, which brings heavy difficulties for the training. To address this issue, we propose a probabilistic framework for unintentional action localization by modeling the uncertainty of annotations. Our framework consists of two main components, including Temporal Label Aggregation (TLA) and Dense Probabilistic Localization (DPL). We first formulate each annotated failure moment as a temporal label distribution. Then we propose a TLA component to aggregate temporal label distributions of different failure moments in an online manner and generate dense probabilistic supervision. Based on TLA, We further develop a DPL component to jointly train three heads (i.e., probabilistic dense classification, probabilistic temporal detection, and probabilistic regression) with different supervision granularities and make them highly collaborative. We evaluate our approach on the largest unintentional action dataset OOPS and demonstrate that our approach can achieve significant improvement over the baseline and state-of-the-art methods.**

***Index Terms*— Probabilistic modeling, action localization, attention model, action intention.**

I. Introduction

EXISTING video understanding techniques have answered many aspects of human action, including what the action content is (i.e., action recognition [1]–[3]), when or where the action occurs (i.e., action localization [4]–[6]), and how well an action is performed (i.e., action quality assessment [7]). However, when an unintentional (failure) action occurs in a video, these methods cannot explain why the action fails. It needs the machine to understand the action intention and localize when the action turns from intentional into unintentional. Understanding the underlying intention of the observed action is of paramount importance for intelligent systems to avoid risks and make decisions, especially in the fields of automatic driving, intelligent robotics, medical service, and public safety.

Psychological researches [8]–[10] show that human has the inherent advantage of understanding action intention, which has been reflected in infancy. However, it is an enormous challenge to train the machine to understand the intention of observed actions, due to the lack of both referable comparisons and reliable annotations. In recent work [11], Epstein *et al.* have collected an annotated video dataset OOPS with various unintentional actions, which annotates each video with multiple timestamps of action transition from intentional to unintentional. With such data, one can train the model to localize unintentional action by classifying the given video clips (or frames) to be intentional or unintentional. For instance, PUAV [11] trains the model as a three-way classifier (intentional, transitional, and unintentional), and utilizes this classifier in a sliding window fashion over the temporal axis to infer whether the action in each temporal location transits from intentional to unintentional.

Despite abundant annotated video datasets, temporal annotations of unintentional action are unreliable due to different annotators being affected by their subjective appraisals and intrinsic ambiguity. As shown in Fig. 1, three different annotators might give three timestamps $\{y_k\}_{k=1}^{3}$ of the action transition for the same unintentional video. Taking a video "a boy falling into the water when throwing a fishing net" as an example, three timestamps respectively are the beginning of throwing (y_1), the moments of turning around (y_2), and falling down (y_3). The uncertainty of the above annotations heavily confuses the optimization direction during the model training. One possible solution is to provide more annotations and calculate the statistical expectation to alleviate the negative effect brought by inherent uncertainty. However, this solution is labor-intensive and expensive.

To address this issue, we propose a probabilistic temporal modeling framework that probabilizes rigid annotations and generates temporal label distributions via Probabilistic Annotation Modeling (PAM). Then, we refine temporal label distributions to obtain reliable supervision via Temporal Label Aggregation (TLA) and further propose Dense Probabilistic Localization (DPL) to utilize this refined temporal label distribution as a supervision signal for training. Specifically, PAM applies a probability distribution (i.e. a Gaussian distribution or

Manuscript received July 26, 2021; revised January 2, 2022 and February 15, 2022; accepted March 18, 2022. Date of publication April 7, 2022; date of current version April 14, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62106124, Grant 62125603, and Grant U1813218; in part by the China Postdoctoral Science Foundation under Grant 2020M680564; and in part by a grant from the Beijing Academy of Artificial Intelligence (BAAI). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ajmal S. Mian. *(Jinglin Xu and Guangyi Chen contributed equally to this work.) (Corresponding author: Jiwen Lu.)*

Jinglin Xu, Guangyi Chen, Nuoxing Zhou, and Jiwen Lu are with the Beijing National Research Center for Information Science and Technology (BNRist) and the Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: xujinglinlove@gmail.com; guangyichen1994@gmail.com; nuoxingzhou@gmail.com; lujiwen@tsinghua.edu.cn).

Wei-Shi Zheng is with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510275, China (e-mail: wszheng@ieee.org).

Digital Object Identifier 10.1109/TIP.2022.3163544

a Laplace distribution) to model the fixed temporal location, which constructs the uncertainty of annotations. The probability of each temporal location in the distribution denotes the possibility of this temporal location occurring unintentional action. Beyond considering the annotation uncertainty, PAM constructs the graduality from the intentional action to unintentional action via approximating the hard signum function (i.e., hard labels) to a sigmoid function. Based on PAM, TLA estimates the reliabilities of different temporal label distributions via a temporal label attention model, and then online aggregates them to generate a final temporal label distribution. The generated final distribution (named dense probabilistic supervision) is used as the supervision signal to train our proposed model. The advantages of TLA are to mine reliable supervision and mitigate the negative effect of noisy annotations. Then, our proposed DPL utilizes dense probabilistic supervision to train the model to localize unintentional action. DPL contains three probabilistic heads with different supervision granularities, respectively, including probabilistic dense classification (Pdc), probabilistic temporal detection (Ptd), and probabilistic regression (Pr).

Specifically, Pdc predicts the probability of each frame occurring the action transition from intentional to unintentional. As a fine-grained supervision manner, Pdc is optimized by calculating cross-entropy loss between predicted probabilities and dense probabilistic supervision. Ptd predicts the temporal boundary of unintentional action, where temporal boundary indicates the temporal locations of the start and the end of unintentional action. As a middle granularity supervision manner, Ptd is optimized by a probabilistic IoU between the predicted and ground truth bounding boxes. Supposed that the predicted bounding box is an indicator function from the predicted temporal location to the end of the video sequence, conventional temporal IoU is calculated by the predicted bounding box and another indicator function from the ground truth location to the end of the video sequence. While probabilistic IoU is computed by the predicted bounding box and the area under the cumulative distribution function of dense probabilistic supervision. Pr predicts a temporal location when the action transits from intentional to unintentional. Different from optimizing the conventional regression with the MSE between ground truth and prediction, Pr calculates the cumulative probability differences in the range of ground truth and predicted temporal locations as the residual term. These three probabilistic heads are highly collaborative via jointly optimizing the framework using supervision signals from fine-grained to coarse granularities.

The contributions of our approach are summarized as:

1) We present a probabilistic framework for unintentional action localization, composed of temporal label aggregation and dense probabilistic localization, providing a new probabilistic perspective to understand human intention.
2) We propose temporal label aggregation to explicitly model the uncertainty of rigid annotations, and mine reliable dense probabilistic supervision by online reliability estimation and temporal label attention model.
3) We design dense probabilistic localization containing three probabilistic heads, which are jointly trained by dense probabilistic supervision in different granularity manners and are highly collaborative.
4) Our approach is evaluated on the largest unintentional action dataset OOPS for both tasks of unintentional action recognition and localization, and significantly outperforms the baseline and the state-of-the-arts.

It is worth mentioning that we have developed a preliminary work [12] named Temporal Label Aggregation for Unintentional Action Localization (UAL-TLA). As an extension, our approach explores how to make full use of dense probabilistic supervision and designs three probabilistic heads, including probabilistic dense classification, probabilistic temporal detection, and probabilistic regression. Our approach jointly trains three probabilistic heads by dense probabilistic supervision in different granularities and further boosts the performance of unintentional action localization.

II. Related Work

In this section, we briefly review related researches, including temporal action localization, unintentional action localization, anomaly detection, and label distribution learning.

A. Temporal Action Localization

Temporal action localization aims to recognize the action category and locates its beginning and end timestamps in the untrimmed video. Compared to action recognition [1], [3], [13]–[20], the challenge of action localization comes from the dramatic changes of the video duration and action instance. Inspired by object detection techniques, the existing anchor-based temporal action localization methods [6], [21]–[26] utilize multiple scales sliding windows to extract the temporal action proposals and identify whether it is an action segment. Other categories of detection methods are also introduced, such as sequential decision-making [27] and single-shot detectors [28]. In addition, some methods apply activity completeness to localize the action temporally, such as structured segment network [24], structured maximal sums [29], modeling sub-actions [30], modeling action dependencies [31], Gaussian temporal awareness networks [32], deep cross-modal hashing [33], iterative-winners-out network [34], multi-scale structure-aware network [35], action unit memory network [36], and combining varying levels of supervision [37]. Besides, the probability distribution curve-based methods [5], [38]–[40] calculate the probability of action for each fixed-length video segment, analyze the probability distribution curves, and extract a high score segment as the final result. However, temporal action localization only tells us the action category and its temporal boundary but cannot recognize the action intentionality variation to explain the reason why the action fails.

B. Anomaly Detection

A common need when analyzing real-world datasets is determining which instances stand out as being dissimilar to all others. Such instances are known as anomalies that may arise from malicious actions, system failures, and intentional

fraud. The goal of anomaly detection is to determine all such instances in a data-driven fashion [41]–[46]. In recent years, deep learning-based anomaly detection methods have become increasingly popular, such as constructing deep generic knowledge [47], designing stacked recurrent neural network [48], cascaded deep network [49], self-training [50], self-supervised learning [51], [52] and plug-and-play CNNs [53], [54]. These deep learning-based methods have been widely applied for a diverse set of tasks, e.g., video surveillance and image analysis for illegal traffic detection [55], health-care for detecting retinal damage [56], networks for cyber-intrusion detection [57], and sensor networks for internet of things big-data anomaly detection [58]. The most challenging problem for anomaly detection is that the solution of some specific anomalous events [59] cannot be generalized to detect other anomalous events due to the diversity and complexity of anomalous events. Differently, the difficulty of unintentional action localization is to learn the knowledge of human action intention and localize unintentional action when only observing the video containing unintentional failures. More specifically, anomaly detection focuses on abnormal behavior patterns, e.g., "retrograde", "fight", and "steal", while unintentional action localization focuses on unintentional failures, e.g., "fall down" and "slip off". Many abnormal behaviors are intentional.

C. Unintentional Action Localization

Different from existing action localization, unintentional action localization (UAL) aims at understanding the intention behind the action and localizing when an intentional action turns into unintentional action. To understand the intention, Epstein *et al.* [11] collect an annotated video dataset and train a three-way classifier to recognize the action as intentional, unintentional, or transitional. It localizes the unintentional action by applying the classifier in a sliding window fashion over the temporal axis and exploring the location with the most confident score. Furthermore, the goals of original intentional action are labeled to improve the quality of the supervision and train the more discriminative video representations [60]. However, the optimizations of these methods are easy to be misled by unreliable rigid annotations due to the intrinsic ambiguity from multiple annotators and their subjective appraisals. To address this problem, we propose to formulate the unintentional action localization as a probabilistic framework. It models the uncertainties of original rigid annotations to mine reliable dense probabilistic supervision and learns to localize unintentional action via developing three probabilistic heads that respectively represent different supervision granularities. The work [61] presented during the same period also proposes to assist action recognition via understanding the causal relationships of "precondition", "action", and "effect", while UAL is a new task (not action recognition) that focuses on the effect of human intention.

D. Label Distribution Learning

Label distribution learning [62]–[66] aims to solve the uncertainty of annotations by replacing a hard label with a probability distribution, which has obtained great success for facial age estimation. For example, Geng *et al.* [62] first proposes to apply an age distribution as the supervision instead of a fixed age label, and extend it into deep learning framework [63], [67]. Recently, label distribution learning has been widely used in different computer vision tasks such as facial landmark detection [68], facial expression recognition [69], pose estimation [70], action quality assessment [71] and crowd counting [72], and demonstrates the effectiveness by mitigating the overfitting of unreliable annotations. In this paper, we apply the label distribution learning method for modeling the temporal location of the action transitioning from intentional to unintentional in a video. We further propose to model the uncertainty of annotations and construct reliable supervision, and then learn to localize unintentional action with dense probabilistic supervision in different supervision granularities via developing three probabilistic heads.

III. Approach

In this section, we introduce the probabilistic framework for unintentional action localization. We first define the task of unintentional action localization mathematically and model each failure moment as a temporal label distribution via Probabilistic Annotation Modeling. We propose Temporal Label Aggregation that online aggregates different temporal label distributions to construct dense probabilistic supervision. We further develop Dense Probabilistic Localization consisting of three probabilistic heads, which utilize different supervision granularities to jointly guide unintentional action localization. Finally, we elaborate on the network architecture of our approach and discuss the differences between the proposed probabilistic framework and conventional methods.

A. Problem Definition

Let $\boldsymbol{X} = \{\boldsymbol{x}_t\}_{t=1}^{T}$ be an input video containing unintentional action, where $\boldsymbol{x}_t$ denotes the t-th video frame. Our goal is to predict the temporal location of the action transition from intentional to unintentional, to localize the beginning of unintentional action. We formulate the task of unintentional action localization as a regression problem that predicts temporal location $\hat{y}$ as:

$$\hat{y} = \mathcal{F}_\theta(\boldsymbol{X}), \tag{1}$$

where $\mathcal{F}_\theta$ denotes a learnable probabilistic framework whose input is a video $\boldsymbol{X}$ and output is the predicted temporal location of the transition. The annotations of $\boldsymbol{X}$ are $Y = \{y_k\}_{k=1}^{K}$, provided by K annotators. Each $y_k \in \mathcal{R}$ denotes a failure timestamp that is the moment of action transition from intentional to unintentional. For convenience, we sort the hard labels in Y as $y_1 < \cdots < y_K$. However, these hard annotations are unreliable since different annotators being affected by subjective appraisals have different perceptions of unintentional action occurrence.

B. Probabilistic Annotation Modeling

It is a challenge to generate reliable supervision from unreliable rigid annotations. Conventional unintentional action

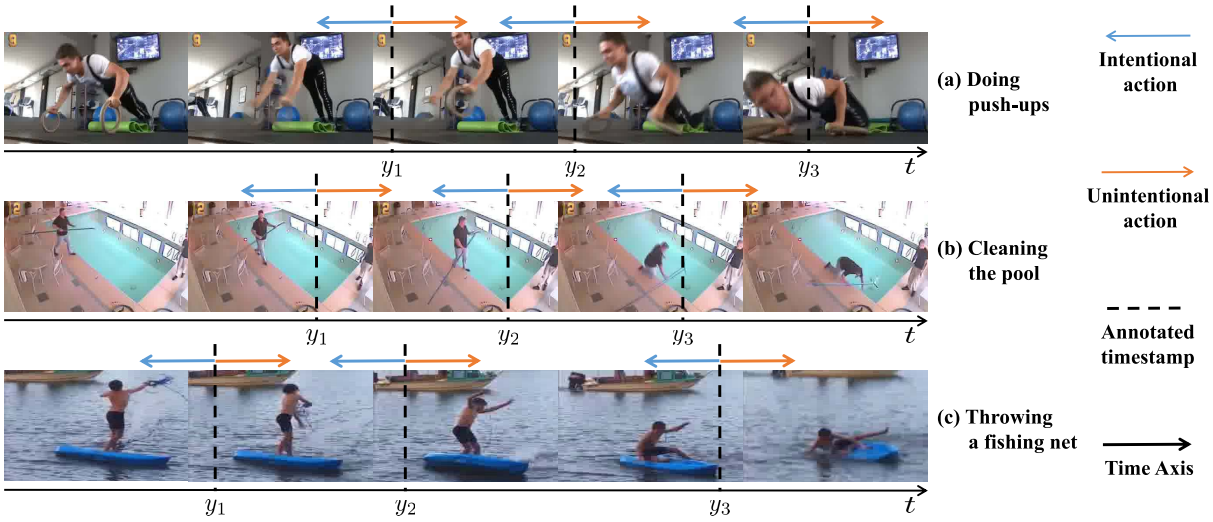


Fig. 1. What are intentional and unintentional actions? The intentional action denotes the person intends for this action to occur, while the unintentional action is always accompanied by the failure or accident. We show some examples of unintentional action in the OOPS dataset including falling when (a) doing push-ups, slipping while (b) cleaning the pool, and falling into the water when (c) throwing a fishing net. We show the ambiguity of temporal locations of unintentional action occurring brought by subjective appraisals of different annotators. In (a)-(c), there are three timestamps $\{y_k\}_{k=1}^{3}$ for each video, where each y_k provided by the k-th annotator denotes the action transition from intentional to unintentional.

location methods directly train the model using unreliable rigid annotations, which brings the negative effect on generalization performance. In this work, we model the uncertainty of rigid annotations and replace them with probabilistic distributions, that is modeling each failure moment as a temporal label distribution. As shown in Fig. 2, we soften each hard label (from signum function to sigmoid function) along the temporal axis and further take the derivative to model each failure moment as a temporal label distribution. This process is named Probabilistic Annotation Modeling.

Given a failure timestamp $y_k \in Y$, we model a hard label into a unimodal distribution (e.g., Gaussian distribution) as:

$$p_{y_k}(t) \sim \gamma \mathcal{N}(\mu = y_k, \sigma^2), \tag{2}$$

where t denotes the temporal location; μ indicates the failure timestamp y_k (raw annotation); σ denotes the degree of deviation; $\gamma = \sqrt{2\pi}\sigma$ is a normalization scalar which makes $p_{y_k}(y_k) = 1$. We use $p_{y_k}(t)$ to represent the probability of being the action transition from intentional to unintentional at the temporal location t. $p_{y_k}(t)$ is higher when the temporal location t is close to the failure timestamp y_k, while the $p_{y_k}(t)$ is lower when t is far away from y_k. The temporal label distribution p_{y_k} becomes sharper with the reduction of σ, while it will degenerate into an one-hot vector when the σ reduces to an extreme. To distinguish the temporal label distribution and the value of the temporal location t of this distribution, we represent the former as p_{y_k} and the latter as $p_{y_k}(t)$.

As shown in Probabilistic Annotation Modeling of Fig. 2, we show the process of constructing temporal label distribution. We first describe the graduality of temporal switch from intentional to unintentional actions via approximating the hard signum function to a sigmoid function, then calculate the derivation of sigmoid to construct temporal label distribution. Besides, we only consider the $p_{y_k}(t)$ in the range of $[1, T]$, where T denotes the length of a video. Though the normalization scalar γ is introduced leading to the integral of p_k not equal to 1, this doesn't affect the optimization during the training. Here, we can replace the Gaussian distribution with any unimodal symmetric distribution, e.g. Laplace distribution.

C. Temporal Label Aggregation

Despite the uncertainty of rigid annotation has been modeled as a temporal label distribution, how to mine a reliable supervision signal from unreliable rigid annotations to guide the model training is still difficult. To construct a reliable temporal label distribution, we propose a temporal label attention model that estimates the reliability α_k of each temporal label distribution p_{y_k} and online aggregates multiple distributions $\{p_{y_k}\}_{k=1}^{K}$ to generate dense probabilistic supervision, as shown in Temporal Label Aggregation of Fig. 2. Different from the conventional self-attention model, our temporal label attention model is online, indicating that the learned reliabilities $\{\alpha_k\}_{k=1}^{K}$ are dependent on the current model predictions.

Given a video $\boldsymbol{X}$ and its labels Y, our approach predicts the temporal location $\hat{y}$ of the action transition and estimates the reliability α_k by calculating the distance between the predicted temporal location $\hat{y}$ and ground truth temporal location y_k,

$$\alpha_k = \Phi(|\hat{y} - y_k|), \tag{3}$$

where Φ is a function that explores the reliabilities $\{\alpha_k\}_{k=1}^{K}$ for different $\{p_{y_k}\}_{k=1}^{K}$ and it is defined as,

$$\Phi(|\hat{y} - y_k|) = \begin{cases} 1, & |\hat{y} - y_k| \geq \xi \\ 2 - \dfrac{|\hat{y} - y_k|}{\xi}, & \text{others}, \end{cases} \tag{4}$$

where ξ indicates a threshold describing the absolute error.

With such a temporal label attention model, α_k becomes larger when $\hat{y}$ is close to y_k. We online aggregate $\{p_{y_k}\}_{k=1}^{K}$ by attention model to construct dense probabilistic

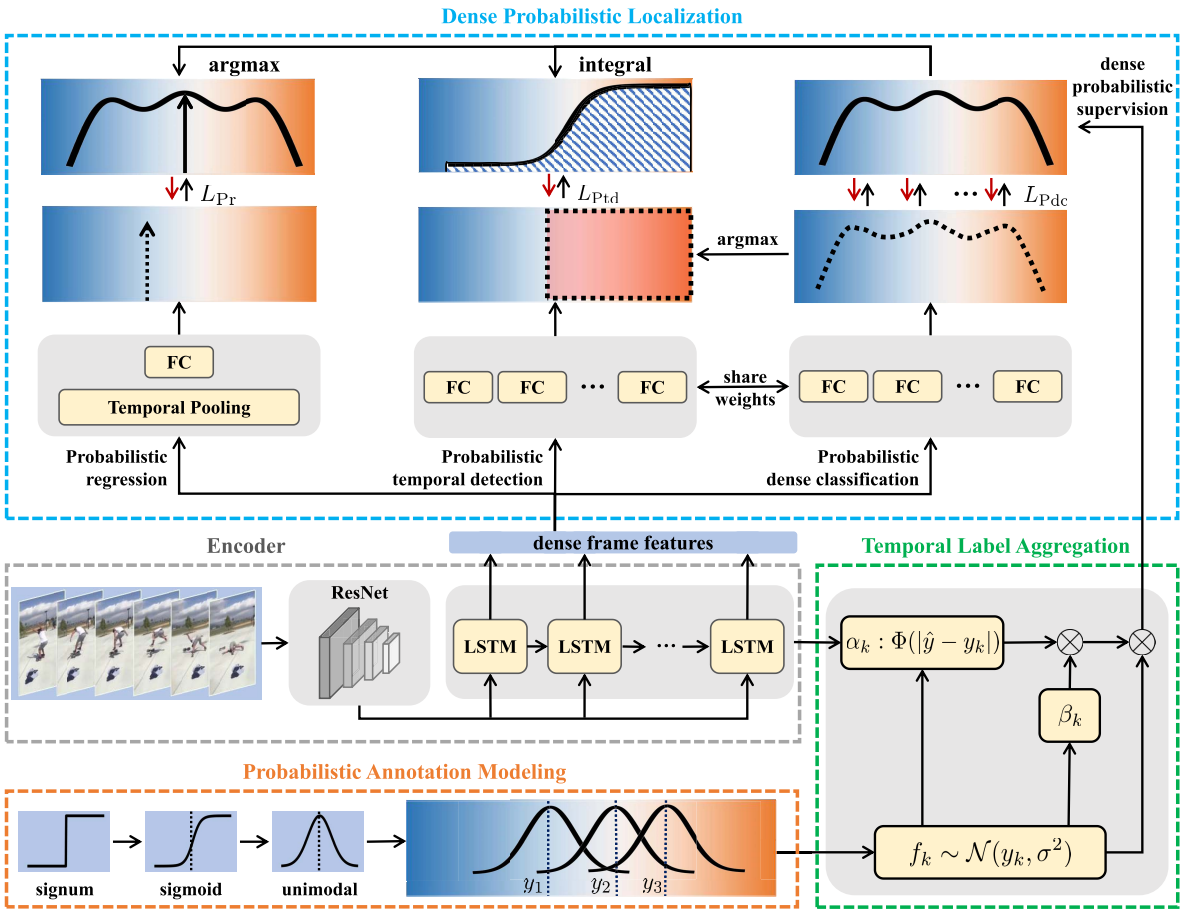


Fig. 2. The architecture of the proposed method. Encoder extracts spatial-temporal features via ResNet combined with a standard LSTM. Probabilistic Annotation Modeling constructs hard labels as temporal label distributions. Then, Temporal Label Aggregation uses an attention model to aggregate multiple temporal label distributions and generate a unified dense probabilistic supervision. Finally, Dense Probabilistic Localization is proposed to utilize dense probabilistic supervision to optimize the framework composed of three heads, i.e., Probabilistic dense classification, Probabilistic temporal detection, and Probabilistic regression. "signum→sigmoid→unimodal" indicates the process of constructing the graduality and uncertainty of annotations. The red arrows indicate the effects of supervision signals, where the red color means the feedback direction.

supervision p_y, which can be formulated as,

$$p_y = \sum_{k=1}^{K} \frac{\alpha_k \beta_k}{\sum_{s=1}^{K} \alpha_s \beta_s} p_{y_k}, \tag{5}$$

where $\frac{\alpha_k \beta_k}{\sum_{s=1}^{K} \alpha_s \beta_s}$ is the posterior weight of p_{y_k} modified by the model prediction. β_k is the prior weight of each p_{y_k}'s reliability, where $\{\beta_k\}_{k=1}^{3}$ is initialized as $\{1/4, 2/4, 1/4\}$ in our experiments since the timestamp y_2 located in the middle of $[y_1, y_3]$ is more likely to be the transition. In our approach, we define p_y as dense probabilistic supervision to train the probabilistic framework that is also inversely applied to learn p_y, which collaboratively optimizes TLA and DPL.

Based on PAM and TLA, we construct dense probabilistic supervision (i.e., a generated temporal label distribution) to model multiple rigid annotations. This label distribution can be understood as a Gaussian mixture distribution which is mixed by multiple sub-distributions. Note that PAM+TLA is slightly different from the Gaussian mixture model (Gmm). TLA aggregates different temporal label distributions by an attention model to generate dense probabilistic supervision, which is an active algorithm for aggregating different temporal label distributions, while Gmm is a formulation of data distribution that can be used for data clustering.

D. Dense Probabilistic Localization

To make full use of dense probabilistic supervision generated by the TLA model, we propose Dense Probabilistic Localization (DPL), which contains three heads, i.e., Probabilistic dense classification (Pdc), probabilistic temporal detection (Ptd), and probabilistic regression (Pr), trained by three supervision granularities.

1) Probabilistic Dense Classification: We predict the probability of occurring action transition at the t-th frame by computing

$$[\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_T] = \mathrm{LSTM}(\boldsymbol{X}) \tag{6}$$

$$\hat{p}_y(t) = \boldsymbol{W}_2 \cdot \mathrm{ReLU}(\boldsymbol{W}_1 \boldsymbol{h}_t), \tag{7}$$

where $\boldsymbol{h}_t \in \mathbb{R}^{d_h}$ is the hidden states of LSTM at the t-th frame; $\hat{p}_y$ is the predicted temporal label distribution where each $\hat{p}_y(t)$ denotes the predicted probability of the t-th frame being the transition from intentional to unintentional, $\hat{y} = \arg\max_t \hat{p}_y(t)$; $\boldsymbol{W}_1 \in \mathbb{R}^{d \times d_h}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{C \times d}$ denote the weight parameters of the network, where $C = 2$ corresponds to the number of classes, i.e., transition or non transition.

With the prediction $\hat{p}_y$ and dense probabilistic supervision p_y, our probabilistic dense classification head predicts the probability of each frame occurring action transition. As a

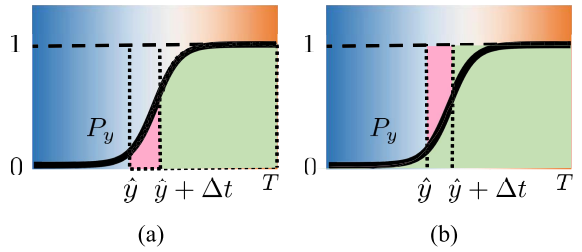


Fig. 3. The descriptions of PIoU. The pink area in (a) shows the change of the intersection when the prediction $\hat{y}$ moves to $\hat{y}+\Delta t$. The pink area in (b) shows the change of the union when the prediction $\hat{y}$ moves to $\hat{y}+\Delta t$.

fine-grained supervision manner, we calculate cross-entropy loss [73] between $\hat{p_y}$ and p_y to optimize the network and find the temporal location with the greatest probability of occurring the transition. The objective function of Pdc is,

$$\mathcal{L}_{\text{Pdc}}(\hat{p_y}, p_y) = -\sum_{t=1}^{T} p_y(t) \log \hat{p_y}(t). \tag{8}$$

2) Probabilistic Temporal Detection: To directly reflect a temporal segment localization quality, we further propose a probabilistic IoU (PIoU) between the predicted and ground truth bounding boxes. Specifically, the predicted bounding box is an indicator function from the beginning to the end temporal locations of unintentional action, where the beginning temporal location is the predicted temporal location $\hat{y}$ and the end temporal location is the last frame of the input video. The ground truth bounding box is defined as the area under the cumulative distribution function P_y of dense probabilistic supervision p_y in the range of $[1, T]$. This is a middle granularity supervision manner.

Since $\hat{y} = \arg\max_t \hat{p_y}(t)$ cannot be directly optimized during the training, we introduce the policy gradient to implement the back-propagation to optimize the network. The objective function of Ptd is formulated as,

$$\mathcal{L}_{\text{Ptd}} = -\mathcal{R}_w \log \hat{p_y}(t = \hat{y}), \tag{9}$$

where $\mathcal{R}_w$ denotes the reward defined in the policy gradient and is computed by,

$$\mathcal{R}_w = (\text{PIoU}(\hat{y}) - \text{PIoU}(\hat{y} + \Delta t))/\Delta t, \tag{10}$$

where Δt is the offset; $\text{PIoU}(\hat{y})$ and $\text{PIoU}(\hat{y}+\Delta t)$ are calculated by,

$$\text{PIoU}(\hat{y}) = \frac{\int_{\hat{y}}^{T} P_y(t)dt}{\int_{0}^{\hat{y}} P_y(t)dt + (T - \hat{y})} \tag{11}$$

$$\text{PIoU}(\hat{y} + \Delta t) = \frac{\int_{\hat{y}+\Delta t}^{T} P_y(t)dt}{\int_{0}^{\hat{y}+\Delta t} P_y(t)dt + (T - \hat{y} - \Delta t)}, \tag{12}$$

As shown in Fig. 3, the sum of green and pink areas in (a) indicates the intersection between the predicted bounding box for $\hat{y}$ and ground truth bounding box; the green area denotes the intersection between the predicted bounding box for $\hat{y}+\Delta t$ and ground truth bounding box; the change of the intersection when the prediction $\hat{y}$ moves to $\hat{y}+\Delta t$ is calculated by the numerator of $\text{PIoU}(\hat{y})$ minus that of $\text{PIoU}(\hat{y} + \Delta t)$. In (b), the sum of green and pink areas indicates the union between the predicted bounding box for $\hat{y}$ and ground truth bounding box; the green area denotes the union between the predicted bounding box for $\hat{y}+\Delta t$ and ground truth bounding box; the change of the union when the prediction $\hat{y}$ moves to $\hat{y}+\Delta t$ is computed by the denominator of $\text{PIoU}(\hat{y})$ minus that of $\text{PIoU}(\hat{y} + \Delta t)$.

The objective function encourages finding the predicted temporal location with higher PIoU. In Fig. 4 (a), supposed that y and $\hat{y}$ are the ground truth and predicted temporal locations, respectively, the conventional punishment for the prediction $\hat{y}$ only depends on the residual $|\hat{y} - y|$, i.e., the pink area in Fig. 4(a). The dynamic punishment for the prediction $\hat{y}$ in Fig. 4 (b) is calculated by the sum of cumulative probabilities in the range of $[y, \hat{y}]$, i.e., $\sum_{t=y}^{\hat{y}} P_y(t)$, which is equivalent to adding a weight on the residual $|\hat{y} - y|$. If $\hat{y}$ is far away from y, the weight becomes bigger and very closed to 1; if $\hat{y}$ is closed to y, the weight becomes smaller than 1. The dynamic punishment is better than the conventional punishment since the former considers the uncertainty of action transition.

3) Probabilistic Regression: The Probabilistic regression (Pr) utilizes global supervision to optimize the probabilistic framework. We predict a temporal location $\hat{y}_r$ by computing

$$\bar{\boldsymbol{h}} = \sum_{t=1}^{T} \lambda_t \boldsymbol{h}_t \tag{13}$$

$$\hat{y}_r = \boldsymbol{W}_3 \cdot \text{ReLU}(\boldsymbol{W}_1 \bar{\boldsymbol{h}}), \tag{14}$$

where $\bar{\boldsymbol{h}} \in \mathbb{R}^{d_h}$ is computed by a weighted average pooling on the hidden states of T video frames, which aggregates all video frame clues for regression. $\boldsymbol{W}_3 \in \mathbb{R}^{1 \times d_h}$ denotes the weight parameters of the network. Note that there are two predicted temporal locations in DPL, including $\hat{y}$ for the Ptd head and $\hat{y}_r$ for the Pr head. Both $\hat{y}$ and $\hat{y}_r$ are used for training, and $\hat{y}$ is used for testing.

With the prediction $\hat{y}_r$ and dense probabilistic supervision p_y, the objective function of Pr is defined as,

$$\mathcal{L}_{\text{Pr}} = \lambda_{\hat{y}_r} |\hat{y}_r - y| \tag{15}$$

where $y = \arg\max_t p_y$ and $\lambda_{\hat{y}_r}$ is a weight of dynamic punishment on the residual $|\hat{y}_r - y|$. $\lambda_{\hat{y}_r}$ is the key difference between our probabilistic regression and conventional regression, which is computed by,

$$\lambda_{\hat{y}_r} = 1 - \frac{1}{|\hat{y}_r - y|} \sum_{t=\min(\hat{y}_r, y)}^{\max(\hat{y}_r, y)} p_y(t). \tag{16}$$

It can be seen that the dynamic punishment in Fig. 4 (d) is more reasonable than conventional punishment in Fig. 4 (c). The conventional punishment for the prediction $\hat{y}_r$ in (c) only depends on the residual $|\hat{y}_r - y|$, i.e., the yellow area in (c). The dynamic punishment for the prediction $\hat{y}_r$ in (d) is calculated by equation (16), i.e., the yellow dashed shadow S_{EAB} in (d), which is equivalent to adding a weight on the residual $|\hat{y}_r - y|$. If $\hat{y}_r$ is far away from y, S_{EAB} becomes bigger; if $\hat{y}$ is closed to y, S_{EAB} becomes smaller. One can imagine an extreme situation that the ground-truth distribution p_y is almost a line parallel to the x-axis (similar to a uniform distribution), which

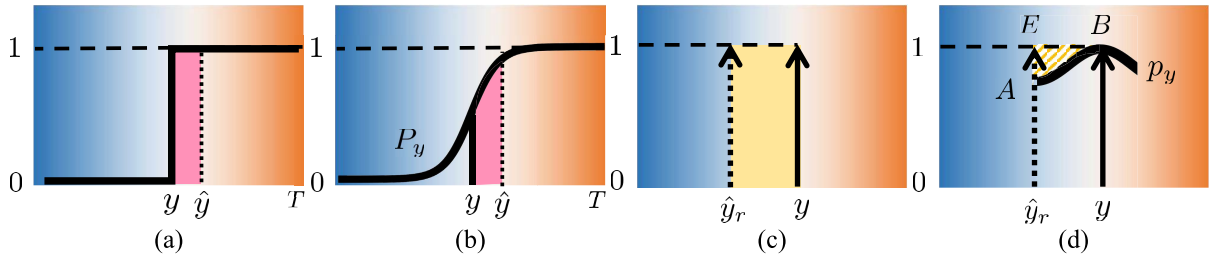


Fig. 4. The benefits of probabilistic temporal detection (Ptd) and probabilistic regression (Pr). (a) and (b) display the difference between conventional temporal detection and Ptd. The dynamic punishment in Ptd (i.e., the pink area in (b)) is better than the conventional punishment in conventional temporal detection (i.e., the pink area in (a)) since Ptd considers the uncertainty of action transition. (c) and (d) show the difference between conventional regression and Pr. The dynamic punishment in Pr (i.e., the yellow dashed shadow in (d)) is more reasonable than the conventional punishment in conventional regression (i.e., the yellow area in (c)) because Pr considers the uncertainty of rigid annotations. Best viewed in color.

denotes that annotators also do not know the temporal location of occurring unintentional action. The conventional regression still punishes the model when the prediction $\hat{y}_r$ is away from the ground truth y obtained by noise, while Pr does not punish the model with any prediction $\hat{y}_r$. The latter is more reasonable since the sample annotated by the noisy ground truth y is useless.

E. Network Architecture

In this subsection, we introduce our network architecture consisting of four parts: Encoder, Probabilistic Annotation Modeling (PAM), Temporal Label Aggregation (TLA), and Dense Probabilistic Localization (DPL), where DPL contains three heads including Probabilistic dense classification (Pdc), Probabilistic temporal detection (Ptd), and Probabilistic regression (Pr). In Fig. 2, Encoder consists of R3D and LSTM, and PAM does not require a network.

- **Encoder.** We extract spatial-temporal visual features of video frames via R3D [15] network that is a widely used video backbone with competitive performance on the action recognition tasks. Based on the R3D features for a video sequence, we learn an LSTM predictor that exploits the current frame feature and the last cell hidden state to generate hidden representations of a video sequence.
- **Probabilistic Annotation Modeling.** The input of PAM is $\{y_k\}_{k=1}^K$ provided by the OOPS dataset. PAM models each y_k as a temporal label distribution p_k via softening the hard label (from signum function to sigmoid function) along the temporal axis and taking derivative to generate temporal label distribution to replace rigid annotations.
- **Temporal Label Aggregation.** The input of TLA is composed of K temporal label distributions $\{p_{y_k}\}_{k=1}^K$ and the predicted temporal location $\hat{y}$ (maximum of Pdc's output). TLA calculates the correlations between $\hat{y}$ and raw rigid annotations (i.e., $\{y_k\}_{k=1}^K$), for updating the reliabilities of $\{p_{y_k}\}_{k=1}^K$. TLA online aggregates $\{p_{y_k}\}_{k=1}^K$ via learnable reliabilities $\{\alpha_k\}_{k=1}^K$ to generate dense probabilistic supervision p_y. This aggregation process is online updated by using current outputs.
- **Dense Probabilistic Localization.**
  (1) **Probabilistic dense classification.** The input of Pdc is the hidden states of LSTM for video sequence $[\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_T]$ and the output is the probabilities of video frames being the transition from intentional to unintentional. A fully connected ReLU network with one hidden layer is trained to predict $\hat{p}_y$ from $[\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_T]$ by minimizing cross-entropy loss between $\hat{p}_y$ and p_y. Pdc not only couples with TLA to generate dense probabilistic supervision p_y but also formulates a dense probabilistic classification problem guided by fine-grained supervision.
  (2) **Probabilistic temporal detection.** The input of Ptd is the hidden states $[\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_T]$ and the output is a predicted temporal location. Ptd shares the same $\hat{p}_y$ with Pdc and gets $\hat{y} = \arg\max_t \hat{p}_y(t)$. The predicted bounding box can be seen as an indicator function from $\hat{y}$ to T, supervised by the ground truth bounding box that is the area under the cumulative distribution function P_y of p_y in the range of $[1, T]$. The supervision granularity of Ptd is coarser than that of Pdc but finer than that of Pr.
  (3) **Probabilistic regression.** The input of Pr is the hidden states $[\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_T]$ and the output is a predicted temporal location $\hat{y}_r$. $\bar{\boldsymbol{h}}$ is calculated by feeding the hidden states into a temporal pooling layer. A fully connected ReLU network with one hidden layer is trained on $\bar{\boldsymbol{h}}$ by minimizing the weighted cumulative probability differences in the range of $[\hat{y}_r, y]$. Pr improves the reasonability of regression residual.

During the training, we optimize the probabilistic framework by the following loss function:

$$\mathcal{L} = \mathcal{L}_{\text{Pdc}} + \lambda_1 \mathcal{L}_{\text{Ptd}} + \lambda_2 \mathcal{L}_{\text{Pr}} \tag{17}$$

where L_{Pdc}, L_{Ptd}, and L_{Pr} are the losses for Pdc, Ptd, and Pr, respectively. λ_1 and λ_2 are the trade-off parameters among different losses.

F. Discussion

In this subsection, we discuss the differences between conventional methods and our DPL. The conceptual comparisons are shown in Fig. 5. The top row shows conventional classification, regression, and detection methods while the bottom row shows the proposed probabilistic dense classification, probabilistic temporal detection, and probabilistic regression.

- ***Probabilistic dense classification head*** predicts the probability of the action transiting from intentional to unintentional at the current temporal location. As shown in Fig. 5 (a), conventional classification is supervised by rigid annotations, while our approach models rigid

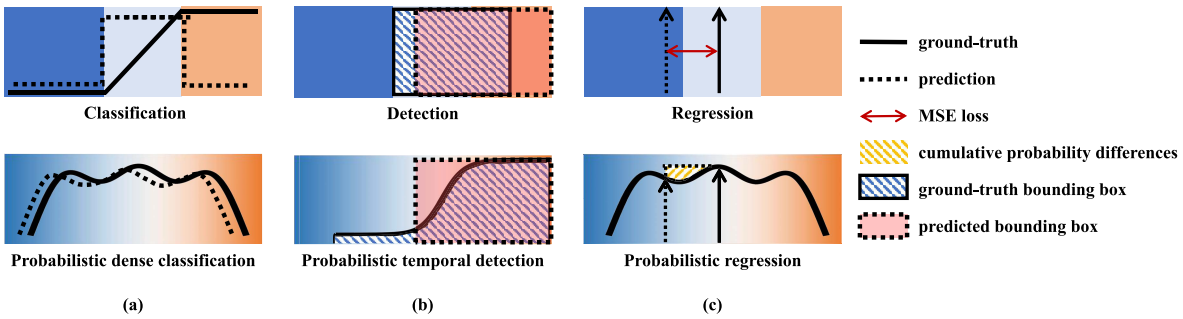


Fig. 5. Conceptual comparison of conventional and probabilistic heads. In (a), conventional classification assigns hard labels to each frame, without considering the uncertainty of rigid annotations. Probabilistic dense classification replaces rigid annotations with temporal label distributions to model the uncertainty of unreliable hard labels and the graduality of the transition from intentional to unintentional actions. In (b), different from conventional detection, probabilistic temporal detection introduces a probabilistic IoU, where the ground truth bounding box is the area under cumulative distribution function of dense probabilistic supervision and the predicted bounding box is an indicator function ranging from the predicted temporal location to the end of the video sequence. In (c), probabilistic regression calculates the cumulative probability differences in the range of ground truth and predicted temporal locations as a residual term, instead of calculating the MSE between them like conventional regression.

annotations as temporal label distributions to generate dense probabilistic supervision. Our approach considers the uncertainty of rigid annotations and constructs the graduality of the transition from intentional action to unintentional action, which alleviates the negative effect brought by over-fitting.

- ***Probabilistic temporal detection head*** predicts the temporal boundary of unintentional action. In Fig. 5 (b), we show the difference between conventional detection and probabilistic temporal detection. For conventional detection, the ground truth bounding box is an indicator function. For probabilistic IoU, the ground truth bounding box is the cumulative distribution function of dense probabilistic supervision. Our approach constructs a dynamic punishment on the residual by considering the uncertainty of action transition.
- ***Probabilistic regression head*** predicts a temporal location when the action transits from intentional to unintentional. Different from the Mean Square Error (MSE) loss of conventional regression, probabilistic regression loss calculates the cumulative probability differences in the range of ground truth and predicted temporal locations as the residual term, as shown in Fig. 5 (c). Our approach introduces a dynamic punishment on the regression residual using dense probabilistic supervision.

IV. Experiments

In this section, our approach is evaluated on the OOPS dataset to demonstrate its effectiveness for the tasks of unintentional action recognition and localization. We also conduct a qualitative comparison with other methods and provided the analysis for the compared results.

A. Dataset

The OOPS dataset contains unintentional actions caused by various errors or factors, e.g., execution errors and unexpected interventions. The OOPS dataset is a big amount of collection of videos consisting of over 20000 videos from the YouTube website. There are 4674 labeled training videos and 3593 testing videos in this dataset, where each video is annotated with the timestamp at the transition from intentional action to unintentional action. Furthermore, according to the statistical information of the OOPS dataset, fifty percent of videos are mainly between the five-second and ten-second, and forty percent of videos start unintentional actions in the middle length of the video. The mean video clip length is 9.4 seconds. We show some examples of the OOPS dataset in Fig. 1, such as " (a) Doing push-ups", " (b) Cleaning the pool"," (c) Throwing a fishing net". We can see that annotations of when the intentional action transitions to the unintentional action are intrinsic ambiguous. For example in Fig. 1 (a), the annotator y_1 argues the transition to unintentional action occurring at the man getting up off the ground, while y_2 argues the failure occurring when the man pushing down, and y_3 argues the failure occurring when the man falling.

B. Experimental Settings

During the training phase, we utilized the 3D ResNet-18 [15] model pre-trained on Kinetics [14] to extract 512-dimension visual features for each video frame at the last convolutional layer. After that, we applied a 2-layer basic LSTM as the backbone and used the Adagrad optimizer to train the model with an initial learning rate of 0.001, where the dimensions of input and hidden state are 512 and 256, respectively. The Pdc head consisted of 256 input units and 128 hidden units with a ReLU activation function, followed by a linear output layer. The Ptd head shared the same network structure as the Pdc head while the output of Ptd was obtained by implementing the operation arg max on the output of Pdc. The Pr head consisted of 256 input units and 128 hidden units with a ReLU activation function, followed by a linear output layer. We set the hyper-parameters as $\lambda_1 = 0.5$, $\lambda_2 = 0.1$.

During the test phase, we directly extracted 512-dimension visual features for original untrimmed video and fed them into our learned model to predict all the video frames whether are the beginning of unintentional action. We followed the experimental setting of recognition and localization in [11] to make evaluations. For a classification task, our probabilistic framework predicted the category of each frame in a testing video. For a temporal localization task, we utilized our model

TABLE I
COMPARISONS OF OUR APPROACH AND OTHER METHODS FOR THE TASK OF UNINTENTIONAL ACTION RECOGNITION ON OOPS

Method	Recognition Accuracy	
	Linear	Fine-tuned
PUAV-Chance [11]	33.3	33.3
PUAV-VideoSort [11]	49.8	60.2
PUAV-VideoContext [11]	50.0	60.3
PUAV-VideoSpeed [11]	53.4	61.6
PUAV-Pretrained [11]	53.6	64.0
LGfF [60]	70.3	77.9
UAL-TLA [12]	72.8	78.6
Ours	**79.2**	**81.6**

TABLE II
COMPARISONS OF OUR APPROACH AND OTHER METHODS FOR THE TASK OF UNINTENTIONAL ACTION LOCALIZATION ON OOPS

Method	Localization Accuracy	
	within 1.0 sec	within 0.25 sec
PUAV-Chance [11]	25.9	6.8
PUAV-VideoSort [11]	43.3	18.3
PUAV-VideoContext [11]	52.0	25.3
PUAV-VideoSpeed [11]	65.3	36.6
PUAV-Pretrained [11]	69.2	37.8
PUAV-Finetuned [11]	75.9	46.7
LGfF [60]	72.4	39.9
UAL-TLA [12]	73.2	40.2
Ours	**76.1**	**47.2**

in a sliding window fashion over the temporal axis and evaluated whether the model can detect the temporal location that the action transits from intentional to unintentional. The predicted boundary was the one with the most confident score of the category "transitional" across all the sliding windows. We considered the prediction correct if the predicted boundary sufficiently overlaps any of the ground truth positions in the dataset, where two thresholds of sufficient overlap were utilized, i.e., within 1 second and within 0.25 second.

C. Results and Analysis

1) Comparison With State-of-the-Art Methods:

a) Unintentional Action Recognition: For the task of unintentional action recognition, we compare with the methods used in [11], including Chance, Video-Speed, VideoContext, VideoSort, Pre-trained model on Kinetics, LGfF method [60] using the extra annotations of action goals, and a preliminary UAL-TLA method [12], where the pre-trained model is trained on the full, annotated Kinetics [14] dataset as feature extractors. Note that, "Linear" denotes that the model is used as a feature extractor, while "Fine-tuned" denotes that the model is fine-tuned with the labeled training set.

Table I shows the experimental results of the task of unintentional action recognition. It can be seen that our approach outperforms other compared methods. For example, by making a comparison between our approach (i.e., Ours) and a competitive fully supervised method (PUAV-Pretrained), we can obtain 25.6 percent and 17.6 percent improvements both on the Linear and Fine-tuned settings, respectively. It indicates that our approach is more appropriate to recognize unintentional action in the case of multiple unreliable annotations. Besides, our approach also outperforms the LGfF method both on the Linear and Fine-tuned settings, and respectively obtain 8.9 percent and 3.7 percent improvements. Different from the LGfF method that utilizes the BERT word embeddings of each action in the recognition, our approach learns the model by jointly training probabilistic heads with different supervision granularities, to achieve action intentionality recognition via assisting dense classification.

b) Unintentional Action Localization: For the task of unintentional action localization, we compare with the methods used in [11], [60], including Chance, Video-Speed, VideoContext, VideoSort, Pretrained/Finetuned model on Kinetics, LGfF, and UAL-TLA, where "Pretrained" denotes the pre-trained model as feature extractors and "Finetuned" denotes the above pre-trained model fine-tuned with the labeled training set of OOPS dataset. Note that, "within 1 second" and "within 0.25 second" denote that two different thresholds of overlap are utilized to judge the prediction was correct or not.

Table II shows the experimental results of the task of unintentional action localization. We see that our approach outperforms other compared methods. For example, comparing our approach with the PUAV-Pretrained method, we can obtain 6.9 percent and 9.4 percent improvements on the settings of within one second and within a one-quarter second, which indicates that probabilistic modeling in our approach is more appropriate to localize unintentional actions utilizing multiple rigid annotations. Making comparisons between our approach and the PUAV-Finetuned method, we still gain 0.2 percent and 0.5 percent improvements on different settings, respectively. Besides, our approach also outperforms the LGfF method both on the settings of within one second and within a one-quarter second, respectively obtains 3.7 percent and 7.3 percent improvements, though we do not use any extra annotation about "action goal" to improve the quality of supervision. It illustrates the effectiveness of our approach supervised by different supervision granularities that can capture more valuable perceptual clues to localize unintentional actions.

2) Ablation Study: To investigate the effectiveness of individual components and different probabilistic heads in our approach, ablation studies with various configurations are conducted for both unintentional action recognition and localization tasks on the OOPS dataset. As shown in Table III, different configurations of our approach are defined as follows:

- **Baseline** stands for the baseline of unintentional action localization, which trains standard LSTM with likelihood correlations between the predictions and annotations, that is a three-way classification task similar to PUAV-Pretrained, where the three categories are intentional, transitional, and unintentional. In **Baseline**, the loss function is a cross-entropy loss without introducing PAM, TLA, and DPL components.
- **Probabilistic** utilizes a temporal label distribution to replace rigid annotations to supervise the learning model, without introducing TLA, Ptd, and Pr. In **Probabilistic**, the loss function is $\mathcal{L}_{\text{Pdc}}$ supervised by p_{y_k} provided in equation (2).

TABLE III
ABLATION STUDIES ON THE OOPS DATASET FOR THE TASKS OF BOTH UNINTENTIONAL ACTION RECOGNITION AND LOCALIZATION

Method	PAM	TLA	DPL			Recognition		Localization	
			Pdc	Ptd	Pr	Linear	Fine-tuned	within 1.0 sec	within 0.25 sec
Baseline						58.5	64.9	68.3	39.2
Probabilistic	✓		✓			59.3	66.1	72.4	39.5
Online	✓	✓	✓			76.9	78.6	73.2	40.2
w/o PR	✓	✓	✓	✓		77.2	79.1	73.7	44.9
w/o PTD	✓	✓	✓		✓	77.5	79.6	73.8	44.8
Ours	✓	✓	✓	✓	✓	**79.2**	**81.6**	**76.1**	**47.2**

- **Online** couples TLA with the Pdc head, where TLA generates dense probabilistic supervision to train Pdc that is also inversely applied to learn dense probabilistic supervision, which collaboratively optimizes TLA and Pdc in an online manner. The loss function is $\mathcal{L}_{\text{Pdc}}$ supervised by p_y provided in equation (5), which is a fine-grained supervision.
- **w/o PR** contains TLA, Pdc, and Ptd, where TLA generates dense probabilistic supervision to train both Pdc and Ptd, where Pdc is inversely applied to update the learning of dense probabilistic supervision. The loss function is $\mathcal{L}_{\text{Pdc}} + \mathcal{L}_{\text{Ptd}}$ supervised by p_y in equation (5) and P_y, which combines two supervision granularities (i.e., fine-grained and middle granularities).
- **w/o PTD** consists of TLA, Pdc, and Pr, where dense probabilistic supervision is generated by TLA to train both Pdc and Pr, where Pdc is also inversely applied to update dense probabilistic supervision. The loss function is $\mathcal{L}_{\text{Pdc}} + \mathcal{L}_{\text{Pr}}$ supervised by p_y in equation (5) and y, which combines two supervision granularities (i.e., fine-grained and coarse granularities).

As shown in Table III, we can draw the following conclusions by comparing experimental results:

- Compared with **Baseline**, our approach achieves 20.7 percent and 16.7 percent improvements in the settings of Linear and Fine-tuned on the recognition task. Our approach also obtains 7.8 percent and 8.0 percent improvements in the settings of within 1 second and within 0.25 second on the localization task. It demonstrates the effectiveness of our approach in probabilistic annotation modeling and dense probabilistic localization.
- Compared with **Baseline**, **Probabilistic** shows that softening a rigid annotation from a one-hot vector into a unimodal distribution is beneficial to model the uncertainty of a rigid annotation. For the localization task, **Probabilistic** outperforms **Baseline** and obtains 4.1 percent and 0.3 percent improvements respectively in the settings of within 1 second and within 0.25 second. However, **Probabilistic** is not as good as **Online** in mining reliable supervision from multiple unreliable rigid annotations. The latter improves the former with 0.8 percentage and 0.7 percentage respectively in the settings of within 1 second and within 0.25 second.
- The performance comparisons between **Online** and **Baseline** show that TLA is effective in modeling the uncertainty of rigid annotations and mining the reliable supervision from multiple unreliable rigid annotations. Compared with **Baseline**, **Online** respectively achieves 18.4 percent and 13.7 percent improvements in the settings of Linear and Fine-tuned on the recognition task and 4.9 percent and 1.0 percent improvements in the settings of within 1 second and within 0.25 second on the localization task.
- Compared with **w/o PR**, our approach shows the superiority of Pr that guides the model training under the coarse granularity supervision, which respectively achieves 2.0 percent and 2.5 percent improvements in the settings of Linear and Fine-tuned on the recognition task and 2.4 percent and 2.3 percent improvements in the settings of within 1 second and within 0.25 second on the localization task.
- The performance comparisons between **w/o PTD** and our approach shows the effectiveness of Ptd that guides the model training using middle granularity supervision. Our approach respectively achieves 1.7 percent and 2.0 percent improvements in the settings of Linear and Fine-tuned on the recognition task and 2.3 percent and 2.4 percent improvements in the settings of within 1 second and within 0.25 second on the localization task. It is because that $\mathcal{L}_{\text{Ptd}}$ and $\mathcal{L}_{\text{Pr}}$ are complementary since their supervision signals have different granularities, where $\mathcal{L}_{\text{Ptd}}$ is used for the temporal segment detection and $\mathcal{L}_{\text{Pr}}$ is used for the temporal location regression.

TABLE IV
STUDY OF THE HYPER-PARAMETER σ ON BOTH UNINTENTIONAL ACTION RECOGNITION AND LOCALIZATION TASKS

σ	Recognition		Localization	
	Linear	Fine-tuned	within 1.0 sec	within 0.25 sec
1	88.1	89.4	25.6	14.0
2	82.3	83.2	55.2	31.7
4	**79.2**	**81.6**	**76.1**	**47.2**
6	74.5	76.9	73.5	44.8
8	69.6	72.8	73.3	44.6
10	65.3	66.0	72.0	42.5

3) Study of the Hyper-Parameter σ: In experiments, we further find that the hyper-parameter σ of TLA in equation (2) affects the performance of unintentional action recognition and localization, where σ depicts the prior uncertainty of rigid annotation. We explore different σ for the tasks of unintentional action recognition and localization. The experimental results are reported in Table IV. We see that when δ is small, temporal label distributions are sharp, which makes the generated dense probabilistic supervision p_y containing

Fig. 6. The result comparisons between our approach and PUAV-Finetuned [11], from top to bottom including "crossing the single-plank bridge", "playing football", "riding the snowmobile", and "doing the splits". We can observe that our approach correctly localizes the unintentional action while the PUAV-Finetuned method fails, which demonstrates the effectiveness of the proposed probabilistic framework for training the model to localize unintentional action.

multiple peaks. It is difficult for unsmooth supervision to guide the model to predict an accurate temporal location. The learned dense predictions are also unsmooth, leading to high recognition accuracy but poor localization performance on unintentional action. When δ is larger than a certain threshold, temporal label distributions are too smooth, which loses important discriminative supervision and results in performance degradation. Therefore, an appropriate δ is necessary to model the uncertainty of rigid annotations. In the experiments, we found that $\delta = 4$ achieves a good trade-off and the peak reaches 76.1 and 47.2 for the settings of "within 1 second" and "within 0.25 second".

4) Visualization: To intuitively show the effectiveness of our approach, we conduct some qualitative analyses, which visualize the performance comparisons between the localization results of PUAV-Finetuned [11] and our approach, i.e., the visualization examples in Fig. 6. Taking the first video "falling down while crossing the single-plank bridge" as an example for comparison, our approach correctly localizes the unintentional action while the PUAV-Finetuned method fails because of localizing the person who falls. It indicates that the model trained with multiple hard labels without a probabilistic framework leads to overfitting the "falling down" results instead of unintentional action (reason) like slipping down. Similarly, in the other examples, the PUAV-Finetuned method still fails since it localizes the person who falls instead of unintentional action.

In addition, we also show some failed results of our approach in Fig. 7 and analyze their reasons. Our probabilistic framework still tends to capture the visual changes in a simple background since a scrambled background easily confuses real unintentional action leading to the failures. For example, the localization result of "walking on the road" in the second video is incorrect since the real unintentional action is the beginning of slipping off.

TABLE V
COMPARISONS OF OUR APPROACH AND OTHER TEMPORAL ACTION LOCALIZATION METHODS ON THE THUMOS14 DATASET

Method	THUMOS14	
	mAP IOU@0.5	mAP IOU@0.7
TAL-Net [6]	42.8	20.8
G-TAD [75]	40.2	/
BSN UNet [40]	36.9	20.0
BMN [74]	32.2	/
R-C3D [39]	28.9	/
Baseline	30.8	19.7
Ours	40.0	24.1

5) General Temporal Action Localization: The task of unintentional action localization is one of the most recent topics. The related work was first published in CVPR 2020 [11]. Based on this work, two methods [12], [60] were published in CVPR 2021 and ICME 2021. As far as we know, OOPS is the current only dataset about unintentional action. To make experiments sufficient, we provide the result of our method on the THUMOS14 dataset in Table V to further demonstrate the effectiveness of our approach on the task of general temporal action localization.

The THUMOS14 dataset contains 2765 trimmed training videos, 1010 untrimmed validation videos, and

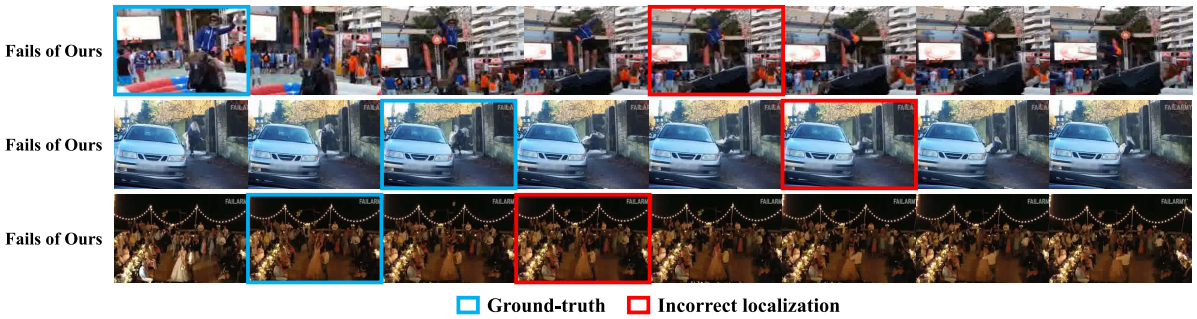


Fig. 7. The fail examples of our approach, i.e., "playing the electric bullfighting machine," "walking on the road," and "throwing bouquets." Our probabilistic framework still tends to capture the visual changes in a simple background since a scrambled background easily confuses the real unintentional action leading to the failures.

1574 untrimmed testing videos, where only 200 validation and 213 testing videos have temporal annotations. Following previous efforts [6], [39], [40], [74], [75], we adopt these 200 validation videos in the training phase and utilize 213 testing videos to evaluate the performance. To make comparisons with previous works [6], [39], [40], [74], [75], we follow their evaluation metrics and report mean Average Precision (mAP) under thresholds tIoU={0.5, 0.7}, as shown in Table V. It can be seen that our method outperforms other recent methods under the metric mAP tIoU@0.7, which demonstrates the effectiveness of our method on the task of general temporal action localization. Baseline is a baseline method that trains standard LSTM with likelihood correlations between the predictions and annotations. The loss function of Baseline is a cross-entropy loss without introducing PAM, TLA, and DPL modules, which has been defined in Section IV.C.2). Compared with Baseline, our method respectively achieves 9.2% and 4.4% improvements under the metrics mAP tIoU@0.5 and mAP tIoU@0.7, which demonstrates the superiority of our probabilistic temporal modeling framework for temporal action localization.

V. Potential Application

Action intention could be one of the cores of human behavior understanding, which has many potential usages in the real world. Different from general action recognition, segmentation, and quality assessment, understanding action intention helps for explaining why the action fails and then making targeted improvements. Specifically, localizing unintentional action via understanding action intention could help to date back to the reason for the action failure, which is useful to make a responsibility determination in the field of smart mobility. In the field of competitive sports, seeking the cause of unintentional action to provide feedback is useful to help athletes' training on targeted parts and improve their competitive skills. Besides, localizing unintentional action via understanding action intention also helps in issuing an early warning. For instance, in the field of smart transportation, action intention understanding can be used for forecasting the onset of pedestrians' unintentional action shortly into the future to make reliable decisions for emergency avoidance to guarantee the safety of pedestrians and avoid traffic injuries [76]. In the field of medical services, localizing unintentional action via understanding action intention can be used for anticipating the onset of older adults' unintentional action to monitor their health, which assists to improve existing health care systems, particularly for the current aging society.

VI. Conclusion

In this paper, we propose a probabilistic framework for unintentional action localization via modeling the uncertainty of rigid annotations and jointly training three probabilistic heads using different granularity supervisions. Our approach formulates unreliable rigid annotations as temporal label distributions to model the uncertainty of hard labels and the graduality of the transition from intentional action to unintentional action. We also propose a temporal label aggregation model to online aggregate temporal label distributions via attention learning to mine dense probabilistic supervision. We further propose a dense probabilistic localization model to jointly train three probabilistic heads with different supervision granularities. Compared to conventional methods, our approach demonstrates a significant improvement in both the recognition and localization tasks.

References

[1] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. NeurIPS*, 2014, pp. 568–576.
[2] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 20–36.
[3] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1933–1941.
[4] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei, "End-to-end learning of action detection from frame glimpses in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2678–2687.
[5] Z. Shou, J. Chan, A. Zareian, K. Miyazawa, and S.-F. Chang, "CDC: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5734–5743.
[6] Y.-W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar, "Rethinking the faster R-CNN architecture for temporal action localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1130–1139.
[7] J.-H. Pan, J. Gao, and W.-S. Zheng, "Action assessment by joint relation graphs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6331–6340.
[8] A. L. Woodward, "Infants' ability to distinguish between purposeful and non-purposeful behaviors," *Infant Behav. Develop.*, vol. 22, no. 2, pp. 145–160, Jan. 1999.
[9] A. L. Woodward, "Infants' grasp of others' intentions," *Current Directions Psychol. Sci.*, vol. 18, no. 1, pp. 53–57, Feb. 2009.
[10] A. C. Brandone and H. M. Wellman, "You can't always get what you want: Infants understand failed goal-directed actions," *Psychol. Sci.*, vol. 20, no. 1, pp. 85–91, Jan. 2009.

[11] D. Epstein, B. Chen, and C. Vondrick, "Oops! Predicting unintentional action in video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 919–929.

[12] N. Zhou, G. Chen, J. Xu, W.-S. Zheng, and J. Lu, "Temporal label aggregation for unintentional action localization," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2021, pp. 1–7.

[13] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, "Towards good practices for very deep two-stream ConvNets," 2015, *arXiv:1507.02159*.

[14] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.

[15] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6546–6555.

[16] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7794–7803.

[17] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proc. CVPR*, 2019, pp. 6202–6211.

[18] D. Li, Z. Qiu, Y. Pan, T. Yao, H. Li, and T. Mei, "Representing videos as discriminative sub-graphs for action Recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3310–3319.

[19] X. Song *et al.*, "Spatio-temporal contrastive domain adaptation for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9787–9795.

[20] C.-F.-R. Chen *et al.*, "Deep analysis of CNN-based spatio-temporal representations for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6165–6175.

[21] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. C. Niebles, "SST: Single-stream temporal action proposals," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2911–2920.

[22] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1049–1058.

[23] X. Dai, B. Singh, G. Zhang, L. S. Davis, and Y. Q. Chen, "Temporal context network for activity localization in videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5793–5802.

[24] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin, "Temporal action detection with structured segment networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2914–2923.

[25] L. Yang, H. Peng, D. Zhang, J. Fu, and J. Han, "Revisiting anchor mechanisms for temporal action localization," *IEEE Trans. Image Process.*, vol. 29, pp. 8535–8548, 2020.

[26] X. Wang, S. Zhang, Z. Qing, Y. Shao, C. Gao, and N. Sang, "Self-supervised learning for semi-supervised temporal action proposal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1905–1914.

[27] H. Alwassel, F. C. Heilbron, and B. Ghanem, "Action search: Spotting actions in videos and its application to temporal action localization," in *Proc. ECCV*, 2018, pp. 251–266.

[28] T. Lin, X. Zhao, and Z. Shou, "Single shot temporal action detection," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 988–996.

[29] Z. Yuan, J. C. Stroud, T. Lu, and J. Deng, "Temporal action localization by structured maximal sums," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3684–3692.

[30] L. Huang, Y. Huang, W. Ouyang, and L. Wang, "Modeling sub-actions for weakly supervised temporal action localization," *IEEE Trans. Image Process.*, vol. 30, pp. 5154–5167, 2021.

[31] P. Tirupattur, K. Duarte, Y. S. Rawat, and M. Shah, "Modeling multi-label action dependencies for temporal action localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1460–1470.

[32] F. Long, T. Yao, Z. Qiu, X. Tian, J. Luo, and T. Mei, "Gaussian temporal awareness networks for action localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 344–353.

[33] Y. Hu, M. Liu, X. Su, Z. Gao, and L. Nie, "Video moment localization via deep cross-modal hashing," *IEEE Trans. Image Process.*, vol. 30, pp. 4667–4677, 2021.

[34] R. Zeng, C. Gan, P. Chen, W. Huang, Q. Wu, and M. Tan, "Breaking winner-takes-all: Iterative-winners-out networks for weakly supervised temporal action localization," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5797–5808, Dec. 2019.

[35] W. Yang, T. Zhang, Z. Mao, Y. Zhang, Q. Tian, and F. Wu, "Multi-scale structure-aware network for weakly supervised temporal action detection," *IEEE Trans. Image Process.*, vol. 30, pp. 5848–5861, 2021.

[36] W. Luo *et al.*, "Action unit memory network for weakly supervised temporal action localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9969–9979.

[37] G. Chéron, J.-B. Alayrac, I. Laptev, and C. Schmid, "A flexible model for training action localization with varying levels of supervision," in *Proc. NeurIPS*, 2018, pp. 1–12.

[38] K. Yang, P. Qiao, D. Li, S. Lv, and Y. Dou, "Exploring temporal preservation networks for precise temporal action localization," in *Proc. AAAI*, 2018, vol. 32, no. 1, pp. 1–8.

[39] H. Xu, A. Das, and K. Saenko, "R-C3D: Region convolutional 3D network for temporal activity detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5783–5792.

[40] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "BSN: Boundary sensitive network for temporal action proposal generation," in *Proc. ECCV*, 2018, pp. 3–19.

[41] V. Chandola, A. Banerjee, and V. Kumar, "Outlier detection: A survey," *ACM Comput. Surv.*, vol. 14, p. 15, Aug. 2007.

[42] E. Jardim, L. A. Thomaz, E. A. B. da Silva, and S. L. Netto, "Domain-transformable sparse representation for anomaly detection in moving-camera videos," *IEEE Trans. Image Process.*, vol. 29, pp. 1329–1343, 2020.

[43] H. Lv, C. Zhou, Z. Cui, C. Xu, Y. Li, and J. Yang, "Localizing anomalies from weakly-labeled videos," *IEEE Trans. Image Process.*, vol. 30, pp. 4505–4515, 2021.

[44] T. Reiss, N. Cohen, L. Bergman, and Y. Hoshen, "PANDA: Adapting pretrained features for anomaly detection and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2806–2814.

[45] S. Wang, L. Wu, L. Cui, and Y. Shen, "Glancing at the patch: Anomaly localization with global and local feature comparison," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 254–263.

[46] M. Salehi, N. Sadjadi, S. Baselizadeh, M. H. Rohban, and H. R. Rabiee, "Multiresolution knowledge distillation for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14902–14912.

[47] R. Hinami, T. Mei, and S. Satoh, "Joint detection and recounting of abnormal events by learning deep generic knowledge," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3619–3627.

[48] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked RNN framework," in *Proc. ICCV*, 2017, pp. 341–349.

[49] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette, "Deep-cascade: Cascading 3D deep neural networks for fast anomaly detection and localization in crowded scenes," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1992–2004, Apr. 2017.

[50] J.-C. Feng, F.-T. Hong, and W.-S. Zheng, "MIST: Multiple instance self-training framework for video anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14009–14018.

[51] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, "CutPaste: Self-supervised learning for anomaly detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9664–9674.

[52] M.-I. Georgescu, A. Barbalau, R. T. Ionescu, F. S. Khan, M. Popescu, and M. Shah, "Anomaly detection in video via self-supervised and multi-task learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12742–12752.

[53] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 733–742.

[54] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection—A new baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6536–6545.

[55] X. Xie, C. Wang, S. Chen, G. Shi, and Z. Zhao, "Real-time illegal parking detection system based on deep learning," in *Proc. Int. Conf. Deep Learn. Technol. (ICDLT)*, 2017, pp. 23–27.

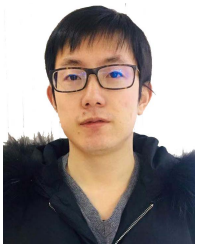
[56] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *Proc. IPMI*, 2017, pp. 146–157.

[57] A. Javaid, Q. Niyaz, W. Sun, and M. Alam, "A deep learning approach for network intrusion detection system," in *Proc. 9th EAI Int. Conf. Bio-Inspired Inf. Commun. Technol.*, 2016, pp. 21–26.

[58] M. Mohammadi, A. Al-Fuqaha, S. Sorour, and M. Guizani, "Deep learning for IoT big data and streaming analytics: A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 2923–2960, 4th Quart., 2018.

[59] G. Di Biase, H. Blum, R. Siegwart, and C. Cadena, "Pixel-wise anomaly detection in complex driving scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16918–16927.
[60] D. Epstein and C. Vondrick, "Learning goals from failure," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11194–11204.
[61] Y. Hongsang, L. Haopeng, K. Qiuhong, L. Liangchen, and Z. Rui, "Precondition and effect reasoning for action recognition," 2021, *arXiv:2112.10057*.
[62] X. Geng, C. Yin, and Z.-H. Zhou, "Facial age estimation by learning from label distributions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 10, pp. 2401–2412, Oct. 2013.
[63] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng, "Deep label distribution learning with label ambiguity," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2825–2838, Jun. 2017.
[64] Z. He *et al.*, "Data-dependent label distribution learning for age estimation," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3846–3858, Aug. 2017.
[65] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *Proc. NeurIPS*, 2019, pp. 1567–1578.
[66] J. Wang and X. Geng, "Label distribution learning machine," in *Proc. ICML*, 2021, pp. 10749–10759.
[67] X. Wen *et al.*, "Adaptive variance based label distribution learning for facial age estimation," in *Proc. ECCV*, 2020, pp. 379–395.
[68] K. Su and X. Geng, "Soft facial landmark detection by label distribution learning," in *Proc. AAAI*, vol. 33, 2019, pp. 5008–5015.
[69] S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, and Y. Rui, "Label distribution learning on auxiliary label space graphs for facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13984–13993.
[70] X. Geng and Y. Xia, "Head pose estimation based on multivariate label distribution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1837–1842.
[71] Y. Tang *et al.*, "Uncertainty-aware score distribution learning for action quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9839–9848.
[72] M. Ling and X. Geng, "Indoor crowd counting by mixture of Gaussians label distribution learning," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5691–5701, Nov. 2019.
[73] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[74] M. Xu, C. Zhao, D. S. Rojas, A. Thabet, and B. Ghanem, "G-TAD: Sub-graph localization for temporal action detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10156–10165.
[75] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen, "BMN: Boundary-matching network for temporal action proposal generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3889–3898.
[76] R. Quan, L. Zhu, Y. Wu, and Y. Yang, "Holistic LSTM for pedestrian trajectory prediction," *IEEE Trans. Image Process.*, vol. 30, pp. 3229–3239, 2021.

Jinglin Xu received the Ph.D. degree in control science and engineering from Northwestern Polytechnical University, Xi'an, China, in 2020. She is currently a Postdoctoral Fellow with the Department of Automation, Tsinghua University, Beijing, China, with a research focus on video understanding. She has broad interests in computer vision, pattern recognition, and machine learning, where she has authored/coauthored 13 scientific papers in these areas, such as IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, CVPR, AAAI, and IJCAI.



Guangyi Chen received the B.S. and Ph.D. degrees from the Department of Automation, Tsinghua University, China, in 2016 and 2021, respectively. His research interests include computer vision and machine learning, with particular expertise in attention learning, causality, and human center vision tasks, such as re-identification, trajectory prediction, and action understanding. He has published more than ten papers on top journals and conferences, such as IEEE TRANSACTIONS ON IMAGE PROCESSING, CVPR, ICCV, and ECCV.

Nuoxing Zhou received the B.S. degree in management science from the Business School, Sichuan University, China, in 2020. He is currently pursuing the M.S. degree in electrical engineering with the McCormick School of Engineering, Northwestern University. From 2020 to 2021, he was a Research Assistant with the Department of Automation, Tsinghua University. His research interests include computer vision, autonomous vehicles, and deep learning.

Wei-Shi Zheng is currently a Professor with Sun Yat-sen University. His research interests include person/object association and activity understanding in visual surveillance, and the related large-scale machine learning algorithm. He has ever joined Microsoft Research Asia Young Faculty Visiting Programme. He was a recipient of the Excellent Young Scientists Fund of the National Natural Science Foundation of China and a recipient of the Royal Society-Newton Advanced Fellowship of the U.K.

Jiwen Lu (Senior Member, IEEE) received the B.Eng. degree in mechanical engineering and the M.Eng. degree in electrical engineering from the Xi'an University of Technology, Xi'an, China, in 2003 and 2006, respectively, and the Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore, in 2012. He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing, China. His current research interests include computer vision and pattern recognition. He was/is a member of the Image, Video and Multidimensional Signal Processing Technical Committee, Multimedia Signal Processing Technical Committee, the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society, the Multimedia Systems and Applications Technical Committee, and the Visual Signal Processing and Communications Technical Committee of the IEEE Circuits and Systems Society. He is a fellow of IAPR. He was a recipient of the National Natural Science Funds for Distinguished Young Scholar. He serves as the General Co-Chair for the International Conference on Multimedia and Expo (ICME) 2022 and the Program Co-Chair for the International Conference on Multimedia and Expo 2020, the International Conference on Automatic Face and Gesture Recognition (FG) 2023, and the International Conference on Visual Communication and Image Processing (VCIP) 2022. He serves as the Co-Editor-of-Chief for *Pattern Recognition Letters*, an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON BIOMETRICS, BEHAVIOR, AND IDENTITY SCIENCE, and *Pattern Recognition*.