Tragedy Plus Time: Capturing Unintended Human Activities from Weakly-labeled Videos

Arnav Chakravarthy, Zhiyuan Fang, Yezhou Yang
Arizona State University
achakr37@asu.edu, zy.fang@asu.edu, yz.yang@asu.edu

Abstract

In videos that contain actions performed unintentionally, agents do not achieve their desired goals. In such videos, it is challenging for computer vision systems to understand high-level concepts such as goal-directed behavior, an ability present in humans from a very early age. Inculcating this ability in artificially intelligent agents would make them better social learners by allowing them to evaluate human action under a teleological lens. To validate this ability of deep learning models to perform this task, we curate the W-Oops dataset, built upon the Oops dataset [11]. W-Oops consists of 2,100 unintentional human action videos, with 44 goal-directed and 30 unintentional video-level activity labels collected through human annotations. Due to the expensive segment annotation procedure, we propose a weakly supervised algorithm for localizing the goal-directed as well as unintentional temporal regions in the video leveraging solely video-level labels. In particular, we employ an attention mechanism based strategy that predicts the temporal regions which contributes the most to a classification task. Meanwhile, our designed overlap regularization allows the model to focus on distinct portions of the video for inferring the goal-directed and unintentional activity, while guaranteeing their temporal ordering. Extensive quantitative experiments verify the validity of our localization method. We further conduct a video captioning experiment which demonstrates that the proposed localization module does indeed assist teleological action understanding. Project website can be found at: https://asu-apg.github.io/TragedyPlusTime.

1. Introduction

Traditional video action recognition [4, 10, 20, 27, 49, 57, 63] focuses on predicting only atomic actions present on the surface appearance of a video. On the other hand, teleological understanding of actions entails understanding the underlying goal of actions and why it was performed.

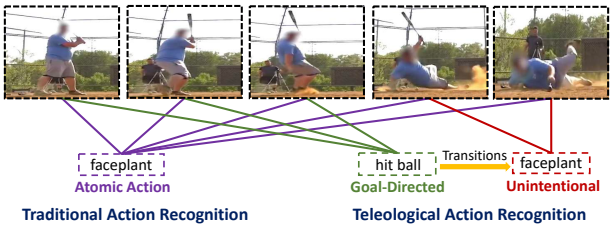

Figure 1. State-of-the-art action recognition models trained on traditional video activity datasets view an unintentional action scene as an atomic action. Although this scene involves a man falling on his face, the man's ultimate goal was to hit the ball. Green lines indicate the regions of the video which indicate the man's goal, red lines indicate the regions where the action deviates from the goal, and purple lines indicate the region the action recognition model focuses on.

These goals can be easily inferred from intentional actions as they are directly defined by their outcome. However, many actions do end up in unintended results where the goal of the action is partially or never achieved. As shown in Fig. 1, an agent tries to hit the ball with a bat, but ends up landing on his face, hence not being able to achieve his goal of hitting the ball. State-of-the-art action recognition models are all trained with vewing the whole scene as "`faceplant`" without paying attention to the goal-directed behavior which was to "`hit the ball`".

Teleological action understanding provides the invaluable ability to explain and justify an action, as well as learn from mistakes in the case the goal was not achieved [8]. For fine-grained understanding of unintentional actions, it is important to know the goal of the action, why was it not fulfilled, and when (in time) did the action start transitioning away from its goal. These abilities are present in humans from a very early age [2, 7, 18, 44, 60]. However, this is a challenging task for deep learning models since it requires the model to understand high level concepts such as goal-directed behavior from unintentional actions which would not be possible without penetrating deeper than the surface appearance of the action [2]. There are few previous works which have taken initial steps towards teleological action

understanding. [11] builds a dataset rich in unintentional human actions, along with single point transition times manually labeled by human annotators which separate the intentional and unintentional regions of the video. They also train models on classification and localization tasks. However, it does not contain well defined classes for the goal-directed action or why this goal gets disrupted. [47] focuses on predicting whether an activity was intentional or unintentional, but not the understanding of the underlying goal of an unintentional action. Previous efforts [12, 24, 61] have tried to speculate about all possible effects of actions but do not explore which effects are undesirable.

In order to build a model which is capable of capturing the unintended activity, it is crucial to first build a dataset containing goal-directed actions and why they get disrupted Additionally, in order to localize their respective regions in time, one may manually label the transition point as in [11] and fully supervise the training. However, these annotations are prohibitively expensive to collect and suffer from human error and bias. Previous works such as [23, 28, 31, 33, 34, 40, 62], which focus on segmenting atomic action scenes from untrimmed videos tackle this problem of expensive manual labelling by training a model in a weakly supervised manner using only video level action labels. Though this task differs from our task (which involves separating the goal-directed action from the region it starts deviating from its goal), it still provides encouragement to solve our task in a weakly supervised manner.

We bring **W**eakly Augmented **Oops** (W-Oops), an augmented human activities dataset which contains "*fail*" videos, building upon Oops [11] but also contains high quality video-level annotations which describe the goal-directed as well as unintentional actions in the video. We further develop an algorithm which allows the model to attend to contextualized visual cues to localize these regions as well as associate them with their respective goal-directed/unintentional class label, leveraging solely video-level labels. Our proposed learning schema includes an encoder to encode the joint representation of the goal-directed and unintentional action in the video as well as temporal attention modules which help the model focus on the respective regions of interest in the video. We also introduce a novel optimization target known as Overlap Regularization which allows the model to pay attention to distinct parts of the video for inferring both types of actions while ensuring their temporal ordering. Finally, we use class-agnostic (bottom-up) as well as class-specific (top-down) attention mechanisms to localize both types of actions. Additionally, we conduct a video captioning experiment leveraging our localization module to demonstrate it's teleological ablity.

To summarize our contributions:

1. We curate W-Oops, an augmented video dataset, built upon Oops [11], containing high quality video level labels for the goal-directed as well as unintentional action. To the best of our knowledge, we are the first to make a step towards such fine-grained understanding of unintentional actions.
2. We propose an attention mechanism based method that highlights relevant temporal regions of the video important to a classification task when inferring the goal-directed and unintentional action while also ensuring their temporal ordering.
3. Finally, we provide in-depth and comprehensive experimental analysis, which validates the effectiveness of our method compared to competitive weakly supervised action localization methods on W-Oops. Additionally, we demonstrate the teleological ability of our localization module through a video captioning experiment.

2. Related Work

Intent Recognition from Visual Input. There has been an increasing research focus on intent recognition of agents in videos. [58] proposes a hierarchical model to predict the intention, as well as the attention of an agent's eye gaze from a RGB-D video. [55] focuses on predicting the action, motivation and scenes from an image by using a third order factor graph built using text. [47] proposes an unsupervised algorithm to discriminate between an intentional and unintentional action performed by an agent. [11] too focuses on discriminating between an intentional and unintentional action, as well as predicting the point in an unintentional video when the action deviates from it's original goal, but does this in a supervised manner. Our work differs from these as we focus on discriminating between the different goal-directed and unintentional action categories in unintentional videos, as well as localizing these action regions in a weakly supervised manner. Action anticipation can also be relevant to predicting an unintentional action or the onset of it. [17, 19, 30, 38, 39] focus on forecasting an event or action based on a small snippet of a video. [51, 56] focus on self supervised learning approaches to predict future action representation using unlabeled videos. [15, 37, 52] focus on predicting a pedestrian's intent to cross the road using the Joint Attention for Autonomous Driving (JAAD) dataset [36]. Our work differs from this as it not only focuses on the past and not on predicting the future, but is also generalized to more diverse environmental settings.

Weakly Supervised Action localization (WSAL) has been drawing increasing attention due to the expensive manual labelling process involved in a fully supervised setting. Previous efforts involve localizing action regions in an untrimmed video by training a model using only video level action labels [3, 21, 33, 35, 41–43, 46, 50, 59], or sentences [5, 13, 14, 29, 32, 45]. In particular, STPN [33] trains a classification model using sum of features weighted by

their class-agnostic attention weights, which it learns using a sparsity loss on the attention weights. It then performs the localization by using both the classification activation as well as these class-agnostic weights, thresholding them to select action locations. WTALC [34] forces the foreground action features from the same action class to be similar and the background features pertaining to an action class to be dissimilar from its foreground feature, and finally localizes the action by thresholding the classification activation. A2CL-PT [31] uses foreground and background features to form triplets and apply the Angular Triplet Center Loss [25] to separate the foreground and background features, as well as use an adversarial branch in order to find supplementary activities from non-localized parts of the video. DGAM [40] propose to separate action frames from context frames by modeling the frame representation conditioned on the bottom-up attention. TSCN [62] fuse the attention sequences from the RGB and optical flow stream and use it as pseudo ground truth to supervise the training.

3. Methodology

We intend to identify the goal-directed and unintended human activities, as well as their corresponding moment of occurrence from an unintentional video in a weakly-supervised manner. To be specific, given the video $\mathcal{V}$ and its categorical labels representing the goal-directed activity, y^{IA}, and the unintended activity, y^{UA}, we expect the model to predict the triplets $\langle s^{\text{IA}}, e^{\text{IA}}, c^{\text{IA}} \rangle$ and $\langle s^{\text{UA}}, e^{\text{UA}}, c^{\text{UA}} \rangle$, containing the starting point, end point and action class associated with this segment by leveraging only the video-level annotations as weak supervision. We formulate this challenge as a weakly supervised action localization (WSAL) task, and address it using an attention mechanism based approach. We start this section by providing an overview of our model, followed by the details of formulations and our proposed learning objective.

3.1. Task Formulation

To encode the videos, pre-trained 3D neural networks are exploited to extract a set of clip-level representations $\mathbf{X}$. We find that in order to encode the goal-directed and unintentional features from the video, directly using static features is not sufficient. Hence, we encode the clip embedding by an encoder network $\mathcal{F}$, which outputs a joint representation for the goal-directed and unintentional action: $\mathbf{O} = \mathcal{F}(\mathbf{X})$, where $\mathbf{O} \in \mathbb{R}^{l \times d}$ denotes the representations in d dimensions for l clips. Here, encoder network $\mathcal{F}$ can either be a bidirectional Gated Recurrent Unit or a Transformer Encoder [53]. On this basis, we introduce two bottom-up attention modules, which outputs the temporal attention weights that reflect the temporal importance of clip representations for the goal-directed/unintentional activity respectively. This is achieved by training the model with a classification loss, *e.g.*, multiple instance learning loss. Note that these attention weights are agnostic to the specific action, and are used to identify generic regions of interest. A stack of 1-D Convolution layers with RELU activation between layers, followed by a Sigmoid function is used to obtain the attention weights λ^{IA}, $\lambda^{\text{UA}} \in \mathbb{R}^{l}$ with a scale between 0 and 1.

In order to obtain goal-directed and unintentional features, we compute a dot product between the joint representation $\mathbf{O}$ and each of the bottom-up attention weights λ^{IA} and λ^{UA}. These features would represent those parts of the joint representation $\mathbf{O}$ which correspond to the goal-directed and unintentional region respectively. Formally,

$$\mathbf{O}^{\text{IA}} = \mathbf{O} \cdot \lambda^{\text{IA}}, \qquad \mathbf{O}^{\text{UA}} = \mathbf{O} \cdot \lambda^{\text{UA}}. \tag{1}$$

We then compute Temporal Class Activation Maps (TCAM) [33], $\mathbf{C}^{\text{IA}} \in \mathbb{R}^{l \times N_{\text{IA}}}$, $\mathbf{C}^{\text{UA}} \in \mathbb{R}^{l \times N_{\text{UA}}}$ for the goal-directed as well as unintentional actions, with N_{IA} and N_{UA} corresponding to the number of goal-directed and unintentional classes, by employing two weight-sharing linear transformation layer on $\mathbf{O}^{\text{IA}}$ and $\mathbf{O}^{\text{UA}}$ respectively. These are one dimensional class-specific activations that signify classification scores over time for both the types of actions for each segment (as illustrated in Fig. 2). These class-specific distributions, along with the class-agnostic distributions are used to predict the triplets $\langle s^{\text{IA}}, e^{\text{IA}}, c^{\text{IA}} \rangle$ and $\langle s^{\text{UA}}, e^{\text{UA}}, c^{\text{UA}} \rangle$ associated with the goal-directed and unintentional activities respectively.

3.2. Video Encoder

In order to learn a joint representation for inferring the goal-directed and unintentional actions, we use a Bidirectional Gated Recurrent Unit [6] as the video encoder. 3D-CNN architectures like R(2+1)D [49] and I3D [4] capture very short clip level information. However, capturing information which helps discriminate between the goal-directed region and an unintentional region requires longer temporal context which can be modeled by a GRU. Specifically, our GRU consists of a reset gate r which controls how much importance to give the previous hidden state h^{t-1} in order to calculate the current hidden state h^{t}, and an update gate u which determines how much of the previous hidden state h^{t-1} should be carried on to the current hidden state h^{t}. Given the backbone feature $\mathbf{X}$, we compute the hidden state at each time-step t using the following equations:

$$\begin{aligned}
z^t &= \sigma(\mathbf{W}^z \mathbf{X}^t + \mathbf{U}^z h^{t-1)}) && \text{Update Gate} \\
r^t &= \sigma(\mathbf{W}^r \mathbf{X}^t + \mathbf{U}^r h^{t-1}) && \text{Reset Gate} \\
\tilde{h}^t &= \tanh(r^t \cdot \mathbf{U} h^{t-1} + \mathbf{W}\mathbf{X}^t) && \text{New Memory} \\
h^t &= (1 - z^t) \cdot \tilde{h}^t + z^t \cdot h^{t-1}, && \text{Hidden State}
\end{aligned} \tag{2}$$

where $\mathbf{U}$ and $\mathbf{W}$ correspond to learnable parameters of this module. In order to capture the forward information flow

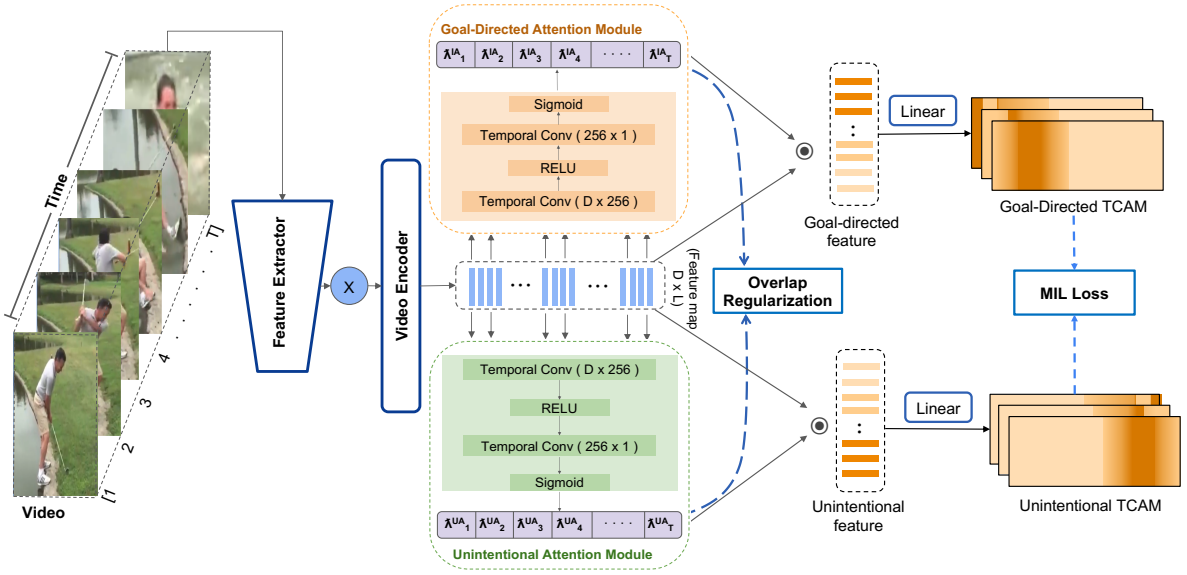

Figure 2. Illustration of our overall architecture. A backbone feature extractor is used to convert raw videos into features, *i.e.*, $\mathbf{X}$ and is kept frozen throughout the training process. $\mathbf{X}$ is then passed to a video encoder which can be either a GRU [6] or a Transformer Encoder [53], to extract high level features $\mathbf{O}$. The two attention modules use $\mathbf{O}$ to predict the bottom-up attention weights λ^{IA} and λ^{UA} for the goal-directed and unintentional action respectively, which are used for the Overlap Regularization. We calculate the goal-directed, *i.e.*, $\mathbf{O}^{\text{IA}}$ and unintentional feature, *i.e.*, $\mathbf{O}^{\text{UA}}$ by computing a dot product between $\mathbf{O}$ and their respective bottom-up attention weights. Finally we pass the goal-directed and unintentional feature through weight-shared linear layers to extract their respective TCAMs $\mathbf{C}^{\text{IA}}$ and $\mathbf{C}^{\text{UA}}$. These TCAMs are used for the MIL Loss.

$\overrightarrow{h^{(t)}}$ as well as backward information flow $\overleftarrow{h^{(t)}}$ we use a Bidirectional-GRU and obtain the final representation $\mathbf{O}$ by concatenating these features from the final hidden layer.

3.3. Multiple Instance Learning Loss

Following previous works in weakly supervised action localization [28,31,34], we use the k-max Multiple Instance Learning (MIL) [64] loss function for classifying the goal-directed and unintentional activities in the video. For each video, we average out the top-k elements of the TCAMs, *i.e.*, $\mathbf{C}^{\text{IA}}$ and $\mathbf{C}^{\text{UA}}$ along the temporal axis for each class to obtain the video-level classification scores $A^{\text{IA}} \in \mathbb{R}^{N_{IA}}$ and $A^{\text{UA}} \in \mathbb{R}^{N_{UA}}$. Here, k is set by $\left[\frac{l}{s}\right]$ where s is a hyper-parameter that regulates the number of clips to consider when making the classification. We then apply a softmax function over class scores, in order to obtain a probability mass function (pmf) over the goal-directed as well as unintentional classes, *i.e.*, p^{IA} and p^{UA}. Let y^{IA} and y^{UA} be the ground truth label vectors for a video. We then l_1-normalize them to obtain ground-truth pmfs q^{IA} and q^{UA}. Finally we conduct cross entropy between these two.

$$\begin{aligned}\mathcal{L}_{cls}^{\text{IA}} &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N_{\text{IA}}} -q_i^{\text{IA}}(j)\log\left(p_i^{\text{IA}}(j)\right)\\ \mathcal{L}_{cls}^{\text{UA}} &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N_{\text{UA}}} -q_i^{\text{UA}}(j)\log\left(p_i^{\text{UA}}(j)\right),\end{aligned} \tag{3}$$

where N corresponds to the total number of training videos, and the final loss is the combination of them: $\mathcal{L}_{cls} = \mathcal{L}_{cls}^{\text{IA}} + \mathcal{L}_{cls}^{\text{UA}}$.

3.4. Overlap Regularization

Let $\lambda_t^{\text{IA}}, \lambda_t^{\text{UA}} \in [0,1]\ \forall\ t \in [1,l]$ be the bottom-up attention weights for the goal-directed actions and unintentional action respectively, obtained from the respective attention modules. λ_t signifies the temporal attention weight for a clip t. During training, a trivial solution which could be learned by the model is to pay attention to the entire video when inferring the goal-directed and unintentional action, *i.e.*, $\lambda_t^{IA}, \lambda_t^{UA} = 1\ \forall\ t \in [1,l]$, though these actions take place at two distinct sections of the video. Simply applying the MIL loss cannot guarantee that such distinctions can be learnt from the data as shown in Section 4.4. We solve this problem by appending an additional regularization term on the overlap of these attention weights:

$$\begin{aligned}\mathcal{L}_{IA} &= \max\Big(0, \frac{\sum_r^{N_{T^{\text{UA}}}} \lambda_{T_r^{\text{UA}}}^{IA}}{N_{T^{\text{UA}}}} - \frac{l}{p}\Big)\\ \mathcal{L}_{\text{UA}} &= \max\Big(0, \frac{\sum_r^{N_{T^{\text{IA}}}} \lambda_{T_r^{\text{IA}}}^{\text{UA}}}{N_{T^{\text{IA}}}} - \frac{l}{p}\Big)\\ \mathcal{L}_{overlap} &= \mathcal{L}_{\text{IA}} + \mathcal{L}_{\text{UA}},\end{aligned} \tag{4}$$

where T^{IA} and T^{UA} are the set of temporal indices of the bottom-up goal-directed and unintentional attention

weights at which they are more than a predefined threshold. $N_{T^{\text{IA}}}$ and $N_{T^{\text{UA}}}$ are the lengths of the sets of activated temporal indices. p is a design parameter which controls the amount of allowed overlap between the attention maps. Lower the value of p, lower the penalization of overlaps.

In the goal-directed as well as unintentional regions of the video, the attention weights should ideally be low at the borders of their respective ground truth region and high towards the center of this region. Hence we view $\lambda_{\text{IA}}, \lambda_{\text{UA}}$ as Gaussian distributions $\mathbf{P}_{\text{IA}} \sim \mathcal{N}(\mu_{\text{IA}}, \sigma^2_{\text{IA}})$ and $\mathbf{P}_{\text{UA}} \sim \mathcal{N}(\mu_{\text{UA}}, \sigma^2_{\text{UA}})$. Every unintentional action begins with an agent performing a goal-directed action in order to achieve it's goal, which then gets disrupted and transitions into an unintentional action. Using this prior that a goal-directed action transitions into an unintentional action, we need to ensure $\mu_{\text{IA}} < \mu_{\text{UA}}$. We approach this by formulating the following regularization:

$$\begin{aligned} \mu_{\text{IA}} &= \frac{\sum_{t=1}^{l} P_t^{\lambda_{\text{IA}}} \cdot t}{\sum_{t=1}^{l} P_t^{\lambda_{\text{IA}}}} \\ \mu_{\text{UA}} &= \frac{\sum_{t=1}^{l} P_t^{\lambda_{\text{UA}}} \cdot t}{\sum_{t=1}^{l} P_t^{\lambda_{\text{UA}}}} \\ \mathcal{L}_{order} &= \texttt{max}(0, \frac{\mu_{\text{IA}} - \mu_{\text{UA}}}{l} + \frac{l}{q}), \end{aligned} \tag{5}$$

where $P^{\lambda_{\text{IA}}}$ and $P^{\lambda_{\text{UA}}}$ are probability distributions obtained by applying softmax over the temporal axis of λ_{IA} and λ_{UA} respectively. q is a design parameter that helps control the margin by which μ_{UA} has to be greater than μ_{IA}. Our model is end-to-end trained with the overall loss as follows:

$$\mathcal{L} = \lambda L_{cls} + (1 - \lambda)(\mathcal{L}_{overlap} + \mathcal{L}_{order}), \tag{6}$$

where λ is the weighting hyper-parameter that controls the trade-off between MIL loss and overlap regularization.

3.5. Classification and Localization

After training our network, we use it to classify goal-directed and unintentional actions as well as localize the regions in which they occur. For a single video, after obtaining the pmf p^{IA} and p^{UA} over each of the classes, as mentioned in Section 3.3, we use mean average precision (mAP) to conduct evaluation for the classification task. For localization of the goal-directed and unintentional regions, we consider only categories having classification scores *i.e.*, A^{IA} and A^{UA} above 0. For each of these categories, we first scale the respective TCAM outputs to [0,1] using a Sigmoid function and weight these using the bottom-up attention weights. This can be formally expressed by:

$$\begin{aligned} \psi^{\text{IA}}(c^{\text{IA}}) &= \lambda_{\text{IA}} \cdot \texttt{Sigmoid}(C^{\text{IA}}(c^{\text{IA}})) \quad c^{\text{IA}} \in [1, N_{\text{IA}}], \\ \psi^{\text{UA}}(c^{\text{UA}}) &= \lambda_{\text{UA}} \cdot \texttt{Sigmoid}(C^{\text{UA}}(c^{\text{UA}})) \quad c^{\text{UA}} \in [1, N_{\text{UA}}], \end{aligned} \tag{7}$$

where $\psi^{\text{IA}}(c^{\text{IA}}), \psi^{\text{UA}}(c^{\text{UA}}) \in \mathbb{R}^l$ are the weighted TCAMs, for the respective classes c^{IA} and c^{UA}. We finally threshold $\psi^{\text{IA}}(c^{\text{IA}})$ and $\psi^{\text{UA}}(c^{\text{UA}})$ to obtain the triplets $\langle s^{\text{IA}}, e^{\text{IA}}, c^{\text{IA}} \rangle$ and $\langle s^{\text{UA}}, e^{\text{UA}}, c^{\text{UA}} \rangle$.

4. Experiments

4.1. Dataset & Implementations

Data Preparation. The original Oops Dataset [11] consists of 20,338 videos containing unintentional human actions obtained by collating "*fail*" videos from different users on YouTube. Amazon Mechanical Turk workers are then asked to label the time at which the video starts transitioning from the goal-directed action to the unintentional action, as well as indicate whether a video does not indicate an unintentional action.

In order to create our dataset, which is built upon the labeled portion of the Oops dataset, we follow a similar pre-processing step as in [11] by removing those videos that 1). Do not contain an unintentional action 2). Are More than 30 seconds which are likely to contain multiple scenes, as well as those less than 3 seconds which are not likely to contain one full scene 3). Where the transition time occurs in the initial/ending 1% of the video, since there would not be enough context to understand the goal-directed/unintentional action respectively. Post this process, we were left with a total of about 7,800 labeled videos.

The authors of [11] also provide annotations in the form of natural language descriptions, which were obtained by asking Amazon Mechanical Turkers to watch the video and answer: "*what was the goal?*" and "*what went wrong?*". Since we want to collect a distinct set of goal-directed and unintentional actions, we followed a technique similar to the Epic Kitchens Dataset [9], by extracting the verbs and associated noun using the SpaCy[1] dependency parser and concatenating them to form an action. We replace all compound nouns by it's second noun: *e.g.*, "*ride mountain bike*" is replaced with "*ride bike*" and so on. Due to the diversity of the worker's vocabulary, we find that the resulting actions are of low quality, with many of them having ambiguous meanings, *i.e.*, "*fly bike*" as well as redundant meanings. In order to overcome this, we manually go over each of these extracted actions and remove those with ambiguous meanings as well as merge the redundant ones, *i.e.*, "*jump over fence*" and "*jump over chair*" into a more general "*jump over obstacle*" category. We finally carry out a human evaluation, going through all the videos manually and ensuring the correctness of the labels, and correcting them if need be. We also give the evaluator an option to discard the video if the goal of the actor was ambiguous. We build an annotation tool in order to make this process easier (refer to the Appendix for details). Finally, we keep a threshold of 15 for the number of videos per goal-directed and unintentional action class, discarding all classes below this threshold, as

[1] https://spacy.io/

Model	Feature	Segment	mAP @ IoU 0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	AVG.
STPN [33]	R(2+1)D	Goal	44.9	41.7	33.0	25.7	18.3	10.0	5.0	3.7	1.2	20.4
		UnInt	30.9	26.6	21.8	15.7	9.9	5.2	1.8	1.00	0.1	12.5
WTALC [34]	R(2+1)D	Goal	45.1	41.8	36.1	28.9	22.8	15.9	10.4	8.1	2.0	23.5
		UnInt	25.5	21.2	15.3	12.6	7.7	4.3	2.3	1.0	0.5	10.1
A2CL-PT [31]	R(2+1)D	Goal	41.1	38.8	34.3	28.4	23.9	16.6	10.9	8.4	2.5	22.8
		UnInt	30.2	24.2	19.8	14.2	8.6	5.0	1.8	0.6	0.1	11.6
Ours	R(2+1)D	Goal	45.3	45.1	44.0	41.8	39.1	29.5	21.9	13.9	3.5	31.6
		UnInt	34.6	33.4	28.4	23.6	19.5	15.0	10.0	3.4	1.0	18.8
STPN [33]	I3D	Goal	44.8	42.8	34.9	27.8	19.9	11.1	6.1	4.0	1.6	21.5
		UnInt	36.3	31.3	26.1	19.5	13.0	6.8	1.7	0.6	0.02	15.0
WTALC [34]	I3D	Goal	38.8	36.4	30.4	26.3	18.6	13.1	7.2	4.5	1.8	19.7
		UnInt	22.9	18.4	14.2	11.0	6.8	3.6	1.2	0.5	0.1	8.8
A2CL-PT [31]	I3D	Goal	38.1	36.7	31.8	26.6	22.7	17.6	12.5	9.0	4.9	22.2
		UnInt	32.4	26.1	21.6	15.3	9.9	5.2	1.6	0.7	0.1	12.5
Ours	I3D	Goal	**51.5**	**51.3**	**49.9**	**44.9**	**41.1**	**32.5**	**24.3**	**14.4**	**5.0**	**35.0**
		UnInt	**39.4**	**39.0**	**36.4**	**32.2**	**30.0**	**26.6**	**17.6**	**10.2**	**2.8**	**26.0**

Table 1. Performance comparison of our model with competitive weakly supervised action localization (WSAL) models. We adjust the WSAL models by attaching two classification heads to compute two TCAMs (for the goal-directed and unintentional action). We then retrain it on our dataset (W-Oops). We can see that our model significantly outperforms the other methods.

Architecture	Feature	GOAL cMAP	UNINT. cMAP
Chance	-	2.7	3.3
STPN	R(2+1)D	44.0	32.6
WTALC	R(2+1)D	48.5	37.5
A2CL-PT	R(2+1)D	46.6	32.6
Ours	R(2+1)D	50.5	38.4
STPN	I3D	45.3	37.5
WTALC	I3D	50.2	38.2
A2CL-PT	I3D	48.5	34.8
Ours	I3D	52.6	41.1

Table 2. Mean average precision of activity classification results using different methods. First row shows the mAP of random chance.

$\mathcal{L}_{cls}$	$\mathcal{L}_{order}$	$\mathcal{L}_{overlap}$	SEG.	mAP @ IoU 0.3	0.5	0.9	AVG.
✓	-	-	Goal	34.7	17.6	0.9	21.2
			UnInt	31.1	14.4	0.1	17.4
✓	✓	-	Goal	46.3	35.2	2.7	30.1
			UnInt	31.7	17.9	0.7	19.0
✓	✓	✓	Goal	49.9	41.1	5.0	35.0
			UnInt	36.4	30.0	2.8	26.0

Table 3. Ablation study on contributions of different losses in our model.

well as the videos associated with these classes. This leaves us with 44 goal-directed and 30 unintentional classes. We provide detailed statistics and analysis of the dataset in the Appendix.

Implementation Details. We extract RGB features by creating chunks of 16 consecutive and non-overlapping frames and using the I3D [4] as well as R(2+1)D [49] pretrained architectures to extract clip-level features from these chunks (details provided in the Appendix). This backbone feature extractor is kept frozen throughout the entire training process. The kernel-size of all the 1-D convolutional layers for the bottom-up attention modules are set to 1. The learning rate and loss weighting function λ is set to 10^{-3} and 0.8 respectively. We set the MIL loss hyper-parameter s to 3. The parameters of the Overlap Regularization, p and q, are set to 1000 and 10 respectively. Finally we set the number of layers of our bidirectional GRU to 3. Our network is implemented and trained on a machine with a single Tesla X Pascal GPU for 10,000 iterations using the Adam Optimizer [22] with a batch size of 16.

4.2. Goal-directed/Unintent. Action Localization

Our model should be able to focus on the correct regions of the video in order to infer the goal-directed and unintentional action segments, hence understanding the transition between these two. In order to evaluate our model on the task of localizing goal-directed as well as unintentional segments, we follow the standard evaluation protocol for temporal localization tasks by calculating the mean average precision (mAP) over different intersection over union (IoU) thresholds for both the types of actions. Since there are no quantitative results reported on our dataset, we use competitive models from the traditional weakly supervised action localization task as baselines. Since these models are trained using only one classification head which is used to identify the atomic actions in the video, we repurpose these models by adding an additional classification head (for the goal-directed and unintentional action) and bottom-up attention module (in the case of STPN [33]) or addi-

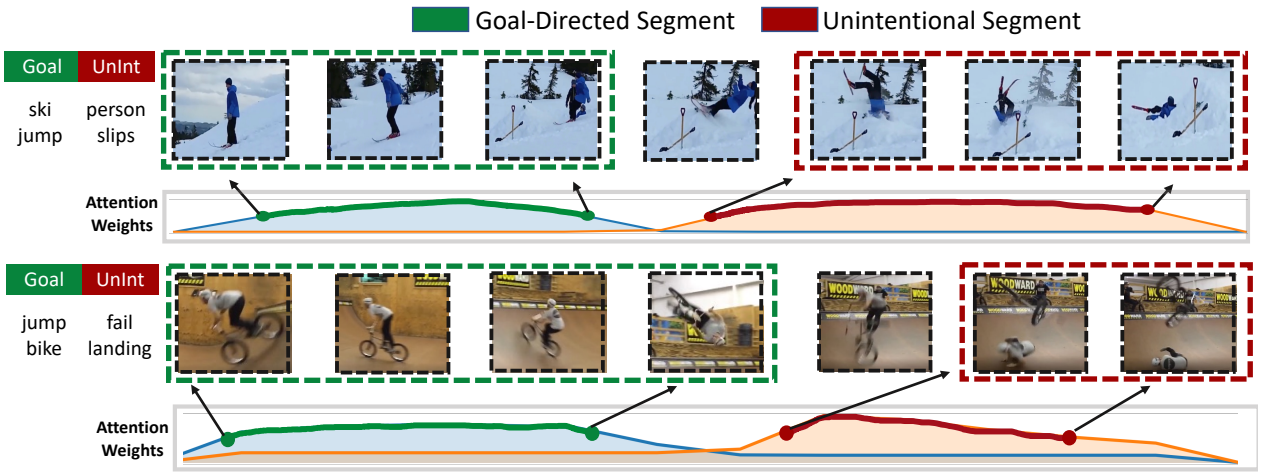

Figure 3. Our method is able to identify the temporal regions that correspond to goal-directed/unintentional activity via the produced weighted TCAMs. Blue and Orange attention maps correspond to the goal-directed action and unintentional action respectively.

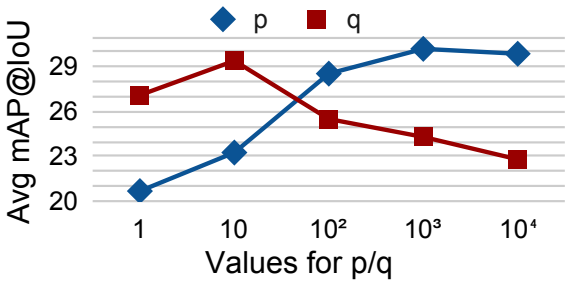

Figure 4. Effect on Average (Goal+UnInt) mAP@IoU for the goal-directed and unintentional action when changing p (blue) and q (red).

tional branch (in case of A2CL-PT [31]) to adapt it to our task. We then retrain these models on our dataset and report quantitative results for comparison in Tab. 1. It may be noted that our method performs significantly better than other weakly supervised methods on this task, when using the same backbone. For example, the average mAP@IoU score of our method outperforms A2CL-PT by 12.8% for the goal-directed action and 13.5% for the unintentional action, when using an I3D backbone. We conjecture that this localization improvement is due to our overlap regularization on the bottom-up attention weights since it enforces the model to focus on distinct portions of the action scene while ensuring the temporal order of the actions, which is a crucial property for solving this task. The qualitative results (see Appendix) show how the WSAL models focus on overlapping regions when inferring the goal-directed/unintentional action which reduces it's localization performance.

4.3. Goal-directed/Unintent. Action Classification

Given any video our model is trained to predict the goal-directed action as well as the unintentional action it eventually transitions into. Following previous works [33, 34], we use mean average precision (mAP) to evaluate the classification performance of our model on predicting the goal-directed action as well as unintentional action. We report our results in Tab. 2. It is interesting to note that our method performs the best on the classification task as well. For example, it performs 4.1% higher on the Goal cMAP and 6.3% higher on the Unintentional cMAP than A2CL-PT when using an I3D backbone.

4.4. Ablation Study

We conduct an ablation study to analyse various components of our model. We analyse the significance of the overlap regularization introduced in Section 3.4. We observe very clearly in Tab. 3 that only using $\mathcal{L}_{cls}$ is not sufficient to localize the goal-directed and unintentional actions, and our final model performs the best. This implies that all components are necessary in order to achieve the best performance and each one is effective. We further analyse the importance of the hyper-parameters p and q used in the overlap regularization in Fig. 4. We can see that increasing p from 1 to 10^3 results in a significant increase in the average mAP@IoU. This shows that the localization performance increases by penalizing the overlap of the bottom-up attentions more, but plateaus after the 10^3 mark. Analysing the q hyper-parameter, we notice that increasing the value of q decreases the performance. Since increasing the value of q results in a lower margin of separation between the expectations of the goal-directed and unintentional bottom-up attention weights, we can conclude that a lower value of q, *i.e.*, higher margins of separation helps achieve a better localization performance. However, $q = 1$ signifies the extreme case when the margin is equal to the length of the clips, forcing the goal-directed and unintentional atten-

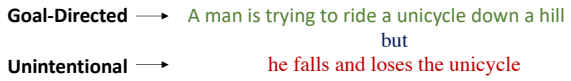

Figure 5. We form the ground truth caption by concatenating the goal-directed and unintentional caption with a *but*.

tion maps to be at two separate ends of the temporal axis, thereby hurting the performance. Fig. 3 shows qualitative examples of localizing the goal-directed and unintentional segments on our W-Oops dataset. We further provide more qualitative examples in the Appendix, which compare our method with previous WSAL methods.

4.5. Video Captioning

Teleological understanding helps explain the action better especially when the goal is partially/not achieved. In order to further verify whether our localization module can assist in teleological understanding, we conduct a captioning experiment which involves explaining the goal-directed and unintentional parts of the video through natural language descriptions. We obtain ground truth captions for the goal-directed and unintentional regions from [11]. The goal-directed and unintentional captions are concatenated (shown in Fig. 5) to form the ground truth caption for the entire video. We use an 80%-20% train/test split obtaining $3,200$ training samples and 800 test samples. Our experimental setup is divided into two parts:

- Train a state-of-the-art video captioner on the entire video and the whole ground truth caption.
- Automatically split the video into the goal-directed and unintentional regions (inferred from their respective attention maps) using our trained localization module. Train two video captioners with distinct weights, to caption the goal-directed and unintentional regions with their respective captions as training signals. Finally, concatenate the two outputted captions by a *but* to form the final predicted caption (during inference). Pipeline shown in Fig. 6.

We use RMN [48] as our captioning module, and extract features in the same manner as mentioned by the authors. Note that the localization module is kept frozen throughout the training process. From Tab. 4, we observe that the results obtained after splitting the video into its goal-directed and unintentional region outperform those when trained on the entire video. In particular, our method with transition localizer significantly outperforms the baseline without localizer by 5.1 SMURF [16] scores and 4.1 CIDEr [54] scores. Notably, SMURF [16] is a caption evaluation algorithm that incorporates diction quality into its evaluation, which demonstrates SOTA correlation with human judgment and improved explainability. Through SMURF, we observe that

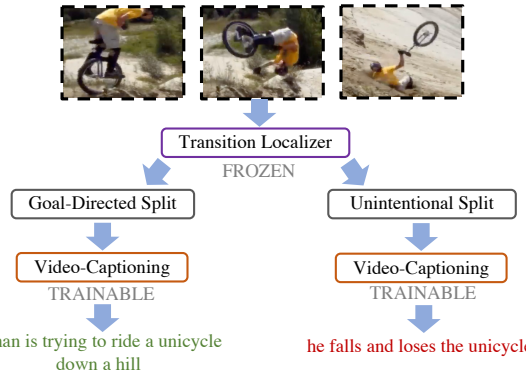

Figure 6. Pipeline for the captioning experiment (using our localization module).

Exp.	R [26]	M [1]	C [54]	S [16]
Without Loc.	16.7	37.7	29.8	14.1
With Loc.	**17.0**	**38.0**	**33.9**	**19.8**

Table 4. Captioning results with or without localization module. R, M, C, S denote ROUGE_L, METEOR, CIDEr and SMURF metrics respectively for video captioning evaluations.

our method improves semantic performance while maintaining the descriptiveness of terms used in the sentence. These metrics clearly show that the teleological ability of our localization module helps a captioning module output more accurate captions on videos containing unintentional actions, which is achieved by the precise capture of unintentional activity in video. We showcase more qualitative examples in the Appendix.

5. Conclusion

In this paper, we propose W-Oops, an augmented unintentional human activity dataset that consists of both goal-directed and unintentional video-level activity annotations, built upon Oops [11]. We consider a weakly supervised task to infer the respective classes as well as the temporal regions in which they occur using only the video-level activity annotations. We further build a neural network architecture which employs a novel overlap regularization on top of the bottom-up attention weights outputted by our attention module, which helps the model focus on distinct parts of the video while maintaining the temporal ordering of these actions when inferring the temporal regions. We conclude from our experiments that out method significantly outperform previous WSAL baselines on our benchmark. The video captioning experiment further verifies the teleological ability of our localization module, which points a promising future research direction for improving captioning quality through teleological analysis.

Acknowledgement. This work was supported by the National Science Foundation under Grant CMMI-1925403, IIS-2132724 and IIS-1750082.

References

[1] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 8

[2] Amanda C Brandone and Henry M Wellman. You can't always get what you want: Infants understand failed goal-directed actions. *Psychological science*, 20(1):85–91, 2009. 1

[3] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. Sst: Single-stream temporal action proposals. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2911–2920, 2017. 2

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1, 3, 6

[5] Zhenfang Chen, Lin Ma, Wenhan Luo, Peng Tang, and Kwan-Yee K Wong. Look closer to ground better: Weakly-supervised temporal grounding of sentence in video. *arXiv preprint arXiv:2001.09308*, 2020. 2

[6] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 3, 4

[7] Gergely Csibra. Goal attribution to inanimate agents by 6.5-month-old infants. *Cognition*, 107(2):705–717, 2008. 1

[8] Gergely Csibra and György Gergely. 'obsessed with goals': Functions and mechanisms of teleological interpretation of actions in humans. *Acta psychologica*, 124(1):60–78, 2007. 1

[9] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 5

[10] Mahdi Davoodikakhki and KangKang Yin. Hierarchical action classification with network pruning. In *International Symposium on Visual Computing*, pages 291–305. Springer, 2020. 1

[11] Dave Epstein, Boyuan Chen, and Carl Vondrick. Oops! predicting unintentional action in video. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 5, 8

[12] Zhiyuan Fang, Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Video2commonsense: Generating commonsense descriptions to enrich video captioning. *Conference on Empirical Methods in Natural Language Processing*, 2020. 2

[13] Zhiyuan Fang, Shu Kong, Charless Fowlkes, and Yezhou Yang. Modularized textual grounding for counterfactual resilience. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6378–6388, 2019. 2

[14] Zhiyuan Fang, Shu Kong, Zhe Wang, Charless Fowlkes, and Yezhou Yang. Weak supervision and referring attention for temporal-textual association learning. *arXiv preprint arXiv:2006.11747*, 2020. 2

[15] Zhijie Fang and Antonio M López. Intention recognition of pedestrians and cyclists by 2d pose estimation. *IEEE Transactions on Intelligent Transportation Systems*, 21(11):4773–4783, 2019. 2

[16] Joshua Feinglass and Yezhou Yang. Smurf: Semantic and linguistic understanding fusion for caption evaluation via typicality analysis. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2021. 8

[17] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. *CoRR*, abs/1905.09035, 2019. 2

[18] György Gergely and Gergely Csibra. Teleological reasoning in infancy: The naıve theory of rational action. *Trends in cognitive sciences*, 7(7):287–292, 2003. 1

[19] M. Hoai and F. De la Torre. Max-margin early event detectors. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2863–2870, 2012. 2

[20] M Esat Kalfaoglu, Sinan Kalkan, and A Aydin Alatan. Late temporal modeling in 3d cnn architectures with bert for action recognition. In *European Conference on Computer Vision*, pages 731–747. Springer, 2020. 1

[21] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4405–4413, 2017. 2

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[23] Pilhyeon Lee, Youngjung Uh, and Hyeran Byun. Background suppression network for weakly-supervised temporal action localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11320–11327, 2020. 2

[24] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. What is more likely to happen next? video-and-language future event prediction. *arXiv preprint arXiv:2010.07999*, 2020. 2

[25] Zhaoqun Li, Cheng Xu, and Biao Leng. Angular triplet-center loss for multi-view 3d shape retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8682–8689, 2019. 3

[26] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 8

[27] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019. 1

[28] Daochang Liu, Tingting Jiang, and Yizhou Wang. Completeness modeling and context separation for weakly supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1298–1307, 2019. 2, 4

[29] Minuk Ma, Sunjae Yoon, Junyeong Kim, Youngjoon Lee, Sunghun Kang, and Chang D Yoo. Vlanet: Video-language alignment network for weakly-supervised video moment retrieval. In *European Conference on Computer Vision*, pages 156–171. Springer, 2020. 2

[30] A. Miech, I. Laptev, J. Sivic, H. Wang, L. Torresani, and D. Tran. Leveraging the present to anticipate the future in videos. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2915–2922, 2019. 2

[31] Kyle Min and Jason J Corso. Adversarial background-aware loss for weakly-supervised temporal activity localization. In *European Conference on Computer Vision*, pages 283–299. Springer, 2020. 2, 3, 4, 6, 7

[32] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K. Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[33] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6752–6761, 2018. 2, 3, 6, 7

[34] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 563–579, 2018. 2, 3, 4, 6, 7

[35] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 563–579, 2018. 2

[36] A. Rasouli, I. Kotseruba, and J. K. Tsotsos. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 206–213, 2017. 2

[37] Amir Rasouli, Mohsen Rohani, and Jun Luo. Pedestrian behavior prediction via multitask learning and categorical interaction modeling. *arXiv preprint arXiv:2012.03298*, 2020. 2

[38] M. S. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *2011 International Conference on Computer Vision*, pages 1036–1043, 2011. 2

[39] Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Basura Fernando, Lars Petersson, and Lars Andersson. Encouraging lstms to anticipate actions very early. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 280–289, 2017. 2

[40] Baifeng Shi, Qi Dai, Yadong Mu, and Jingdong Wang. Weakly-supervised action localization by generative attention modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1009–1019, 2020. 2, 3

[41] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5734–5743, 2017. 2

[42] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 154–171, 2018. 2

[43] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1049–1058, 2016. 2

[44] Jessica A Sommerville and Amanda L Woodward. Pulling out the intentional structure of action: the relation between action processing and action production in infancy. *Cognition*, 95(1):1–30, 2005. 1

[45] Yijun Song, Jingwen Wang, Lin Ma, Zhou Yu, and Jun Yu. Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos. *arXiv preprint arXiv:2003.07048*, 2020. 2

[46] Chen Sun, Sanketh Shetty, Rahul Sukthankar, and Ram Nevatia. Temporal localization of fine-grained actions in videos by domain transfer from web images. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 371–380. ACM, 2015. 2

[47] Stuart Synakowski, Qianli Feng, and Aleix Martinez. Adding knowledge to unsupervised algorithms for the recognition of intent. *International Journal of Computer Vision*, Jan 2021. 2

[48] Ganchao Tan, Daqing Liu, Meng Wang, and Zheng-Jun Zha. Learning to discretely compose reasoning module networks for video captioning. *arXiv preprint arXiv:2007.09049*, 2020. 8

[49] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 1, 3, 6

[50] Du Tran and Junsong Yuan. Max-margin structured output regression for spatio-temporal action localization. In *Advances in neural information processing systems*, pages 350–358, 2012. 2

[51] Vinh Tran, Yang Wang, and Minh Hoai. Back to the future: Knowledge distillation for human action anticipation. *CoRR*, abs/1904.04868, 2019. 2

[52] D. Varytimidis, F. Alonso-Fernandez, B. Duran, and C. Englund. Action and intention recognition of pedestrians in urban traffic. In *2018 14th International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*, pages 676–682, 2018. 2

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 3, 4

[54] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 8

[55] Carl Vondrick, Deniz Oktay, Hamed Pirsiavash, and Antonio Torralba. Predicting motivations of actions by leveraging text. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2997–3005, 2016. 2

[56] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating the future by watching unlabeled video. *CoRR*, abs/1504.08023, 2015. 2

[57] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 1

[58] P. Wei, Y. Liu, T. Shu, N. Zheng, and S. Zhu. Where and why are they looking? jointly inferring human attention and intentions in complex tasks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6801–6809, 2018. 2

[59] Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Learning to track for spatio-temporal action localization. In *Proceedings of the IEEE international conference on computer vision*, pages 3164–3172, 2015. 2

[60] Amanda L Woodward, Jessica A Sommerville, and Jose J Guajardo. How infants make sense of intentional action. *Intentions and intentionality: Foundations of social cognition*, pages 149–169, 2001. 1

[61] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[62] Yuanhao Zhai, Le Wang, Wei Tang, Qilin Zhang, Junsong Yuan, and Gang Hua. Two-stream consensus network for weakly-supervised temporal action localization. In *European Conference on Computer Vision*, pages 37–54. Springer, 2020. 2, 3

[63] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018. 1

[64] Zhi-Hua Zhou. Multi-instance learning: A survey. *Department of Computer Science & Technology, Nanjing University, Tech. Rep*, 2, 2004. 4