Guess Future Anomalies from Normalcy: Forecasting Abnormal Behavior in Real-World Videos

Snehashis Majhi[1,2,*], Mohammed Guermal[1,2,*], Antitza Dantcheva[1,2], Quan Kong[3], Lorenzo Garattoni[4], Gianpiero Francesca[4], François Brémond[1,2]

[1] INRIA [2] Côte d'Azur University [3] Woven by Toyota [4] Toyota Motor Europe

* Joint first authors.

Abstract

Forecasting Abnormal Human Behavior (AHB) aims to predict unusual behavior in advance by analyzing early patterns of normal human interactions. Unlike typical action prediction methods, this task focuses on observing only normal interactions to predict both, short and long term future abnormal behavior. Despite its affirmative impact on society, AHB prediction remains under-explored in current research. This is primarily due to the challenges involved in anticipating complex human behaviors and interactions with surrounding agents in real-world situations. Further, there exists an underlying uncertainty between the early normal patterns and the future abnormal behaviour, thereby making the prediction harder. To address these challenges, we introduce a novel transformer model that improves early interaction modeling by accounting for uncertainties in both, observations and future outcomes. To the best of our knowledge, we are the first to explore the task. Therefore, we present a new comprehensive dataset referred to as "AHB-F"[†] , which features real-world scenarios with complex human interactions. The AHB-F has a deterministic evaluation protocol that ensures only normal frames to be observed for long- and short-term future prediction. We extensively evaluate and compare competitive action anticipation methods on our benchmark. Our results show that our method consistently outperforms existing action anticipation approaches, both in quantitative and qualitative evaluations.

1. Introduction

Abnormal human behavior (AHB) has the ability to cause harm for humans and associated property. Such behavior is rare and often comprises complex patterns, making them difficult to understand. Recently, video analysis methods [5, 34, 41, 46] have focused on detecting anomalies either in offline (after they occur) or online (as they happen) mode to assist with investigations or provide alerts.

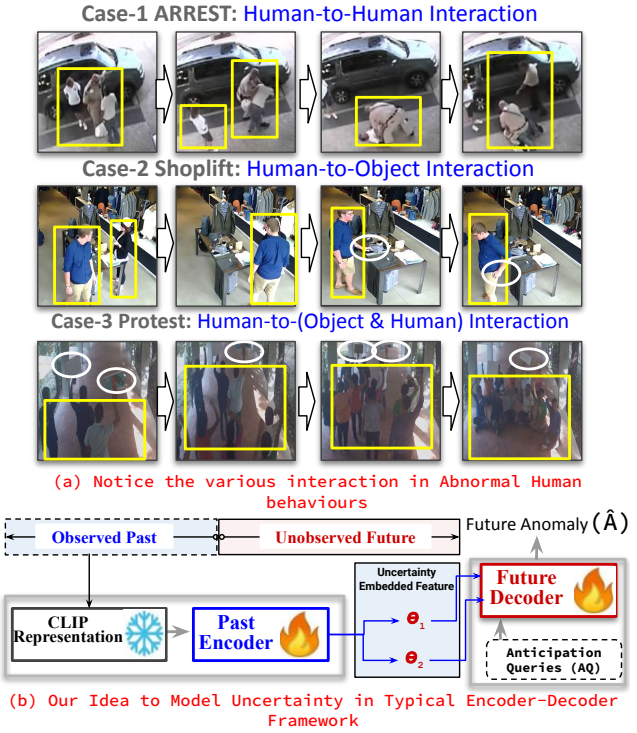

Figure 1. **In (a):** Illustrates three interaction cases in abnormal human behavior that can be challenging due to divergent cues. Abnormal human-to-human interactions (e.g. arrest), usually has significant changes in appearance and motion, while human-to-object interactions (e.g. shoplifting) tend to be more subtle. However, some interactions (e.g.protest) have both humans and objects with a unique spatio-temporal blend. **In (b):** Our typical framework that accounts the uncertainty while predicting the future anomaly.

However, these methods do not predict the anomalies before they happen, thereby fail to facilitate any anomaly preventive measures. Therefore, AHB forecasting/anticipating in real-world scenarios has a high societal impact, carrying the premise to minimize casualties and damages through mitigatory measures.

Forecasting abnormal human behavior (AHB) incorporates two main challenges: (I) understanding complex AHB, and (II) overcoming uncertainty between current observations and future events. First, real-world scenarios are

[†]Code, Models, Dataset: https://github.com/snehashismajhi/AHB-F

unpredictable and ever-changing, rendering it difficult to analyze complex human behavior and interactions for predicting abnormal behavior. These situations differ significantly from everyday life, where interactions are simpler. In abnormal scenarios, there is a wide range of interactions shown in Figure 1a, from subtle cues (such as shoplifting, which involves a few human-object interactions) to more intuitive ones (such as protests, where there occur dense human-to-human and human-to-object interactions). Abnormal behavior is often accompanied by normal activities and occurs rarely, which challenges prediction. Second, predicting future AHB based on normal interactions is naturally uncertain in both, short and long-term behavior. For instance, the same normal observation can result in multiple plausible continuations. Despite the importance of forecasting AHB, few methods exist that can handle these uncertainties and complexities in real-world, dynamic situations.

Motivated by this, we use an encoder-decoder framework shown in Figure 1b to process past interactions and predict future AHB. Our past encoder introduces a new transformer model, referred to as "Space-time Interaction aware Transformer (SIaT)", with two key components: (i) Interaction Modules (TIM/OIM) and (ii) Normalcy Uncertainty Latent Learner (NULL). These components helps for capturing early human interaction patterns and handle the uncertainty between normal interactions to future abnormal behavior. Deviating from previous methods [42, 45], our interaction modules separately encode scene-level temporal context (TIM) and object-level spatial interactions (OIM), providing a detailed understanding from broader scene dynamics to fine object interactions. We use panoptic object masks and raw RGB frames to represent objects and scenes level agents, thereby making both spatially coherent. This enables SIaT to effectively capture correlations between scene and object interactions. Next, the correlation encoded normal scene and object semantics are associated via NULL by considering the uncertainties associated with normal observations. The NULL module handles the uncertainty by distinguishing between relevant and non-relevant scene-object associations. Further, it adjusts the flow of information from the past encoder to the future decoder by learning latent features that is aligned with future predictions.

Towards validating our method, we find that widely adopted benchmark datasets [6, 16] entail activities of daily living and there is no benchmark that has real-world AHB. Addressing this limitation, we provide a larger-scale diversified dataset, denoted Abnormal Human Behaviour-Forecast (AHB-F) with a dedicated evaluation protocol for long and short-term anticipation. Thanks to the latter, our SIaT can portray its robustness towards AHB anticipation and related results suggest that our model consistently outperforms state-of-the-art action anticipation approaches.

To summarize, our contributions are in three-folds.

- We introduce a novel affirmative task named "Abnormal Human Behavior Anticipation" in real-world videos to promote mitigatory measures in serious crimes. Towards this, we present a benchmark dataset, "AHB-F" with dedicated evaluation protocol for long and short term anticipation.
- We propose a novel transformer "SIaT" that effectively encodes the early human interactions by considering uncertainty of observation and future predictions.
- We provide exhaustive experimental analysis to corroborate the robustness of SIaT in AHB-F dataset. The obtained results outline that SIaT outperforms previous approaches in many scenarios considered.

2. Related Work

Action Anticipation: This task has been investigated in both third-person videos [1, 2, 12, 30, 37] and egocentric videos [6, 7, 11, 14, 28]. Standard approaches are generally categorized into LSTM/RNN-based methods [7, 33] and transformer-based methods. LSTM-based approaches [10, 24] typically utilize a rolling LSTM to encode the observed video and maintain an updated summary. For inference, an unrolling LSTM is initialized with the hidden and cell states of the rolling LSTM to predict the next action. Often LSTM struggle with capturing long-horizon temporal dependencies. Recently, transformer-based approaches [14, 28] have gained attention, leveraging global attention mechanisms [13], modeling appearance changes in human-object interactions [32], conditioning on intention [23], and hierarchical feature aggregation [23]. Further, due to their ability to capture long-range dependencies, LSTR [42], TesTra [45], FUTR [13], OADTR [40], JOADAA [15] have benefited from the transformer backbones to address the tasks of action anticipation. However, these architectures are suitable only for simple activities and simple datasets, which is not applicable to real-world scenarios that have multiple actions occurring at the same time.

Video Anomaly Understanding: This task is majorly studied along the horizon of online and offline anomaly detection. Prominent methods are based on weakly-supervised multiple instance learning. The video anomaly detection (VAD) [5, 9, 18, 20–22, 26, 34, 38, 41, 43, 44, 46, 48] methods widely use UCF-Crime [34], XD-Violence [41], ShanghaiTech [19], IITB-Corridor [31] datasets. One crucial aspect in real-world VAD is to learn discriminative features for normal and anomaly segments. As a classical approach, authors in [44] and [48] utilize TCN [17] and optical flow motion [35] cues to capture the sharp anomalies only. In contrast, authors in [36] proposed an MTN

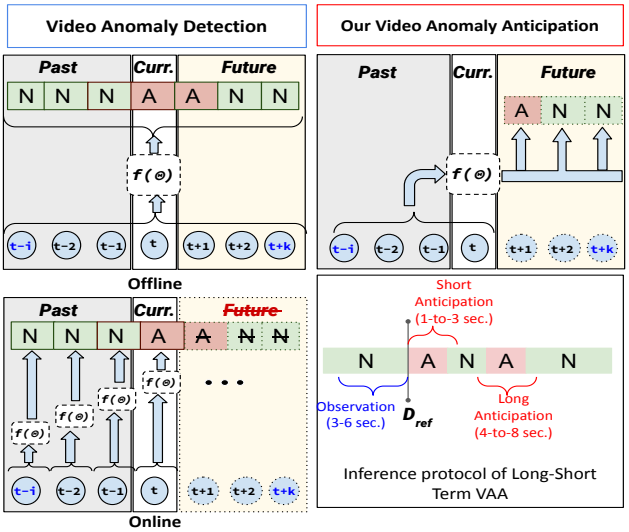

Figure 2. Illustration VAD Vs. VAA: Suppose the current time step is t. For **online VAD**, a parametrized model $f(\theta)$ can predict normal (N) or anomaly (A) for the current t based on observed time stamps $t-i \ldots t-1, t$, where i represents the observed duration. However, for **our VAA** we predict what kind of anomaly will occur in the future in a range of $[t+1, t+2, \ldots, t+k]$ where k represents anticipation duration. Further, we comprehend the short and long-term anticipation to identify the potential re-occurrence of an anomaly in the long future.

network for global context relation between normal and anomaly segments. Recently, Zhou *et al.* [47] and Chen *et al.* [3] adopt transformer-based global-local and focus-glance blocks, respectively, to capture long and short-term temporal dependencies in normal and anomalous videos. However, as [3, 36, 47] follow a magnitude-based optimization, they mostly encourage the sharp abnormal cues of scene anomalies. Thus, they tend to overlook the subtle cues of human anomalies and hence fail to detect them.

Despite decent progresses on action anticipation and video anomaly understanding tasks, prior methods have not explored real-world video anomaly anticipation. Motivated by this, in this work we aim to provide a new transformer method that extends the ability of traditional methods to anticipate anomalies in real-world scenarios. Further, to promote research in this domain, we provide a benchmark dataset and evaluation protocol to predict long and short future events.

3. Preliminaries

The objective of the video anomaly anticipation (VAA) task differs significantly from offline and online video anomaly detection (VAD) tasks. This is because either offline or online VAD can only provide anomaly prediction probability for a snippet that has already appeared or appearing currently. In contrast, VAA can answer uncertainties like: **(i) Whether an anomaly will occur in the near future? (Short Anticipation), (ii) If yes, What kind of anomaly is likely to occur? (anomaly class), (iii) Is there a chance of re-occurrence of the same anomaly in a future time window? (long Anticipation)**. We illustrate and compare VAD and our VAA tasks in Figure 2.

Note that classical daily living action anticipation methods can be applicable to VAA tasks. However, their approach to model the early/past trend may not directly benefit VAA due to the unique characteristics of real-world AHB. Therefore, in the next section we propose a novel transformer that can effectively encode the early trends of AHB and there by predicts the future abnormalities.

4. Proposed Method

In this section, we present our Space-time Interaction aware Transformer (SIaT) shown in Figure 3a that acts as a past-encoder to learn the early behavioral trends of humans via their interaction with the other environmental agents for anticipating AHB. This is accomplished by introducing scene and object-based vision-language representations into the SIaT via Feature Encoder. The SIaT has two key building blocks: (I) Interaction Module that constitutes two identical modules with different functionalities, namely Temporal Interaction Module (TIM) and Object Interaction Module (OIM) to dissociatively capture the scene-level global temporal interactions and object-level local spatial interactions respectively; (ii) Normal Uncertainty Latent Learner (NULL) associates the interaction encoded scene and object semantics by exploiting the inherent uncertainty associated with normal observation to future AHB. Next, we proceed to provide a concise description on the feature representation and then each building block of SIaT.

4.1. Feature Representation

For a given temporal observation duration (t), we extract scene and object features from the Scene Encoder (SE) and Object Encoder (OE). The **OE** first extracts the frame-level panoptic masks with the corresponding text labels of k objects from Mask2former [4] and stacks them along the temporal dimension (t). Then we extract object-level $d0$ dimensional vision language features from CLIP [29] Image and Text encoder to obtain $F_O \in \mathbb{R}^{t \times k \times d0}$ and $F_{txt} \in \mathbb{R}^{t \times k \times d0}$ respectively. Then the **SE** extracts $d0$ dimensional frame-level spatial features from CLIP [29] Image encoder and stacks them along the t dimension to obtain a global scene feature map $F_S \in \mathbb{R}^{t \times d0}$.

4.2. Interaction Modules (T/O-IM)

The aim of interaction modules (T/O-IM) (shown in Figure 3a) is to learn low-level discriminative representations for future AHB w.r.t normal ongoing events by effectively encoding the global and local interactions in the temporal interaction module (T-IM) and object interaction module

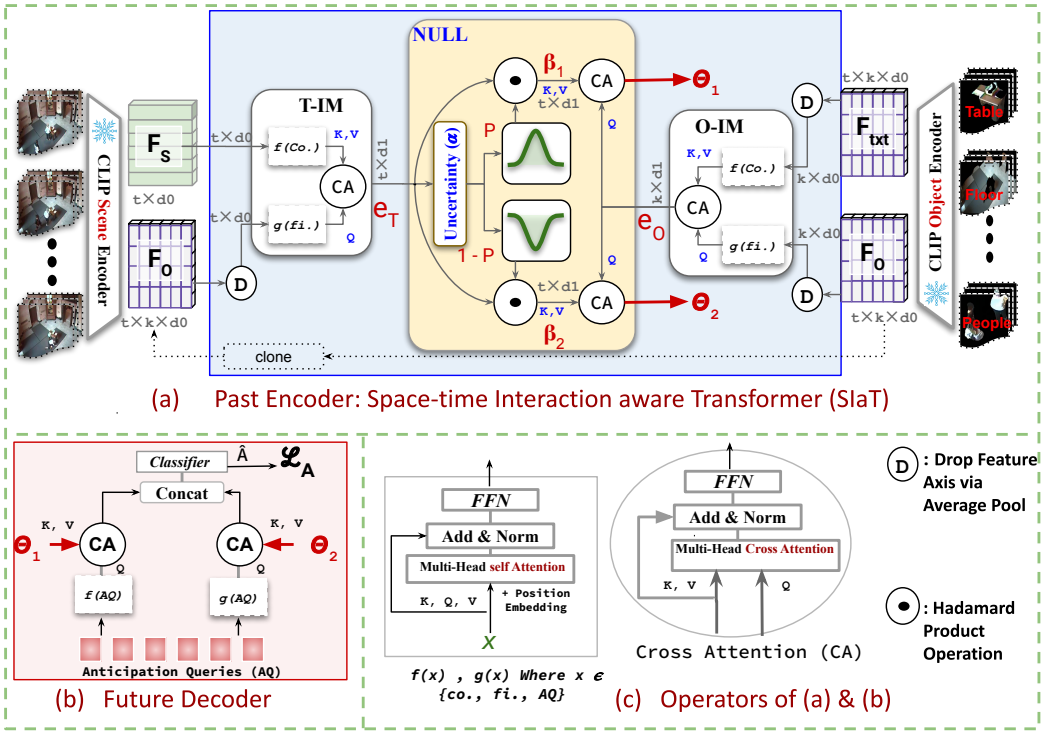

Figure 3. **In (a) An overview of proposed Spatial Interaction aware Transformer (SIaT):** It has two key components: (i) Interaction Modules (TIM/OIM) and (ii) Normalcy Uncertainty Latent Learner (NULL) to capture early human interaction patterns and handle the uncertainty between normal interactions to future abnormal behavior. **In (b) Functional diagram of Future Decoder** is portrayed that inputs the uncertainty encoded latent representation of the past (*i.e.* θ_1 and θ_2) is taken as input along with the anticipation queries (AQ) to predict the future event. **In (c), operators** used in (a) and (b) are defined.

(O-IM) respectively. This is enforced by dissociatively encoding scene and object-relevant sharp and subtle precursor clues via T-IM and O-IM. For this, T-IM first aims to highlight the temporal saliencies of the observation by encoding the cross temporal interactions among coarse-grained scene ($F_S \in \mathbb{R}^{t\times d0}$) and fine-grained object $F_O \in \mathbb{R}^{t\times k\times d0}$ level feature maps. While processing the $F_O \in \mathbb{R}^{t\times k\times d0}$ in T-IM, a spatial-pooling operation is applied on k dimension of F_O to suppress the object appearance features and encourage the object-specific fine motion features. Next, O-IM aims to promote the salient object features out of many irrelevant ones by encoding their spatial interactions with the surroundings. Although individual object mask features ($F_O \in \mathbb{R}^{t\times k\times d0}$) are empowered with fine-grained representations, they are contextually sparse. Further, encoding the object interaction with sparse context leads to a partial understanding of the complex interactions (e.g. ambiguity between arrest and fighting w/o a policeman as context). Due to this, CLIP pre-trained object-level textual feature $F_{txt} \in \mathbb{R}^{t\times k\times d0}$ is taken into consideration for infusing rich contextual information while encoding critical object interactions in O-IM. When processing both fine-grained F_O and contextual F_{txt} in O-IM, a temporal-pooling operation is applied on t dimension of F_O and F_{txt} to suppress the object motion features and focus on the object appearance and spatial location features. Although T-IM and O-IM encodes two distinct features, they are functionally identical.

Functionality of T/O-IM: Primarily, the T/O-IM learns the temporal and object level spatial interaction by encoding the cross-correlation between the respective fine-grained($fi.$) and contextual ($co.$),representations. First, individual fine-grained and contextual feature maps (*i.e.* $fi.$ and $co.$) are processed in parallel via $g(fi.)$ and $f(co.)$ for all pair self-correlation-encoding. In practice, $f()$ and $g()$ are the standard multi-head self attention layers. Next, for computing the cross-interactions between coarse and fine latent features obtained from $f(co.)$ and $g(fi.)$ cross-attention operator is applied by treating the contextual latent features as the key and $value$ and the fine-grained latent features as the $query$. The temporal and object interaction encoded outputs of T-IM and O-IM is represented by $e_T \in \mathbb{R}^{t\times d1}$ and $e_O \in \mathbb{R}^{k\times d1}$, where t is the observation length k is the associated objects and $d1$ is the embedding dimension.

4.3. Normalcy Uncertainty Latent Learner (NULL)

The goal of Normalcy Uncertainty Latent Learner (NULL) is to exploit the uncertainty in normal observation while associating the object-centric spatial interaction embedding $e_O \in \mathbb{R}^{k\times d1}$ to each scene-centric temporal interaction in $e_T \in \mathbb{R}^{t\times d1}$. The exploration of observa-

tional uncertainty begins with computing two kinds of latent weights w.r.t. future events, *i.e.* (i) relevant (P) and (ii) opponent ($1-P$) for the normal temporal regions of the past $e_T \in \mathbb{R}^{t\times d1}$ through an `uncertainty` estimator (α). α can be seen as a multi-layer perceptron where the final layer has single unit with `sigmoid` activation to compute relevant (P) latent weights independently across t. The opponent ($1-P$) latent weights are computed by simple `minus` operation with P. Here, by uncertainty, we mean **if** P **and** $1-P$ **are equivalent,** there exist no corelation between the observation o the future event and **if** P **and** $1-P$ **are divergent,** ateleast some portion of the normal has strong correlation to the future events. Next, the two types of latent weights are multiplied separately by `Hadamard-product` $(\cdot)$ with the input e_T to generate the relevant and opponent embedding of normal observation **i.e.** $\beta_1 \in \mathbb{R}^{t\times d1}$ and $\beta_2 \in \mathbb{R}^{t\times d1}$ respectively. Now the idea is to associate the object spatial interaction embedding $e_O \in \mathbb{R}^{k\times d1}$ to β_1 and β_2 to identify the relevant and opponent associations, out of which one association encouraged for anticipation. For this, two `cross attention` layers are used distinctively to obtain $\theta_1 \in \mathbb{R}^{t\times d1}$ and $\theta_2 \in \mathbb{R}^{t\times d1}$ for corresponding β_1 and β_2.

Further, to ensure better calibration among the relevant and opponent associations (*i.e.* θ_1 and θ_2) while considering the uncertainties, we aim to produce smaller gap between the NULL outcome embedding and actual predicted action of the future $\hat{A}$ (obtained from (2)). This is enforced by employing (1) as in below.

$$\mathcal{L}_U(\theta_1,\theta_2,\hat{A}) = \underbrace{\|\sum_{i=1}^{t}(\theta_1) - \hat{A}\|}_{L_{relevent}} + \underbrace{\|\sum_{i=1}^{t}(\theta_2) - \hat{A}\|}_{L_{opponent}} \quad (1)$$

Thus, the perfect calibration occurs when one of the estimation θ_1 or θ_2 perfectly matches the actual action prediction.

4.4. Anticipation Decoder

The decoder takes learnable tokens as input, referred to as *anticipation queries (AQ)* and the outputs of NULL *i.e.* θ_1, θ_2 (shown in) to predict the future labels. It also learns the long-term action relation between the observed and future anomaly via self attention and cross attention. The anomaly queries are embedded with M learnable queries $AQ \in \mathbb{R}^{M\times d1}$. The temporal orders of the queries are fixed to be equivalent to that of the future anomalies, i.e., the i^{th} query corresponds to the i^{th} future anomaly. The decoder consists of two parallel multi-head cross attention (CA), and MLP (as classifer). The final output of decoder is computed by following (2) and the output logits $\hat{A}$ are then *softmax* activated.

$$\hat{A} = MLP(concat(CA(\theta_2, g(AQ), CA(\theta_1, f(AQ))))) \quad (2)$$

Training Objective: For future decoder, he M number of AQ are matched to the N number of ground-truth actions to apply action anticipation loss $\mathcal{L}^A$. The $\mathcal{L}^A$ loss is defined with standard `cross-entropy` between action A and logits $\hat{A}$. The SIaT transformer along with the future decoder are trained jointly with $\mathcal{L}_{total} = \lambda_1\mathcal{L}_A(\hat{A}, A) + \lambda_2\mathcal{L}_U(\theta_1,\theta_2,\hat{A})$, where λ_1 and λ_2 are weighting factors.

5. Our Benchmark Dataset and Evaluation

The experiments are explicitly conducted on our Abnormal Human Behaviour Forecast dataset (AHB-F) only as this is the only dataset depicting such task. We also propose a deterministic evaluation protocol to evaluate our method.

5.1. AHB-F Dataset

We collect AHB-F with a aim address the common limitations of video anomaly datasets, *i.e.* : **(I)** Events performed by actors with simple background, **(II)** Single-type anomaly datasets such as only fighting, **(III)** lack of temporally annotated videos. Thus to combat (I), AHB-F collects real-world videos with actual anomaly occurrences with dynamic backgrounds (such as a streets, shops, corridor, banks etc.). Further, to handle (II), AHB-F includes videos containing 26 categories of human anomalies recorded in CCTV scenarios. In order to accumulate those types of anomaly categories, AHB combines 5 major real-world CCTV scenario datasets (*UCF-Crime [34], LAD-2000 [39], UCF-CrimeV2 [25], UBI [8], CCTV-Fight [27]*) and selects those abnormal videos where the anomaly is triggered by (or impacts) humans. As a result, AHB-F consists of 1470 untrimmed videos in total. Next, to ensure an unbiased train-test protocol, 75% and 25% of each category are reserved for training and testing, which is collectively 1091 and 379 videos respectively. In term of size, AHB-F is as big as the commonly used action anticipation datasets (like Breakfast) and AHB-F has 2.5× times larger action categories than the existing Breakfast dataset [16]. In AHB-F, the average video length is 102.4 and 94.13 seconds in train and test sets. Similarly, the average anomaly length is around 43.8 and 42.5 seconds with an average number of instances 1.6 and 1.8 in train and test sets. Since our objective is to observe the normal frames and anticipate the future abnormal frames, we almost balance the train and test sets in term of average anomaly lengths and instances. In order to address (III), we combine the annotations made by [25, 39] and rectified noisy annotations to provide the complete temporal labels for all videos of AHB-F. Further, as shown in Figure 4c, the abnormal videos are marked with three kinds of abnormal interactions, such as : Human-to-Human (H-H), Human-to-Object (H-O), and Human-to-(Human and Object) (H-HO). These interactions are marked by carefully observing the abnormal attributes (such as human, object trajectories, their associated interactions). The introduction of AHB-F will promote anomaly anticipation for ine-grained and subtle human anomalies.

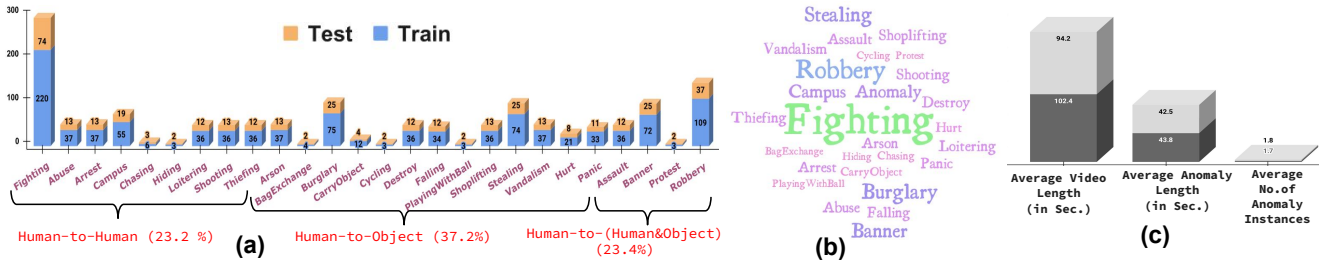

Figure 4. **AHB-F dataset properties:** (a) Category-wise video distribution in train and test set, (b) Diversity visualization of abnormal categories in word cloud, (c) Analysis of average video lengths, anomaly length and no of anomaly instances.

5.2. Evaluation Metrics

In order to provide a more robust evaluation that suffices our objective (*i.e. observe normal to anticipate anomaly*), we provide a dedicated evaluation protocol different from traditional action anticipation methods [1]. Briefly, action anticipation methods are evaluated based on $\alpha_1 \in [0.2, 0.3]$ and $\alpha_2 \in [0.1, 0.2, 0.3, 0.5]$, where α_1, α_2 denote the percentages of frames of a video that are used as observation and future prediction, respectively. However, in our case such percentage of frames division may include part of anomaly frames in the observation set, which will violate our evaluation objective. Thus, we provide an evaluation protocol which utilizes deterministic reference points ($\mathbf{D_{ref}}$) in a video to separate the observation and future prediction frames. Here, all frames prior and post to the $\mathbf{D_{ref}}$ can be defined as observation and future prediction frames. Further, the $\mathbf{D_{ref}}$ in a video is placed in such away that the observation frames contain only normal frames and the future prediction frames contain a mixture of both anomaly and normal frames that ensures our evaluation objective. Kindly not that, in test set a video can have **at least 1** and **at max 3** $\mathbf{D_{ref}}$ points to cover all the anomaly instances in the videos. Following previous work, we aim to utilize frame-level *Top-1* mean average precision (mAP) for the entire test-set. Moreover, another crucial aspect of evaluation lies in covering varying anticipation duration. For this, we provide 2 **long-short performance** indicators to explicitly evaluate the anticipation performance for long and short duration of future. The "short indicator" reports the mean mAP for future 1st to 3rd seconds, where as "long indicator" computes the mean mAP for future 4th to 8th seconds. By this, the long-short indicators cover all the anomaly instances with varying length.

6. Ablation Study

In this section, we conduct a series of ablation studies on the AHB-F dataset to assess the robustness of our approach.

Effectiveness of SIaT: As shown in Table 1b, we verify the impact of each component in SIaT by evaluating them in terms of AHB long-short anticipation performances. As a baseline set of experiments, we report initial anticipation performances by first stacking the future decoder on top of scene and object encoders. Then we incorporated the TIM and OIM modules independently to the respective baseline. By doing this, we obtained approx $+2\%$ and $+3\%$ on short and long metric with TIM which shows the relevance of temporal interaction in long-term AHB anticipation. Similarly, we obtained approx $+4.5\%$ and $+1.5\%$ on short and long metric which corroborate the necessity of object interaction encoding in short AHB prediction. Now to, complement both long and short AHB prediction, TIM and OIM modules are utilized together and obtained at least $+8.5\%$ performance gain from initial baseline experiments. Next, to take the normalcy uncertainty into account, we integrated NULL module along with TIM and OIM and obtained at least $+3.5\%$ performance gain in long and short AHB prediction. The significant performance gain by individual components have demonstrated their robustness in AHB anticipation.

How much of observation frames are required?: To answer this, we conduct experiments in Table 1c using various number of observation frames in SIaT. We start with 50 observation frames and linearly increase it to see the impact on long-short performance metric. We observe that the performance grows up till 200 frames and then the it tend to degrade slight to moderate by further increasing the number of observation frames. This can be intuitive as we accumulate long past information with more increased frame numbers and the information from long past may be noisy for future AHB, thereby the performance tend to decline.

Relevance of TEXT based semantics: From Table 1d, we observe the impact of infusing CLIP based text semantics into SIaT. From Table 1b, we verified that incorporation of OIM significantly improves the short term anticipation performance and it is possible due to the text based semantic injection to improve object interaction representation. As a result, when comparing to with out text semantics, the text feature improves both short and long term performance by at least $+1.5\%$ which corroborates it's relevance to the task.

7. State-of-the-art (SoTA) Comparison

To investigate the advantage of our proposed method in the abnormal behavior anticipation task, we extensively compare our approach (SIaT) with existing action anticipation methods in Table 1a and Figure 5. To validate the effectiveness of SIaT, we replace our past encoder and future

Methods	Short (Top-1 mAP %)				Long (Top-1 mAP %)		
	1 sec.	2 sec.	3 sec.	Avg	4 sec.	8 sec.	Avg
*Exp-1: SoTA with **Scene** Feature*							
OADTR [40]	62.37	61.58	62.11	62.02	61.58	56.10	58.84
FUTR [13]	62.53	59.89	60.42	60.94	61.21	55.67	58.44
LSTR [42]	62.16	60.94	61.46	61.52	62.47	58.16	60.31
JOADAA [15]	61.21	61.21	61.47	61.29	60.02	55.67	57.84
TesTra [45]	62.53	61.21	63.32	62.35	62.00	58.00	60.00
*Exp-2: SoTA with **Object** Feature*							
OADTR [40]	50.65	51.18	51.18	51.00	50.65	46.96	48.80
FUTR [13]	59.10	58.31	55.40	57.60	55.40	50.65	53.02
LSTR [42]	54.96	56.05	54.21	55.06	56.74	51.96	54.35
JOADAA [15]	55.93	56.46	55.40	55.93	55.14	49.07	52.10
TesTra [45]	55.40	56.20	54.35	55.31	55.14	51.48	53.31
*Exp-3: SoTA with **concat(Scene,Object)** Feature*							
OADTR [40]	63.07	62.40	62.58	62.68	62.06	59.13	60.59
FUTR [13]	61.47	61.21	61.74	61.47	62.79	56.46	59.62
LSTR [42]	63.06	62.00	63.06	62.70	63.06	59.63	61.34
JOADAA [15]	62.79	62.79	62.53	62.70	62.00	57.51	59.75
TesTra [45]	63.85	63.32	62.80	63.32	62.53	59.10	60.81
SIaT (ours)	65.96	64.64	63.85	64.81	63.85	59.63	61.74

(a) State-of-the-art comparison of method with benchmark methods in AHB-F datasets.

Scene		Object		NULL	Top-1 mAP %	
SE	TIM	OE	OIM		Short	Long
✓	-	-	-	-	54.32	50.71
✓	✓	-	-	-	56.21	53.54
-	-	✓	-	-	51.33	48.63
-	-	✓	✓	-	55.82	50.11
✓	✓	✓	✓	-	59.47	58.06
✓	✓	✓	✓	✓	64.81	61.74

(b) Impact of each component in SIaT framework on long short anticipation.

Observation Frames	Top-1 mAP %	
	Short	Long
50	56.11	47.30
100	61.26	55.65
150	63.93	59.97
200	64.81	61.74
250	63.52	60.11
300	61.06	58.11
350	61.06	56.24

(c) Study of observation frames required for AHB prediction.

Text	w/o Text	Top-1 mAP %	
		Short	Long
✓	-	64.81	61.47
-	✓	62.53	60.97

(d) Study to showcase the relevance of text in anticipation performance.

Table 1. Experimental results to showcase quantitative superiority compared to state-of-the-art methods and to portray the robustness in ablation studies.

decoder with previous methods. For fair comparison we kept the network optimization of SoTA unchanged and the backbone feature encoder CLIP [29] are kept same for all the SoTA methods. CLIP has the ability to encode object-centric features which is essential for our assumptions (*i.e.* encoding normal interactions to predict future anomaly).

Baseline Implementations and Quantitative Comparison: For the baseline set of experiments, we select widely used action anticipation SoTA methods: FUTR [13], OADTR [40], LSTR [42], TesTra [45], and JOADAA [15]. As shown in Table 1a, we perform three sequential experiments with different input features to analyse and compare SoTA methods, *i.e.* **Exp-1:** with only scene feature obtained from CLIP, **Exp-2:** with only object feature obtained from Mask2Former followed by CLIP, **Exp-3:** with concatenated scene and object features.

Observation from Exp-1: Methods like OADTR [40], LSTR [42], TesTra [45] gives decent initial performance in short term prediction as they reasonably encode the observation via complex transformer blocks. Further, methods like LSTR, TesTra have additional memory units to retain few abnormal precursors. As a result, they perform better w.r.t other SoTA in short term anticipation. However, in long-term anticipation their results are quite low. This is potentially due to: **(I)** existing methods ignore the underlying uncertainty between the observation and long future, **(II)** the scene features F_S do not carry enough fine-grained object representation that may be involved in the abnormality in the long future.

Observation from Exp-2: Although object features carries the individual fine-grained object representations, but it majorly lacks the collective context of the event happening in the scene. As a result, in short-term indicator, SoTA methods (OADTR [40], JOADAA [15]) completely relying on transformer for observation context understanding severely fail with object only features. However, SoTA methods (LSTR [42], TesTra [45]) with precursor memory units have relatively less performance drop w.r.t. others in short-term anticipation.

Observation from Exp-3: Here, methods relying on large transformer encoder have moderate gain on both long and short term prediction due to coarse and fine-grained representations of the observation. In contrast, we find that methods (LSTR [42], TesTra [45]) preforming decent in **Exp-1:** has little performance gain in long-term predictions. It means they have almost saturated performances with scene-only features. Because LSTR and TesTra gives more priority to the precursor memory units embedding over the fine-grained object representations. However, having additional precursor memory units can benefit only short-term anticipation, but may not contribute to long-term future events. Thus, modeling the uncertainty between the past observation to long future and enhancing the fine-grained object representation by embedding the interaction among them may be crucial, which is missing in previous SoTA.

Comparison with Exp-3: With the above observation,

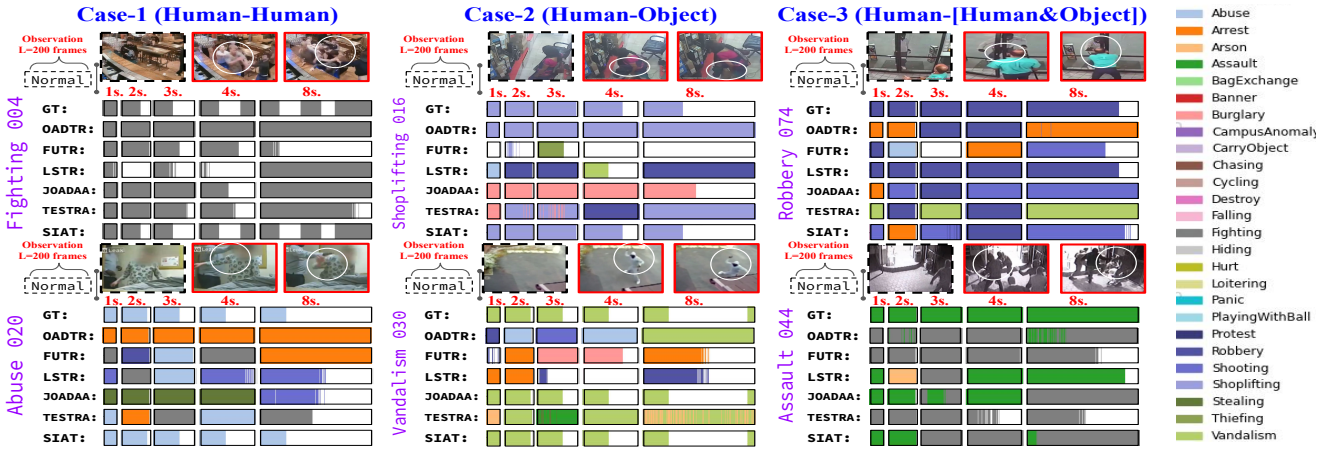

Figure 5. Qualitative results visualization and comparison of our proposed SIaT with five popular stat-of-the-art methods on various cases. Best viewed in color and with zoomed in.

our SIaT provides a dedicated mechanism to handle uncertainty between observation to the future and dissociatively encodes the temporal and object interactions. As a result, SIaT achieves relatively higher performance (at least $+1.15\%$) than pure-transformer based SoTA (FUTR [13], OADTR [40], and JOADAA [15]) as in **Exp-3**. This improvement is significant as it covers 26 abnormal categories with multiple instances. However, without any additional memory units, our method is still able to surpass memory based methods (LSTR [42] and TesTra [45]) marginally (at least $+0.4\%$) which shows the efficacy of our method in long-term prediction. Further, comparing with short-term anticipations, our method significantly (at least $+2.11\%$) improves from FUTR [13], OADTR [40], JOADAA [15] and moderately surpasses the memory based methods (at least $+1.49\%$). This corroborate the efficacy of our method in long and short term anticipation over SoTA methods.

Qualitative Comparison and Analysis: To bring additional analytical insights to SoTA performance comparison, Figure 5 provides a qualitative performance comparison of our method w.r.t. five selected SoTA in three major interaction cases (**case-1: Human-Human, case-2: Human-Object, and case-3:Human-[Human&Object]**). From Figure 5, it is observed that our method is accurate across all future duration (*i.e. 1st to 8th seconds*) for human-human and human-object interactions (*case-1,2*) anomaly types. For all video of case-1 and 2, "Fighting-04" (*a group of people fighting inside a bar*),"Abuse-020" (*a patient is being physically abused by the care taker*), "Shoplifting-016" (*a woman customer stealing a laptop*) and "Vandalism-033" (*a man throwing stone repetitively to a house*) SIaT has obtained most similar heat map prediction as in GT. But, pure transformer based methods like FUTR [13], OADTR [40], JOADAA [15] lack the temporal preciseness in the prediction of "Fighting-04" and to some extends give ambiguous prediction for "Abuse-020". However, LSTR [42] and TesTra [45] encounter similar issues despite additional precursor memory. This could be due to the limited understanding of the observation by the previous SoTA. Next, the criticality lies in case-3 *i.e.* Human-[Human&Object] interactions where all entities are interacting together to cause anomaly. In such a case, all the methods including ours have some ambiguous prediction and temporal impreciseness due to overlapping cues between the interaction categories (*i.e.* Assault may look like fighting, robbery may look like stealing) and heavy occlusion caused by too much entity involvement. Along this direction, we believe that there exist enough scope for further improvements and our proposed task will serve as a open problem statement to attract wide range of research in this domain.

8. Conclusion

In this work, an affirmative task named "Abnormal Human Behavior Anticipation" in real-world videos is introduced to promote mitigatory measures in serious crimes. We aim to accomplish this task by observing normal past frames only. In pursuit of this, we propose a novel past encoder, namely SIaT, that facilitates to capture the early human interaction patterns and model the uncertainty between normal interactions and future abnormal behavior predictions. Additionally, we provide a larger-scale diversified dataset, namely "AHB-F" with a dedicated evaluation protocol to promote long and short-term anticipation. This task opens up new directions to analyse complex human abnormal behaviours prior to there occurrence in real-world videos, which is majorly missing in previous methods. By extensive benchmark evaluation and comparison, our SIaT achieves competitive performance w.r.t. prior arts.

Acknowledgements: This work was supported by Toyota Motor Europe (TME) and the French government, through the 3IA Cote d'Azur Investments managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002

References

[1] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what?-anticipating temporal occurrences of activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5343–5352, 2018. 2, 6

[2] Junwen Chen, Gaurav Mittal, Ye Yu, Yu Kong, and Mei Chen. Gatehub: Gated history unit with background suppression for online action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19925–19934, 2022. 2

[3] Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 387–395, 2023. 3

[4] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. 2021. 3

[5] MyeongAh Cho, Minjung Kim, Sangwon Hwang, Chaewon Park, Kyungjae Lee, and Sangyoun Lee. Look around for anomalies: Weakly-supervised anomaly detection via context-motion relational learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12137–12146, 2023. 1, 2

[6] 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. 2

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022. 2

[8] Bruno Degardin and Hugo Proença. Human activity analysis: Iterative weak/self-supervised learning frameworks for detecting abnormal events. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–7. IEEE. 5

[9] Shikha Dubey, Abhijeet Boragule, and Moongu Jeon. 3d resnet with ranking loss function for abnormal activity detection in videos. *arXiv preprint arXiv:2002.01132*, 2020. 2

[10] Antonino Furnari and Giovanni Maria Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4021–4036, 2020. 2

[11] Antonino Furnari and Giovanni Maria Farinella. Towards streaming egocentric action anticipation. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1250–1257. IEEE, 2022. 2

[12] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. Red: Reinforced encoder-decoder networks for action anticipation. *arXiv preprint arXiv:1707.04818*, 2017. 2

[13] Dayoung Gong, Joonseok Lee, Manjin Kim, Seong Jong Ha, and Minsu Cho. Future transformer for long-term action anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3052–3061, 2022. 2, 7, 8

[14] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2

[15] Mohammed Guermal, Abid Ali, Rui Dai, and François Brémond. Joadaa: Joint online action detection and action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6889–6898, January 2024. 2, 7, 8

[16] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014. 2, 5

[17] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks: A unified approach to action segmentation. In *European conference on computer vision*, pages 47–54. Springer, 2016. 2

[18] Shuheng Lin, Hua Yang, Xianchao Tang, Tianqi Shi, and Lin Chen. Social mil: Interaction-aware for crowd anomaly detection. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2019. 2

[19] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013. 2

[20] Snehashis Majhi, Rui Dai, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. Human-scene network: A novel baseline with self-rectifying loss for weakly supervised video anomaly detection. *Computer Vision and Image Understanding*, 241:103955, 2024. 2

[21] Snehashis Majhi, Srijan Das, François Brémond, Ratnakar Dash, and Pankaj Kumar Sa. Weakly-supervised joint anomaly detection and classification. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–7. IEEE, 2021. 2

[22] Snehashis Majhi, Srijan Das, and François Brémond. Dam: Dissimilarity attention module for weakly-supervised video anomaly detection. In *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8, 2021. 2

[23] Esteve Valls Mascaró, Hyemin Ahn, and Dongheui Lee. Intention-conditioned long-term human egocentric action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6048–6057, 2023. 2

[24] Nada Osman, Guglielmo Camporese, Pasquale Coscia, and Lamberto Ballan. Slowfast rolling-unrolling lstms for action anticipation in egocentric videos. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, pages 3437–3445, 2021. 2

[25] Halil İbrahim Öztürk and Ahmet Burak Can. Adnet: Temporal anomaly detection in surveillance videos. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part IV*, pages 88–101. Springer, 2021. 5

[26] Seongheon Park, Hanjae Kim, Minsu Kim, Dahye Kim, and Kwanghoon Sohn. Normality guided multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2665–2674, 2023. 2

[27] Mauricio Perez, Alex C Kot, and Anderson Rocha. Detection of real-world fights in surveillance videos. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2662–2666. IEEE, 2019. 5

[28] Zhaobo Qi, Shuhui Wang, Chi Su, Li Su, Qingming Huang, and Qi Tian. Self-regulated learning for egocentric video activity anticipation. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):6715–6730, 2021. 2

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 7

[30] Mamshad Nayeem Rizve, Gaurav Mittal, Ye Yu, Matthew Hall, Sandra Sajeev, Mubarak Shah, and Mei Chen. Pivotal: Prior-driven supervision for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22992–23002, 2023. 2

[31] Royston Rodrigues, Neha Bhargava, Rajbabu Velmurugan, and Subhasis Chaudhuri. Multi-timescale trajectory prediction for abnormal human activity detection. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 2

[32] Debaditya Roy, Ramanathan Rajendiran, and Basura Fernando. Interaction visual transformer for egocentric action anticipation. *arXiv preprint arXiv:2211.14154*, 2022. 2

[33] Yuge Shi, Basura Fernando, and Richard Hartley. Action anticipation with rbf kernelized feature mapping rnn. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 301–317, 2018. 2

[34] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2018. 1, 2, 5

[35] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018. 2

[36] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4975–4986, 2021. 2, 3

[37] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 98–106, 2016. 2

[38] Boyang Wan, Yuming Fang, Xue Xia, and Jiajie Mei. Weakly supervised video anomaly detection via center-guided discriminative learning. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020. 2

[39] Boyang Wan, Wenhui Jiang, Yuming Fang, Zhiyuan Luo, and Guanqun Ding. Anomaly detection in video sequences: A benchmark and computational model. *IET Image Processing*, 15(14):3454–3465, 2021. 5

[40] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Zhengrong Zuo, Changxin Gao, and Nong Sang. Oadtr: Online action detection with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7565–7575, 2021. 2, 7, 8

[41] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European Conference on Computer Vision*, pages 322–339. Springer, 2020. 1, 2

[42] Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu, and Stefano Soatto. Long short-term transformer for online action detection. *Advances in Neural Information Processing Systems*, 34:1086–1099, 2021. 2, 7, 8

[43] Muhammad Zaigham Zaheer, Arif Mahmood, Hochul Shin, and Seung-Ik Lee. A self-reasoning framework for anomaly detection using video-level labels. *IEEE Signal Processing Letters*, 27:1705–1709, 2020. 2

[44] Jiangong Zhang, Laiyun Qing, and Jun Miao. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4030–4034. IEEE, 2019. 2

[45] Yue Zhao and Philipp Krähenbühl. Real-time online video detection with temporal smoothing transformers. In *European Conference on Computer Vision*, pages 485–502. Springer, 2022. 2, 7, 8

[46] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H. Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2

[47] Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. *arXiv preprint arXiv:2302.05160*, 2023. 3

[48] Yi Zhu and Shawn Newsam. Motion-aware feature for improved video anomaly detection. *arXiv preprint arXiv:1907.10211*, 2019. 2