Weakly-Supervised Spatio-Temporal Anomaly Detection in Surveillance Video

Jie Wu[1,3], **Wei Zhang**[2], **Guanbin Li**[1 *], **Wenhao Wu**[2], **Xiao Tan**[2], **Yingying Li**[2], **Errui Ding**[2] and **Liang Lin**[1]
[1]Sun Yat-sen University
[2]Baidu Inc.
[3]ByteDance Inc.
wujie.10@bytedance.com, {zhangwei99, wuwenhao01, tanxiao01,liyingying05,dingerrui}@baidu.com, liguanbin@mail.sysu.edu.cn, linliang@ieee.org

Abstract

In this paper, we introduce a novel task, referred to as *Weakly-Supervised Spatio-Temporal Anomaly Detection (WSSTAD)* in surveillance video. Specifically, given an untrimmed video, WSSTAD aims to localize a spatio-temporal tube (i.e., a sequence of bounding boxes at consecutive times) that encloses the abnormal event, with only coarse video-level annotations as supervision during training. To address this challenging task, we propose a dual-branch network which takes as input the proposals with multi-granularities in both spatial-temporal domains. Each branch employs a relationship reasoning module to capture the correlation between tubes/videolets, which can provide rich contextual information and complex entity relationships for the concept learning of abnormal behaviors. *Mutually-guided Progressive Refinement* framework is set up to employ dual-path mutual guidance in a recurrent manner, iteratively sharing auxiliary supervision information across branches. It impels the learned concepts of each branch to serve as a guide for its counterpart, which progressively refines the corresponding branch and the whole framework. Furthermore, we contribute two datasets, i.e., *ST-UCF-Crime* and *STRA*, consisting of videos containing spatio-temporal abnormal annotations to serve as the benchmarks for WSSTAD. We conduct extensive qualitative and quantitative evaluations to demonstrate the effectiveness of the proposed approach and analyze the key factors that contribute more to handle this task.

1 Introduction

Anomaly detection in the surveillance video is a fundamental computer vision task and plays a critical role in video structure analysis and potential down-stream applications, e.g., accident forecasting, urban traffic analysis, evidence investigation. Although it has attracted intense attention in recent years, it remains a very challenging problem due to the serious imbalance between normal and abnormal samples, the lack of fine-grained abnormal labeling data and the ambiguity of the concept of abnormal behavior. Previous studies [Luo *et al.*, 2017; Hasan *et al.*, 2016; Xu *et al.*, 2015] generally leverage normal training samples to model abnormal concepts, and identify the distinctive behaviors that deviate from normal patterns as anomalies. However, these works are not accessible to the abnormal videos, which may incorrectly classify some normal behaviors with abrupt action as abnormal ones, i.e., car acceleration or pedestrian intrusion, resulting in a high false alarm rate. Recently, some studies [Sultani *et al.*, 2018; Zhong *et al.*, 2019; Zhang *et al.*, 2019; Zhu and Newsam, 2019] have attempted to introduce videos of abnormal behavior during the training process to better predict which frames or snippets contain abnormal behavior. However, these works can be regarded as coarse-level detection since they can only predict anomaly in the temporal dimension, but fail to provide more critical region-level abnormal details. In fact, fine-grained anomaly detection that simultaneously predicts when and where the anomaly happens in the video is more in line with the requirements of the actual application scenario. For example, during the accident investigation, the traffic polices pay more attention to the objects involved in the accident and their corresponding trajectories, rather than the moment when the abnormality occurred. Moreover, fine-grained spatio-temporal location of abnormal behaviors can also provide more reliable explanatory guarantees for the anomaly classification.

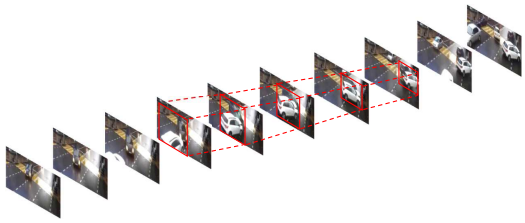

Figure 1: Our proposed WSSTAD task is to localize the spatio-temporal tube of abnormal event (as shown in red) using only video-level label during training.

These observations motivate us to introduce a novel task, referred to as *Weakly-Supervised Spatio-Temporal Anomaly*

*Most of the work is done when Jie Wu was a research intern at Baidu. Corresponding author is Guanbin Li.

Detection (WSSTAD). WSSTAD aims to localize a spatio-temporal tube (i.e., a sequence of bounding boxes at consecutive times), which encloses the trajectory of an abnormal event in an untrimmed video. We follow [Sultani *et al.*, 2018] to address this task in a weakly supervised setting that it does not rely on any spatio-temporal annotations during the training process, refer to an example (two cars in the collision) as shown in Figure 1. Compared with existing anomaly detection problems, our proposed task poses three additional challenges. 1) The weakly supervised nature of this problem is that both the temporally segment-level labels and spatially region-wise labels are not available during training. 2) This localization task spans spatial and temporal dimensions while the spatial details and temporal correlation can be viewed as cues at different granularity levels. How to leverage such multi-granularity information to jointly facilitate model training remains to be studied. 3) Some anomalies such as "road accident" involve the interaction of objects, thus an inherent challenge is to automatically infer the latent relationship between objects in the videos.

In order to tackle this task, we formulate it as a Multiple Instance Learning problem [Sultani *et al.*, 2018; Chen *et al.*, 2019; Yamaguchi *et al.*, 2017]. We extract two kinds of tube-level instance proposals and feed them into a tube branch to capture spatial cues. It is non-trivial to distinguish the anomaly from a single instance, so we propagate information among the instances to make a more comprehensive prediction. Concretely, each branch employs a relationship modeling module that adopts the multi-head self-attention mechanism to capture the relationships between video objects, and thus incorporate the contextual information and complex entity behavior relationships for anomaly inference. As each branch helps to capture abnormal abstractions at different granularity level, we can transfer the learned concepts from one branch to the other intuitively. To this end, we present a novel *Mutually-guided Progressive Refinement* (MGPR) framework, which involves a dual-path mutual guidance mechanism in a recurrent manner to iteratively facilitate the optimization procedure. Our experiments show that dual-path recurrent guidance coordinates to mutually reinforce two training procedures and boost the performance progressively.

The contributions of this work are summarized into four folds: 1) We present a new task WSSTAD to localize a spatio-temporal tube that semantically corresponds to abnormal event, with no reliance on any spatio-temporal annotations during training. 2) To address this task, MGPR framework is designed to transfer learned abstractions across branches, encouraging mutual guidance and progressive refinement in the whole framework. 3) We contribute two datasets that provide fine-grained tube-level annotations for abnormal videos to serve as the benchmarks. 4) In-depth analyses are conducted to demonstrate the effectiveness of the proposed framework over some competitive methods and discuss the key factors that contribute more to handle this task.

2 Related Work

Anomaly Detection. As a long-lasting task in the computer vision field, anomaly detection has been extensively studied for a long time [Nallaivarothayan *et al.*, 2014; Luo *et al.*, 2017; Hasan *et al.*, 2016; Xu *et al.*, 2015; Sultani *et al.*, 2018; Zhong *et al.*, 2019; Zhang *et al.*, 2019; Zhu and Newsam, 2019; Li *et al.*, 2020; Wu *et al.*, 2020b; Zhao *et al.*, 2021; Wu *et al.*, 2021]. Sultani *et al.* [Sultani *et al.*, 2018] propose to learn anomaly in the weakly supervised setting that merely resort to video-level labels instead of snippet-level during the training process. However, these works [Sultani *et al.*, 2018; Zhong *et al.*, 2019; Zhang *et al.*, 2019; Zhu and Newsam, 2019] regard anomaly detection as a frame/segment-level anomaly ranking task and employ AUC metric to evaluate the model performance. In fact, we believe that abnormal behavior detection should be defined as a fine-grained video understanding task that involves spatio-temporal location, because in actual applications, early warning and subsequent analysis of an abnormal behavior need to include the time and location of specific events. To this end, we formulate the weakly-supervised spatio-temporal anomaly detection task to predict tube-level anomalies. compared to unary classification based works [Sabokrou *et al.*, 2018; Sabokrou *et al.*, 2017], our proposed WSSTAD setting is novel and meaningful as it first realizes spatial-temporal anomaly detection with only video-level labels under the binary-classification setting.

Spatio-Temporal Weakly Supervised Learning. Weakly-supervised learning aims at optimizing a model without substantial manual labeling works. This learning paradigm generally resorts to MIL and applies in many AI tasks such as object detection, captioning and language grounding [Wu *et al.*, 2020a]. Recently, some work have conducted a deep exploration of spatio-temporal weakly supervised learning [Chen *et al.*, 2019; Yamaguchi *et al.*, 2017; Escorcia *et al.*, 2020; Mettes and Snoek, 2018; Wu *et al.*, 2019], which is studied to localize a tube that corresponds to the task requirement. Chen *et al.* [Chen *et al.*, 2019] exploit fine-grained relationships from cross-modal and address the task of grounding natural sentence in the video. Comparing with weakly supervised s.t. action localization, our task differs significantly and is more challenging due to 1) our proposed task is category agnostic while category info can be used in existing s.t. action localization studies; 2) anomaly essentially differs from human activities; 3) the temporal boundaries of the anomaly (e.g. fight, burglary, accident) are fuzzier than human activities (e.g. surfing, play guitar); 4) the anomalous segment accounts for a much smaller proportion in the entire video (e.g. only 9.28% for STRA) compared with popular datasets (e.g. UCF101-24 and J-HMDB-21 about 40%).

3 Methodology

In this paper, we cast the WSSTAD task as the multiple instance learning (MIL) problem. We first illustrate the tube-level and videolet-level instances in section 3.1. Then we describe two relationship modeling module based prediction branches (i.e., tube branch and temporal branch) in section 3.2. Section 3.3 introduces the proposed mutually-guided progressive refinement framework and the inference procedure is described in Section 3.4.

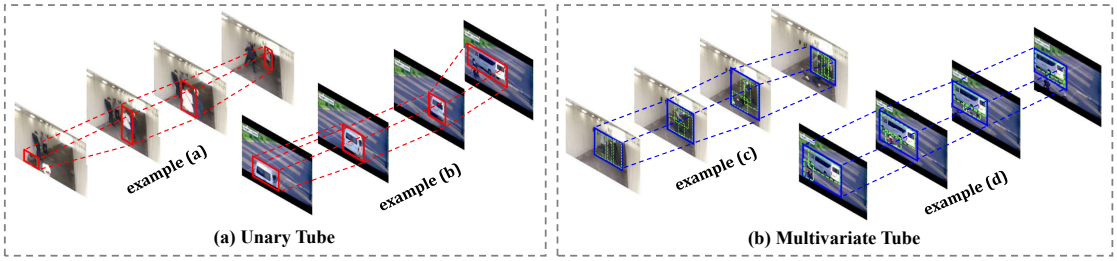

Figure 2: Instances visualizations for "fighting" (example (a), (c)) and "road accident" (example (b), (d)). In multivariate tubes, each union box (as shown in the blue boxes) may contain several objects (as shown the green boxes).

3.1 Spatio-Temporal Instance Generation

We first extract candidate proposals as the instances of MIL. Conventional anomaly detection works [Sultani *et al.*, 2018; Zhong *et al.*, 2019; Zhang *et al.*, 2019] merely take video snippets as instances, which fail to locate where the anomaly happens spatially. To this end, we introduce the tube-level instance that links the objects bounding boxes along time into an tube to capture spatio-temporal abnormal cues. Furthermore, considering that an abnormal event may involve the behavior of a single object or multiple objects, we carefully design two kinds of tube-level instance from different gradations. As shown in the Figure 2, unary tube encloses an individual object and multivariate tube contains multiple intersecting objects. Multivariate tube is particularly important for capturing the associations between objects which contributes to the localization of abnormal events. In the experiments, we observe that such multi-gradation instance setting is robust to cope with diverse anomalies. The generation of the tube-level instance is detailed in the supplementary material. Besides extracting tube-level instances, we also follow [Sultani *et al.*, 2018] to utilize temporal correlation to construct videolet-level instances, which is helpful to capture temporal dependencies.

3.2 Relationship Modeling Prediction

As shown in Figure 3 (a), we accordingly design a dual-branch network architecture, which contains one tube branch that employs tube-level instances to capture spatial cues, and another temporal branch that leverages videolet-level instances for exploiting temporal correlation.

In each branch, the instance features $\mathcal{F}$ are first extracted by a pre-trained C3D model [Tran *et al.*, 2015]. Previous works [Sultani *et al.*, 2018; Zhang *et al.*, 2019; Zhu and Newsam, 2019] generally fail to take into account the relationship between instances, which ignores the contextual information can provide more reliable cues for comprehensive inference. To address this issue, we creatively introduce a relationship modeling module that employs the multi-head self-attention [Vaswani *et al.*, 2017] to model the correlations between instances, where the dependency between any two instances is learnable as the attention weight. The multi-head self-attention [Vaswani *et al.*, 2017] is established on the basis of the scaled dot-product attention g_a, which calculates the weights by scaled dot-products of query $\mathbf{Q}$ and keys $\mathbf{K}$. Then g_a computes the output as a weighted sum of values $\mathbf{V}$, formulated as:

$$g_a(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}^\top\mathbf{K}}{\sqrt{d_q}}\right)\mathbf{V}^\top, \tag{1}$$

where d_q are the dimension of $\mathbf{Q}$ and the Softmax $(\cdot)$ function is performed along rows. The multi-head setting deploys H ($H = 8$) paralleled attention layers to aggregate information from different representation subspaces and sufficiently decompose the complicated dependencies:

$$g_{ma}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathop{\Big\|}_{j=1}^{H} h_j^\top, h_j = g_a(\mathbf{W_j^Q Q}, \mathbf{W_j^K K}, \mathbf{W_j^V V}), \tag{2}$$

where $\|$ denotes the concatenation operation on the column dimension, $\mathbf{W_j^Q} \in \mathbb{R}^{\frac{d_k}{H}\times d_k}$,$\mathbf{W_j^K} \in \mathbb{R}^{\frac{d_k}{H}\times d_k}$, $\mathbf{W_j^V} \in \mathbb{R}^{\frac{d_k}{H}\times d_k}$ are linear projection matrices in j^{th} head. Multi-head self-attention is a special case of the scaled dot-product attention where the $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ are set to the same video feature matrix $\mathcal{F} = (\mathcal{F}_1;...;\mathcal{F}_i;...;\mathcal{F}_n) \in \mathbb{R}^{d_k\times n}$. $\mathcal{F}_i \in \mathbb{R}^{d_k}$ denotes the feature of the i^{th} instance and n is the number of instances in the corresponding video. Then the relationship modeling based self-attentive representation $\tilde{\mathcal{F}} = (\tilde{\mathcal{F}}_1,.., \tilde{\mathcal{F}}_n)$ is computed via the residual connection:

$$\tilde{\mathcal{F}} = g_{ma}(\mathcal{F},\mathcal{F},\mathcal{F}) + \mathcal{F}. \tag{3}$$

This residual connection setting maintains original concepts and aggregate contextual information from other instances. The obtained $\tilde{\mathcal{F}}_i$ is then fed into a three-layer fully connected neural network (abnormal score predictor) [Sultani *et al.*, 2018] to predict the abnormal score.

3.3 Mutually-guided Progressive Refinement

Multiple Instance Learning. Following the setting of MIL, we divide the instances belonging to the anomalous video into the positive bag and the instances for normal videos are putted into the negative bag. The optimization goal of MIL is to push abnormal instances apart from normal instances in terms of the abnormal score. Considering that there may be only one abnormal instance in a positive bag, we take the instance that obtains the max abnormal score over all instances in each bag (named as *max instance*) to compute the loss function. Additionally, we first explain some abbreviations used in the following section. $\mathcal{V}^a$ and $\mathcal{V}^n$ denote the videolet-level instances in the anomalous and normal video. For the tube instances, the sequences of RGB images can be obtained from the inside

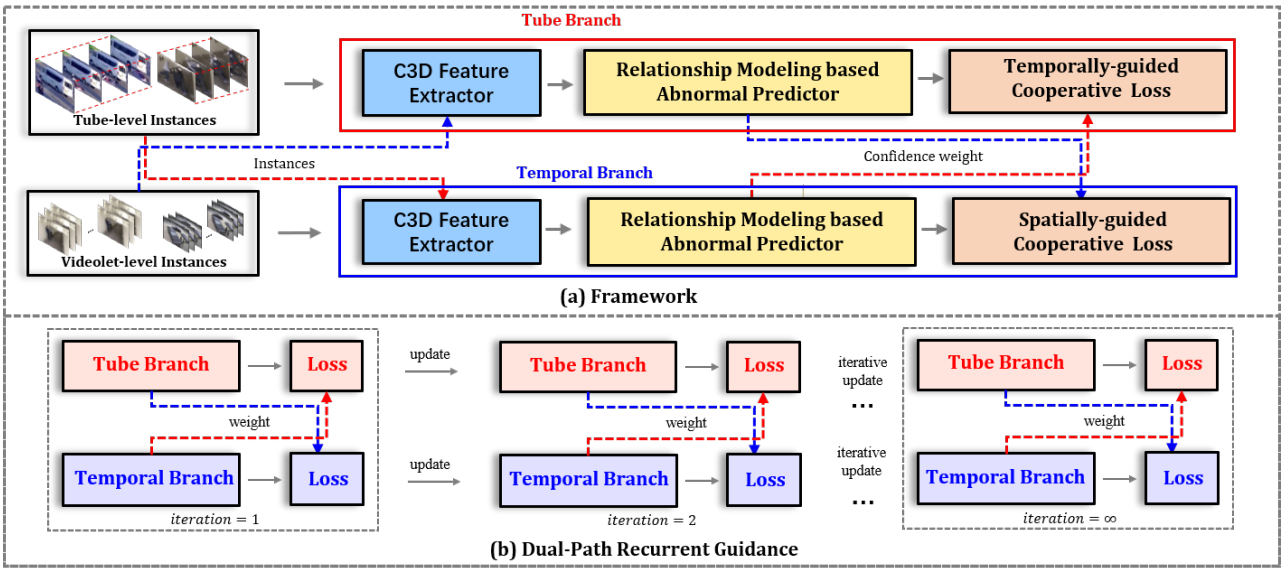

Figure 3: The illustration of *Mutually-guided Progressive Refinement* framework.

of tubes (region-level) or from the entire image (image-level). So we use $\mathcal{T}^{a,r}$, $\mathcal{T}^{a,g}$ to respectively denote the region-level based tube instances and image-level based tube instances in the anomalous video. $\mathcal{T}^{n,r}$ is the region-level based tube instances in the normal video. Furthermore, we use p_t and p_v to represent the scores predicted by the tube and temporal branch, respectively.

Dual-Path Mutual Guidance. Our ranking loss in MIL builds upon the existing observations that the learned concepts from the tube and temporal branch are often complementary to each other in action recognition and localization task [Saha *et al.*, 2016; Simonyan and Zisserman, 2014]. In the tube-temporal dual branches, we can transfer the learned concepts from one branch to the other branch intuitively, and the concepts can be leveraged as auxiliary supervision to encourage the other branch to learn anomaly from other granularity level. These observations motivate us to propose the *Mutually-guided Progressive Refinement* framework, which exploits dual-path mutual guidance across branches and refines tube and temporal branches in a progressive manner. As shown in the Figure 3, a novel mutually-guided ranking loss is designed to leverage the feedback of each branch to serve as a guidance for its counterpart, making both branches mutually guided. Specifically, we feed the max instance of the tube branch into the temporal branch and outputs an abnormal score, which is treated as a confidence weight of the ranking loss that guides the training of the tube branch:

$$\mathcal{L}^{tube}_{MG\text{-}Rank} = \max(0, p_v(\mathcal{T}^{a,g}_m) \times (1 - p_t(\mathcal{T}^{a,r}_m) + p_t(\mathcal{T}^{n,r}_m))), \tag{4}$$

where $\mathcal{T}^{a,r}_m$, $\mathcal{T}^{a,g}_m$, $\mathcal{T}^{n,r}_m$ denotes the corresponding max instance obtained in the tube branch. We feed $\mathcal{T}^{a,g}_m$ instead of $\mathcal{T}^{a,r}_m$ into the temporal branch, which ensures the input adapts to the characteristics of the temporal branch. From the equation we can see that the confidence of the ranking loss is adaptively adjusted by the abnormal concepts already learned by the other branch, instead of fixing it to 1 gruffly in the conventional MIL paradigm [Sultani *et al.*, 2018]. If the temporal branch holds that the max instance is less related to the anomaly (the returned score is lower), it will penalize the ranking loss of the tube branch. Similarly, the temporal branch is optimized by the guidance from the tube branch:

$$\mathcal{L}^{tem}_{MG\text{-}Rank} = \max(0, p_t(\mathcal{V}^{a}_m) \times (1 - p_v(\mathcal{V}^{a}_m) + p_v(\mathcal{V}^{n}_m))), \tag{5}$$

where $\mathcal{V}^{a}_m$, $\mathcal{V}^{n}_m$ denotes the max instance in temporal branch.

In order to obtain the weight from the tube branch, we simply regard $\mathcal{V}^{a}_m$ as tubes and feed $\mathcal{V}^{a}_m$ into the tube branch. By introducing dual-path mutual guidance, the diversity of abnormal abstraction from one branch will be enhanced by using the complementary granularity concepts from the others, which intrinsically improves the anomaly learning ability of both branches. In each training iteration, we first freeze the tube branch and provide loss weights to optimize the temporal branch. Then we switch to training the tube branch and obtain weights from the temporal branch. This alternate optimization procedure is repeated iteratively during training:

$$\mathcal{L}_{MG\text{-}Rank} = \psi \times \mathcal{L}^{tube}_{MG\text{-}Rank} + (1 - \psi) \times \mathcal{L}^{tem}_{MG\text{-}Rank}, \tag{6}$$

where ψ is a binary variable indicating the selection of the training branch. With the training going on, as depicted in Figure 3 (b), the dual-path mutual guidance can be regraded as a recurrent guidance scheme. Concretely, the outputs from the tube branch gradually provide accurate guidance for the temporal branch. Meanwhile, with the increasingly informative guidance from the temporal branch, tube branch is refined in the loop. This recurrent guidance scheme contributes to iteratively sharing auxiliary supervision across branches and progressively enhance each branch and the whole framework.

Cross Entropy Loss. During training, we enforce that the predictor can produce high scores for the abnormal instances, while the normal instances get low scores. Hence the cross-entropy loss is adopted to encourage the score of max instances aligning to the video-level label. Formally, the score of the max instance in the anomalous/normal should get close

	ST-UCF-Crime				STRA			
Baseline	VAUC	IoU@0.3	IoU@0.1	MIoU	VAUC	IoU@0.2	IoU@0.1	MIoU
Random	44.69	2.52	10.08	2.64	43.80	3.17	4.76	1.33
Upper Bound	100.00	31.93	78.99	25.63	100.00	37.40	63.56	18.62
DMRM [Sultani *et al.*, 2018]	84.28	9.24	21.01	6.47	89.37	7.94	9.52	4.14
GCLNC [Zhong *et al.*, 2019]	85.53	10.92	**25.21**	8.63	91.60	12.70	17.46	5.41
STIL [Mettes and Snoek, 2018]	85.16	9.24	21.84	8.07	91.23	12.70	15.87	5.19
ASA [Escorcia *et al.*, 2020]	86.33	**11.76**	23.52	8.41	92.17	14.29	18.49	6.54
Ours	**87.65**	**11.76**	24.37	**8.98**	**92.88**	**15.87**	**20.63**	**7.23**

Table 1: The performance (in %) with state-of-the-art methods. The top entry of all the methods except the upper bound is highlighted.

to target 1/0:

$$\begin{aligned}\mathcal{L}_{CE} = &-[\log(p_t(\mathcal{T}_m^{a,r})) + \log(1 - p_t(\mathcal{T}_m^{n,r}))] \\ &-[\log(p_v(\mathcal{V}_m^{a})) + \log(1 - p_v(\mathcal{V}_m^{n}))].\end{aligned} \tag{7}$$

Cross entropy loss also ensures that the model does not optimize $\mathcal{L}_{Rank}$ by predicting all the scores as zero. To sum up, mutually-guided ranking loss and cross-entropy loss are combined to optimize the proposed framework jointly:

$$\mathcal{L} = \mathcal{L}_{MG\text{-}Rank} + \mathcal{L}_{CE}. \tag{8}$$

3.4 Inference

In the inference stage, we treat the tube-level instances as the hypothetical instances. Considering that a tube-level instance may have abnormal and normal events, we evenly divide each instance into M hypothetical tube instances $\{\mathcal{T}_i\}_{i=1}^{M}$. Then we input the region-level instances into the tube branch to predict an abnormal score $p_t(\mathcal{T}_i^r)$. Subsequently, the global-level instances are fed into the temporal branch to get the corresponding score $p_v(\mathcal{T}_i^g)$. Finally, we calculate the prediction scores and retrieve the top-1 tube $\mathcal{T}_{pred}$ via employing an aggregate function that averages scores from two branches:

$$\mathcal{T}_{pred} = \arg\max_{\mathcal{T}_i} \frac{p_t(\mathcal{T}_i^r) + p_v(\mathcal{T}_i^g)}{2}. \tag{9}$$

4 Spatio-Temporal Anomaly Dataset

A major challenge for the proposed WSSTAD task is the lack of appropriate datasets. UCF-Crime and other anomaly detection datasets [Luo *et al.*, 2017; Li *et al.*, 2013] do not provide spatio-temporal annotations for the ground truth instances, which are necessary for evaluation. To this end, we build a new dataset (denoted as ST-UCF-Crime) that annotates spatio-temporal bounding boxes for abnormal events in UCF-Crime [Sultani *et al.*, 2018], which contains anomaly videos of diverse categories in complicated surveillance scenarios. Furthermore, we contribute a new dataset, namely Spatio-Temporal Road Accident (abbreviated as STRA) that consists of various road accidents videos, such as motorcycles crash into cars, cars crash into people, etc.. STRA contributes to fine-grained anomaly detection in actual traffic accident scenarios and promoting the development of intelligent transportation. We provide more details in the appendix.

5 Experiments

5.1 Experimental Setup

Implementation Details. We feed each 16-frame image sequence within the instance into C3D [Tran *et al.*, 2015] and extract the features from the *fc6* layer. Then we take a mean pooling layer to average these features and obtain the instance feature. We randomly choose 30 positive and 30 negative bags to construct a mini-batch, and the number of instances per bag is limited to 200. The total loss is optimized via Adam optimizer with the learning rate of 0.0005.

Evaluation Metric. 1) Video-level AUC. The abnormal score of top-1 tube $\mathcal{T}_{pred}$ is regarded as the abnormal score of the whole video. We use abnormal score to perform receiver operating characteristic curve and area under the curve is viewed as video-level AUC (VAUC). VAUC is adopted to evaluate the model's ability to distinguish between normal and abnormal video. 2) We utilize spatio-temporal localization score [Yamaguchi *et al.*, 2017] to measure the overlap of $\mathcal{T}_{pred}$ and ground-truth tube $\mathcal{T}_{gt}$:

$$S_{loc}(\mathcal{T}_{gt}, \mathcal{T}_{pred}) = \frac{\sum_{f\in\Gamma} IoU(\mathcal{T}_{gt}^f, \mathcal{T}_{pred}^f)}{|\Gamma|}, \tag{10}$$

where IoU denotes the intersection over union between those two bounding boxes. Γ is the intersection of two set of frames f. The first set contains the frames have bounding boxes provided by detector. The second set includes the frames in which $\mathcal{T}_{gt}$ or $\mathcal{T}_{pred}$ has any bounding box. We define two metrics to evaluate spatio-temporal localization accuracy for the abnormal testing videos. "IoU@ϵ" means the percentage of the videos that have S_{loc} larger than ϵ. "MIoU" denotes the average IoU for all abnormal testing videos.

5.2 Comparison with the State-of-the-art

We compare our approach with some state-of-the-art weakly-supervised anomaly detection approaches, DMRM [Sultani *et al.*, 2018] and GCLNC [Zhong *et al.*, 2019] in Table 1. We also compare with some weakly-supervised spatio-temporal action localization approaches, STLA[Mettes and Snoek, 2018] and ASA [Escorcia *et al.*, 2020]. We additionally show the performance of randomly selecting a candidate tube and the upper bound performance of choosing the tube of the largest overlap with the ground-truth.

As shown in the Table 1, our approach significantly exceeds the performance of random selection in VAUC or MIoU. Furthermore, our approach outperforms [Sultani *et al.*, 2018; Zhang *et al.*, 2019; Zhong *et al.*, 2019; Mettes and Snoek, 2018; Escorcia *et al.*, 2020] and achieves best performance on both datasets. From video-level metric, our method improves VAUC by 1.32% and 0.71% compared with the previous best [Escorcia *et al.*, 2020] on two datasets, respectively. It reveals that our approach helps to better determine

		ST-UCF-Crime		STRA	
Temporal	Tube	VAUC	MIoU	VAUC	MIoU
✓		86.41	8.26	90.04	5.39
	✓	87.15	8.43	92.40	6.17
✓	✓	**87.65**	**8.98**	**92.88**	**7.23**

Table 2: The performance of different learning branches.

			ST-UCF-Crime		STRA	
$\mathcal{L}_{Rank}$	$\mathcal{L}_{MG\text{-}Rank}$	$\mathcal{L}_{CE}$	VAUC	MIoU	VAUC	MIoU
✓			87.31	8.22	92.17	6.61
	✓		76.83	8.10	92.26	6.53
		✓	87.57	8.39	91.94	5.88
✓		✓	87.60	8.55	92.78	6.77
	✓	✓	**87.65**	**8.98**	**92.88**	**7.23**

Table 3: The performance of different loss functions.

whether the video is abnormal. For spatio-temporal localization accuracy, the MIoU of our approach achieves 7.23% (8.98%) in STRA (ST-UCF-Crime), obtaining comparative enhancement over [Escorcia *et al.*, 2020] by 10.55% (7.02%). In the contrast experiment with other methods [Sultani *et al.*, 2018; Zhang *et al.*, 2019; Zhong *et al.*, 2019], our approach shows a more significant performance advantage on STRA than ST-UCF-Crime. It may be due to the fact that STRA is more demanding in understanding the relationship between objects, while other methods do not pay much attention to the relationship between objects with interaction.

5.3 Ablation Study

The Effectiveness of Dual Branches. To validate the significance of dual branches for modeling abnormal concepts, we design three variants that use a separate branch from our framework to predict anomaly and the results are reported in Table 2. "Temporal" is the temporal branch and "Tube" denotes the tube branch that utilizes region-level RGB images. As shown in Table 2, the performance of "Tube" is superior to "Temporal", which shows that tube-level abnormal cues are more effective than the temporal correlation of the abnormal segments. Furthermore, getting rid of temporal or tube branch will cause performance degradation. It indicates that exploiting the complementary of abnormal abstractions from dual branches contributes to further boosting the performance.

The Effectiveness of Loss Function. To investigate the setting of the loss function, we further design four baselines with different loss functions and summarize the results in Table 3. As shown in Table 3, the baseline with the ranking loss $\mathcal{L}_{Rank}$ [Sultani *et al.*, 2018] gets an unimpressive result. Employing mutual guidance to the ranking loss further declines the performance. It is beacuse the model learns to reduce the loss function via predicting all scores to approach 0 trickly. $\mathcal{L}_{CE}$ helps to alleviate this issue and cooperate with $\mathcal{L}_{MG\text{-}Rank}$ to obtain promising results. We can see that our approach with these two losses can obtain 8.98 (7.23) in ST-UCF-Crime (STRA) in terms of MIOU, which performs best among the baselines in Table 3. The variant discards the mutual guidance mechanism based on our approach suffers from performance degradation. The above observations reveal that our approach helps to take full advantage of auxiliary supervision information to achieve accurate anomaly inference.

			ST-UCF-Crime		STRA	
MVT	RMM	GCN	VAUC	MIoU	VAUC	MIoU
	✓		84.08	8.29	91.42	5.66
✓			86.59	8.17	92.08	6.55
✓		✓	87.42	8.32	92.33	6.44
✓	✓		**87.65**	**8.98**	**92.88**	**7.23**

Table 4: The performance of different interaction methods.

The Effectiveness of Interaction Methods. It may be difficult to detect anomaly from a single object, hence the interaction between objects has an important impact on learning anomalies cues. To explore different interaction methods in our model, we design several variants and present the results in Table 4. We observe that VAUC declines from 87.65% to 84.08% in ST-UCF-Crime without multivariate tube (MVT). A similar performance degradation can be observed in the absence of relationship modeling module (RMM). It reveals the effectiveness of the MVT and the RMM. In fact, MVT and RMM can be viewed as the interaction of the objects at different levels. Furthermore, we try to replace RMM with graph convolutional network (GCN) designed in [Zhong *et al.*, 2019]. From Table 4, we find that although the baseline with GCN can achieve promising results, our model with RMM gains 0.66% and 0.79% improvement w.r.t. MIoU in the two datasets respectively. It may be due to the fact that the adjacency matrix of GCN in is prior and hand-crafted, while the corresponding relationship weights in our approach are obtained via feature learning.

6 Conclusions

This work explores WSSTAD, a novel task that aims to localize a tube that semantically corresponds to the abnormal event. To handle this task, we design two prediction branches based on the relationship modeling module to exploit multi-granularity abnormal concepts and establish instances relationship for comprehensive inference. Mutually-guided Progressive Refinement framework is designed to transfer the learned abnormal concepts of each branch to the other, employing a dual-path recurrent guidance scheme to facilitate mutual guidance across branches and refine the optimization process progressively. To evaluate this task, we contribute two datasets and conduct extensive experiments to analyze the key factors that contribute more to this task.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under-Grant 2018YFC0830103, in part by the National Natural Science Foundation of China under Grant No.61976250 and No.U1811463, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant No.2020B1515020048, in part by the Guangzhou Science and technology project under Grant No.202102020633. It was also sponsored by CCF-Tencent Open Research Fund.

References

[Chen *et al.*, 2019] Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee Kenneth Wong. Weakly-supervised spatio-temporally grounding natural sentence in video. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1884–1894, 2019.

[Escorcia *et al.*, 2020] Victor Escorcia, Cuong Duc Dao, Mihir Jain, Bernard Ghanem, and Cees G M Snoek. Guess where? actor-supervision for spatiotemporal action localization. *Computer Vision and Image Understanding*, 2020.

[Hasan *et al.*, 2016] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proc. CVPR*, pages 733–742, 2016.

[Li *et al.*, 2013] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 36(1):18–32, 2013.

[Li *et al.*, 2020] Yingying Li, Jie Wu, Xue Bai, Xipeng Yang, Xiao Tan, Guanbin Li, Shilei Wen, Hongwu Zhang, and Errui Ding. Multi-granularity tracking with modularlized components for unsupervised vehicles anomaly detection. In *Proc. CVPRW*, pages 586–587, 2020.

[Luo *et al.*, 2017] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proc. ICCV*, pages 341–349, 2017.

[Mettes and Snoek, 2018] Pascal Mettes and Cees GM Snoek. Spatio-temporal instance learning: Action tubes from class supervision. *arXiv preprint arXiv:1807.02800*, 2018.

[Nallaivarothayan *et al.*, 2014] Hajananth Nallaivarothayan, Clinton Fookes, Simon Denman, and Sridha Sridharan. An mrf based abnormal event detection approach using motion and appearance features. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 343–348. IEEE, 2014.

[Sabokrou *et al.*, 2017] Mohammad Sabokrou, Mohsen Fayyaz, Mahmood Fathy, and Reinhard Klette. Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. *IEEE Trans. on Image Processing*, 26(4):1992–2004, 2017.

[Sabokrou *et al.*, 2018] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. In *Proc. CVPR*, pages 3379–3388, 2018.

[Saha *et al.*, 2016] Suman Saha, Gurkirt Singh, Michael Sapienza, Philip HS Torr, and Fabio Cuzzolin. Deep learning for detecting multiple space-time action tubes in videos. *arXiv preprint arXiv:1608.01529*, 2016.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Neurips*, pages 568–576, 2014.

[Sultani *et al.*, 2018] Waqas Sultani, Chen Chen, Mubarak Shah, and Waqas Sultani. Real-world anomaly detection in surveillance videos. In *Proc. CVPR*, pages 6479–6488, 2018.

[Tran *et al.*, 2015] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proc. ICCV*, pages 4489–4497, 2015.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neurips*, pages 5998–6008, 2017.

[Wu *et al.*, 2019] Wenhao Wu, Dongliang He, Xiao Tan, Shifeng Chen, and Shilei Wen. Multi-agent reinforcement learning based frame sampling for effective untrimmed video recognition. In *Proc. ICCV*, 2019.

[Wu *et al.*, 2020a] Jie Wu, Guanbin Li, Xiaoguang Han, and Liang Lin. Reinforcement learning for weakly supervised temporal grounding of natural language in untrimmed videos. In *Proc. ACMMM*, pages 1283–1291, 2020.

[Wu *et al.*, 2020b] Jie Wu, Yingying Li, Wei Zhang, Yi Wu, Xiao Tan, Hongwu Zhang, Shilei Wen, Errui Ding, and Guanbin Li. Modularized framework with category-sensitive abnormal filter for city anomaly detection. In *Proc. ACMMM*, pages 4669–4673, 2020.

[Wu *et al.*, 2021] Jie Wu, Xionghui Wang, Xuefeng Xiao, and Yitong Wang. Box-level tube tracking and refinement for vehicles anomaly detection. In *Proc. CVPRW*, 2021.

[Xu *et al.*, 2015] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. Learning deep representations of appearance and motion for anomalous event detection. *arXiv preprint arXiv:1510.01553*, 2015.

[Yamaguchi *et al.*, 2017] Masataka Yamaguchi, Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Spatio-temporal person retrieval via natural language queries. In *Proc. ICCV*, pages 1453–1462, 2017.

[Zhang *et al.*, 2019] Jiangong Zhang, Laiyun Qing, Jun Miao, and Jiangong Zhang. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In *Proc. ICIP*, pages 4030–4034. IEEE, 2019.

[Zhao *et al.*, 2021] Yuxiang Zhao, Wenhao Wu, Yue He, Yingying Li, Xiao Tan, and Shifeng Chen. Good practices and a strong baseline for traffic anomaly detection. In *Proc. CVPRW*, 2021.

[Zhong *et al.*, 2019] 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 *Proc. CVPR*, pages 1237–1246, 2019.

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