Weakly Supervised Video Anomaly Detection: From MIL Ranking to Vision-Language Models

2026. 07. 10

Data Mining & Quality Analytics Lab.

강동훈

발표자 소개

About Me

❖ **강동훈 (Donghun Kang)**

- 고려대학교 산업경영공학과 석사과정 (2026.03~Present)
- Data Mining & Quality Analysis Lab (김성범 교수님)

❖ **Research Interest**

- Video Anomaly Detection
- Vision-Language Models (VLM)

❖ **Contact**

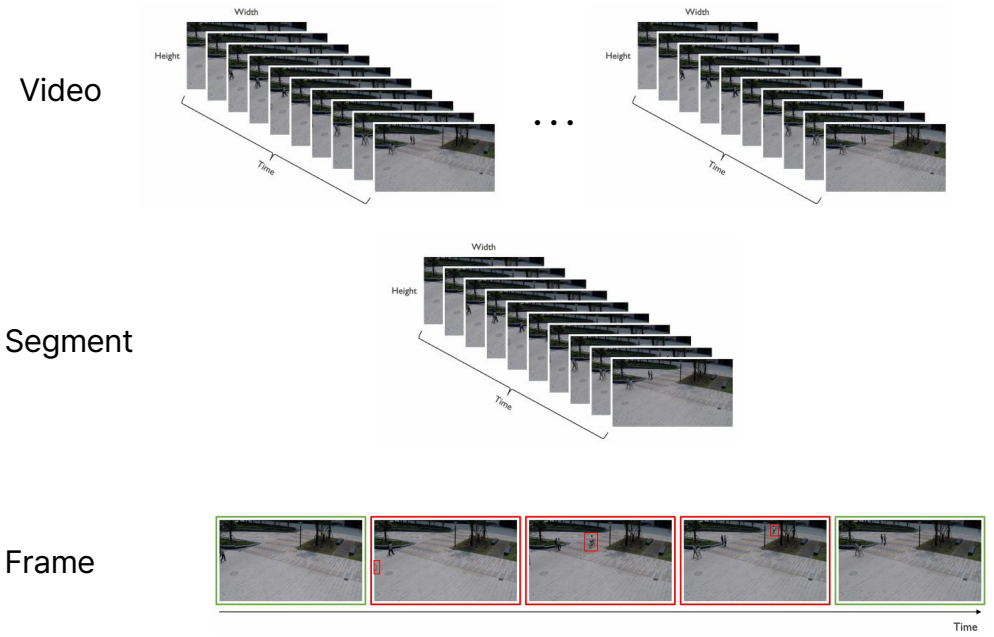
- dhkang00@korea.ac.kr

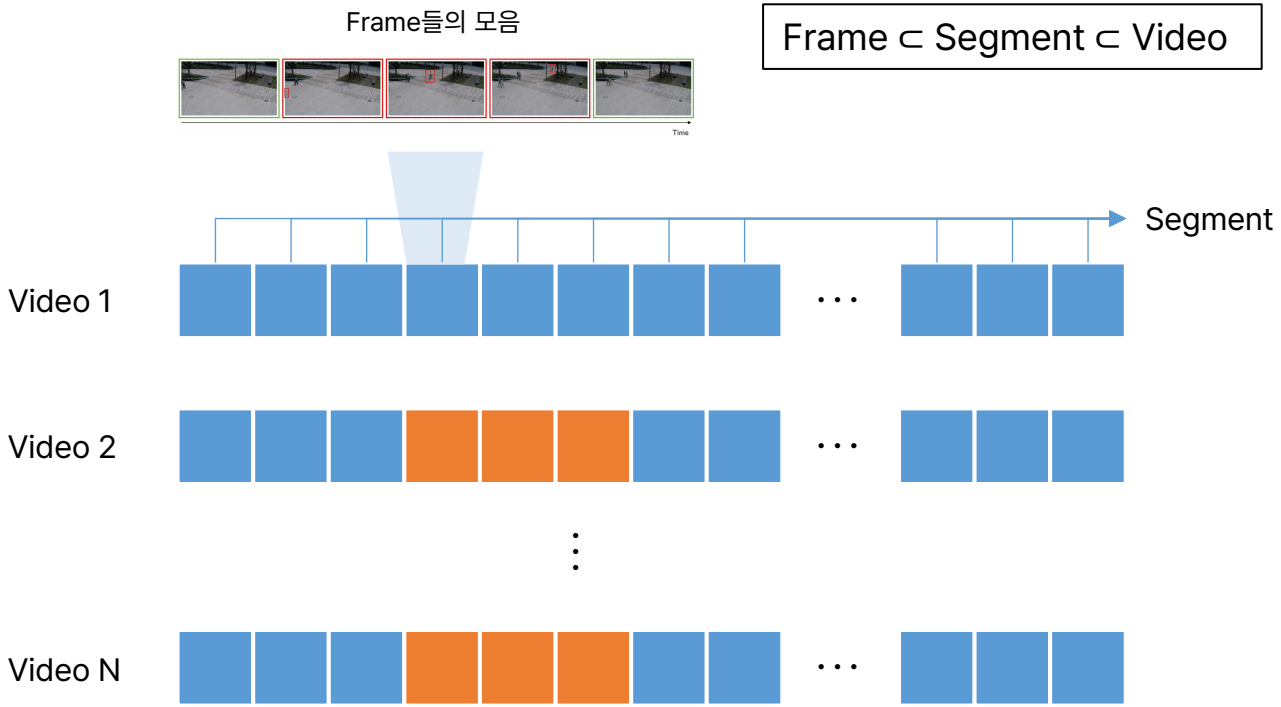
Introduction

Basic Terminology

❖ Video, Segment, Frame 이란?

- **Video**: 여러 segment 또는 frame들이 시간 순서대로 이어진 전체 영상
- **Segment**: 여러 개의 Frame을 일정한 시간 단위로 묶은 짧은 비디오 구간
- **Frame**: 비디오를 구성하는 한 장의 이미지



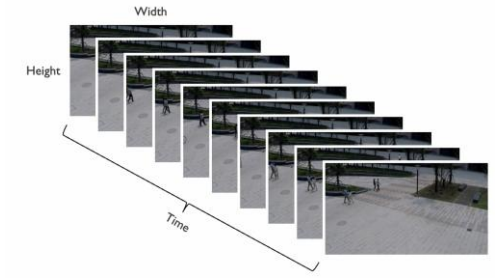


Introduction

Video Anomaly Detection

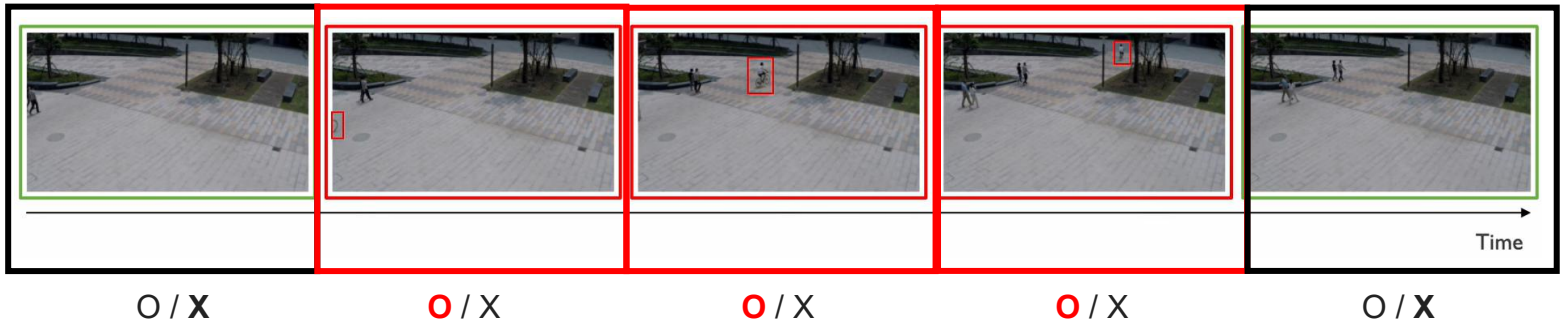
❖ VAD(Video Anomaly Detection)란?

- 영상 속에서 정상적인 행동이나 장면과 다르게 보이는 이상 사건을 탐지하는 task



- Video Anomaly Detection

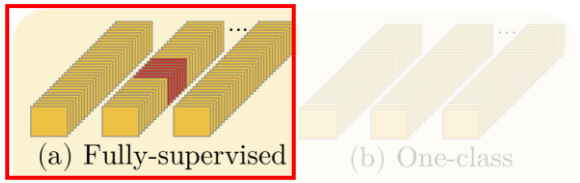
: 시간 정보가 함께 고려되어 *Frame, Segment*-level로 이상 현상 탐지 가능

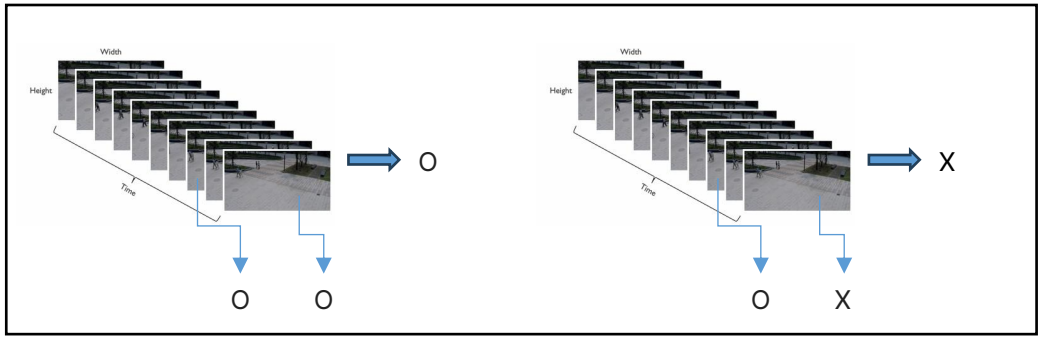


Introduction

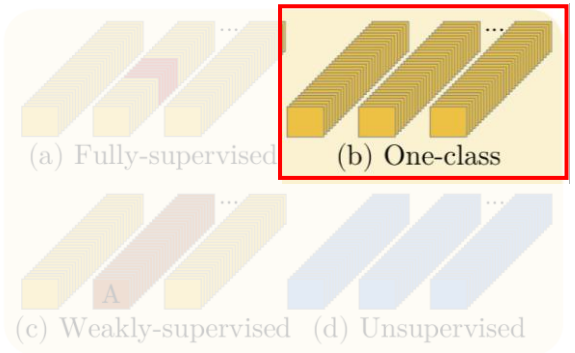
Training Method

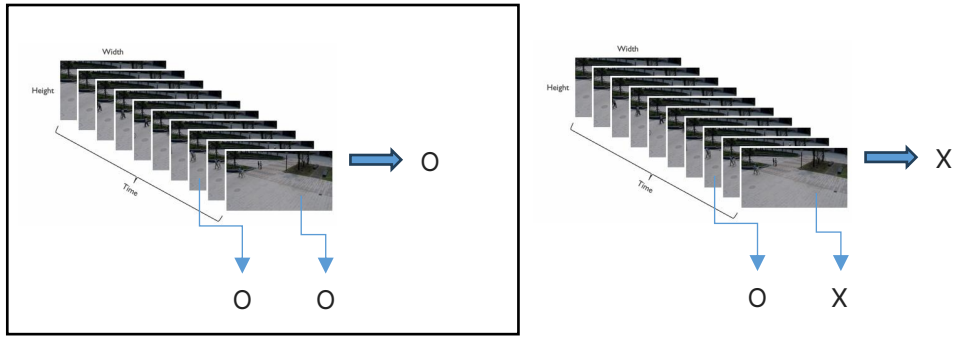
❖ **Fully-Supervised:** 정상 프레임과 이상 프레임에 대한 frame-level-label을 활용





❖ **One-class Supervised:** 정상 비디오만을 사용하여 정상 패턴을 학습





[1] Zanella, Luca, et al. "Harnessing large language models for training-free video anomaly detection." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

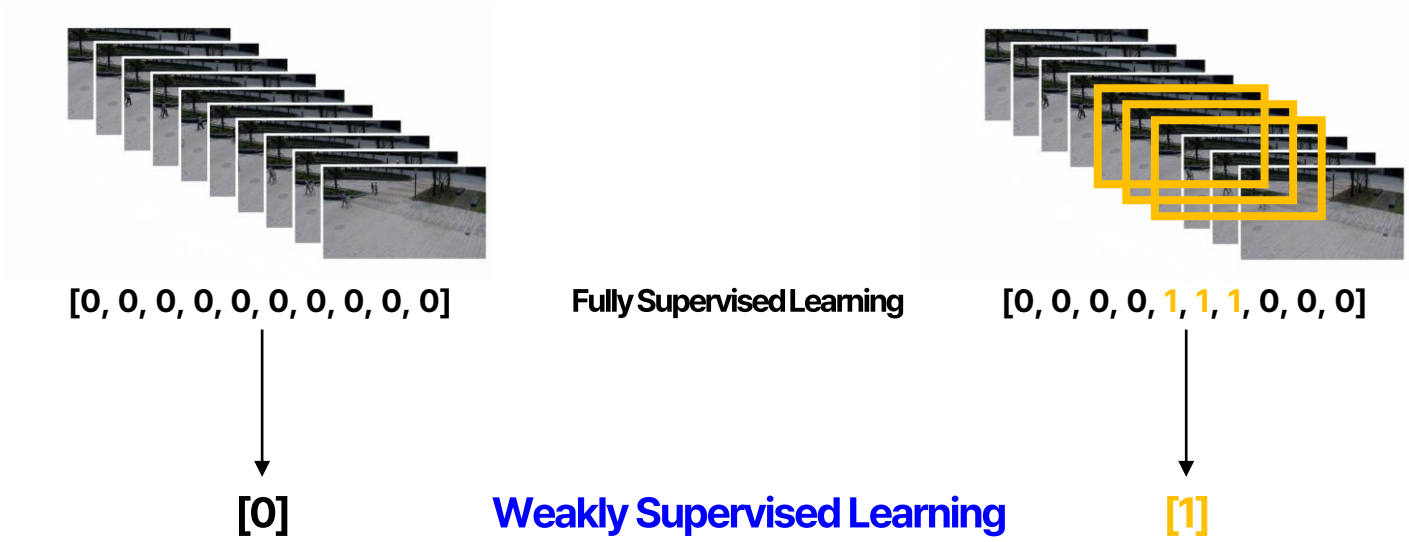
Introduction

Training Method

❖ **기존 학습 방식의 문제점은 무엇일까?**

- 프레임 단위의 라벨링은 시간과 비용이 많이 든다. (Fully Supervised)
- 정상 영상만 학습하므로 이상 패턴을 반영하기 어렵다. (One-Class)

➢ 비디오 전체의 이상/정상 여부만 라벨링하자!

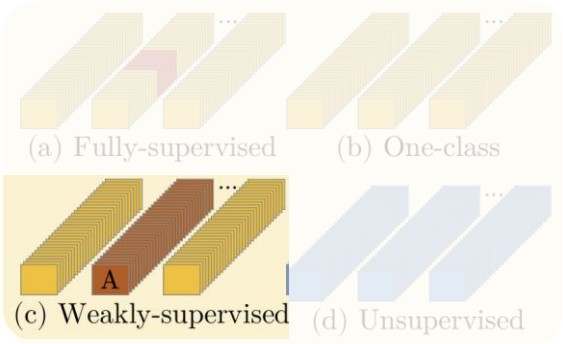


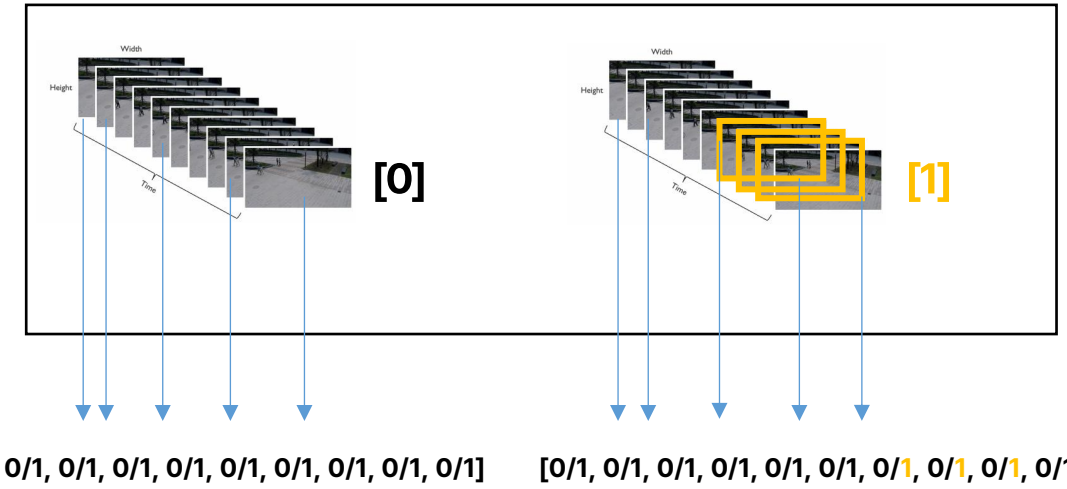
[1] Zanella, Luca, et al. "Harnessing large language models for training-free video anomaly detection." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

Introduction

Training Method (Weakly-Supervised)

❖ **Weakly-Supervised:** 각 프레임에 대한 라벨 없이, 비디오 전체에 대한 라벨만 있는 데이터로 학습





Weakly Supervised Video Anomaly Detection

"비디오의 이상/정상 여부만 알 수 있을 때 프레임 단위의 이상 징후를 찾는다."

[1] Zanella, Luca, et al. "Harnessing large language models for training-free video anomaly detection." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

Related Works

Video 단위 라벨만으로 이상 구간을 어떻게 학습할 수 있을까?

❖ **Real-world Anomaly Detection in Surveillance Videos (CVPR 2018)**

- WSVAD 문제를 MIL(Multi Instance Learning) 기반으로 정식화
- MIL Ranking Loss 제안

Real-world Anomaly Detection in Surveillance Videos

Waqas Sultani[1], Chen Chen[2], Mubarak Shah[2]
[1]Department of Computer Science, Information Technology University, Pakistan
[2]Center for Research in Computer Vision (CRCV), University of Central Florida (UCF)
waqas5163@gmail.com, chenchen870713@gmail.com, shah@crcv.ucf.edu

Abstract

Surveillance videos are able to capture a variety of real-tic anomalies. In this paper, we propose to learn anoma-s by exploiting both normal and anomalous videos. To oid annotating the anomalous segments or clips in train-g videos, which is very time consuming, we propose to arn anomaly through the deep multiple instance ranking amework by leveraging weakly labeled training videos, . the training labels (anomalous or normal) are at video-vel instead of clip-level. In our approach, we consider rmal and anomalous videos as bags and video segments instances in multiple instance learning (MIL), and auto-atically learn a deep anomaly ranking model that predicts gh anomaly scores for anomalous video segments. Fur-ermore, we introduce sparsity and temporal smoothness nstraints in the ranking loss function to better localize nomaly during training.

We also introduce a new large-scale first of its kind taset of 128 hours of videos. It consists of 1900 long d untrimmed real-world surveillance videos, with 13 re-istic anomalies such as fighting, road accident, burglary, bbery, etc. as well as normal activities. This dataset n be used for two tasks. First, general anomaly de-ction considering all anomalies in one group and all rmal activities in another group. Second, for recog-zing each of 13 anomalous activities. Our experimen-l results show that our MIL method for anomaly detec-on achieves significant improvement on anomaly detec-on performance as compared to the state-of-the-art ap-oaches. We provide the results of several recent deep arning baselines on anomalous activity recognition. The w recognition performance of these baselines reveals at our dataset is very challenging and opens more op-rtunities for future work. The dataset is available at: ttps://webpages.uncc.edu/cchen62/dataset.html

1. Introduction

Surveillance cameras are increasingly being used in pub lic places *e.g.* streets, intersections, banks, shopping malls etc. to increase public safety. However, the monitoring ca pability of law enforcement agencies has not kept pace. Th result is that there is a glaring deficiency in the utilization o surveillance cameras and an unworkable ratio of cameras t human monitors. One critical task in video surveillance i detecting anomalous events such as traffic accidents, crime or illegal activities. Generally, anomalous events rarely oc cur as compared to normal activities. Therefore, to allevi ate the waste of labor and time, developing intelligent com puter vision algorithms for automatic video anomaly detec tion is a pressing need. The goal of a practical anomal detection system is to timely signal an activity that deviate normal patterns and identify the time window of the occu ring anomaly. Therefore, anomaly detection can be consid ered as coarse level video understanding, which filters ou anomalies from normal patterns. Once an anomaly is de tected, it can further be categorized into one of the specifi activities using classification techniques.

A small step towards addressing anomaly detection is t develop algorithms to detect a specific anomalous event, fo example violence detector [30] and traffic accident detecto [23, 34]. However, it is obvious that such solutions canno be generalized to detect other anomalous events, therefor they render a limited use in practice.

Real-world anomalous events are complicated and di verse. It is difficult to list all of the possible anomalou events. Therefore, it is desirable that the anomaly detec tion algorithm does not rely on any prior information abou the events. In other words, anomaly detection should b done with minimum supervision. Sparse-coding based ap proaches [28, 41] are considered as representative meth ods that achieve state-of-the-art anomaly detection results These methods assume that only a small initial portion of video contains normal events, and therefore the initial por tion is used to build the normal event dictionary. Then, th main idea for anomaly detection is that anomalous event

[2] Sultani, Waqas, Chen Chen, and Mubarak Shah. "Real-world anomaly detection in surveillance videos." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

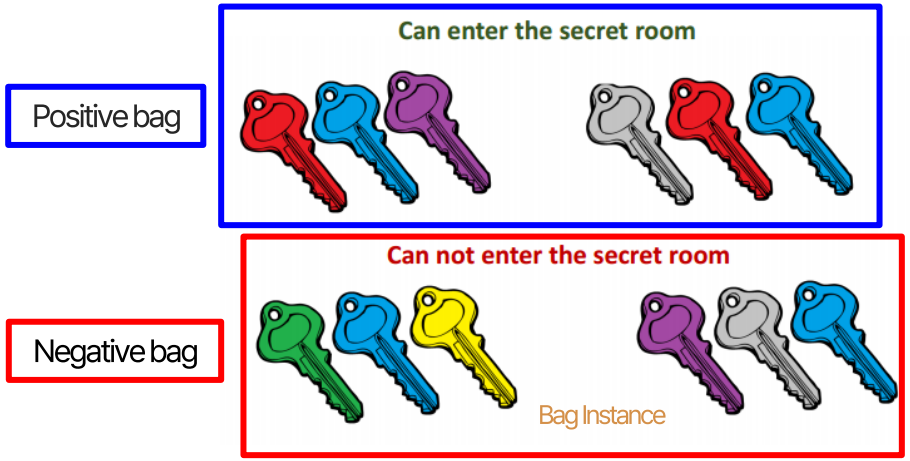
Related Works

Real-world Anomaly Detection in Surveillance Videos (CVPR 2018)

❖ **MIL(Multi Instance Learning)이란?**

- 비디오 전체 라벨만으로 각 segment의 이상 점수를 학습

EX) Instance 하나하나에 대해서는 라벨이 없고, Bag 전체에는 라벨이 있다.



Q) 새로운 Set이 들어왔을 때 문을 열 수 있을까?

[2] Sultani, Waqas, Chen Chen, and Mubarak Shah. "Real-world anomaly detection in surveillance videos." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

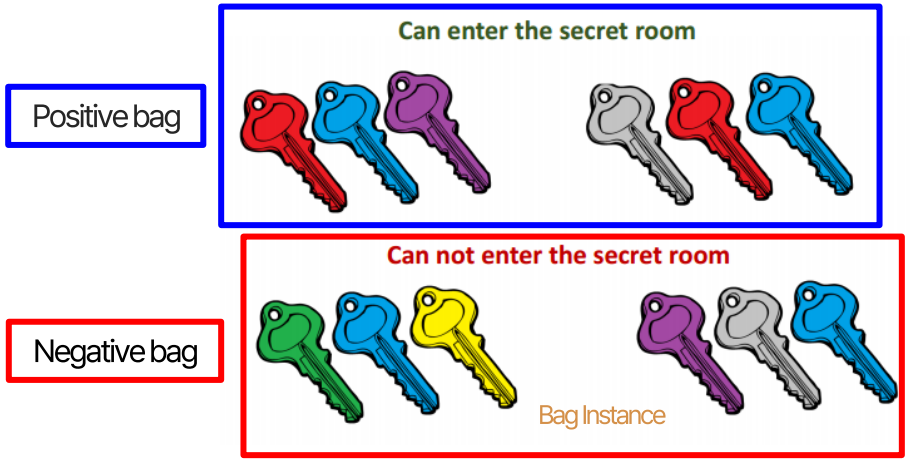
Related Works

Real-world Anomaly Detection in Surveillance Videos (CVPR 2018)

❖ **MIL(Multi Instance Learning)이란?**

- 비디오 전체 라벨만으로 각 segment의 이상 점수를 학습할 수 있게 하는 방법

EX) Instance(**Segment**) 하나하나에 대해서는 라벨이 없고, Bag(**Video**) 전체에는 라벨이 있다.



Q) 새로운 Set이 들어왔을 때 문을 열 수 있을까?

A) NO! (Key = Red)

→ Bag(Video)의 라벨 정보만으로 Instance(Segment) 의 결과 판단이 가능

[2] Sultani, Waqas, Chen Chen, and Mubarak Shah. "Real-world anomaly detection in surveillance videos." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

Related Works

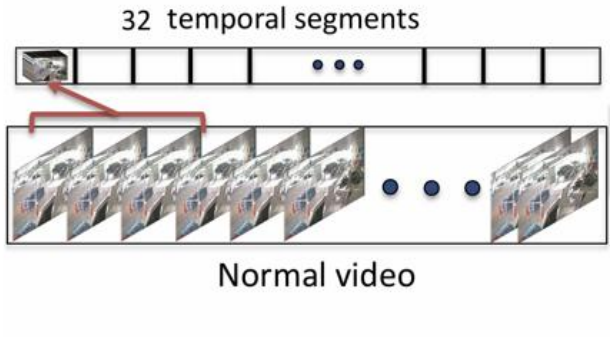
Real-world Anomaly Detection in Surveillance Videos (CVPR 2018)

❖ **MIL을 WSVAD에 적용한다면?**

- MIL Ranking 프레임워크 제시

32개의 Segment

128개의 Frame



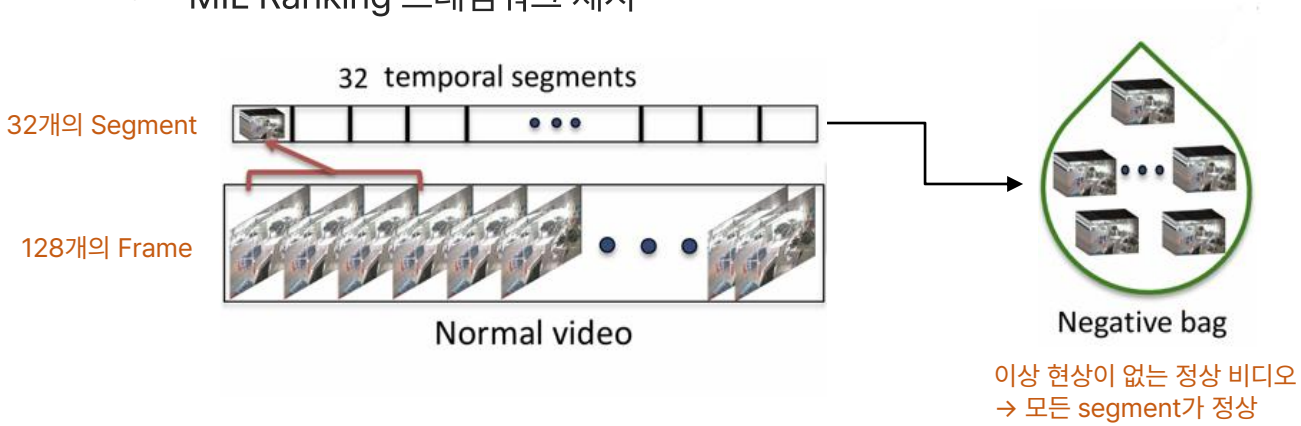
[2] Sultani, Waqas, Chen Chen, and Mubarak Shah. "Real-world anomaly detection in surveillance videos." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

Related Works

Real-world Anomaly Detection in Surveillance Videos (CVPR 2018)

❖ **MIL을 WSVAD에 적용한다면?**

- MIL Ranking 프레임워크 제시



이상 현상이 없는 정상 비디오
→ 모든 segment가 정상

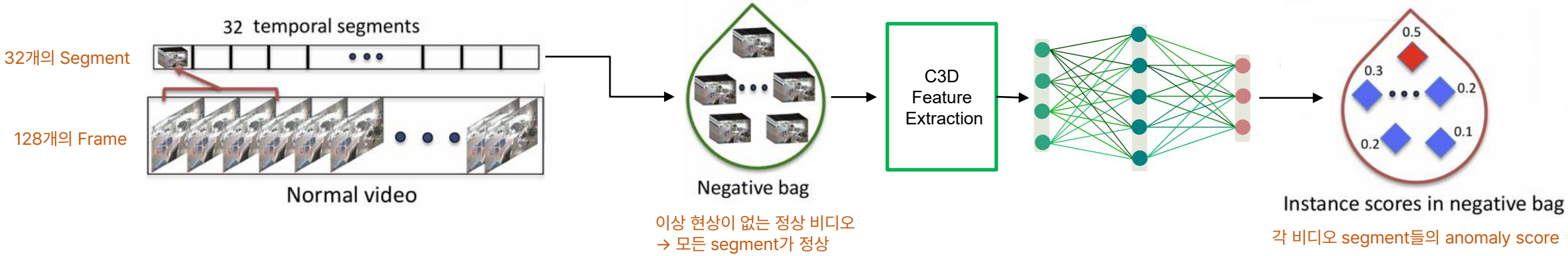
[2] Sultani, Waqas, Chen Chen, and Mubarak Shah. "Real-world anomaly detection in surveillance videos." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

Related Works

Real-world Anomaly Detection in Surveillance Videos (CVPR 2018)

❖ **MIL을 WSVAD에 적용한다면?**

- MIL Ranking 프레임워크 제시



이상 현상이 없는 정상 비디오
→ 모든 segment가 정상

각 비디오 segment들의 anomaly score

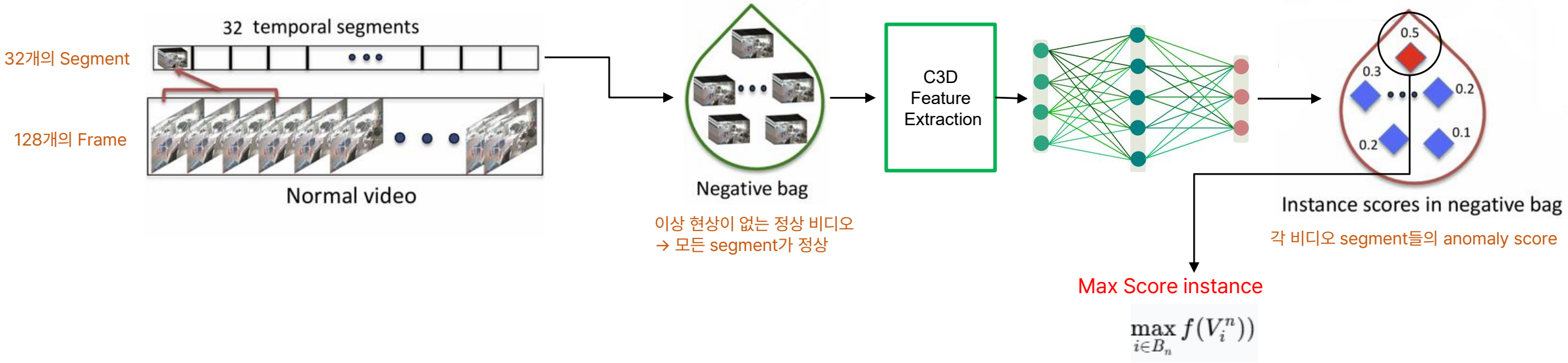
[2] Sultani, Waqas, Chen Chen, and Mubarak Shah. "Real-world anomaly detection in surveillance videos." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

Related Works

Real-world Anomaly Detection in Surveillance Videos (CVPR 2018)

❖ **MIL을 WSVAD에 적용한다면?**

- MIL Ranking 프레임워크 제시



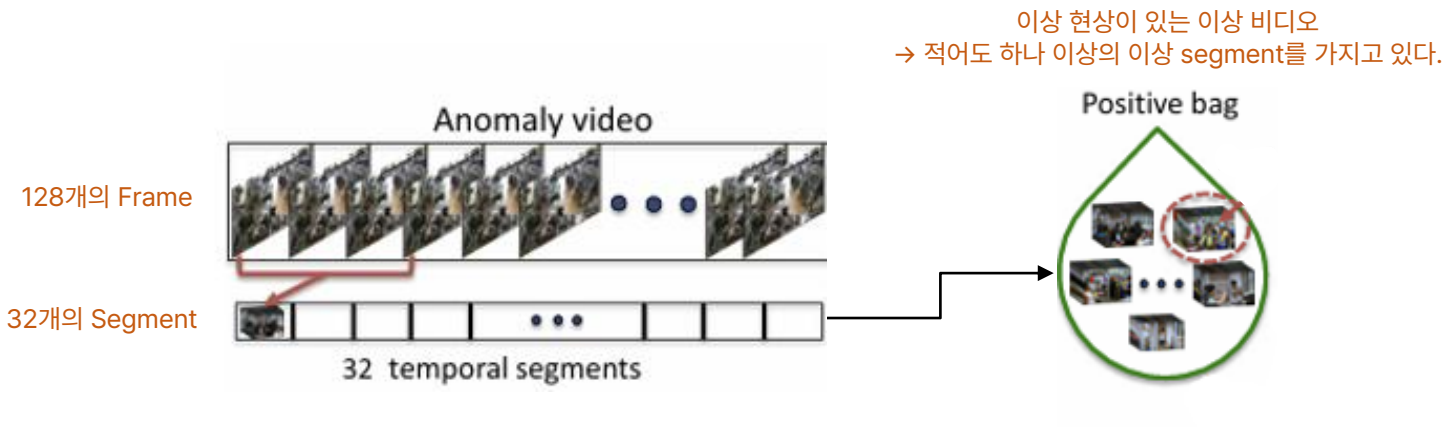
[2] Sultani, Waqas, Chen Chen, and Mubarak Shah. "Real-world anomaly detection in surveillance videos." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

Related Works

Real-world Anomaly Detection in Surveillance Videos (CVPR 2018)

❖ **MIL을 WSVAD에 적용한다면?**

- MIL Ranking 프레임워크 제시



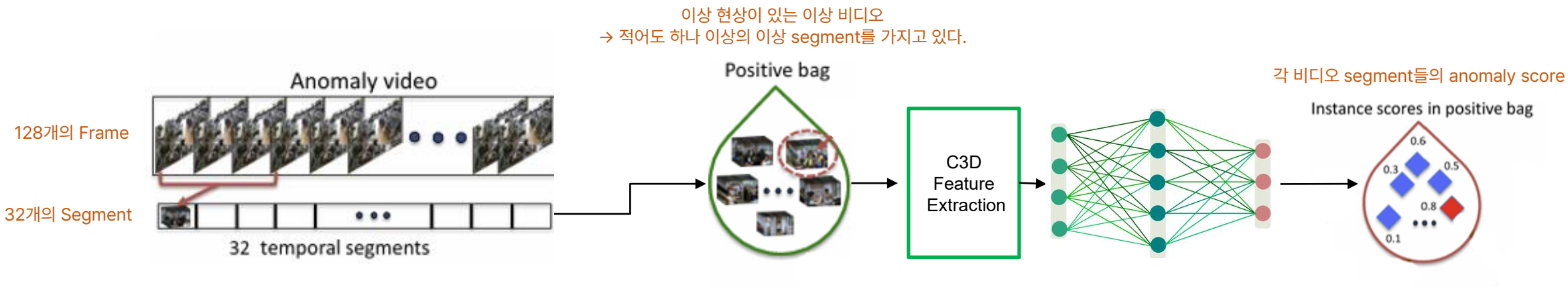
[2] Sultani, Waqas, Chen Chen, and Mubarak Shah. "Real-world anomaly detection in surveillance videos." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

Related Works

Real-world Anomaly Detection in Surveillance Videos (CVPR 2018)

❖ **MIL을 WSVAD에 적용한다면?**

- MIL Ranking 프레임워크 제시



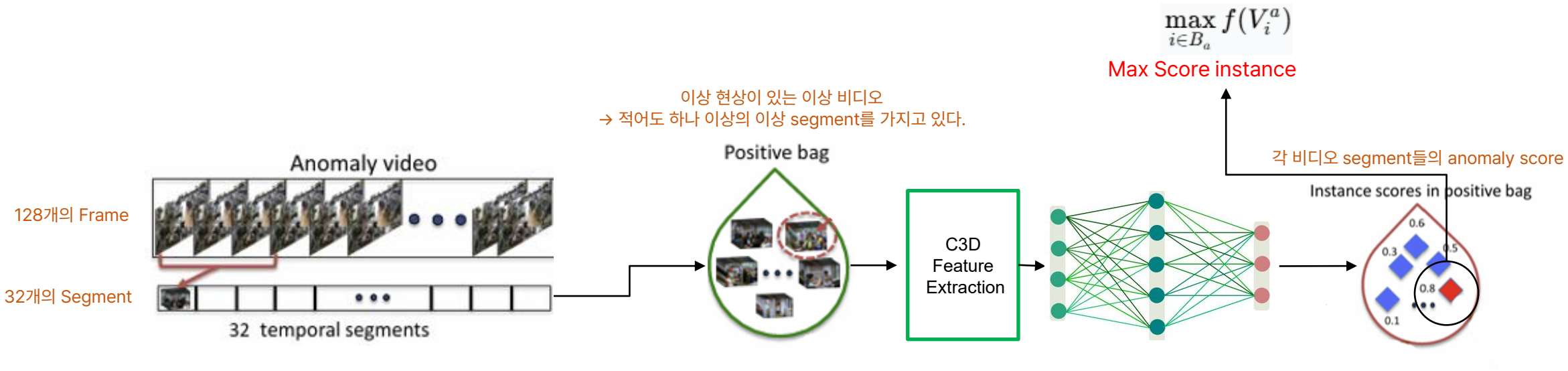
[2] Sultani, Waqas, Chen Chen, and Mubarak Shah. "Real-world anomaly detection in surveillance videos." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

Related Works

Real-world Anomaly Detection in Surveillance Videos (CVPR 2018)

❖ **MIL을 WSVAD에 적용한다면?**

- MIL Ranking 프레임워크 제시



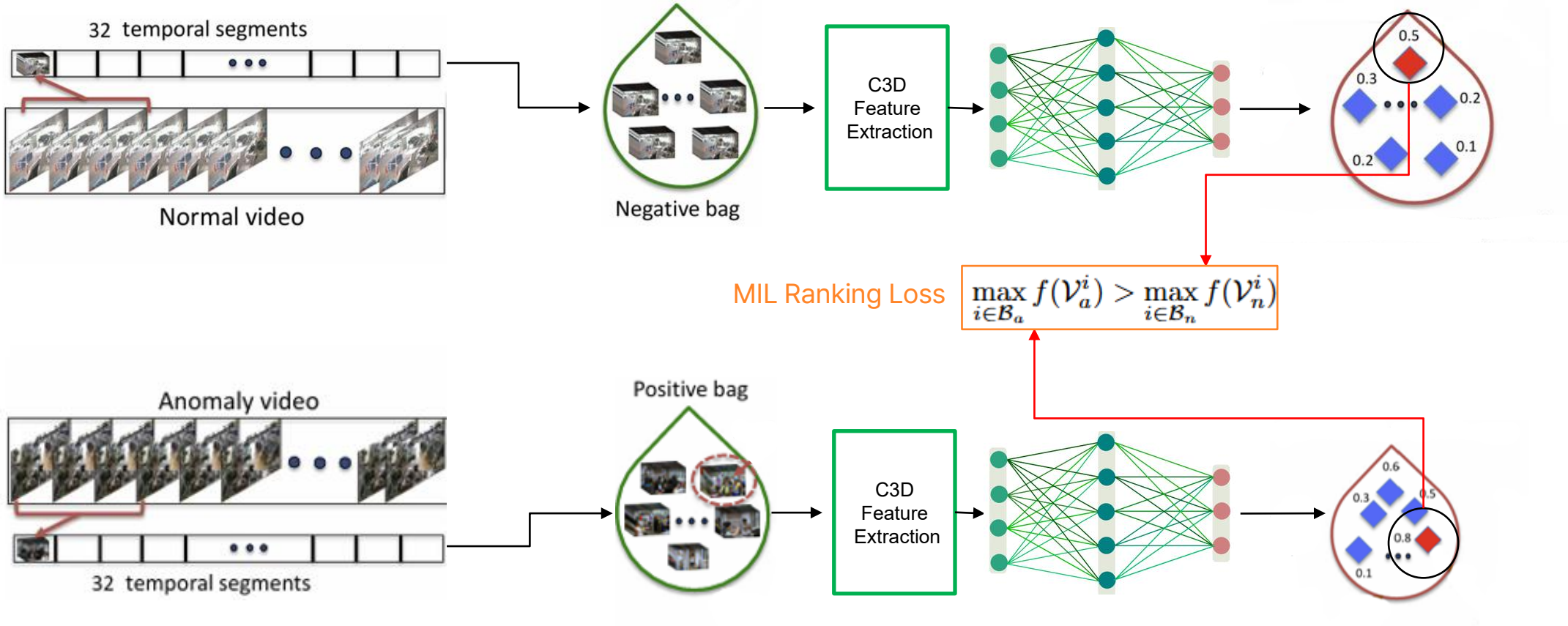
[2] Sultani, Waqas, Chen Chen, and Mubarak Shah. "Real-world anomaly detection in surveillance videos." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

Related Works

Real-world Anomaly Detection in Surveillance Videos (CVPR 2018)

❖ MIL Ranking Loss란?

- 이상 비디오 내의 가장 anomaly score가 높은 segment > 정상 비디오 내의 가장 anomaly score가 높은 segment



[2] Sultani, Waqas, Chen Chen, and Mubarak Shah. "Real-world anomaly detection in surveillance videos." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

Related Works

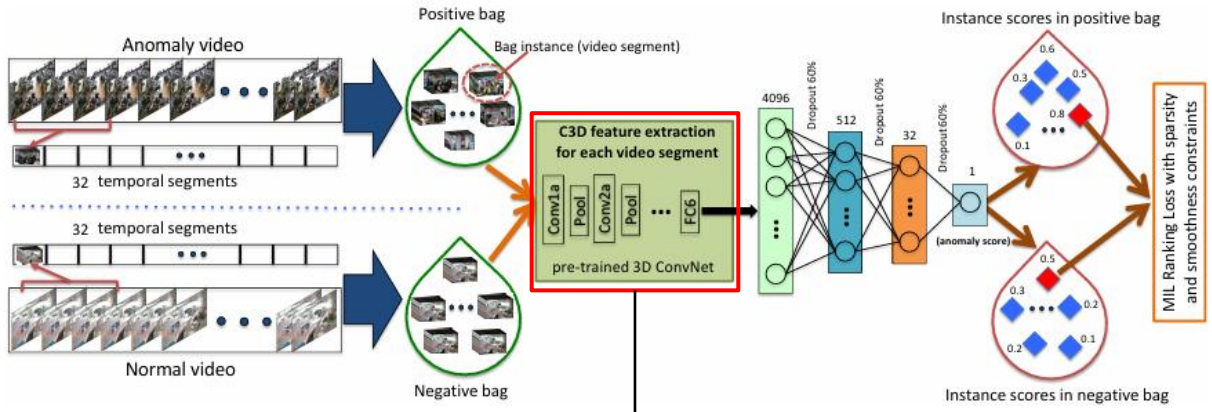
Real-world Anomaly Detection in Surveillance Videos (CVPR 2018)

❖ **해당 논문의 한계점은 무엇일까?**

① C3D 기반 feature의 표현력 한계

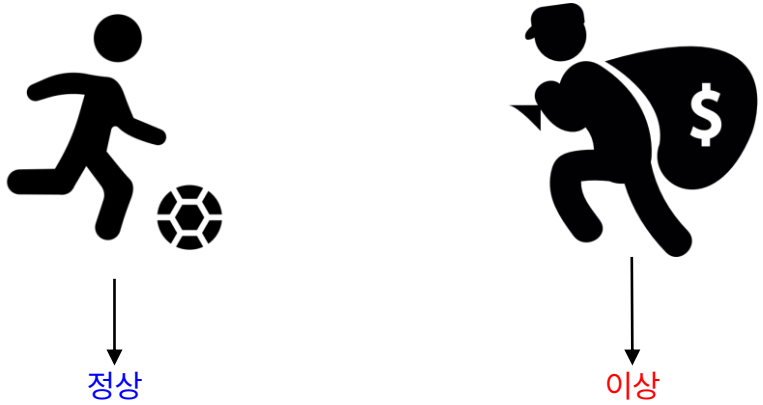
→ 기존 backbone은 행동 인식(Action Recognition) 중심으로 학습

→ 사건의 비정상적 맥락을 표현하기 어렵다.



Anomaly detection 성능이 C3D feature 표현력에 크게 의존

"Running"



같은 행동도 맥락에 따라 다르게 해석된다!

[2] Sultani, Waqas, Chen Chen, and Mubarak Shah. "Real-world anomaly detection in surveillance videos." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

Related Works

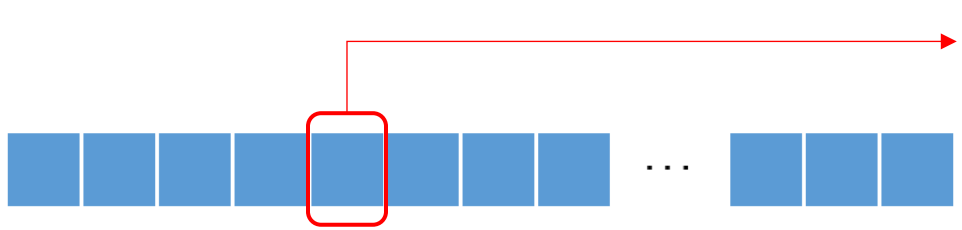
Real-world Anomaly Detection in Surveillance Videos (CVPR 2018)

❖ **해당 논문의 한계점은 무엇일까?**

② Max-score segment 학습에 집중된다.

→ Anomaly 전체 구간을 충분히 학습하기 어렵다.

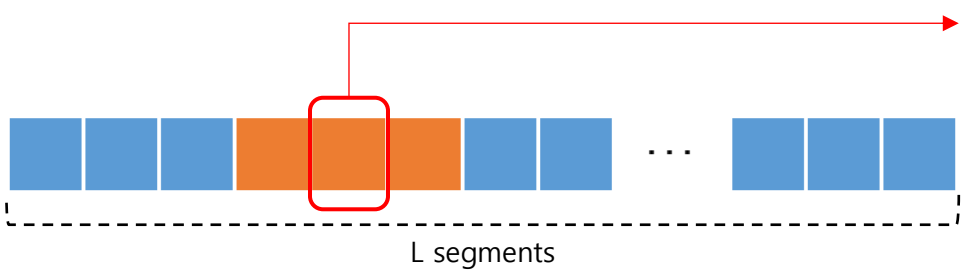
✓ **정상 비디오**



가장 이상해 보이는
정상 비디오 내 segment → $max_{i \in B_n} f(V_n^i)$

∧

✓ **이상 비디오**



가장 이상해 보이는
이상 비디오 내 segment → $max_{i \in B_a} f(V_a^i)$

[2] Sultani, Waqas, Chen Chen, and Mubarak Shah. "Real-world anomaly detection in surveillance videos." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

Related Works

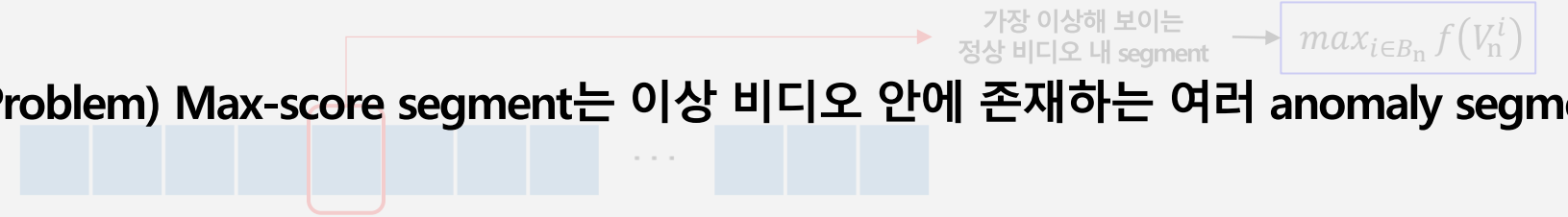
Real-world Anomaly Detection in Surveillance Videos (CVPR 2018)

❖ **해당 논문의 한계점은 무엇일까?**

② Max-score segment 학습에 집중된다.

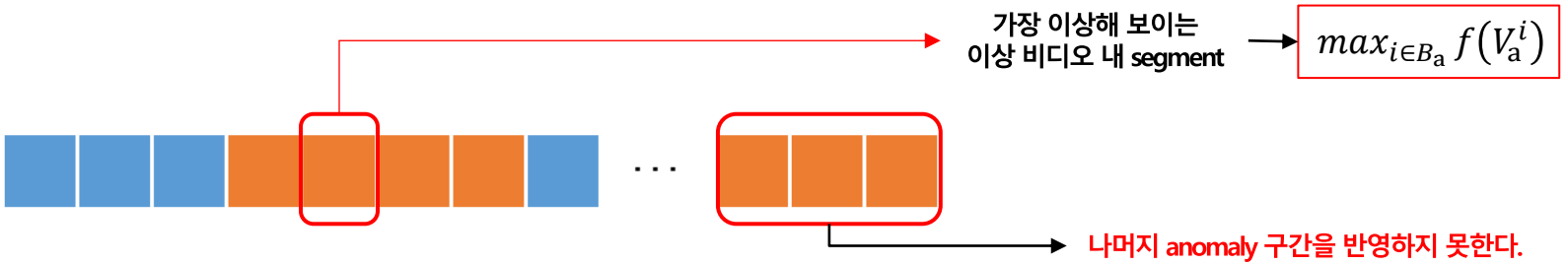
→ Anomaly 전체 구간을 충분히 학습하기 어렵다.

✓ 정상 비디오



Problem) Max-score segment는 이상 비디오 안에 존재하는 여러 anomaly segment를 반영하지 못한다.

✓ **이상 비디오**



[2] Sultani, Waqas, Chen Chen, and Mubarak Shah. "Real-world anomaly detection in surveillance videos." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

Related Works

이상 상황의 맥락과 여러 이상 구간을 함께 고려할 수 있을까?

❖ **CLIP-TSA: CLIP-Assisted Temporal Self-Attention for Weakly Supervised Video Anomaly Detection (ICIP 2023)**

- CLIP의 Vit 기반 visual feature를 통해 이상 상황 맥락을 파악
- TSA(Temporal Self-Attention)로 앞뒤 segment의 흐름을 고려
- Top-k function을 통한 여러 anomaly segment 선택 방법 도입

CLIP-TSA: CLIP-ASSISTED TEMPORAL SELF-ATTENTION FOR WEAKLY-SUPERVISED VIDEO ANOMALY DETECTION

Hyekang Kevin Joo†, Khoa Vo‡, Kashu Yamazaki‡, Ngan Le‡

Dept. of Computer Science, University of Maryland, College Park, MD, USA†
Dept. of Computer Science and Computer Engineering, University of Arkansas, Fayetteville, AR, USA‡

ABSTRACT

ideo anomaly detection (VAD) – commonly formulated as multiple-instance learning problem in a weakly-supervised anner due to its labor-intensive nature – is a challenging roblem in video surveillance where the frames of anomaly eed to be localized in an untrimmed video. In this paper, we rst propose to utilize the ViT-encoded visual features from LIP, in contrast with the conventional C3D or I3D features the domain, to efficiently extract discriminative represen-tions in the novel technique. We then model temporal de-endencies and nominate the snippets of interest by leverag-g our proposed Temporal Self-Attention (TSA). The abla-on study confirms the effectiveness of TSA and ViT feature. he extensive experiments show that our proposed CLIP-TSA utperforms the existing state-of-the-art (SOTA) methods by large margin on three commonly-used benchmark datasets the VAD problem (UCF-Crime, ShanghaiTech Campus and D-Violence). Our source code is available at https://ithub.com/joos2010kj/CLIP-TSA.

***Index Terms*—** video anomaly detection, temporal self-ttention, weakly supervised, multimodal model, subtlety

1. INTRODUCTION

ideo action understanding is an active research field with any applications, e.g., action localization [1, 2, 3, 4] action cognition [5, 6, 7, 8, 9, 10], video captioning [11, 12, 13, 4], etc [15]. Video anomaly detection (VAD) is the task of calizing anomalous events in a given video with three main aradigms, *i.e.*, Fully-supervised (Sup.) [16] , Un-Sup [17] and Weakly-Sup [18]. While it generally yields high per-ormance, Fully-Sup VAD requires fine-grained anomaly la-els (*i.e.*, *frame-level* normal/abnormal annotations), and the roblem has traditionally suffered from the laborious nature f data annotation. In Un-Sup VAD, one-class classification OCC) [19] is a common approach, where the model is trained n only normal class samples with the assumption that unseen bnormal videos have high reconstruction errors. However, e performance of Un-Sup VAD is usually poor as it lacks

Fig. 1. Overall chart of CLIP-TSA in train time, with $\mathcal{X}^{2*B}$ as input. Each video $\mathcal{X}$ consisting of N frames (*i.e.*, $\mathcal{X} = \{x_j\}_1^N$) is divided into a set of δ-frame snippets $\{s_i\}_1^T$. Each δ-frame snippet s_i is represented by a V-L feature $f_i \in \mathbb{R}^d$. Then, the video is represented by $\mathcal{F} = \{f_i\}_1^T$. Our TSA is then applied to obtain anomaly attention feature $\tilde{\mathcal{F}} = \{\tilde{f}_i\}_1^T$. The features $\tilde{\mathcal{F}}$ from $2 * B$ videos then undergo difference maximization trainer to weakly-train anomaly classifier.

prior knowledge of abnormality and from its inability to cap-ture all normality variations [20]. Compared to both Un-Sup and Sup VAD, Weakly-Sup VAD is considered the most prac-tical approach because of its competitive performance and an-notation efficiency by employing *video-level* labels to reduce the cost of manual fine-grained annotations [21].

In the Weakly-Sup VAD, there exist two fundamental problems. First, anomalous-labeled frames tend to be domi-nated by normal-labeled frames, as the videos are untrimmed and there is no strict length requirement for the anomalies in the video. Second, the anomaly may not necessarily stand out against normality. As a result, it occasionally becomes challenging to localize anomaly frames. Thus, the prob-lem is commonly designed with multiple instance learning (MIL) framework [18], where a video is treated as a bag containing multiple instances, each instance being a video snippet. The video is labeled as anomalous if any of its snippets are anomalous, and normal if all of its snippets are normal. Anomalous-labeled videos belong to the positive bag and normal-labeled videos belong to the negative bag. How-ever, the existing MIL-based Weakly-Sup VAD approaches are limited in dealing with an arbitrary number of abnormal snippets in an abnormal video. To address such an issue, we are inspired by the differentiable top-K operator [22] and introduce a novel technique, termed top-κ function, that lo-calizes κ snippets of interest in the video with differentiable

[3] Joo, Hyekang Kevin, et al. "Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection." *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023.

Related Works

CLIP-TSA: CLIP-Assisted Temporal Self-Attention for Weakly Supervised Video Anomaly Detection (ICIP 2023)

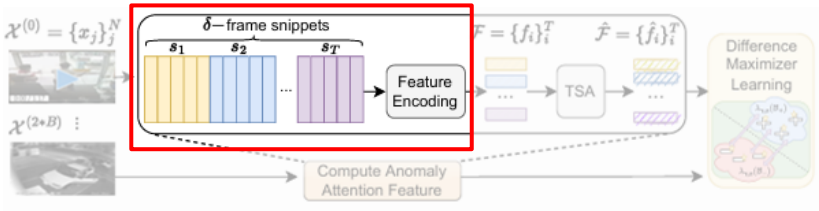
❖ **이상 상황의 맥락을 더 잘 표현할 수 있는 feature는 무엇일까?**

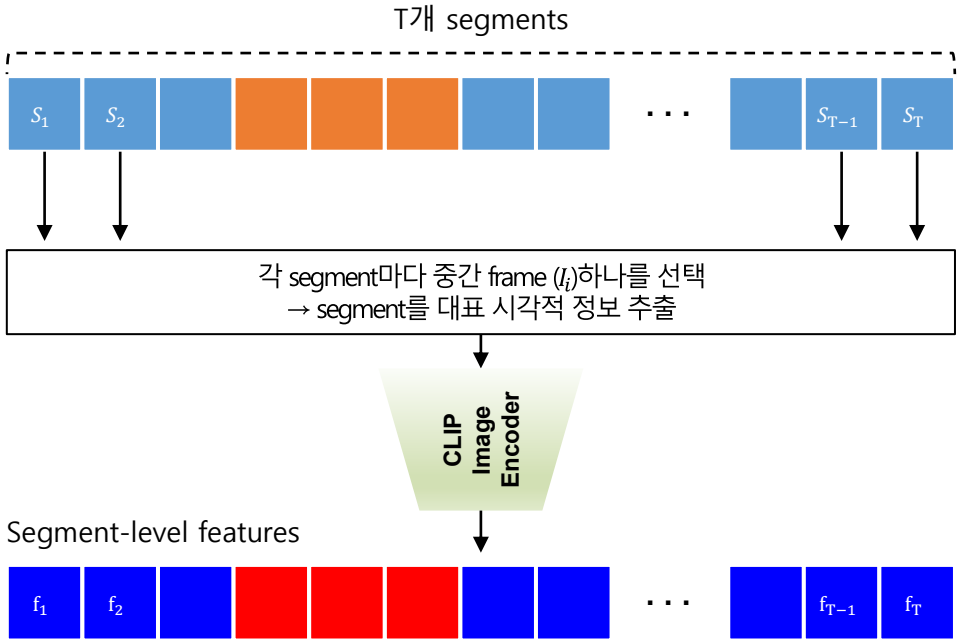
- 기존 C3D feature는 action recognition 중심으로 학습

→ 행동 자체는 잘 표현하지만, 이상 상황의 의미적 맥락은 반영하는데 한계가 있다.

- CLIP은 Image-Text 의미 관계를 학습

→ Image Encoder가 장면, 객체, 맥락 정보를 풍부하게 표현





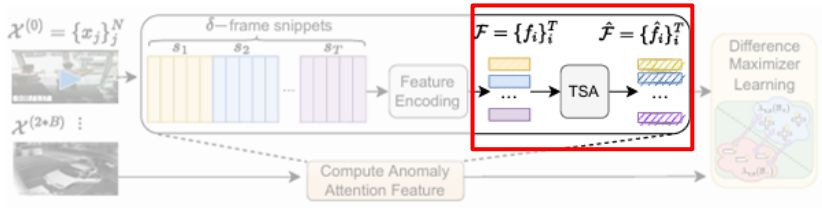
[3] Joo, Hyekang Kevin, et al. "Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection." *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023.

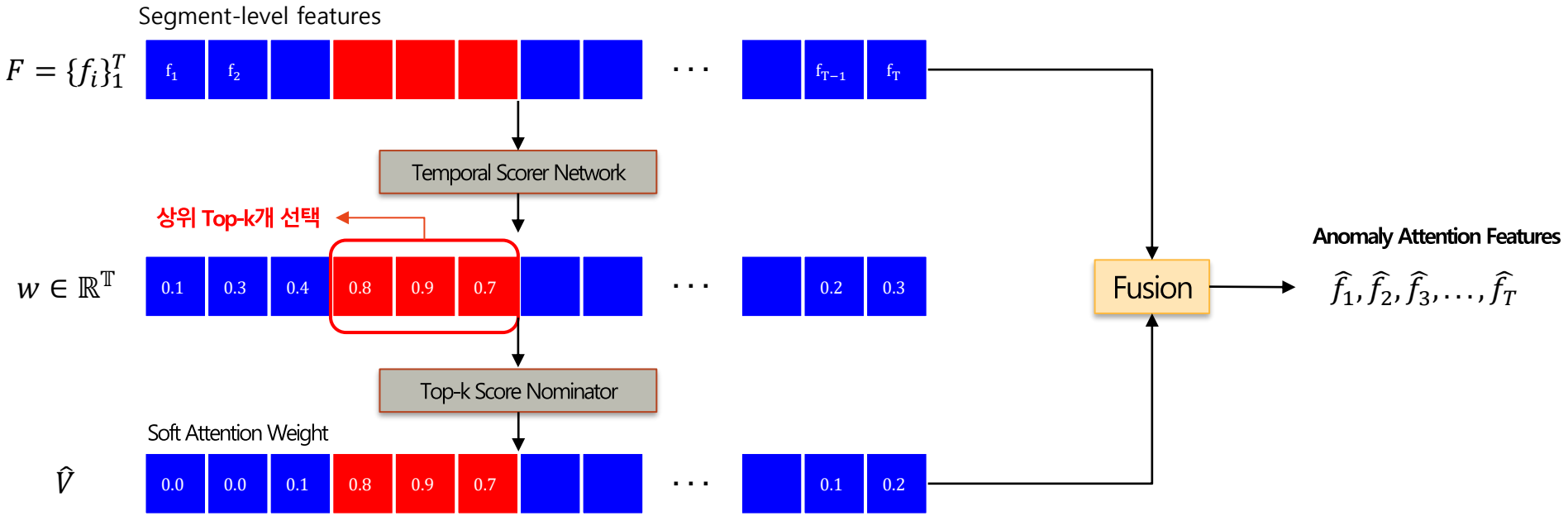
Related Works

CLIP-TSA: CLIP-Assisted Temporal Self-Attention for Weakly Supervised Video Anomaly Detection (ICIP 2023)

❖ **어느 구간이 이상 징후를 담고 있을까?**

- TSA(Temporal Self-Attention): 비디오 내의 이상 징후가 강한 구간에 더 집중
- Top-k 방식으로 이상과 관련성이 높은 여러 구간을 선택





[3] Joo, Hyekang Kevin, et al. "Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection." *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023.

Related Works

CLIP-TSA: CLIP-Assisted Temporal Self-Attention for Weakly Supervised Video Anomaly Detection (ICIP 2023)

❖ Experiments

- 2가지 데이터셋(UCF-Crime, XD-Violence)에서 성능 평가
- 기존 C3D, I3D 기반 방법 대비 큰 성능 향상

✓ AUC (↑) (UCF-Crime): 정상과 이상을 잘 구분하는가?

✓ AP (↑) (XD-Violence): 이상이라고 판단한 것들이 실제 이상인가?

Table 2. Comparisons on UCF-Crime Dataset [18]

Sup.	Method	Venue	Feature	AUC@ROC ↑
Un-	Lu et al. [31]	ICCV'13	C3D	65.51
	Hasan [32]	CVPR'16	-	50.60
	BODS [33]	ICCV'19	I3D	68.26
	GODS [33]	ICCV'19	I3D	70.46
	GCL [34]	CVPR'22	ResNext	71.04
Fully-	Liu & Ma [16]	MM'19	NLN	82.0
Weakly-	GCN [21]	CVPR'19	TSN	82.12
	GCL_{WS} [34]	CVPR'21	ResNext	79.84
	Purwanto et al. [35]	ICCV'21	TRN	85.00
	Thakare et al. [36]	ExpSys'22	C3D+I3D	84.48
	Sultani et al. [18]	CVPR'18	C3D	75.41
	Zhang et al. [37]	ICIP'19		78.70
	GCN [21]	CVPR'19		81.08
	CLAWS [19]	ECCV'20		83.03
	RTFM [38]	ICCV'21		83.28
	Sultani et al. [18]	CVPR'18	I3D	77.92
	Wu et al. [39]	ECCV'20		82.44
	DAM [40]	AVSS'21		82.67
	RTFM [38]	ICCV'21		84.30
	Wu & Liu [41]	TIP'21		84.89
	BN-SVP [42]	CVPR'22		83.39
	Ours: CLIP-TSA		CLIP	**87.58**

Table 3. Comparisons on XD-Violence Dataset [39]

Sup.	Modality	Method	Venue	Feature	AUC@PR ↑
Un-	–	OCSVM [43]	NeurIPS'00	–	27.25
		Hasan et al. [32]	CVPR'16	–	30.77
Weakly-	Vision & Audio	Wu et al. [39]	ECCV'20	I3D(V) + VGGish(A)	78.64
		Wu & Liu [41]	TIP'21	I3D(V) + VGGish(A)	75.90
		Pang et al. [44]	ICASSP'21	I3D(V) + VGGish(A)	81.69
		MACIL-SD [45]	MM'22	I3D(V) + VGGish(A)	83.40
		DDL [46]	ICCECE'22	I3D(V) + VGGish(A)	**83.54**
	Vision	Sultani et al. [18]	CVPR'18	C3D(V)	73.20
		RTFM [38]	ICCV'21	C3D(V)	75.89
		RTFM [38]	ICCV'21	I3D(V)	77.81
		Ours: CLIP-TSA		CLIP(V)	82.19

[3] Joo, Hyekang Kevin, et al. "Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection." *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023.

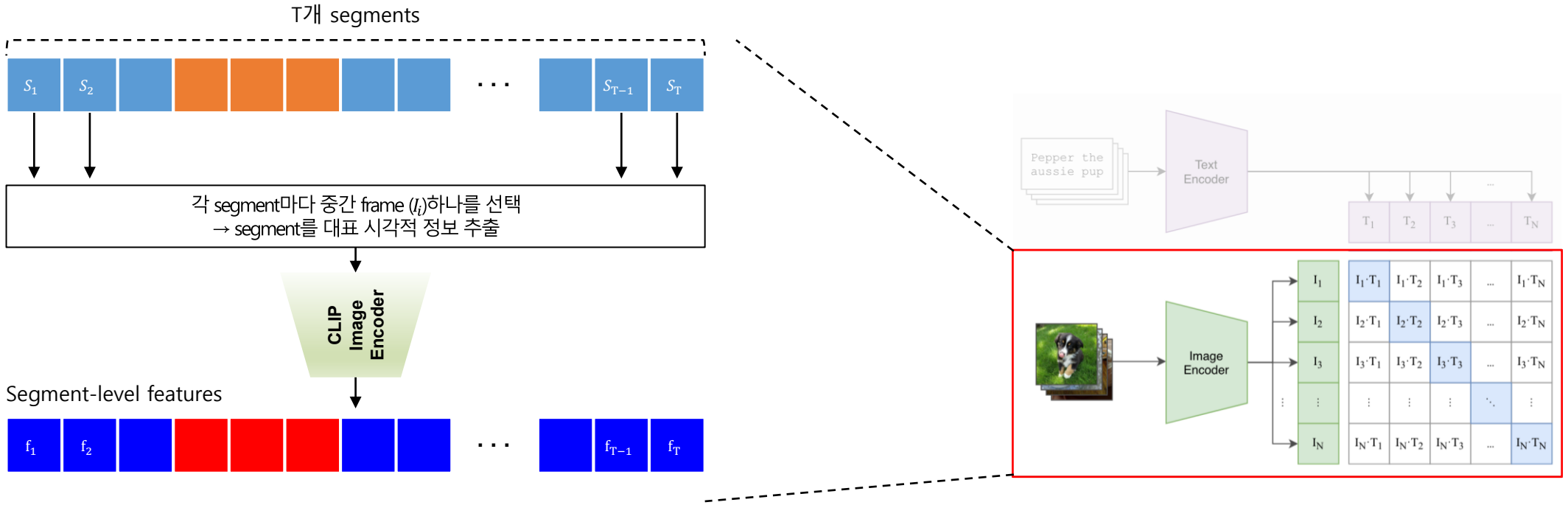
Related Works

CLIP-TSA: CLIP-Assisted Temporal Self-Attention for Weakly Supervised Video Anomaly Detection (ICIP 2023)

❖ **해당 논문의 한계점은 무엇일까?**

① CLIP을 Visual Feature Extractor로만 사용

→ CLIP의 Text Encoder나 언어 지식을 활용하지 못한다.



[3] Joo, Hyekang Kevin, et al. "Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection." *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023.

Related Works

CLIP-TSA: CLIP-Assisted Temporal Self-Attention for Weakly Supervised Video Anomaly Detection (ICIP 2023)

❖ **해당 논문의 한계점은 무엇일까?**

② Anomaly score를 기준으로 "이 구간이 정상인지, 이상인지"를 판단한다.

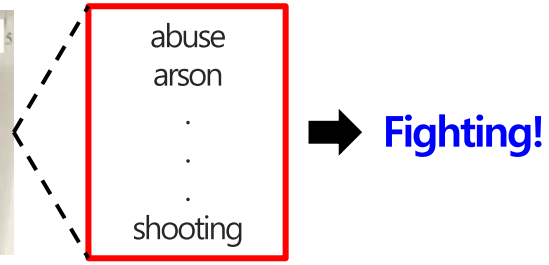
→ 정상/ 이상 여부 외의 해당 segment가 어떤 종류의 이상 현상인지는 구분하지 못한다.

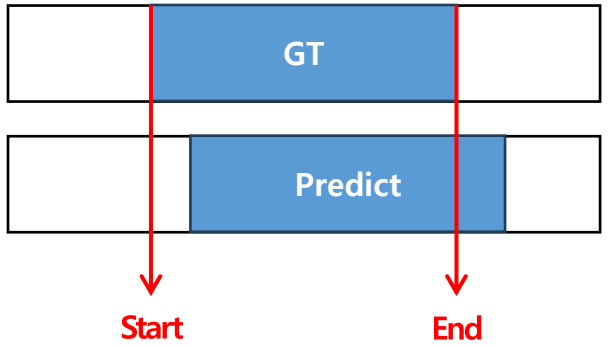
➢ **Coarse-grained:** "이 장면이 정상인지, 이상인지"를 판별 (Binary Classification)

- 정상 비디오 (y=0)
: 비디오 내의 모든 프레임에 이상 징후가 없는 경우

- 이상 비디오 (y=1)
: 비디오 내에 최소 하나 이상의 프레임에 이상 징후가 포함된 경우

➢ **Fine-grained:** "어떤 종류의 이상이 발생했는가?" (Category) + "이상 현상은 언제 발생했는가?" (Precision)





[3] Joo, Hyekang Kevin, et al. "Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection." *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023.

Related Works

CLIP의 사전 학습된 언어-시각 지식을 WSVAD에 활용한다면?

❖ **VadCLIP: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection (AAAI 2024)**

- Pretrained & Frozen CLIP 모델을 활용한 WSVAD
- Dual-branch를 통해 Coarse와 Fine-grained 이상 탐지를 동시에 수행

VadCLIP: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection

Peng Wu[1], Xuerong Zhou[1], Guansong Pang[2*], Lingru Zhou[1], Qingsen Yan[1], Peng Wang[1*], Yanning Zhang[1]
[1]ASGO, School of Computer Science, Northwestern Polytechnical University, China
[2]School of Computing and Information Systems, Singapore Management University, Singapore
{xdwupeng, zxr2333}@gmail.com, gspang@smu.edu.sg, {lingruzhou, yqs}@mail.nwpu.edu.cn, {peng.wang, ynzhang}@nwpu.edu.cn

Abstract

The recent contrastive language-image pre-training (CLIP) model has shown great success in a wide range of image-level tasks, revealing remarkable ability for learning powerful visual representations with rich semantics. An open and worthwhile problem is efficiently adapting such a strong model to the video domain and designing a robust video anomaly detector. In this work, we propose VadCLIP, a new paradigm for weakly supervised video anomaly detection (WSVAD) by leveraging the frozen CLIP model directly without any pre-training and fine-tuning process. Unlike current works that directly feed extracted features into the weakly supervised classifier for frame-level binary classification, VadCLIP makes full use of fine-grained associations between vision and language on the strength of CLIP and involves dual branch. One branch simply utilizes visual features for coarse-grained binary classification, while the other fully leverages the fine-grained language-image alignment. With the benefit of dual branch, VadCLIP achieves both coarse-grained and fine-grained video anomaly detection by transferring pre-trained knowledge from CLIP to WSVAD task. We conduct extensive experiments on two commonly-used benchmarks, demonstrating that VadCLIP achieves the best performance on both coarse-grained and fine-grained WSVAD, surpassing the state-of-the-art methods by a large margin. Specifically, VadCLIP achieves 84.51% AP and 88.02% AUC on XD-Violence and UCF-Crime, respectively. Code and features are released at https://github.com/nwpu-zxr/VadCLIP.

Introduction

n recent years, weakly supervised video anomaly detection WSVAD, VAD) has received growing concerns due to its road application prospects. For instance, with the aid of WSVAD, it is convenient to develop more powerful intelligent video surveillance systems and video content review ystems. In WSVAD, the anomaly detector is expected to enerate frame-level anomaly confidences with only video-evel annotations provided. The majority of current research n this field follows a systematic process, wherein the ini-ial step is to extract frame-level features using pre-trained isual models, e.g., C3D (Tran et al. 2015; Sultani, Chen, nd Shah 2018), I3D (Carreira and Zisserman 2017; Wu

Figure 1: Comparisons of different paradigms for WSVAD.

et al. 2020), and ViT (Dosovitskiy et al. 2020; Li, Liu, and Jiao 2022), followed by feeding these features into multiple instance learning (MIL) based binary classifiers for the purpose of model training, and the final step is to detect abnormal events based on predicted anomaly confidences. Despite their simple schemes and promising results, such a classification-based paradigm fails to take full advantage of cross-modal relationships, e.g, vision-language associations.

During the past two years, we have witnessed great progress in the development of vision-language pre-training (VLP) models (Kim, Son, and Kim 2021; Jia et al. 2021; Wang et al. 2021; Chen et al. 2023a), e.g., CLIP (Radford et al. 2021), for learning more generalized visual representations with semantic concepts. The main idea of CLIP is to align images and texts by contrastive learning, that is, pull together images and matched textual descriptions while pushing away unmatched pairs in the joint embedding space. Thanks to hundreds of million noisy image-text pairs crawled from the web, such models pre-trained at a large scale really demonstrate their strong representation learning as well as associations between vision and language. In view of the breakthrough performance of CLIP, recently, building task-specific models on top of CLIP is becoming emerging research topics and applied to a broad range of vision tasks, and these models achieve unprecedented performance.

Although CLIP and its affiliated models demonstrate the great potential on various vision tasks, these methods mainly focus on the image domain. Therefore, how to efficiently adapt such a model learned from image-text pairs to more

*Corresponding Authors

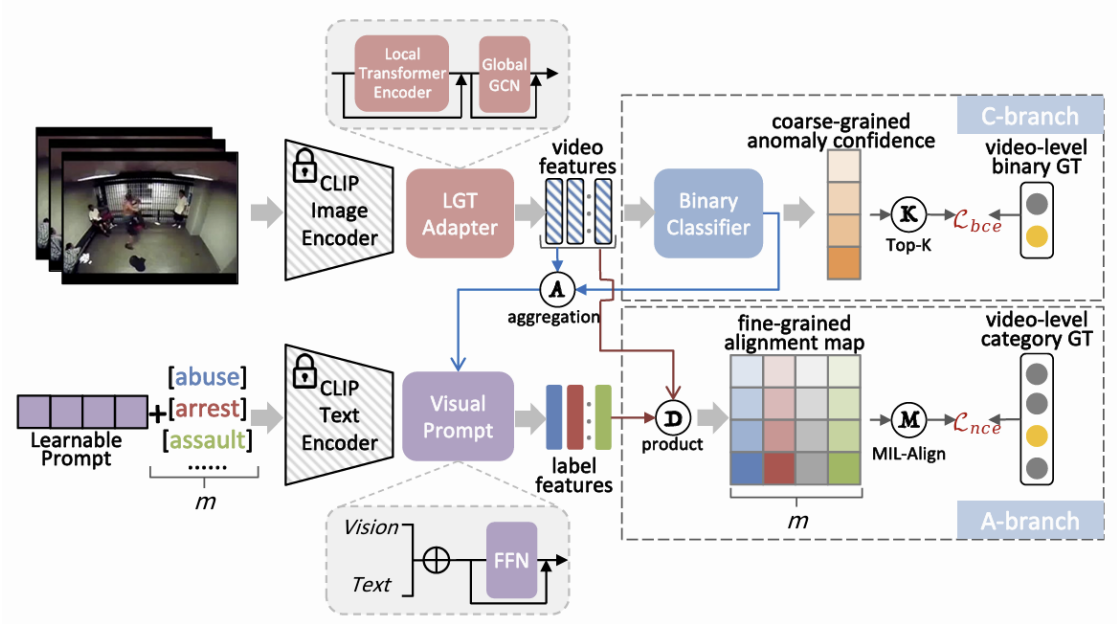
[4] Wu, Peng, et al. "Vadclip: Adapting vision-language models for weakly supervised video anomaly detection." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 38. No. 6. 2024.

Related Works

VadCLIP: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection (AAAI 2024)

❖ **해당 논문이 해결하고자 했던 문제점은?**

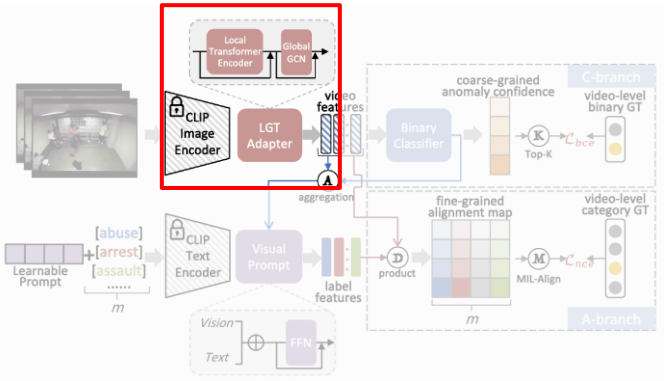
- Challenge
  - ① Temporal Dependency를 효과적으로 포착할 것인가? **LGT-Adapter**
  - ② Pretrained CLIP을 어떻게 활용할 것인가? **Dual-branch fashion**
  - ③ 각 segment가 어떤 종류의 이상 상황인지 어떻게 알 수 있는가? **MIL-Align**



[4] Wu, Peng, et al. "Vadclip: Adapting vision-language models for weakly supervised video anomaly detection." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 38. No. 6. 2024.

Related Works

VadCLIP: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection (AAAI 2024)

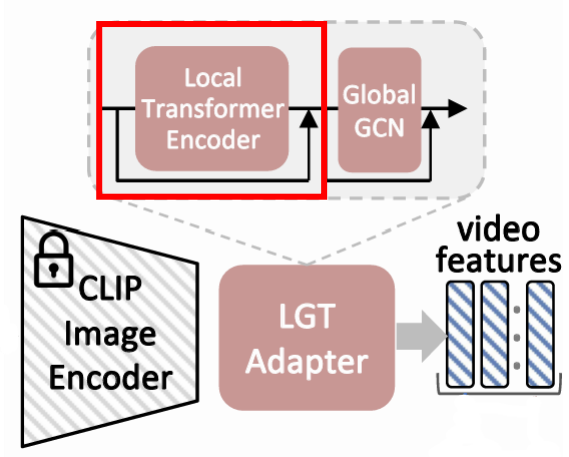


❖ **LGT-Adapter (Local-Global Temporal Adapter)**

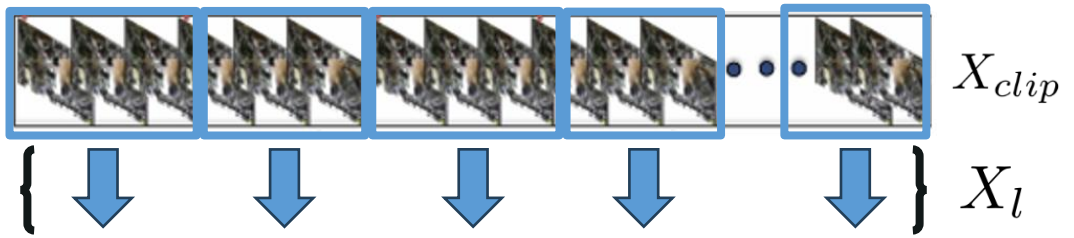
- Challenge 1: Temporal Dependency를 어떻게 효과적으로 포착할 것인가?

  CLIP은 이미지에 대해서 pretrained → 시간적 순서를 반영하지 못한다.

  1) **Local Temporal Adapter**: 가까운 시간적 관계(Local Temporal Dependencies) 포착



Local Self-Attention

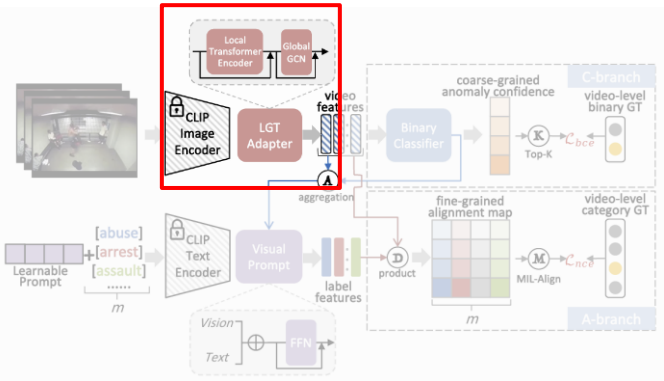


- **Shooting**
시간이 짧은 Event

[4] Wu, Peng, et al. "Vadclip: Adapting vision-language models for weakly supervised video anomaly detection." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 38. No. 6. 2024.

Related Works

VadCLIP: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection (AAAI 2024)

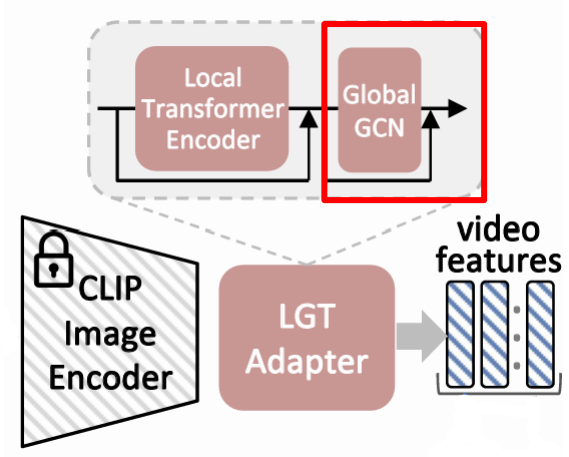


❖ **LGT-Adapter (Local-Global Temporal Adapter)**

- Challenge 1: Temporal Dependency를 어떻게 효과적으로 포착할 것인가?

  CLIP은 이미지에 대해서 pretrained → 시간적 순서를 반영하지 못한다.

  2) **Global Temporal Adapter**: 장기적인 시간적 관계(Global Temporal Dependencies) 포착



Feature Similarity & Position Distance

X_l

모든 프레임끼리 1:1로 매칭

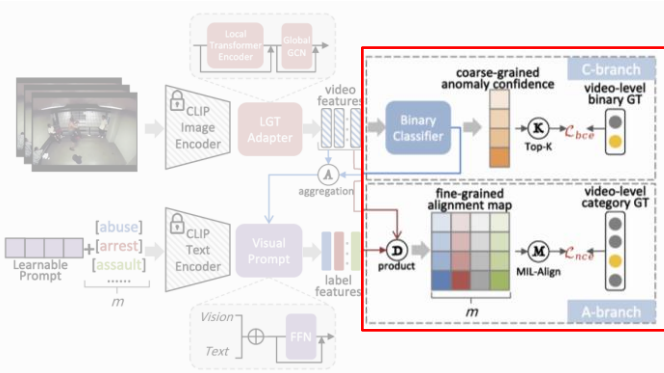
➢ 두 Frame 간의 cosine 유사도 & 시간적 거리 측정

- **Arson**
  시간이 긴 Event

[4] Wu, Peng, et al. "Vadclip: Adapting vision-language models for weakly supervised video anomaly detection." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 38. No. 6. 2024.

Related Works

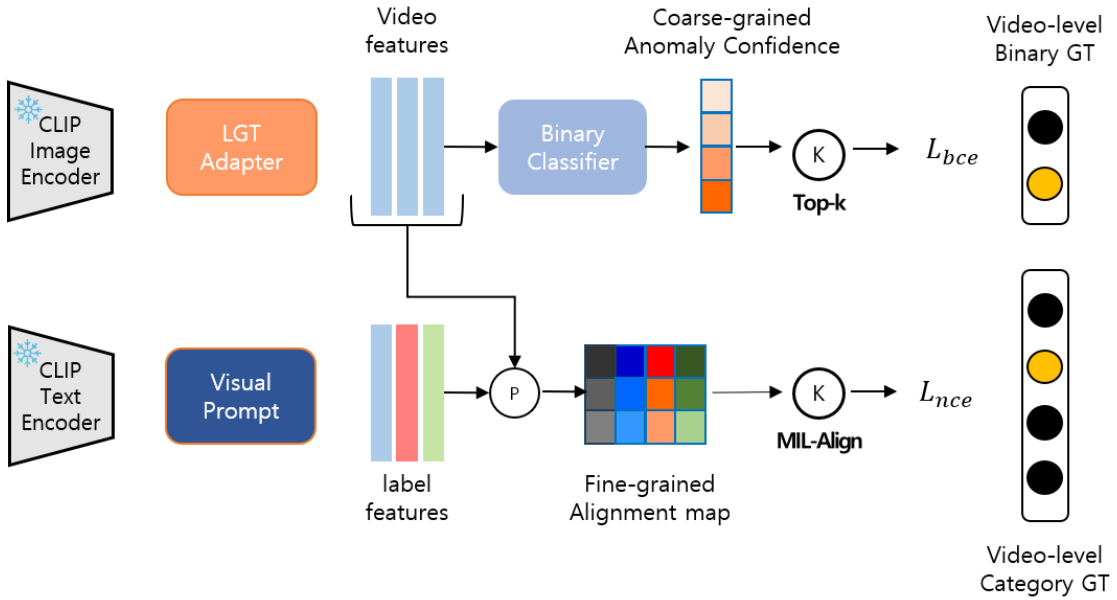
VadCLIP: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection (AAAI 2024)



❖ **Dual-branch fashion**

- Challenge 2: Pretrained CLIP을 어떻게 활용할 것인가?

  C(Classification)-branch + A(Alignment)-branch



[4] Wu, Peng, et al. "Vadclip: Adapting vision-language models for weakly supervised video anomaly detection." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 38. No. 6. 2024.

Related Works

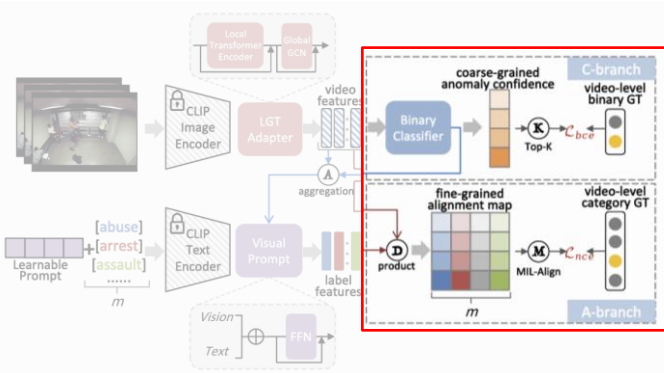
VadCLIP: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection (AAAI 2024)

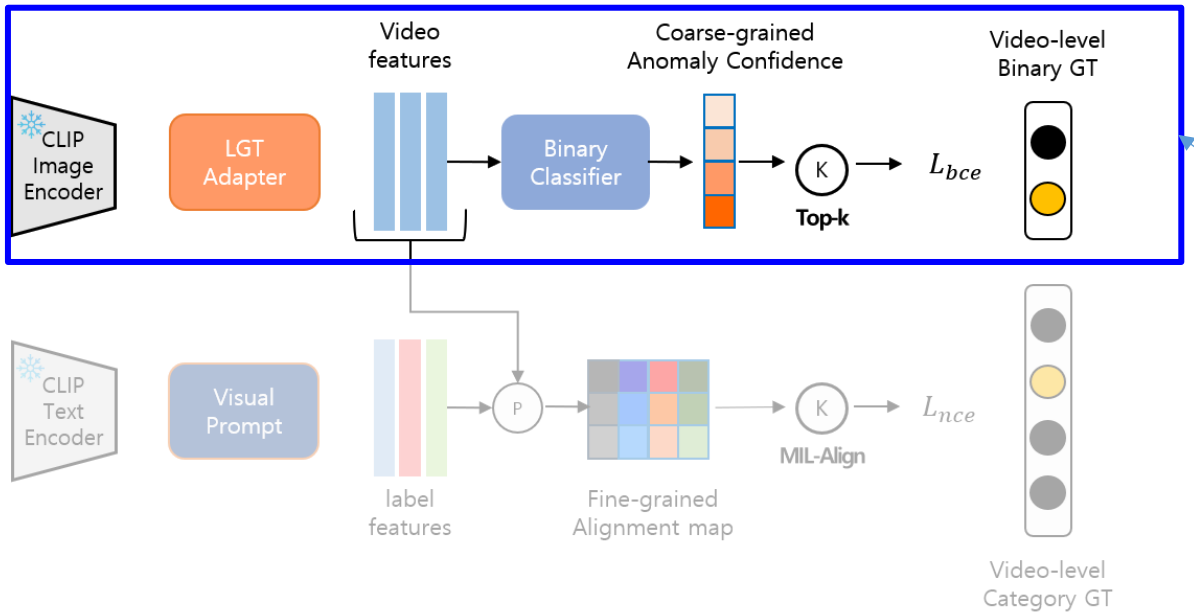
❖ **Dual-branch fashion**

- Challenge 2: Pretrained CLIP을 어떻게 활용할 것인가?

✓ **C-branch**: Video 전체에 대한 이상 여부

→ visual feature만을 사용하여 Binary Classification





Coarse-grained

Normal / Anomaly? ➡ *Anomaly!*

[4] Wu, Peng, et al. "Vadclip: Adapting vision-language models for weakly supervised video anomaly detection." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 38. No. 6. 2024.

Related Works

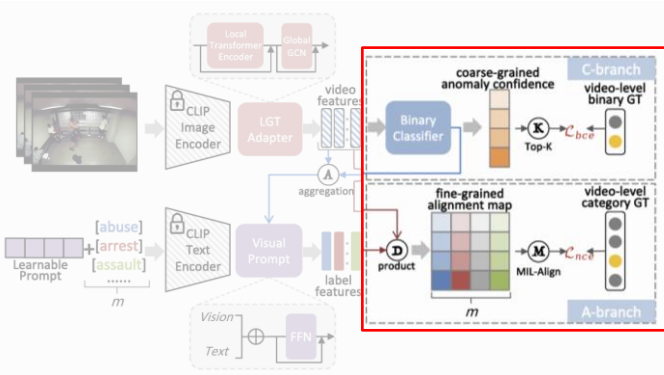
VadCLIP: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection (AAAI 2024)

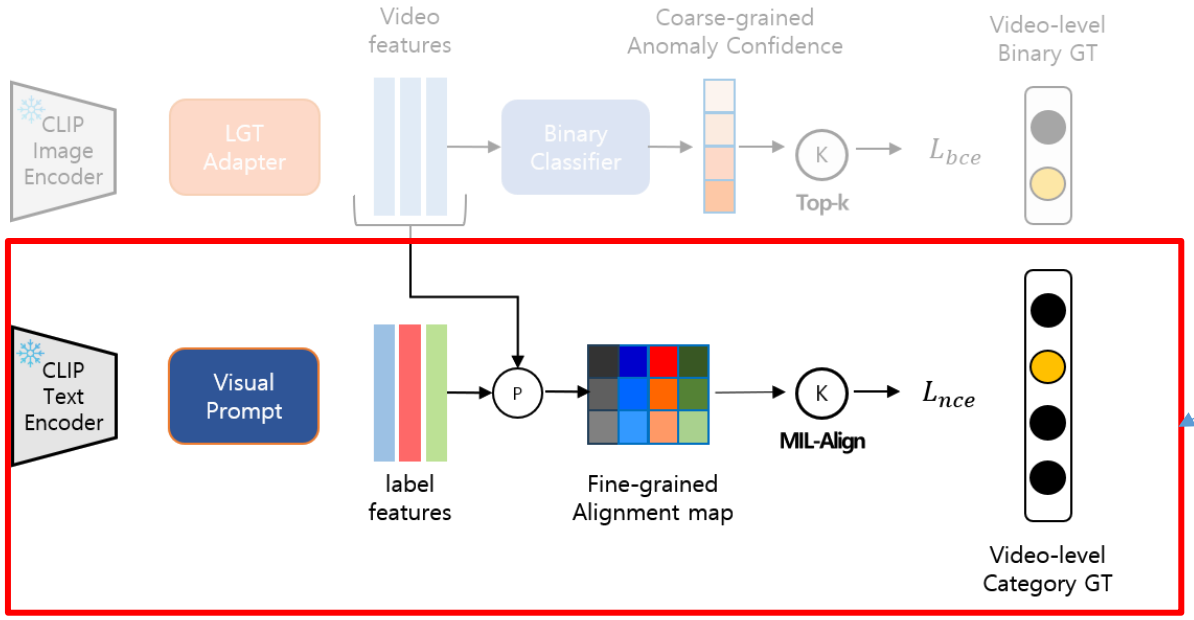
❖ **Dual-branch fashion**

- Challenge 2: Pretrained CLIP을 어떻게 활용할 것인가?

✓ **A-branch**: 어떤 종류의 이상 현상인지 구분

→ visual context + Text label





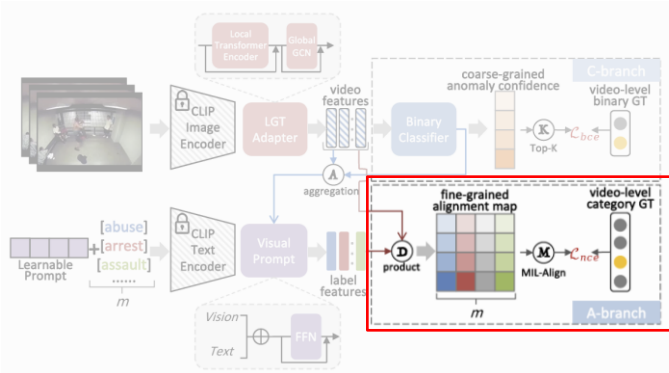
Abuse / Shooting / Stealing,.. ➡ *Stealing!*

Fine-grained

[4] Wu, Peng, et al. "Vadclip: Adapting vision-language models for weakly supervised video anomaly detection." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 38. No. 6. 2024.

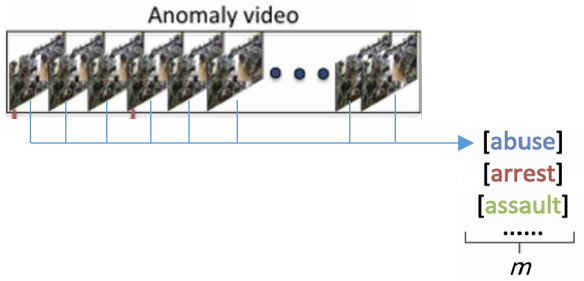
Related Works

VadCLIP: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection (AAAI 2024)

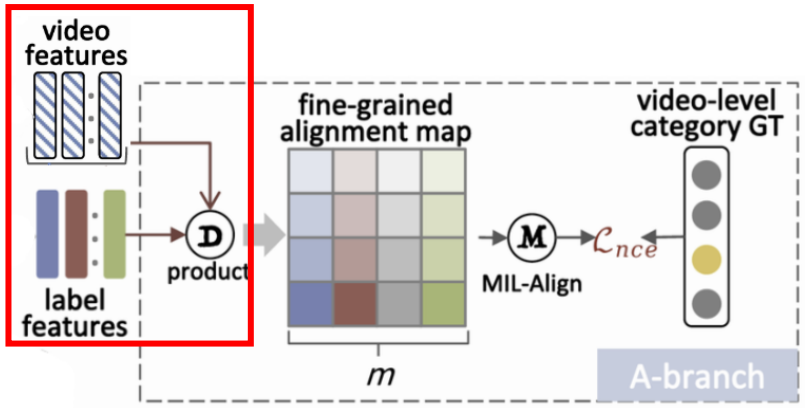


❖ **MIL-Align**

- Challenge 3: 각 segment가 어떤 종류의 이상 상황인지 어떻게 알 수 있을까?
  - ✓ **Alignment Map "M"**



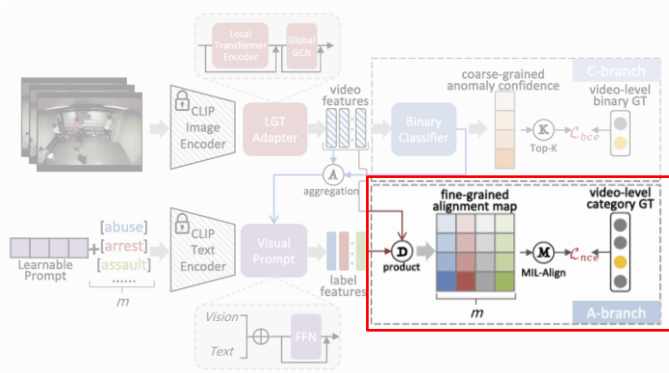
각 Frame 마다 "abuse"와 얼마나 유사한가 계산



[4] Wu, Peng, et al. "Vadclip: Adapting vision-language models for weakly supervised video anomaly detection." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 38. No. 6. 2024.

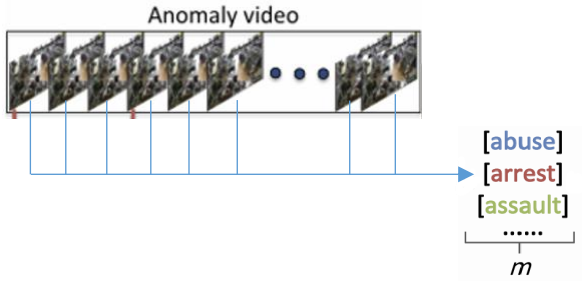
Related Works

VadCLIP: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection (AAAI 2024)

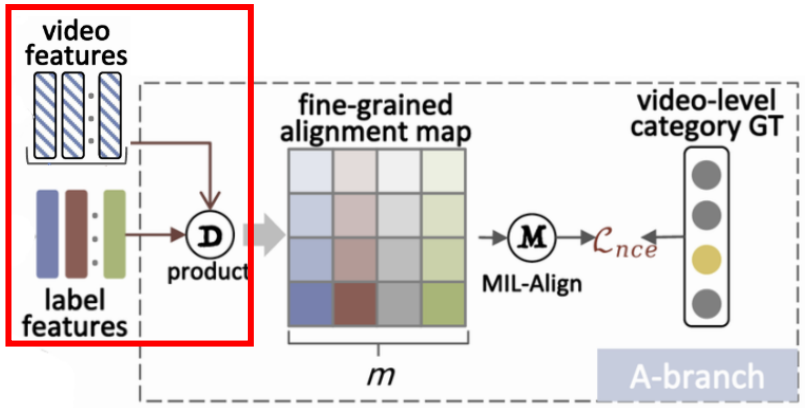


❖ **MIL-Align**

- Challenge 3: 각 segment가 어떤 종류의 이상 상황인지 어떻게 알 수 있을까?
  - ✓ **Alignment Map "M"**



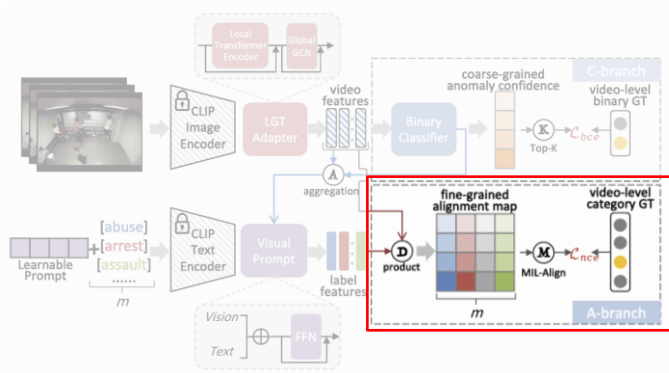
각 Frame 마다 "arrest"와 얼마나 유사한가 계산



[4] Wu, Peng, et al. "Vadclip: Adapting vision-language models for weakly supervised video anomaly detection." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 38. No. 6. 2024.

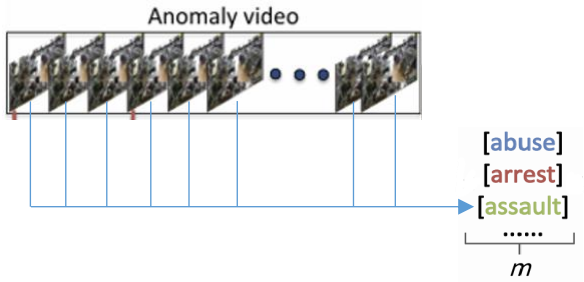
Related Works

VadCLIP: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection (AAAI 2024)

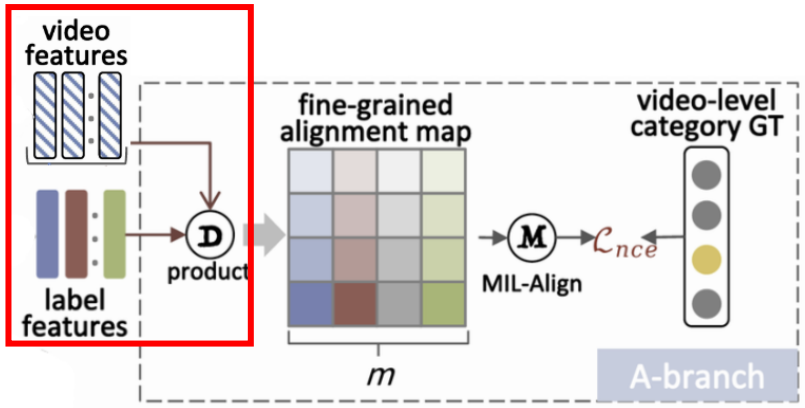


❖ **MIL-Align**

- Challenge 3: 각 segment가 어떤 종류의 이상 상황인지 어떻게 알 수 있을까?
  - ✓ **Alignment Map "M"**



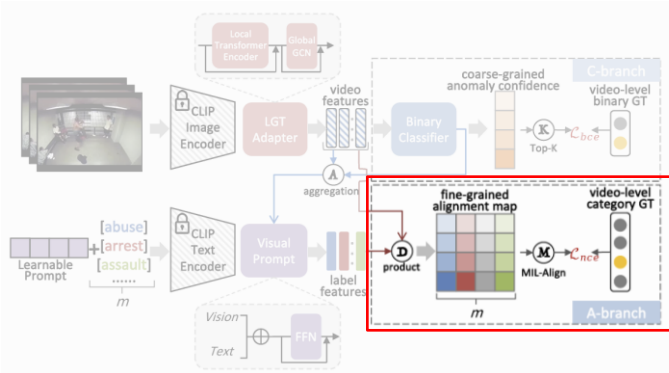
각 Frame 마다 "assault"와 얼마나 유사한가 계산



[4] Wu, Peng, et al. "Vadclip: Adapting vision-language models for weakly supervised video anomaly detection." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 38. No. 6. 2024.

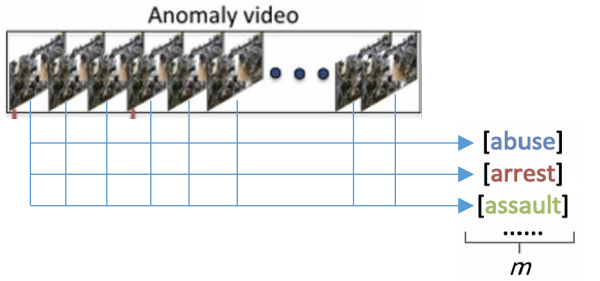
Related Works

VadCLIP: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection (AAAI 2024)

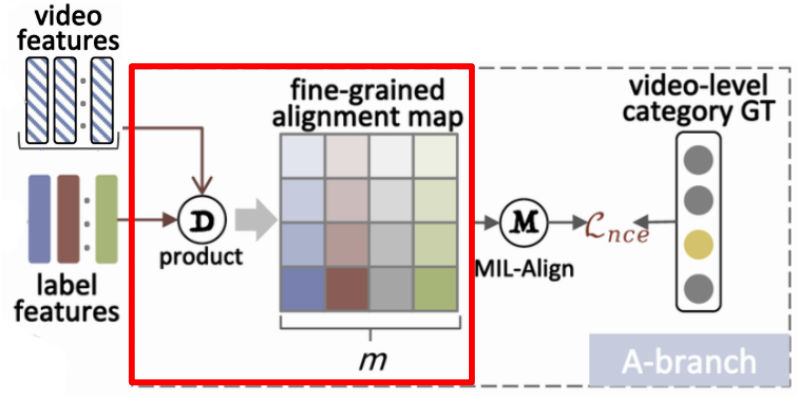


❖ **MIL-Align**

- Challenge 3: 각 segment가 어떤 종류의 이상 상황인지 어떻게 알 수 있을까?
  - ✓ **Alignment Map "M"**



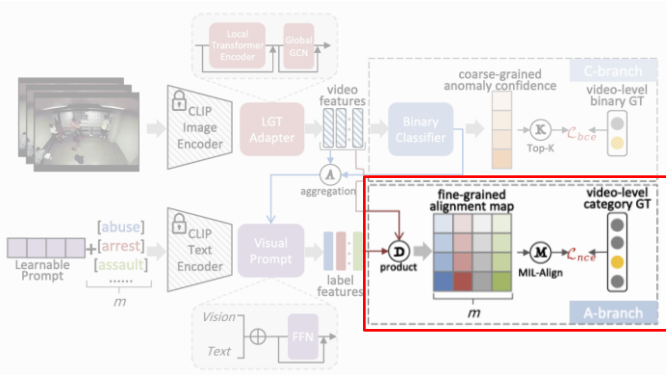
➢ Alignment map M (n x m)
- n: 총 frame 수
- m: Text label 개수



[4] Wu, Peng, et al. "Vadclip: Adapting vision-language models for weakly supervised video anomaly detection." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 38. No. 6. 2024.

Related Works

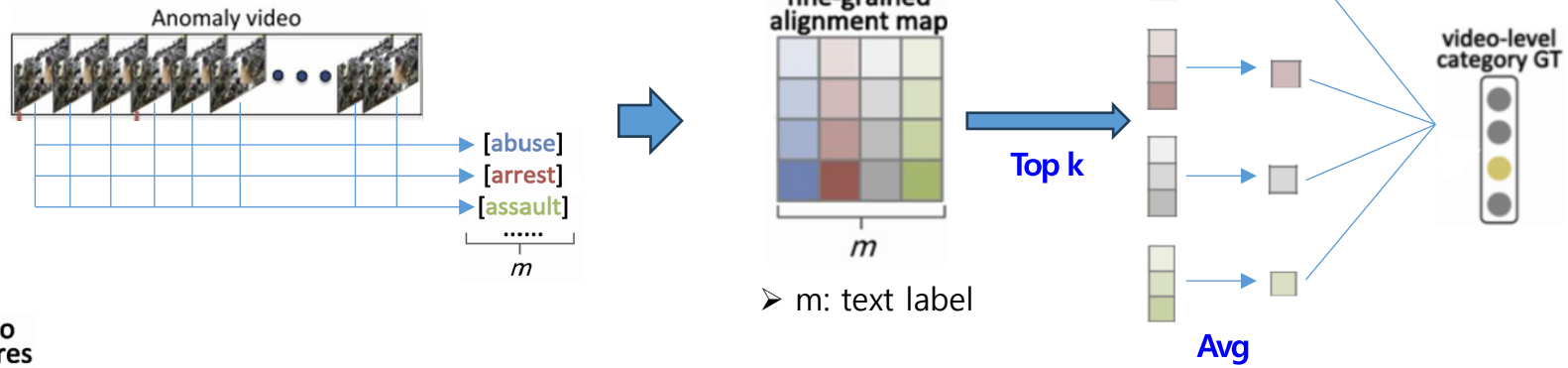
VadCLIP: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection (AAAI 2024)



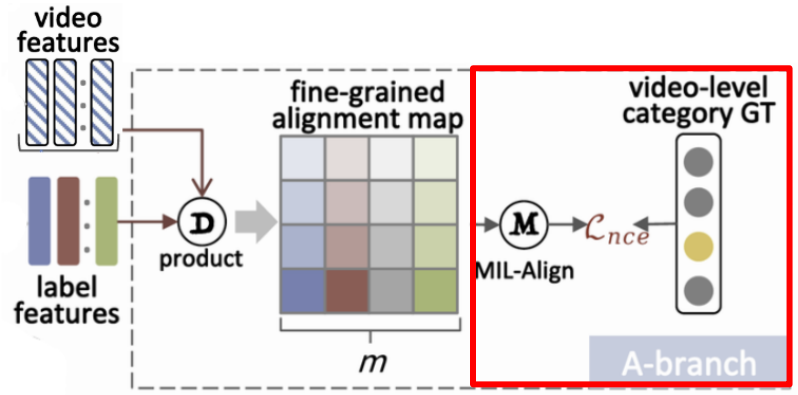
❖ **MIL-Align**

- Challenge 3: 각 segment가 어떤 종류의 이상 상황인지 어떻게 알 수 있을까?

✓ **Alignment Map "M"**



➢ m: text label



➢ MIL-Align
- Top K
- Avg

[4] Wu, Peng, et al. "Vadclip: Adapting vision-language models for weakly supervised video anomaly detection." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 38. No. 6. 2024.

Related Works

VadCLIP: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection (AAAI 2024)

❖ Experiments

- 2가지 데이터셋(UCF-Crime, XD-Violence)에서 성능 평가
- CLIP-TSA에 대비 Coarse-grained detection 성능 향상

Category	Method	AP(%)
Semi	SVM baseline	50.80
	OCSVM (Schölkopf et al. 1999)	28.63
	Hasan et al. (Hasan et al. 2016)	31.25
Weak	Ju et al. (Ju et al. 2022)	76.57
	Sultani et al. (Sultani, Chen, and Shah 2018)	75.18
	Wu et al. (Wu et al. 2020)	80.00
	RTFM (Tian et al. 2021)	78.27
	AVVD (Wu, Liu, and Liu 2022)	78.10
	DMU (Zhou, Yu, and Yang 2023)	82.41
	CLIP-TSA (Joo et al. 2023)	82.17
	VadCLIP (Ours)	**84.51**

Table 1: Coarse-grained comparisons on XD-Violence.

Category	Method	AUC(%)	Ano-AUC(%)
Semi	SVM baseline	50.10	50.00
	OCSVM (1999)	63.20	51.06
	Hasan et al. (2016)	51.20	39.43
Weak	Ju et al. (2022)	84.72	62.60
	Sultani et al. (2018)	84.14	63.29
	Wu et al. (2020)	84.57	62.21
	AVVD (2022)	82.45	60.27
	RTFM (2021)	85.66	63.86
	DMU (2023)	86.75	68.62
	UMIL (2023)	86.75	68.68
	CLIP-TSA (2023)	87.58	N/A
	VadCLIP (Ours)	**88.02**	**70.23**

Table 2: Coarse-grained comparisons on UCF-Crime.

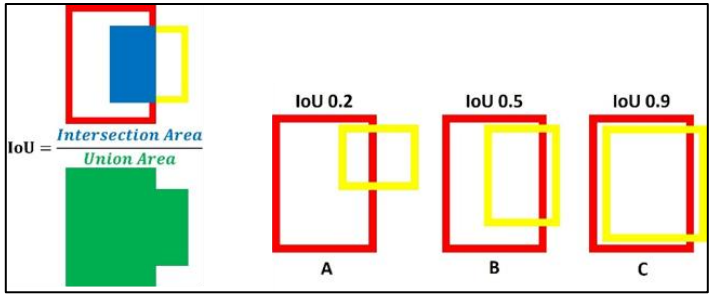
[4] Wu, Peng, et al. "Vadclip: Adapting vision-language models for weakly supervised video anomaly detection." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 38. No. 6. 2024.

Related Works

VadCLIP: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection (AAAI 2024)

❖ Experiments

- VadCLIP은 Fine-grained detection 성능도 함께 평가
- XD-Violence와 UCF-Crime에서 mAP@IOU로 성능 비교



Method	mAP@IOU(%)					
	0.1	0.2	0.3	0.4	0.5	AVG
Random Baseline	1.82	0.92	0.48	0.23	0.09	0.71
Sultani et al. (2018)	22.72	15.57	9.98	6.20	3.78	11.65
AVVD (2022)	30.51	25.75	20.18	14.83	9.79	20.21
VadCLIP (Ours)	**37.03**	**30.84**	**23.38**	**17.90**	**14.31**	**24.70**

Table 3: Fine-grained comparisons on XD-Violence.

Method	mAP@IOU(%)					
	0.1	0.2	0.3	0.4	0.5	AVG
Random Baseline	0.21	0.14	0.04	0.02	0.01	0.08
Sultani et al. (2018)	5.73	4.41	2.69	1.93	1.44	3.24
AVVD (2022)	10.27	7.01	6.25	3.42	**3.29**	6.05
VadCLIP (Ours)	**11.72**	**7.83**	**6.40**	**4.53**	2.93	**6.68**

Table 4: Fine-grained comparisons on UCF-Crime.

[4] Wu, Peng, et al. "Vadclip: Adapting vision-language models for weakly supervised video anomaly detection." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 38. No. 6. 2024.

Related Works

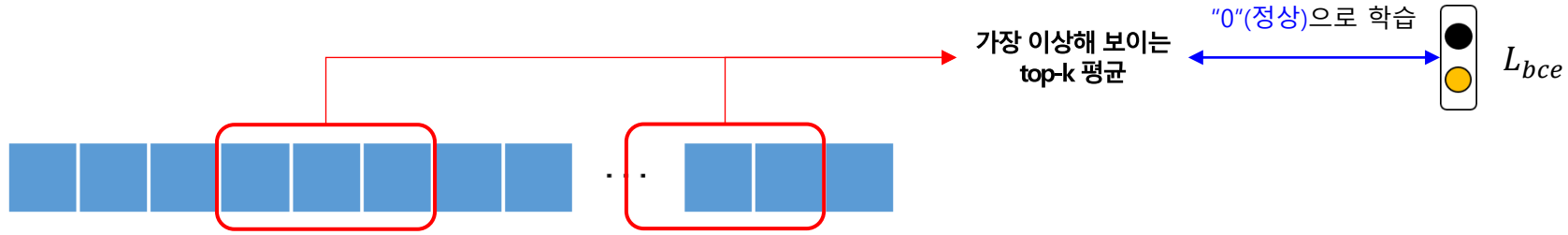
VadCLIP: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection (AAAI 2024)

❖ **해당 논문의 한계점은 무엇일까?**

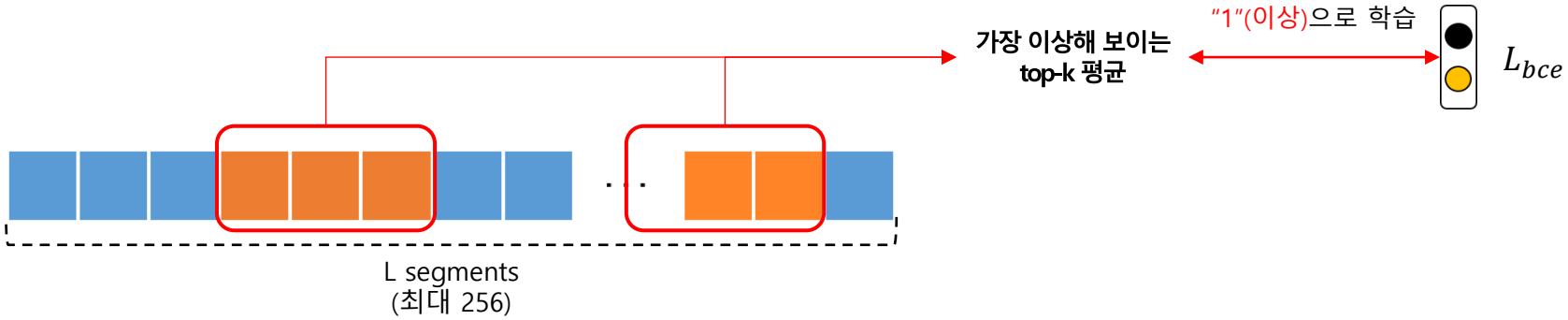
① 기존 MIL 방식은 가장 두드러진 이상 구간에만 집중한다.

→ 비디오 내부의 다양한 정상 패턴을 학습하지 못한다.

✓ **정상 비디오**



✓ **이상 비디오**



[4] Wu, Peng, et al. "Vadclip: Adapting vision-language models for weakly supervised video anomaly detection." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 38. No. 6. 2024.

Related Works

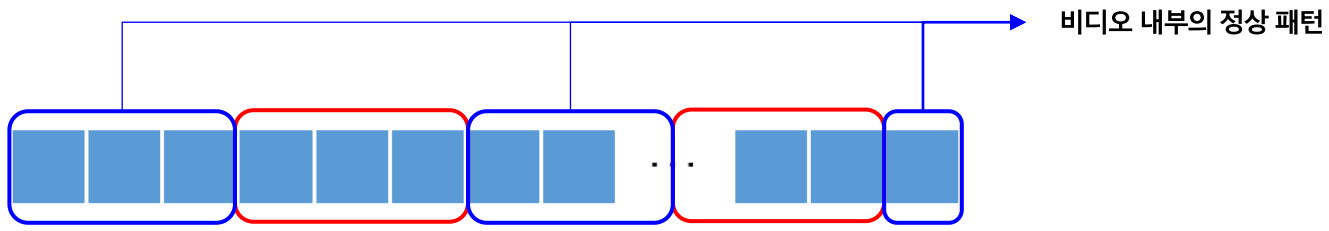
VadCLIP: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection (AAAI 2024)

❖ **해당 논문의 한계점은 무엇일까?**

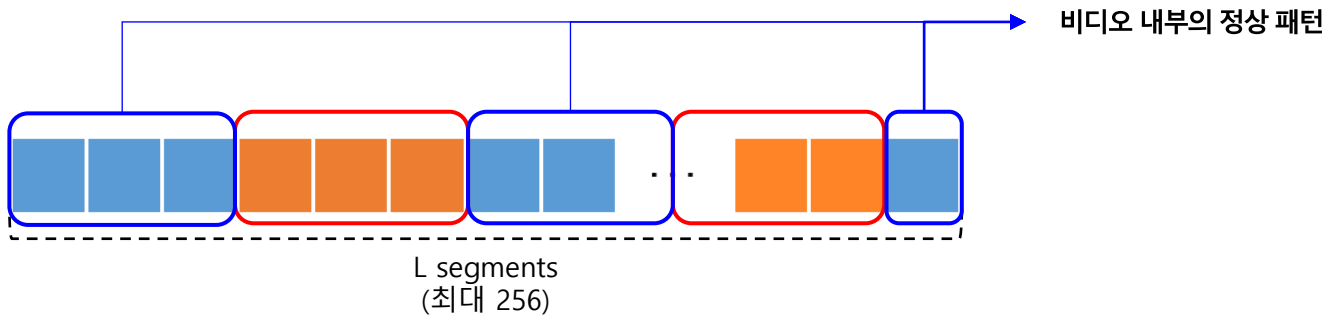
① 기존 MIL 방식은 가장 두드러진 이상 구간에만 집중한다.

→ 비디오 내부의 다양한 정상 패턴을 학습하지 못한다.

✓ **정상 비디오**



✓ **이상 비디오**



[4] Wu, Peng, et al. "Vadclip: Adapting vision-language models for weakly supervised video anomaly detection." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 38. No. 6. 2024.

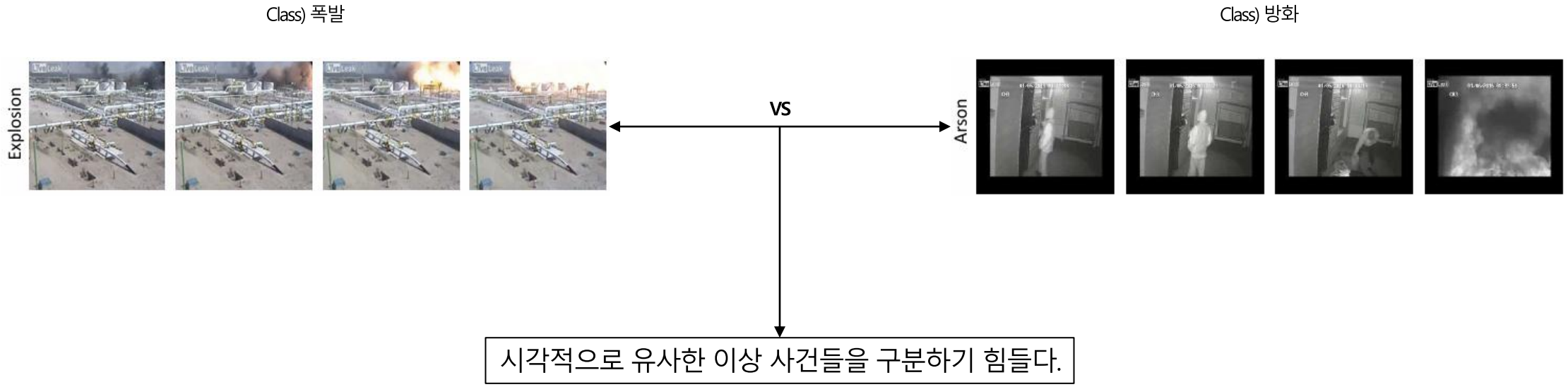
Related Works

VadCLIP: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection (AAAI 2024)

❖ **해당 논문의 한계점은 무엇일까?**

② 시각적으로 비슷한 이상 사건들이 서로 혼동될 수 있다.

→ Anomaly Event와 주변 Background Context가 함께 섞여 학습된다.



[4] Wu, Peng, et al. "Vadclip: Adapting vision-language models for weakly supervised video anomaly detection." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 38. No. 6. 2024.

Related Works

Learning to tell apart: Weakly supervised video anomaly detection via disentangled semantic alignment (AAAI 2026)

❖ **DSANet**

- 기존 MIL 기반 WSVAD의 두가지 한계에 주목
- 정상 패턴에 대한 이해 부족
- 유사 anomaly category 간 혼동
- SG-NM branch를 도입하여 영상 내에서 정상 패턴을 학습
- DCSA를 설계하여 비디오 특징을 이벤트와 배경을 구분

Learning to Tell Apart: Weakly Supervised Video Anomaly Detection via Disentangled Semantic Alignment

Wenti Yin[1], Huaxin Zhang[1], Xiang Wang[1], Yuqing Lu[1], Yicheng Zhang[1], Bingquan Gong[1], Jialong Zuo[1], Li Yu[2], Changxin Gao[1], Nong Sang[1*]

[1]Key Laboratory of Image Processing and Intelligent Control, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology

[2]School of Electronic Information and Communications, Huazhong University of Science and Technology

{yinwt, nsang}@hust.edu.cn

Abstract

Recent advancements in weakly-supervised video anomaly detection have achieved remarkable performance by applying the multiple instance learning paradigm based on multi-modal foundation models such as CLIP to highlight anomalous instances and classify categories. However, their objectives may tend to detect the most salient response segments, while neglecting to mine diverse normal patterns separated from anomalies, and are prone to category confusion due to similar appearance, leading to unsatisfactory fine-grained classification results. Therefore, we propose a novel Disentangled Semantic Alignment Network (DSANet) to explicitly separate abnormal and normal features from coarse-grained and fine-grained aspects, enhancing the distinguishability. Specifically, at the coarse-grained level, we introduce a self-guided normality modeling branch that reconstructs input video features under the guidance of learned normal prototypes, encouraging the model to exploit normality cues inherent in the video, thereby improving the temporal separation of normal patterns and anomalous events. At the fine-grained level, we present a decoupled contrastive semantic alignment mechanism, which first temporally decomposes each video into event-centric and background-centric components using frame-level anomaly scores and then applies visual-language contrastive learning to enhance class-discriminative representations. Comprehensive experiments on two standard benchmarks, namely XD-Violence and UCF-Crime, demonstrate that DSANet outperforms existing state-of-the-art methods.

ode — https://github.com/lessiYin/DSANet

1 Introduction

eakly Supervised Video Anomaly Detection (WS-AD) (Sultani, Chen, and Shah 2018) aims to temporally de-ct anomaly segments in a long untrimmed video with only deo-level labels (*i.e.*, indicating whether a video contains ı anomaly), drastically reducing annotation costs com-ured to its fully supervised counterparts (Wu et al. 2024a; bdalla et al. 2024; Nayak, Pati, and Das 2021), and has ceived considerable attention in recent years (Wang et al.)21, 2022; Shi et al. 2025; Wang et al. 2025a; Zhu et al.

*Corresponding author.

opyright © 2026, Association for the Advancement of Artificial

Figure 1: Schematic diagram about motivation. We identif two main issues: *1) limited understanding of normality*, an *2) category confusion*. We address them through normalit modeling and decoupled contrastive semantic alignment.

2024; Liang et al. 2023). The predominant approach in WS VAD is built upon the multiple instance learning (MIL framework (Tian et al. 2021; Lv et al. 2023; Chen et a 2024). The general pipeline involves first extracting dee features for each video using a pre-trained backbone lik I3D (Carreira and Zisserman 2017) and CLIP (Radford et a 2021), and then feeding the obtained features to a binar classifier to generate instance-level anomaly scores (Yu et a 2025; Wang et al. 2025b,c). For example, CLIP-TSA (Jo et al. 2023) uses CLIP's visual encoder with multi-scale tem poral aggregation and a multiple instance learning branc for detection. VadCLIP (Wu et al. 2024b) employs a bi nary classifier for anomaly detection and text prompts fo anomaly types identification. PEMIL (Pu et al. 2024) de signs anomaly- and context-aware prompts to model com plex event boundaries. ITC (Liu, Lam, and Bao 2024) intro duces learnable textual cues in a dual-branch framework fo robust cross-modal anomaly recognition.

Despite its recent success, the prevailing WS-VAD ap proaches based on multiple instance learning still suffe from two fundamental limitations. At the coarse-graine level, the discriminative nature of MIL results in an incom plete understanding of normality. By focusing exclusivel on identifying the most salient anomalous segments, suc models fail to construct a robust and explicit representatio of the diverse normal patterns present within a video. Thi deficiency compromises the model's ability to distinguis

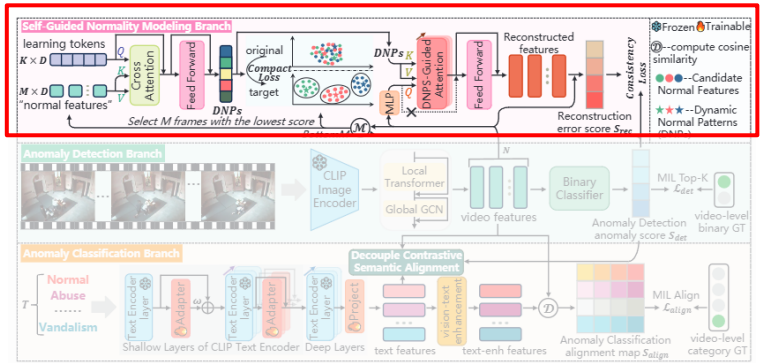
[5] Yin, Wenti, et al. "Learning to tell apart: Weakly supervised video anomaly detection via disentangled semantic alignment." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 40. No. 14. 2026.

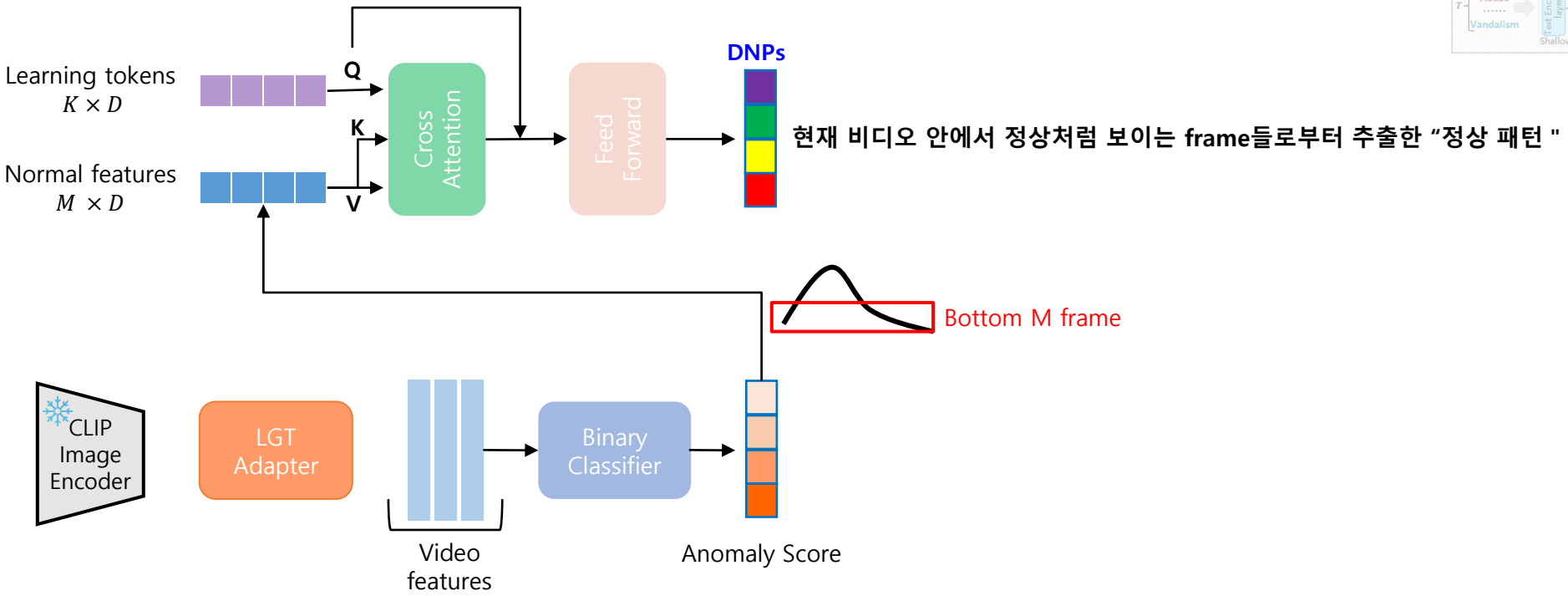
Related Works

Learning to tell apart: Weakly supervised video anomaly detection via disentangled semantic alignment (AAAI 2026)

❖ **"무엇이 정상인가? " 를 모델이 직접 학습할 수 있을까?**

- 기존 MIL 기반 방식은 가장 두드러진 anomaly segment에 집중
- SG-NM: (1) 비디오 내부에서 정상처럼 보이는 anomaly score 낮은 Bottom-M개의 frame 선택





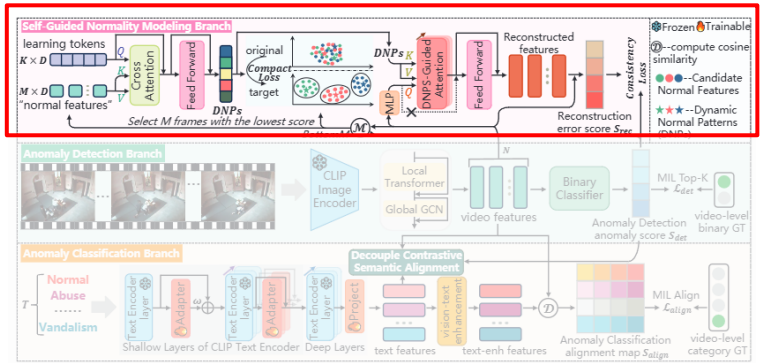
[5] Yin, Wenti, et al. "Learning to tell apart: Weakly supervised video anomaly detection via disentangled semantic alignment." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 40. No. 14. 2026.

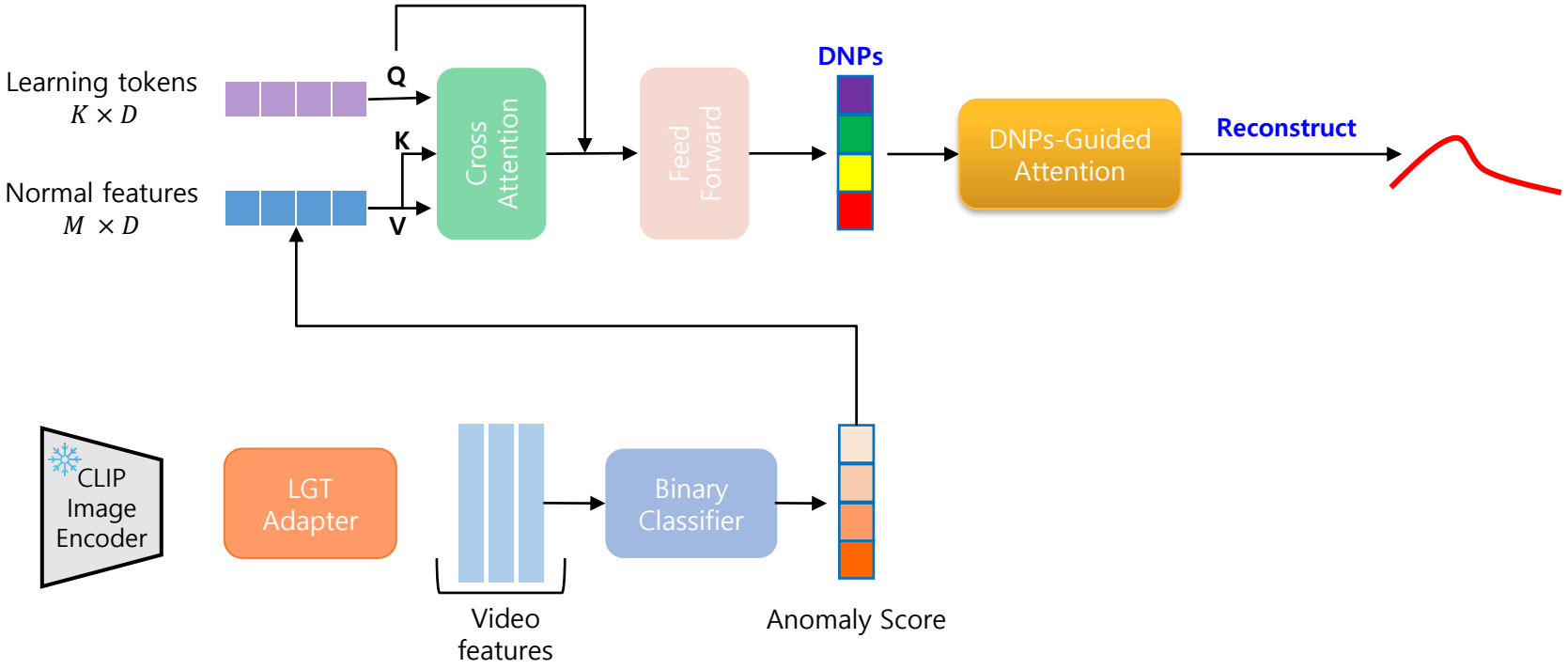
Related Works

Learning to tell apart: Weakly supervised video anomaly detection via disentangled semantic alignment (AAAI 2026)

❖ **"무엇이 정상인가? " 를 모델이 직접 학습할 수 있을까?**

- 기존 MIL 기반 방식은 가장 두드러진 anomaly segment에 집중
- SG-NM: (2) DNPs를 이용해 원래 Video features를 복원





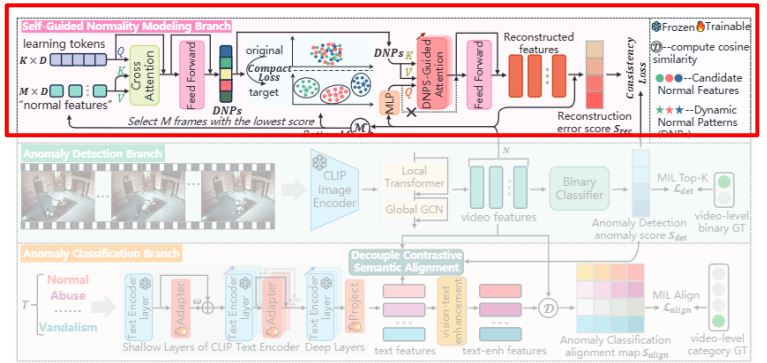
[5] Yin, Wenti, et al. "Learning to tell apart: Weakly supervised video anomaly detection via disentangled semantic alignment." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 40. No. 14. 2026.

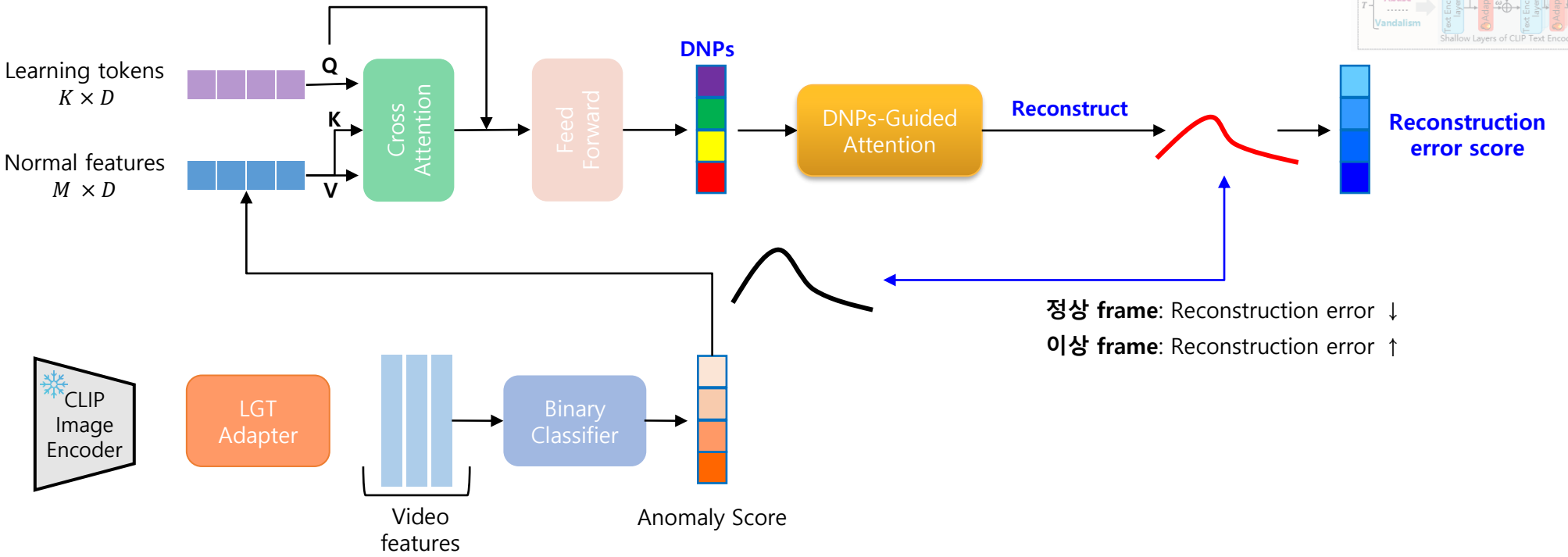
Related Works

Learning to tell apart: Weakly supervised video anomaly detection via disentangled semantic alignment (AAAI 2026)

❖ **"무엇이 정상인가? " 를 모델이 직접 학습할 수 있을까?**

- 기존 MIL 기반 방식은 가장 두드러진 anomaly segment에 집중
- SG-NM: (3) 원본 feature와 Reconstruction feature를 비교





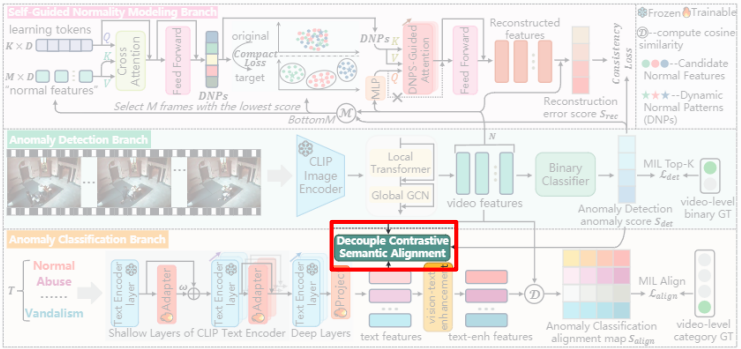
[5] Yin, Wenti, et al. "Learning to tell apart: Weakly supervised video anomaly detection via disentangled semantic alignment." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 40. No. 14. 2026.

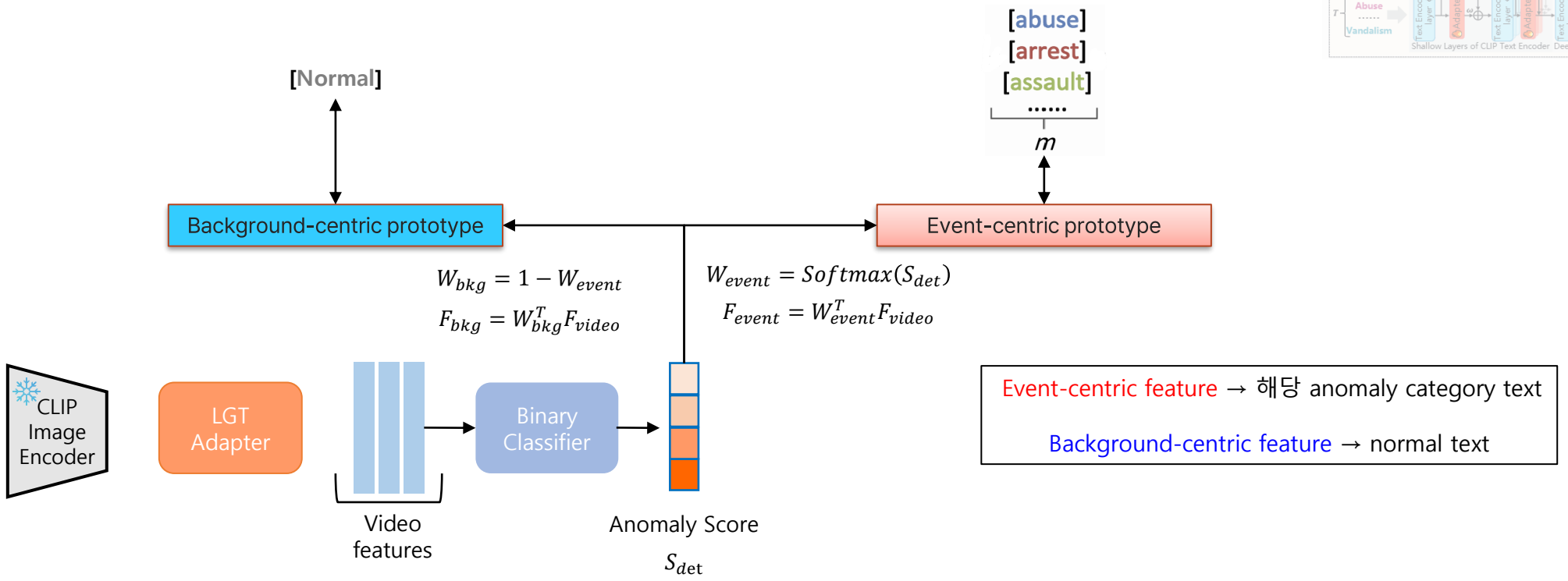
Related Works

Learning to tell apart: Weakly supervised video anomaly detection via disentangled semantic alignment (AAAI 2026)

❖ **비슷한 이상 사건을 더 정확히 구분할 수 있을까?**

- DCSA: video feature를 두가지 prototype으로 분리
  - ➢ Event-centric prototype: 이상 사건을 대표하는 feature
  - ➢ Background-centric prototype: 배경/정상 맥락을 대표하는 feature





Event-centric feature → 해당 anomaly category text

Background-centric feature → normal text

[5] Yin, Wenti, et al. "Learning to tell apart: Weakly supervised video anomaly detection via disentangled semantic alignment." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 40. No. 14. 2026.

Related Works

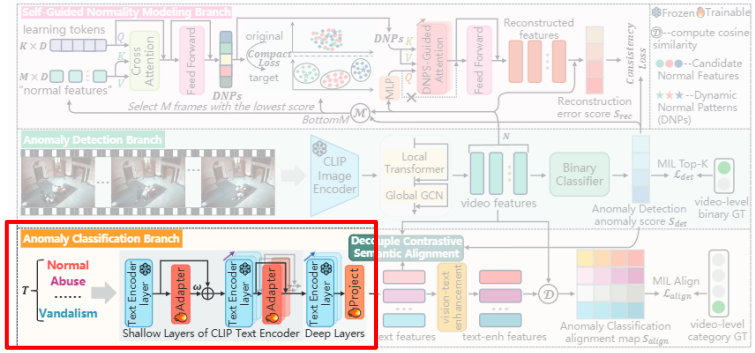
Learning to tell apart: Weakly supervised video anomaly detection via disentangled semantic alignment (AAAI 2026)

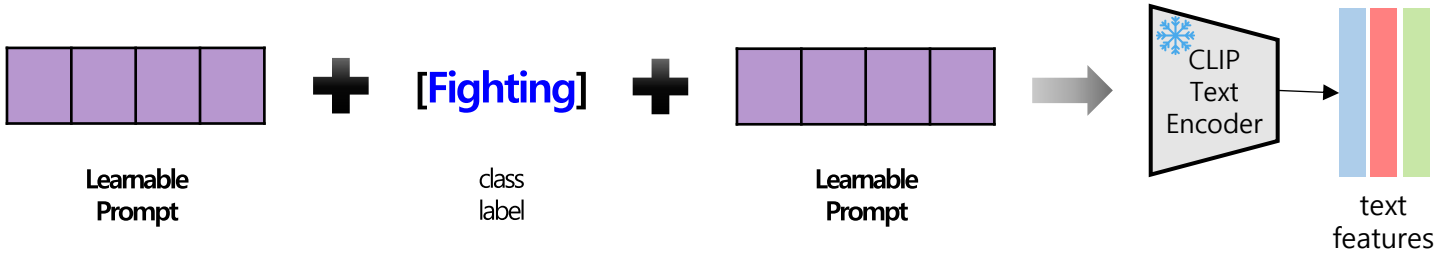
❖ **CLIP의 text feature를 WSVAD에 맞게 보정할 수 있을까?**

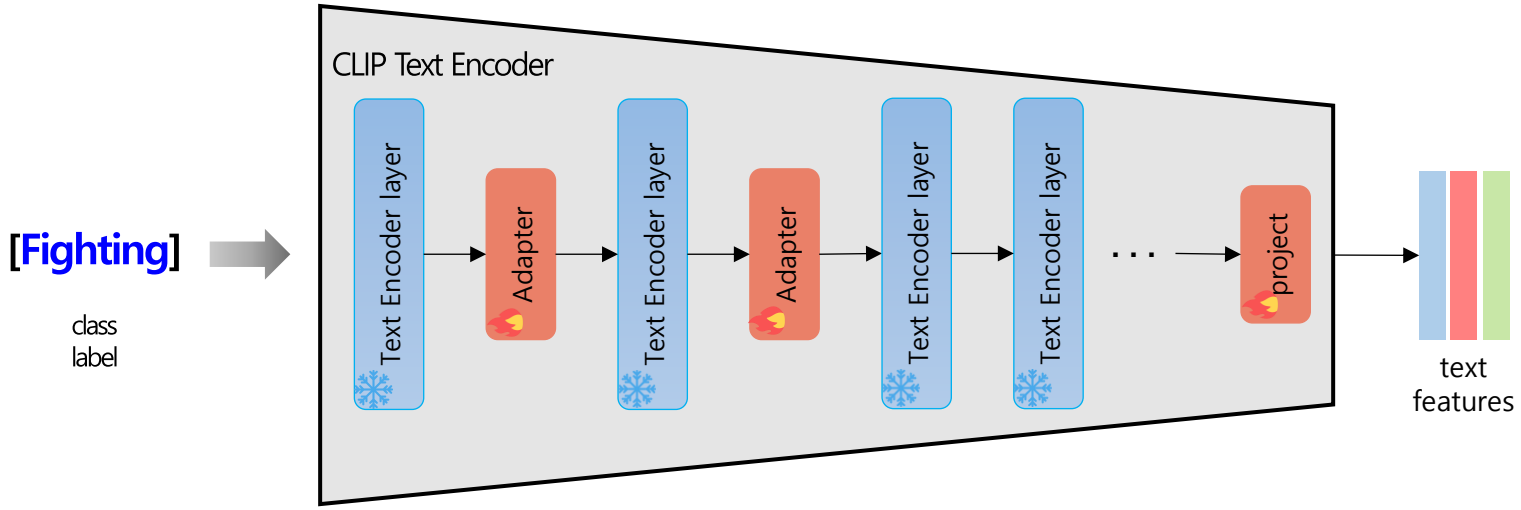
- 기존 VadCLIP은 Learnable Prompt로 짧은 class label의 의미를 보완

→ CLIP Text Encoder 내부 표현 자체를 조정하지는 못한다.

- DSANet은 CLIP Text Encoder 일부 layer에 Lightweight Text Adapter 삽입







[5] Yin, Wenti, et al. "Learning to tell apart: Weakly supervised video anomaly detection via disentangled semantic alignment." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 40. No. 14. 2026.

Related Works

Learning to tell apart: Weakly supervised video anomaly detection via disentangled semantic alignment (AAAI 2026)

❖ Experiments

- 2가지 데이터셋(UCF-Crime, XD-Violence)에서 성능 평가
- VadCLIP에 비해 두 데이터셋에서 모두 Coarse 성능 향상

Category	Method	Features	AP(%)
Un	LTR(Hasan et al. 2016)	-	30.77
Weak	RAD(Sultani, Chen, and Shah 2018)	C3D	73.20
	RTFML(Tian et al. 2021)	I3D	77.81
	ST-MSL(Li, Liu, and Jiao 2022)	I3D	78.28
	LA-Net(Pu and Wu 2022)	I3D	80.72
	DMU(Zhou, Yu, and Yang 2023)	I3D	82.41
	PEL4VAD(Pu et al. 2024)	I3D	85.59
	CLIP-TSA(Joo et al. 2023)	CLIP	82.19
	TPWNG(Yang, Liu, and Wu 2024)	CLIP	83.68
	VadCLIP(Wu et al. 2024b)	CLIP	84.51
	ITC(Liu, Lam, and Bao 2024)	CLIP	85.45
	ReFLIP(Dev, Hazari, and Das 2024)	CLIP	85.81
	DSANet(Ours)	**CLIP**	**86.95**

Table 1: Coarse-grained comparisons on XD-Violence.

Category	Method	Features	AUC(%)
Un	LTR(Hasan et al. 2016)	-	50.60
Weak	RAD(Sultani, Chen, and Shah 2018)	I3D	77.92
	RTFML(Tian et al. 2021)	I3D	84.30
	LA-Net(Pu and Wu 2022)	I3D	85.12
	ST-MSL(Li, Liu, and Jiao 2022)	I3D	85.30
	DMU(Zhou, Yu, and Yang 2023)	I3D	86.75
	PEL4VAD(Pu et al. 2024)	I3D	86.76
	CLIP-TSA(Joo et al. 2023)	CLIP	87.58
	TPWNG(Yang, Liu, and Wu 2024)	CLIP	87.79
	VadCLIP(Wu et al. 2024b)	CLIP	88.02
	ReFLIP(Dev, Hazari, and Das 2024)	CLIP	88.57
	ITC(Liu, Lam, and Bao 2024)	CLIP	89.04
	DSANet(Ours)	**CLIP**	**89.44**

Table 2: Coarse-grained comparisons on UCF-Crime.

[5] Yin, Wenti, et al. "Learning to tell apart: Weakly supervised video anomaly detection via disentangled semantic alignment." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 40. No. 14. 2026.

Related Works

Learning to tell apart: Weakly supervised video anomaly detection via disentangled semantic alignment (AAAI 2026)

❖ Experiments

- 2가지 데이터셋(UCF-Crime, XD-Violence)에서 성능 평가
- VadCLIP에 비해 두 데이터셋에서 모두 Fine 성능 향상

Adapter	SG-NM	DCSA	AP(%)	AVG(%)
Baseline			84.51	24.70
✓			85.00	28.15
✓	✓		85.94	28.39
✓		✓	85.67	28.25
✓	✓	✓	**86.95**	**28.87**

Table 5: Ablation studies on model components. "SG-NM" denotes Self-Guided Normality Modeling, and "DCSA" denotes Decoupled Contrastive Semantic Alignment.

Method	mAP@IOU(%)					
	0.1	0.2	0.3	0.4	0.5	AVG
RAD(2018)	22.72	15.57	9.98	6.20	3.78	11.65
AVVD(2022)	30.51	25.75	20.18	14.83	9.79	20.21
VadCLIP(2024)	37.03	30.84	23.38	17.90	14.31	24.70
ITC(2024)	40.83	32.80	25.42	19.65	15.47	26.83
ReFLIP(2024)	39.24	33.45	27.71	20.86	17.22	27.36
DSANet(Ours)	**40.93**	**34.63**	**28.21**	**22.70**	**17.89**	**28.87**

Table 3: Fine-grained comparisons on XD-Violence.

Method	mAP@IOU(%)					
	0.1	0.2	0.3	0.4	0.5	AVG
RAD(2018)	5.73	4.41	2.69	1.93	1.44	3.24
AVVD(2022)	10.27	7.01	6.25	3.42	3.29	6.05
VadCLIP(2024)	11.72	7.83	6.40	4.53	2.93	6.68
ITC(2024)	13.54	9.24	7.45	5.46	3.79	7.90
ReFLIP(2024)	14.23	10.34	9.32	7.54	6.81	9.62
DSANet(Ours)	**21.39**	**14.96**	**11.74**	**8.98**	**8.00**	**13.01**

Table 4: Fine-grained comparisons on UCF-Crime.

[5] Yin, Wenti, et al. "Learning to tell apart: Weakly supervised video anomaly detection via disentangled semantic alignment." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 40. No. 14. 2026.

Conclusion

- **WSVAD**는 비디오 전체가 정상인지 이상인지만 알고 있을 때, 실제 이상이 발생한 시간 구간을 찾는 task

- **MIL Ranking**
  → 이상 비디오와 정상 비디오를 비교하여 segment 별 anomaly score를 학습
- **CLIP-Based WSVAD**
  → CLIP feature를 사용해 이상 상황의 맥락을 더 잘 표현
- **Vision-Language Alignment**
  → visual feature와 text label을 연결해 이상 사건의 종류까지 예측
- **Normality & Semantic Disentanglement**
  → "무엇이 정상인지"를 학습 & 이상 사건과 배경을 분리하여 더 정확히 구분

Reference

[1] Zanella, Luca, et al. "Harnessing large language models for training-free video anomaly detection." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

[2] Sultani, Waqas, Chen Chen, and Mubarak Shah. "Real-world anomaly detection in surveillance videos." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[3] Joo, Hyekang Kevin, et al. "Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection." *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023.

[4] Wu, Peng, et al. "Vadclip: Adapting vision-language models for weakly supervised video anomaly detection." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 38. No. 6. 2024.

[5] Yin, Wenti, et al. "Learning to tell apart: Weakly supervised video anomaly detection via disentangled semantic alignment." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 40. No. 14. 2026.

고맙습니다