OE-CTST: Outlier-Embedded Cross Temporal Scale Transformer for Weakly-supervised Video Anomaly Detection
Supplementary Material

Snehashis Majhi[1,2], Rui Dai[1,2], Quan Kong[3], Lorenzo Garattoni[4], Gianpiero Francesca[4],
François Brémond[1,2]
[1] INRIA [2] Côte d'Azur University [3] Woven by Toyota [4] Toyota Motor Europe

Table 1. Overview of Supplementary Material

Section	Content
1	Dataset Description
2	Implementation Details
3	Network Complexity Analysis
4	Added Ablation study
5	Necessity of K-Fold Evaluation
6	Category Wise Performance Analysis

1. Dataset Description

In this section, we provide a detailed description of the three datasets considered to evaluate our method. Further, we provide a statistics of anomaly instances in terms of duration and frequency in Figure 1.

UCF-Crime (UCF-C) [8] : It is a diverse and large-scale dataset containing 1900 real-world surveillance videos from 13 types of anomaly activities. In this dataset anomaly activities may occur for a long or short duration, which makes the detection problem more challenging. It has 1610 videos for training, out of which 810 and 800 videos belong to anomaly and normal classes respectively. Similarly, for testing there are 290 videos containing 140 anomalies and 150 normal videos.

XD-Violence (XD-V) [10] : It is a diverse and large-scale audio-visual dataset collected from movies, games, CCTV cameras to cover 6 types of anomalies. It contains 4754 untrimmed videos with a total of 217 hours, out of which 2349 videos are normal and 2405 videos are anomalies. The official training set contains 3954 videos and the test set contains 800 videos. To support the setting of weak supervision, only the test set has temporal labels for performance evaluation.

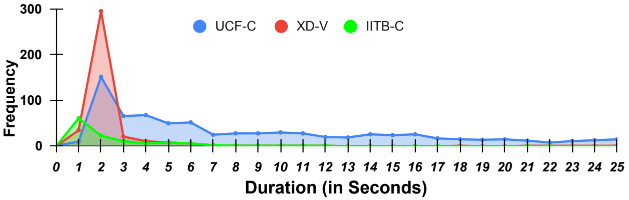

Figure 1. Visualization of length or duration of anomaly instances and their corresponding frequencies in UCF-C, XD-V, and IITB-C. To obtain these plots, we parse the whole datasets of UCF-C and IITB-C, but for XD-V dataset we consider only the test set. Based on this plot, we set the *threshold (th)* to 2 second for defining short and long anomalies in all three datasets.

IITB-Corridor (IITB-C) [7] : It is also a medium-scale dataset recorded in a corridor of IIT Bombay campus with a single camera setup. It contains a total of 358 videos in standard protocol [7] where 208 videos in the training set are normal videos only and the test set contains 10 normal videos and 140 anomaly videos. This standard-setting considering only normal videos in the train set is not suitable for WSVAD. For this, [6] reorganizes the dataset to meet the requirement of WSVAD. The new training split contains both normal and anomaly classes and hence it is prepared by randomly moving 71 anomaly videos from the standard test set to the new anomaly class of the train set followed by 147 normal videos from the standard train set to the new normal class of the train set. Similarly, the new test split is a collection of 69 remaining anomaly videos of the standard test set and 71 normal videos from the standard train, test set. In summary, the new training and testing split contain 218 and 140 videos respectively.

2. Implementation Details

We consider two popular general purpose backbone (*i.e.* I3D-ResNet50 and Video-Swin) for spatio-temporal feature extraction. For each 16-frame snippet, 2048D features from *'mixed_5c'* layer of I3D-ResNet50 and 1024D features from *'stage-4'* layer of Video-Swin are extracted where each backbone is pre-trained on Kinetics [4] dataset. Following previous works [8, 9], we divide each video into 32 non-overlapping temporal segments. In outlier embedder (ED), we use three types of one-class learner (OC-L)

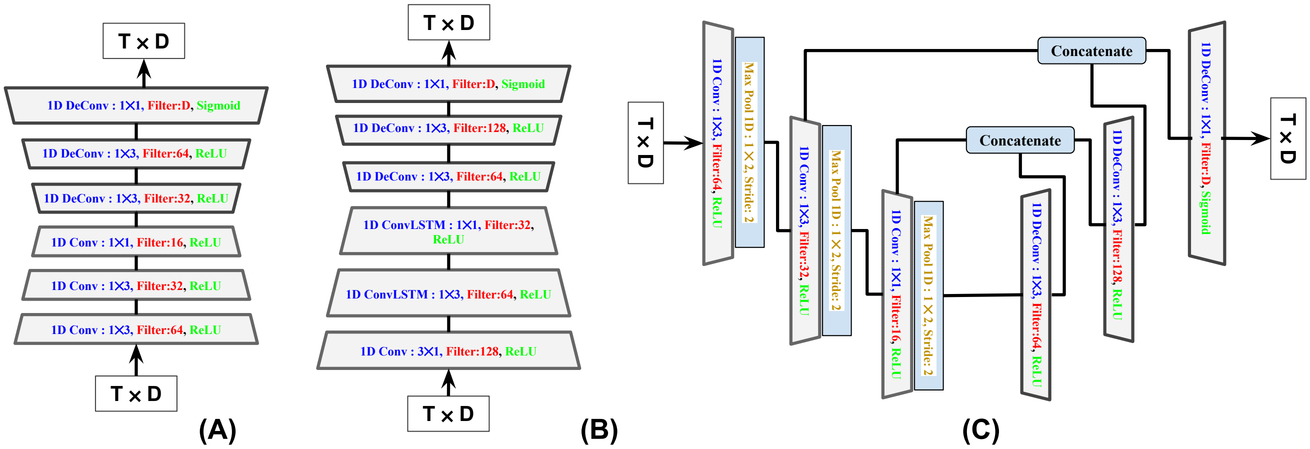
T × D
1D DeConv : 1×1, Filter:D, Sigmoid
1D DeConv : 1×3, Filter:64, ReLU
1D DeConv : 1×3, Filter:32, ReLU
1D Conv : 1×1, Filter:16, ReLU
1D Conv : 1×3, Filter:32, ReLU
1D Conv : 1×3, Filter:64, ReLU
T × D
(A)
T × D
1D DeConv : 1×1, Filter:D, Sigmoid
1D DeConv : 1×3, Filter:128, ReLU
1D DeConv : 1×3, Filter:64, ReLU
1D ConvLSTM : 1×1, Filter:32, ReLU
1D ConvLSTM : 1×3, Filter:64, ReLU
1D Conv : 3×1, Filter:128, ReLU
T × D
(B)
T × D
1D Conv : 1×3, Filter:64, ReLU
Max Pool 1D : 1×2, Stride: 2
1D Conv : 1×3, Filter:32, ReLU
Max Pool 1D : 1×2, Stride: 2
1D Conv : 1×1, Filter:16, ReLU
Max Pool 1D : 1×2, Stride: 2
1D DeConv : 1×3, Filter:64, ReLU
Concatenate
1D DeConv : 1×3, Filter:128, ReLU
Concatenate
1D DeConv : 1×1, Filter:D, Sigmoid
T × D
(C)

Figure 2. Architectural design of three One-class learner modules: (a) Temporal Auto encoder, (b) Spatio-temporal Auto-encoder, (C) UNet. Here, blue text denotes the 1D conv or 1D Deconv operation with respective kernel size, red text denotes the number of filters and green text denotes the activation function used.

and the detailed architectural design is shown in Figure 2. In CTST, the down scaler performs the downscaling by applying a `max-pool` operator with *stride* $= 2^{\tau}$ and the up scaler duplicates the temporal token to upsample the temporal resolutions. In CTFA we use 128 filters in each temporal convolution (TC) of K, Q, V, and the number of heads of multi-head attention is set to 4, 4, 2 for UCF-Crime, XD-Violence, and IITB-Corridor dataset respectively. The number of units in the linear layer of CTFA block is set to 256 for the UCF-Crime and IITB-Corridor datasets and 512 for the XD-Violence dataset. In the detector, the number of neurons for three FC layers is set to 128,32 and 1 respectively. Each FC layer in the detector is followed by a *ReLU* activation and a *dropout* function with dropout rate = 0.6 for all three datasets. The OE-CTST is end-to-end trainable excluding the visual backbone. We train OE-CTST using Adam optimizer with a learning rate of 0.0001 for UCF-Crime and IITB-corridor datasets and 0.001 for the XD-Violence dataset. The loss weighting factors are set to $\lambda_1 = \lambda_2 = 0.5$ and $\beta_1 = 0.3$,$\beta_2 = 0.7$ for all three datasets. We also randomly select 30 anomaly and 30 normal videos as a mini-batch and compute the gradient using reverse mode automatic differentiation on computation graph using Tensorflow. Then the loss is computed and back-propagated for the whole batch. In UCF-Crime, XD-Violence and IITB-Corridor datasets we train up to 1050, 1700, and 450 epochs respectively four parallel 2080Ti GPUs. For all three datasets, we use ten crops for training and one crop for testing to reduce the inference time.

3. Network Complexity Analysis

This section performs a complexity analysis of our OE-CTST to meet real-world applicability. In contrast to traditional transformers, which are usually expensive in computation, our method can work on real-time scenarios on a single 2080Ti GPU. As shown in Table 2, our method is computationally competitive in terms of FLOPs and speed w.r.t. [9] while boosting the detection performance significantly. Unlike [6], our method does not rely on people detection and tracking thus, ours is computationally much more efficient and applicable for real-time applications.

Methods	FLOPs(G)	Speed(FPS)
Tian *et al.* [9]*(I3D-Res)	153.2	211
Majhi *et al.* [6](I3D-Res)	435.6	29.7
Chen *et al.* [1]* (Video-Swin)	234.5	194
OE-CTST (I3D-Res)	153.6	206
OE-CTST (Video-Swin)	292.8	130

Table 2. Complexity comparison of OE-CTST with competitive methods. Here, G: Giga, FPS: Frames-per-second and * means the methods use ten crops for testing which can even increase the FLOPs and lower the speed mentioned in the table by a factor of 10.

4. Added Ablation Study

In this section, we provide an additional ablation study of our method to show the necessity and effectiveness of using temporal regularity features in normality learning by one-class learner (OC-L). Since the overall performance of our method has a direct dependency on the normalcy learning of the OC-L, we study the impact of various input features to OC-L as shown in Table 3. It is visible that the usage of temporal regularity features has stronger superiority over the classical appearance features in terms of overall performance gain in the official test set of the three data sets. In UCF-C and XD-V datasets, there exists a significant performance gap between appearance and temporal-regularity features, *i.e.* *on average* 4.06% and 5.02% respectively. But for the IITB-C dataset, the performance gap is reduced to 0.4% (*on average*). This phenomenon arises in UCF-C and XV-D datasets because there are large intra-

OC-L	Appearance Feature			Temporal-Regularity Feature		
	UCF-C	XD-V	IITB-C	UCF-C	XD-V	IITB-C
T-AE [3]	82.54	76.21	88.62	**86.99**	**81.78**	**89.04**
ST-AE [2]	83.01	76.53	88.89	**86.99**	**81.54**	**89.26**
UNet [5]	83.18	76.81	88.79	**86.94**	**81.31**	**89.18**

Table 3. Ablation study to show the impact on overall anomaly detection performance for different input features used in OC-L.

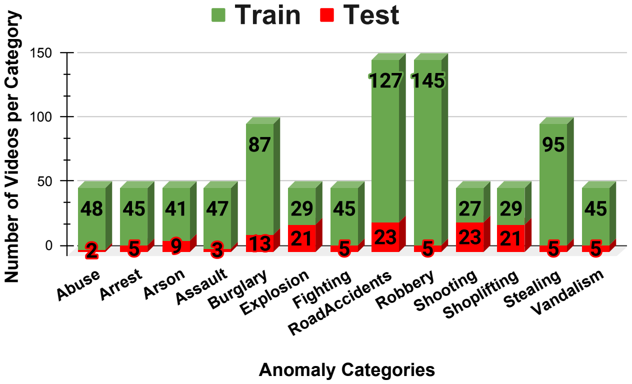

Figure 3. Visualization of the number of videos used for training and testing for each abnormal category in the official split of the UCF-Crime dataset.

class appearance variations in the normal distribution since the videos are collected from multiple sources (like CCTV, YouTube, video games, and movies) with different lighting conditions. Thus, the normal representation learned from appearance features is less compact than that of temporal regularity features. However, IITB-C is recorded in a fixed scenario (in a corridor) without many variations of appearance cues in the normal distribution and hence the performance gap between appearance and temporal regularity feature is relatively less important. For a complex and real-world event distribution, we found that the temporal regularity features are more robust and salient, and enable OC-L to generate more effective temporal position embeddings.

5. Necessity of K-Fold Evaluation

The anomaly detection performance obtained from official test-split [8] of UCF-Crime has a strong bias towards the easy and short anomalies (contains sharp changes in the appearance and motion cues) like *explosion, road accidents, shooting* as they have significantly more samples for testing. Due to this, some methods [8–10] which are capable of capturing sharp changes tend to perform well on the official test split. From Figure 3, it can be seen that anomalies like *abuse, arrest, arson, assault, burglary, fighting, robbery, stealing vandalism* have fewer test samples compared to their training counterparts and these anomalies are characterized by both subtle and progressive spatio-temporal cues which are difficult to detect in a real-world complex scenario. For this, we adopt the K-fold test evaluation (*where, K=5*) in our work which covers the entire dataset for an unbiased evaluation of both simple/short and complex/long anomalies.

6. Category Wise Performance Analysis

In this section, we provide an anomaly category wise performance analysis of our method on UCF-C [8] dataset. Since the official test-set of UCF-C has a non-uniform number of anomaly samples, we focus on the K-folds (K= 5) evaluation, and the category-wise performance is shown in Figure 4. It can be observed from Figure 4 that the average AUC of *explosion, burglary, arson, assault, anomalies* is relatively higher than the others. This is due to the presence of spatio-temporally more salient abnormal patterns in these categories which is easy to detect. But for anomalies like *abuse, arrest, stealing, shoplifting, robbery* the average AUC is relatively lower due to the existence of the subtle and less discriminative cues w.r.t normal counterpart which makes it difficult to detect. However, our method has obtained better performance than Tian *et al.* [9] and Chen *et al.* [1] in most of these abnormal categories as shown in Figure 5. Further, our method has gained a significant performance boost in long anomaly categories like *stealing, arrest, robbery, shooting, shoplifting, burglary, fighting, assault* compared to Tian *et al.* [9] and Chen *et al.* [1]. But there exists a few exceptions like road-accidents and vandalism where our method is less better than Tian *et al.* [9] and Chen *et al.* [1]. To further investigate, we visualize the videos and corresponding predictions in road-accidents and vandalism categories. We observe that our method produces more false positives after the occurrence of an anomaly where the scene continues to be panic and this situation is quite similar to an abnormal situation.

References

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

[2] Yong Shean Chong and Yong Haur Tay. Abnormal event detection in videos using spatiotemporal autoencoder. In *International symposium on neural networks*, pages 189–196. Springer, 2017.

[3] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. Learning temporal regularity in video sequences. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[4] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

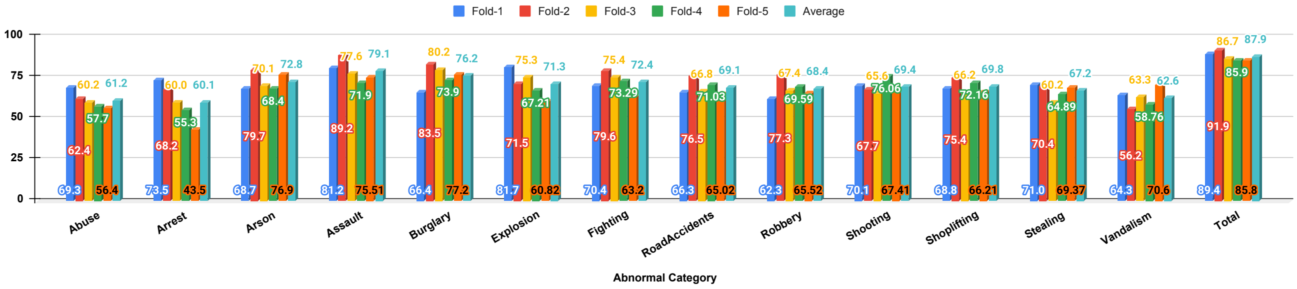

Figure 4. Category-wise detection performance of UCF-Crime dataset in 5-Folds evaluation.

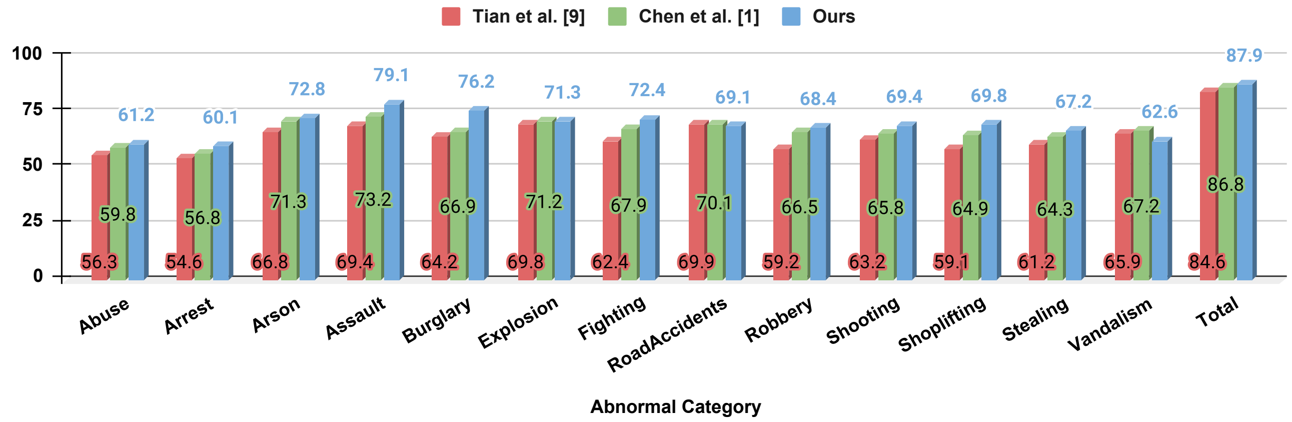

Figure 5. Category wise performance comparison of our method with Tian *et al.* [9] and Chen *et al.* [1] in 5-Folds evaluation.

[5] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6536–6545, 2018.

[6] Snehashis Majhi, Rui Dai, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. Human-scene network: A novel baseline with self-rectifying loss for weakly supervised video anomaly detection. *arXiv preprint arXiv:2301.07923*, 2023.

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

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

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

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