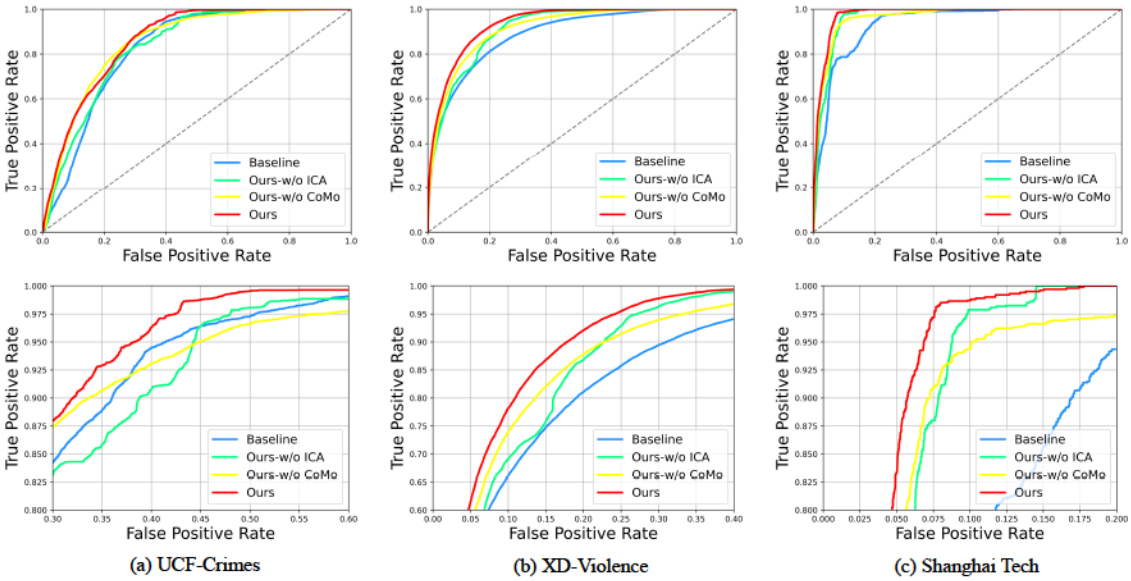
[Supplementary Material] Look Around for Anomalies: Weakly-supervised Anomaly Detection via Context-Motion Relational Learning

MyeongAh Cho[1] Minjung Kim[1] Sangwon Hwang[2] Chaewon Park[1] Kyungjae Lee[3] Sangyoun Lee[1]
[1]*Yonsei University* [2] *Hyundai Motor Company* [3]*Yong In University*
t{maycho0305,mjkima,chaewon28,syleee}@yonsei.ac.kr tsangwonH@hyundai.com tkjlee@yongin.ac.kr



(a) UCF-Crimes (b) XD-Violence (c) Shanghai Tech

Figure I. ROC curves of baseline, proposed framework without ICA or CoMo modules, and proposed framework score in the three benchmarks. The first and second rows of the figure show curves in the range [0, 1] and curves zoomed in at the top left, respectively.

1. ROC Curves on Three Benchmarks

Fig. I shows the Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) curve of the proposed method for the large-scale WVAD databases, i.e., UCF-Crimes [8] and XD-Violence [10], and the unsupervised VAD ShanghaiTech [5] database that is split and reconstructed for WVAD. The blue line is the curve of the baseline, which is the backbone network with FC layers trained by MIL loss; the green and yellow lines indicate the proposed network without the ICA module or CoMo module, respectively; the red line is the curve of the proposed method. On UCF-Crimes, the AUC score of each network is 82.43%, 83.26%, 85.22%, and 86.07%, respectively. On XD-Violence, the AP scores are 73.1%, 75.4%, 76.99%, and 81.31%, respectively, and on ShanghaiTech, the AUC scores are 93.46%, 96.19%, 96.33%, and 97.3%, respectively.

The performance is considerably improved when the proposed modules are added to the baseline, and there is a noticeable performance gap between the red and blue lines. When comparing the green and yellow lines, better performance is shown with ICA than CoMo in all three benchmarks. This demonstrates that it is difficult for a feature obtained through a single backbone branch to have sufficient representative information about normal and abnormal classes, but ICA helps features to have discrepancies. Furthermore, it shows that relational modeling of CoMo with the ICA feature (red line) is much more effective than utilizing the backbone feature (green line).

2. Detailed Implementations

The input frame size of the video data is 256×256, and 10-crop augmentation is preprocessed as described in the following previous papers [3, 6, 9]. For UCF-Crimes, we extract $D = 2048$ dimensional features using ResNet50-I3D, the same backbone as RTFM, and use $D = 1024$ dimensional features of Inception-v1 I3D for other databases. We utilize only RGB features without optical flow features.

Table I. Instantiation of framework. Output size is in the order of (batch×snippet×dimension) size.

Module			Layers	Output Size
Backbone			I3D-'*mix_5c*'	$B \times T \times D$
CLAV	ICA		$Conv1d(3,1,2D)$ $MaxActivation$	$B \times T \times 2D$ $B \times T \times D$
	FC1		$FC(512)$	$B \times T \times 512$
	FC2		$FC(128)$	$B \times T \times 128$
	FC3		$FC(1)$	$B \times T \times 1$
	CSN/ CSA		$FC(D)$	$B \times K \times D$
CoMo	Dynamic path		$Conv1d(1,1,512)$ $Conv1d(1,1,1)$	$B \times T \times 512$ $B \times T \times 1$
	Context path		$Conv1d(3,1,D)$ $Conv1d(1,1,512)$ $Conv1d(1,1,C)$	$B \times N \times D$ $B \times N \times 512$ $B \times N \times C$
	GCN	projection	$Conv1d(1,1,32)$	$B \times T \times 32$
		channel reduction	$Conv1d(1,1,128)$	$B \times T \times 128$
		node propagte	$Conv1d(1,1,32)$	$B \times 32 \times 128$
		state update	$Conv1d(1,1,128)$	$B \times 32 \times 128$
		interrelation	$Conv2d(1,1,1)$	$B \times 32 \times 128$
	FC		$FC(1)$	$B \times T \times 1$

Video frames are stacked in groups of 16 to become a single snippet, and among all snippets, T number of snippets are uniformly selected and become input data.

The layer information for each module of the entire framework is presented in Table I. In $ConvNd(k,s,c)$ and $FC(c)$, k, s, and c indicate the kernel, stride, and output channel size, respectively. ReLU activation function and batch normalization are followed between each layer, and dropout with $p = 0.7$ is applied between FC layers. Input snippets pass through the backbone to become $(B \times T \times D)$ size of feature B, and through ICA, the channel size doubles and splits in half, followed by a max operation that activates differently depending on the class, resulting in feature $\boldsymbol{F}$. Then, $\boldsymbol{F}$ passes through FC1 and FC2 to become $\boldsymbol{F}_{FC2}$ of $(B \times T \times 128)$ shape, which is input into FC3 and CS module (CSN or CSA). The anomaly score $\boldsymbol{S}$ is calculated through FC3, which learns using loss function L_{mil}. In addition, the CS module reconstructs $\boldsymbol{F}'^{n}_{topk}$ and $\boldsymbol{F}'^{a}_{topk}$ to predict $\boldsymbol{F}^{n}_{topk} = \{\boldsymbol{f}^{n}{}_{i}\}_{i=topk}$ and $\boldsymbol{F}^{a}_{topk} = \{\boldsymbol{f}^{a}{}_{i}\}_{i=topk}$ of $(B \times K \times D)$ shape and learns patterns specific to normal and abnormal classes through auxiliary loss L_{cs}. Then, in CoMo, dynamic path outputs motion intensity scores for T snippets through Conv1d layers with a temporal kernel size of 1 by temporal independently. Through this value, the bottom-N indices of static features $\{\boldsymbol{F}_i\}_{i=bottomN}$ with low motion intensity are selected to become inputs of the context path, and aggregation between static features is performed through a layer with kernel size 3 and become $(B \times N \times D)$ size of features which N-mean becomes $(B \times 1 \times D)$ shape of context feature $\boldsymbol{F}_{cont}$. In order to focus on the appearance information of the static scene, context path predicts the object class score $\boldsymbol{S}^{obj}$ with $(B \times N \times C)$ size within the N snippets trained with L_{obj}. In GCN, first, the size of class-activated feature $\boldsymbol{F}$ and context feature F_{cont} becomes $(B \times T \times 128)$ and $(B \times 1 \times 128)$, respectively, through channel reduction, and the features are projected from the temporal to interaction space by projection matrix P to become a $(B \times n \times st)$-size node and state matrix where number of node and state is 32 and 128, respectively. In interaction space, each relation is explored through node propagation followed by state updation and become $\boldsymbol{V}$ and $\boldsymbol{V}_{cont}$. In the interrelation process, these two relational information are concatenated and fused to become R. Through reprojection with $\boldsymbol{P}$, a final relation vector $\boldsymbol{F}_R$ in temporal space is obtained, which becomes $\boldsymbol{S}_R$ with an FC layer.

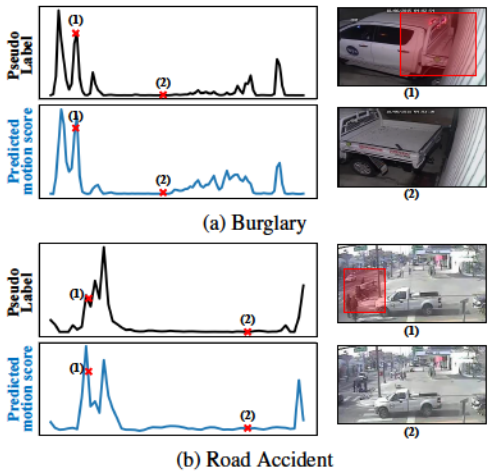


Figure II. Pseudo label and predicted motion score within the test video on UCF-Crimes [8]. For each video, motion score is (1) high in the dynamic scene and (2) low in the static scene. For better understanding, an abnormal event is marked using a red box.

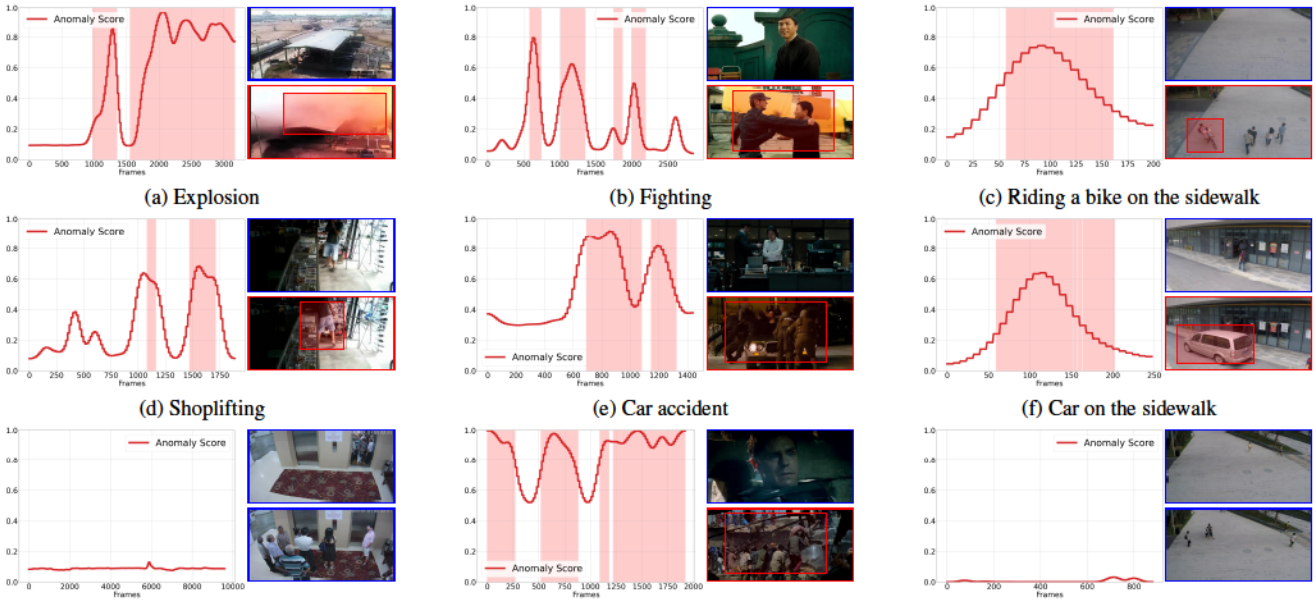
The hyper parameters are determined experimentally, and the batch consists of normal and abnormal videos in equal proportions for class balance, which is set within the range of [16, 64] for each database. For input snippets, we set to $T = 16$ for large-scale and untrimmed UCF-Crimes and XD-Violence database and 8 for ShanghaiTech and CUHK Avenue, which are relatively small datasets. For testing, the final score is calculated as a weighted sum of the anomaly score $\boldsymbol{S}$ and the relational score $\boldsymbol{S}_R$, where λ is set in the [0, 1] range. In the largest-scale XD-Violence dataset composed of complex and diverse scenes, λ is set to 1, which relies on the relational score the most. All models are trained in an end-to-end manner using PyTorch [7] with an Nvidia TITAN GPU.

3. Pseudo Labeling for Auxiliary Tasks

Dynamic Path. In CoMo, dynamic path predicts motion scores to select static features. To learn motion information, the ground truth of the motion score becomes the optical flow intensity I. We compute the optical flow of each frame using the TV-L1 algorithm [11] and average the intensity

Table II. AUC scores by hyper-parameters K, λ_{cs}, and λ_d on UCF-Crimes [8].

K				λ_{cs}				λ_d			
1	3	5	7	0	0.5	1	2	0	1	10	20
84.04	86.07	84.55	84.61	83.83	84.99	86.07	83.87	85.07	84.93	86.07	84.8



(a) Explosion (b) Fighting (c) Riding a bike on the sidewalk

(d) Shoplifting (e) Car accident (f) Car on the sidewalk

(g) Normal (h) Riot (i) Normal

Figure III. Anomaly score within the test videos. Each column is plot on the UCF-Crimes [8], XD-Violence [10], and ShanghaiTech [5] dataset. The x-axis is the frame range, and the red highlighted ranges are ground-truth abnormal frames. The blue and red border lines of the frame represent normal and abnormal scene, respectively, and for better understanding, an abnormal event is marked using a red box.

within the frame. As T frames are stacked for each snippet, the mean value of intensity becomes the motion score of every snippet. In Fig. II, the first row of each subplot shows the intensity score of every snippet, which is pseudo label I. The second row is the motion score S^{int} predicted by the dynamic path in the test video of UCF-Crimes, which shows that the prediction is similar to the optical flow intensity value. In addition, Fig. II (1) a high motion score in a dynamic scene and (2) a low motion score in a static scene are shown. The static feature corresponding to the bottom-N motion score becomes the input of the context path.

Context Path. In order to focus on the appearance and context information of surroundings rather than motion information, we select the features of the static snippet and predict the object class score S^{obj} appearing in the frames within the snippet. For ground truth, we utilize the output object classes' confidence score of the YOLOv5 [2] pretrained model on the MS COCO [4] object detection dataset. As the video anomaly detection (VAD) task is used for surveillance systems in the real world, only the person, car, motorcycle, truck, chair, and TV classes, which are subjective classes appearing in normal and abnormal situations, are considered (the average number of objects is 7 per frame and 119 per snippet on UCF-Crimes). A pseudo label O is generated by the mean score of each object class within the T frames.

Table III. AUC score on XD dataset

Scenario	w/o CoMo	w CoMo
Movies	76.54	76.42
Non-movies	77.31	83.42
Total	76.99	81.31

4. Context Feature in Other Scenarios

Considering the XD dataset with high motion intensity and changing background: As snippets corresponding to the bottom-N of the motion intensity score predicted in CoMo's dynamic path are selected and input to the context path, context features are extracted by exploring relatively static scenes within the video; the background may change due to mixed setups, such as handheld, sports, and movies, but it is mostly a single place (in a train, on the road, etc.). Although the camera moves, each video has similar contexts, and to compensate for this, general features

Table IV. Comparison of computational complexity with other models.

Model	#params	GFLOPs
Noise-C [12]-C3D	78M	386.2G
RTFM [9]-C3D	110M	101.1G
RTFM [9]-I3D	60M	56.5G
MIST [1]-C3D	85M	39.3G
MIST [1]-I3D	31M	45.7G
Ours-I3D	76M	64.4G
Ours-I3D (Inception-v1)	24M	34.7G

are extracted using the N static snippets. However, unlike real-world scenarios, for movie clips, context capturing is difficult when the scenes are switched. Tab. III shows that the total AUC score improves with CoMo, but we observe a large gap between the AUC of movies and non-movies which is because of the confusion of context.

5. Experimental Results by Hyper-parameters

We experimentally set the parameters used for training, and the results are presented in Table II. All settings of the experiment are the same; only K, λ_{cs}, or λ_d is different. While training, we assume that the top-K snippets are abnormal for the weakly labeled abnormal video, and $K = 3$ shows better performance than learning only with the highest score of $K = 1$. When the snippet with the highest score in an abnormal video is not an abnormal snippet, this error brings a huge impact on training. Furthermore, when there is more than one abnormal snippet, the chance to learn about the remaining data is missed [9]. Therefore, learning with the mean value of K snippets is effective and shows the best performance when $K = 3$. There is an optimal K value depending on the length of the abnormal interval in the video of each dataset, but to reduce the dependence on the data, we used the same K value in all datasets.

λ_{cs} and λ_d are the weights of the class-specific loss and relative distance loss, respectively, which indicate the importance to the total training loss value. When λ_{cs} is 0 and 1, it shows a large performance difference of 2.24%, indicating that the CS module helps the framework in extracting normal and abnormal class-representative features. In addition, the best result is achieved when $\lambda_d = 10$, which addresses relative distance loss complements MIL-based score learning with feature learning to good effect.

6. Computational Complexity

We compare the number of parameters and FLOPs of Noise-C [12], RTFM [9], and MIST [1] with the proposed model in Table IV. The complexity is computed according to the backbone which the layers and feature dimension used for each model are different. We utilize the 'mix 5c' layer of 2048-dim and 1024-dim features from the ResNet-50 I3D and Inception-v1 I3D backbone, respectively; for UCF-Crimes dataset, same as RTFM, the backbone of our model is the ResNet-50 I3D. In this case, our model with 76M number of parameters and 64.4G FLOPs has a higher complexity than RTFM, but the performance is 2.1% higher which is competitive. Furthermore, for other datasets, we utilize Inception-v1 I3D backbone and shows SOTA performance with low model complexity; compared with MIST, our model shows 3.2% higher performance in ShanghaiTech dataset with lower complexity.

7. Qualitative Results

Fig. III shows the abnormal score in the test video predicted by our framework. Columns in the figure give the results of the UCF-Crimes, XD-Violence, and ShanghaiTech videos. As shown in the examples, UCF-Crimes and XD-Violence databases consist of various real-world scenes that are more complex than those in the ShanghaiTech database. In score plots, high scores are shown not only in (a) explosion and (e) car accident, which are abnormal events with large motion, but also in (d) shoplifting, where anomalies need to be inferred through relational information. Furthermore, comparing (h) and (g), there are high scores for riot events but low scores without false alarms for the normal event where a crowded group is waiting for an elevator and boarding all at once. These examples address the importance of focusing on the relationship between motion and context information for VAD.

References

[1] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021. 4

[2] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, TaoXie, Kalen Michael, Jiacong Fang, imyhxy, Lorna, Colin Wong, (Zeng Yifu), Abhiram V, Diego Montes, Zhiqiang Wang, Cristi Fati, Jebastin Nadar, Laughing, UnglvKitDe, tkianai, yxNONG, Piotr Skalski, Adam Hogan, Max Strobel, Mrinal Jain, Lorenzo Mammana, and xylieong. ultralytics/yolov5: v6.2 - YOLOv5 Classification Models, Apple M1, Reproducibility, ClearML and Deci.ai integrations, Aug. 2022. 3

[3] Shuo Li, Fang Liu, and Licheng Jiao. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. *AAAI*, 2022. 1

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[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. 1, 3

[6] Zhen Ma, José JM Machado, and João Manuel RS Tavares. Weakly supervised video anomaly detection based on 3d convolution and lstm. *Sensors*, 21(22):7508, 2021. 1

[7] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 2

[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. 1, 2, 3

[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*, 2021. 1, 4

[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. 1, 3

[11] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint pattern recognition symposium*, pages 214–223. Springer, 2007. 2

[12] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1237–1246, 2019. 4