Supplementary: A Coarse-to-Fine Pseudo-Labeling (C2FPL) Framework for Unsupervised Video Anomaly Detection

Anas Al-lahham Nurbek Tastan Muhammad Zaigham Zaheer Karthik Nandakumar
Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)
Abu Dhabi, UAE
{anas.al-lahham, nurbek.tastan, zaigham.zaheer, karthik.nandakumar}@mbzuai.ac.ae

This supplementary material is for the WACV Submission 1501, providing an extended overview of our A Coarse-to-Fine Pseudo-Labeling (C2FPL) framework for unsupervised video anomaly detection. This document provides further qualitative visualizations of our proposed model, as well as relevant discussions. In addition, we discuss the design choice of our self-attention and analyze its effectiveness in comparison with the other state-of-the-art (SOTA) anomaly detection methods using similar/comparable designs. Finally, we present a convergence analysis of our model's performance.

A. Self Attention

Figure 6 shows the detailed architecture of our proposed C2FPL network. The FC layers described in manuscript: Section 3.3 have 512 and 32 neurons where each is followed by a ReLU activation function and a dropout layer with a dropout rate of 0.6. In addition, we add two self-attention layers. In this section, we will discuss the choice design as well as the aim of using this layer.

The aim of self-attention (SA) in our proposed C2FPL framework is to highlight parts of feature vectors critical in detecting anomalies. Our configuration applies self-attention over each feature vector (feature dimension) independently without requiring temporal order. This is unlike a compareable existing architecture by Zaheer *et al*. [3] where the Normalcy Suppression Module (NSM) aims to learn attention based on the temporally consistent feature vectors in the input batch (Figure 8(a)) and the attention is calculated along the batch dimension (temporal axis).

To study this in details, we define several possible configurations of the self-attention used in our C2FPL and report their performances in this section. Through thorough analysis, we verify the effectiveness of our design choices within the framework.

A.1. Residual vs Multiplicative Self-Attention (SA)

Zaheer *et al*. [3], in CLAWS Net, formulate the problem of self-attention in terms of suppressing certain features which are achieved by multiplicative attention. To provide a comparison, we discuss two different SA configurations as shown in Figure 7. First, following Zaheer *et al*. [3], given an input batch b we calculate the output $H(b)$ by performing an element-wise multiplication $\otimes$ between SA output $S(b)$ and backbone output $FC(b)$ as:

$$H(b) = S(b) \otimes FC(B)$$

Although such multiplication has been helpful in CLAWS Net, generally it has been shown to have the unfavorable result of dissipating model representations [1, 2]. It's because attention generates probabilities that, when multiplied by the features directly, can drastically lower the values.

In our framework, we utilize residual SA in which attention-applied features are added back to the original features. Therefore, The output $H(b)$ is calculated as:

$$H(b) = (FC(b) \otimes S(b)) \oplus FC(b)$$

where $\oplus$ is an addition operation.

Table 5 shows the performance difference between multiplication and residual attention approaches. We can observe that the use of multiplication negatively affects our model's AUC performance (63.5%). We attribute this to the suppression nature of multiplication [1, 2]. The specifically

Framework	SA configuration	AUC (%)
CPL → FPL → AD	Multiplicative	63.5
	Residual (Ours)	80.6

Table 5. Area under the curve (AUC) comparison of two SA configurations configurations on the UCF-Crime dataset. (The framework configuration is the same as shown in manuscript: Table 3).

Framework	SA Dimension	AUC (%)
CPL → FPL → AD	Batch Dimension	76.5
	Feature Dimension (Ours)	80.6

Table 6. Area under the curve (AUC) comparison of two SA types on UCF-Crime dataset. (The framework is the same as shown in manuscript: Table 3).

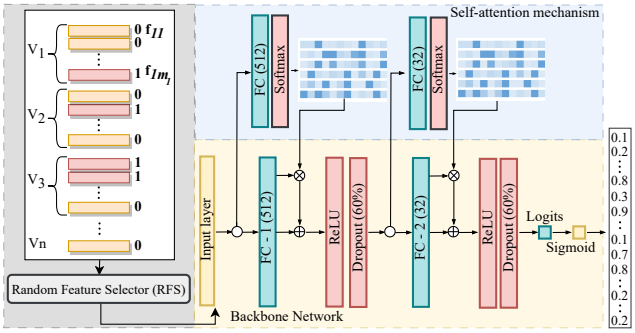

Figure 6. Detailed architecture of our proposed learning network: The training batch containing pseudo-labeled feature vectors is the input to the FC backbone network (lower). In addition to the backbone network, we add two self-attention layers (upper).

designed NSM of CLAWS Net [3] aims to dissipate normal portions of the temporally consistent input batches that help the backbone network produce low anomaly scores. However, the nature of our training is not suitable for this formulation. Therefore, using residual attention, which only highlights individual parts of each feature vector in a given batch, the performance of our model increases to 80.6% on the UCF-crime dataset.

A.2. Types of self-attention

In conjunction with Zaheer *et al*. [3], We discuss two different types of self-attentions depending on the dimensions along which Softmax probabilities are computed in an element-wise fashion.

Softmax probabilities over the batch dimension (BD). As mentioned, Zaheer *et al*. [3] calculates the probabilities temporally to make use of the temporal information preserved within a batch (Figure 8 (a)). However, we have argued and demonstrated in our presented C2FPL framework that preserving temporal information is not necessary for improved anomaly detection performance. Therefore, using temporal attention along the batch dimension may not be as effective in our framework as it has been proven in CLAWS Net by Zaheer *et al*. [3]. Nevertheless, we utilize their proposed self-attention and compare it with our design of self-attention.

Softmax probabilities over the feature dimension

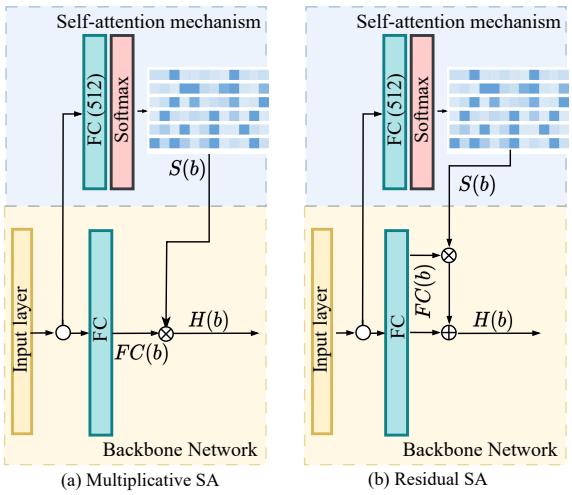

Figure 7. Visualization of the two self-attention configurations including (a) Multiplicative SA and our proposed (b) Residual SA.

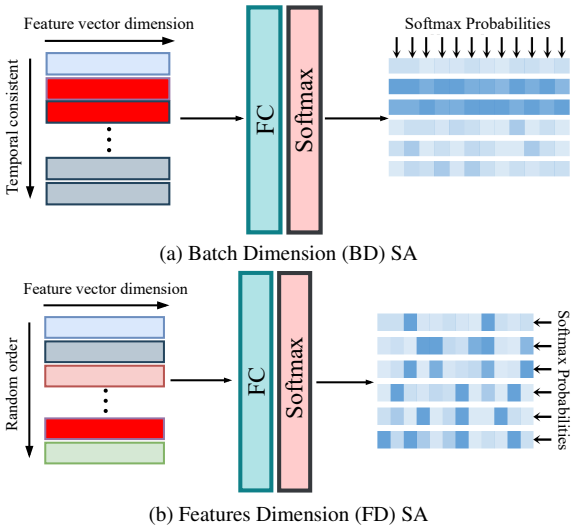

Figure 8. Visualization of the two types of self-attention: (a) Batch Dimension (BD): Softmax probabilities are calculated along the Batch dimension (temporal axis). (b) Features Dimension (FD): Softmax probabilities are calculated along the feature vector dimension.

(FD). This self-attention over feature dimension (FD) is the configuration used in our C2FPL framework, as explained in manuscript: Section 3.3 (lines 494-500). Since we assume no temporal consistency among batches, the probabilities are computed over the feature dimension (Figure 8 (b)).

Table 6 summarizes the frame-level AUC performance of the two types. It can be seen that the FD type (ours) outperforms the BD type attention by a margin of 4.1%. This verifies the importance of using self-attention along

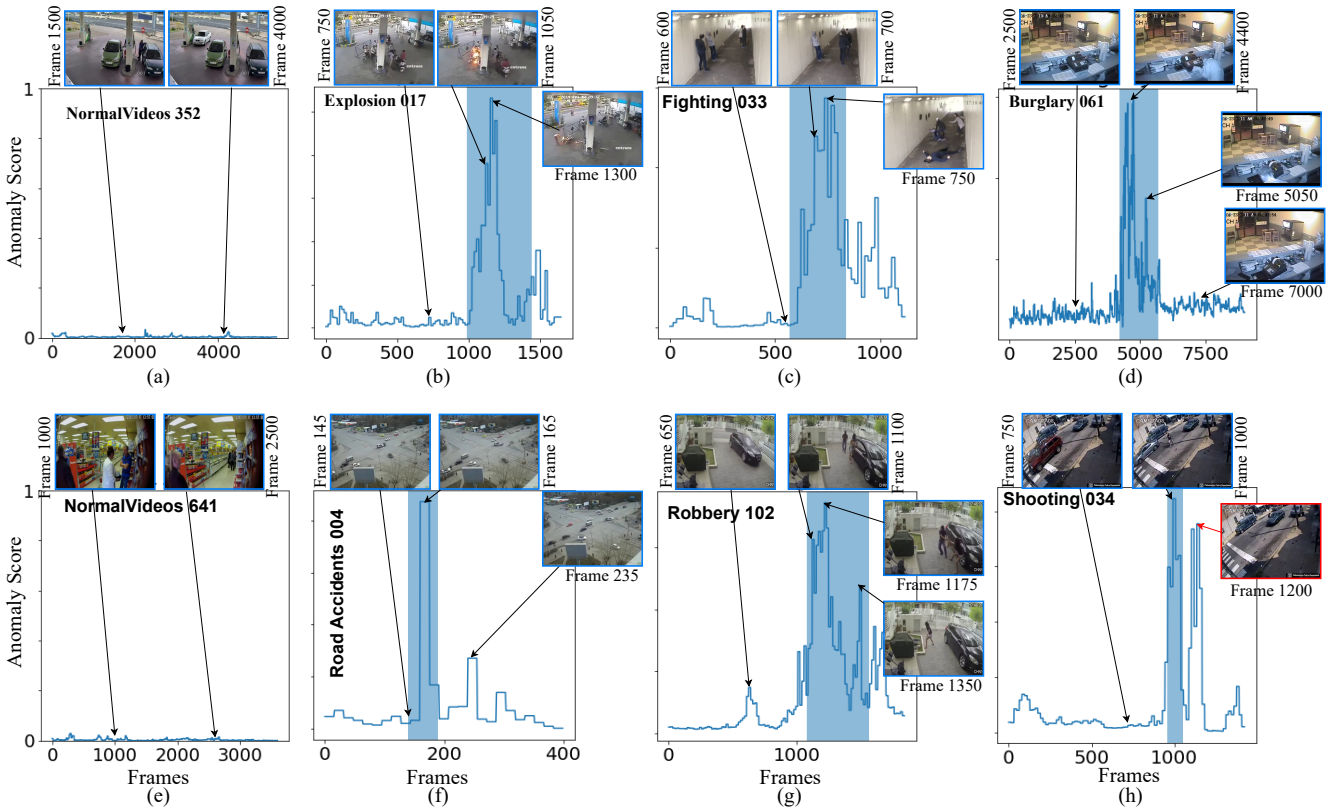

Figure 9. Anomaly scores of the proposed C2FPL framework on different videos from the UCF-Crime Dataset.

the feature vector dimension, achieving significant performance gains.

B. Qualitative Results

We also provide additional qualitative results in Figure 9, where anomaly scores predicted by our C2FPL approach are visualized for other classes of anomalous videos from the UCF-Crime dataset. In some cases, the anomalous frames in certain videos might exceed the annotated ones because the annotations only cover a portion of the event. For instance, the abnormal event in the RoadAccidents004 video begins at about frame 145 and lasts significantly longer than the annotated window, which only shows the accident impact event.

An additional **failure case**, shooting034 video (UCF-Crime), is also visualized in Figure 9(h). Our proposed model correctly predicts the ground-truth anomalous window. However, later frames (1200) of the video show one of the occupants involved in the shooting quickly entering his car before speeding off, which our detector marks as an anomalous event while that event is annotated as a normal event.

C. Convergence Analysis

As our approach is an unsupervised anomaly detection method, we empirically analyze its convergence using 10 random seed runs as shown in Figure 10. For all experiments, our C2FPL model attains an average AUC of 80.14% ± 0.31%. This demonstrate that our proposed framework not only achieves excellent anomaly detection but also demonstrates good convergence.

References

[1] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 1

[2] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017. 1

[3] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Pro-*

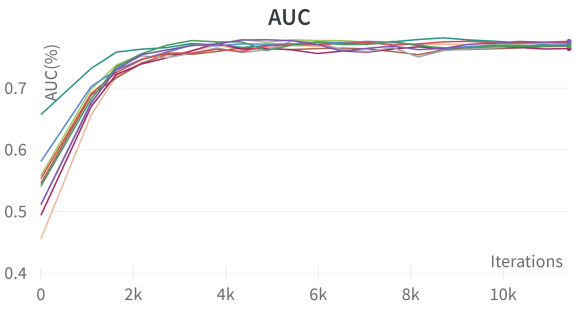

Figure 10. Convergence of our proposed model using multiple random seed experiments.

ceedings, Part XXII 16, pages 358–376. Springer, 2020. 1, 2