Collaborative Learning of Anomalies with Privacy (CLAP) for Unsupervised Video Anomaly Detection: A New Baseline

Supplementary Material

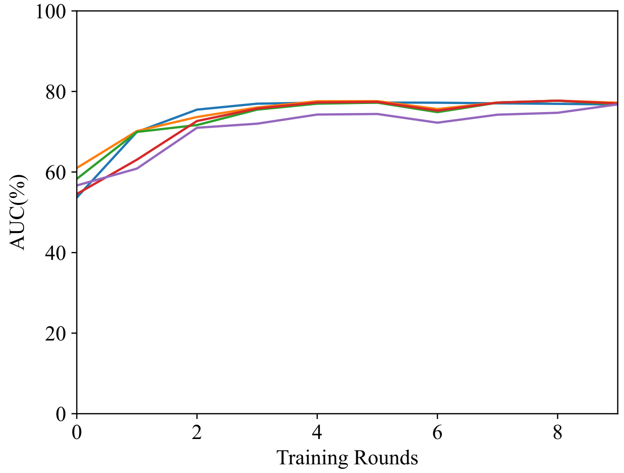


Figure 5. Empirical training convergence. Experiments are run using 5 different seeds enabling randomized data splits across participants. CLAP achieves an average AUC of 77.32 ± 0.189.

6. Collabroative algorithms Study

In addition to the main results in the manuscript where FedAVG is used as the main FL method, we conduct experiments using other FL methods including FedProx [16] and SCAFFOLD [13]. In these experiments, CLAP achievs 73.4% & 73.7% AUC respectively on **scene based split**. Overall, the performance is comparable with the 73.99% AUC when using FedAVG.

7. Training convergence

As CLAPis an unsupervised learning model where each participant uses its share of the data to collaborate towards training a joint model, we empirically validate its convergence by repeating the training on UCF-Crime using 5 different random seeds. These random seeds also enable random data splits between the participants. As seen in Figure 5, CLAPachieves an average AUC of 77.32% ± 0.189%. This shows the robustness of CLAPin yielding good performance with small variation even with significant variation in the dataset splits across participants.

8. Bandwidth Consumption

In real-world surveillance applications, network bandwidth allowing data communication between the training server and the participants can be limited due to several factors such as remote locations, large number of participants, etc. Given the involvement of lengthy surveillance videos for anomaly detection, a collaborative learning approach such as CLAPshould preferably communicate a limited amount of data per training round. As shown in Algorithm 3, the server receives the Gaussian parameters from each participant in addition to receiving the gradients of each local model (2.1 M parameters) during the training rounds. Therefore, on each communication round, CLAPcommunicates an average of 6.07 Mega Bytes (MB) from each participant. Given 10 training rounds, the overall data transfer remains around 60.7 MB which is significantly lower than the case of central training where all data is transferred to the central server for training.

9. Dataset Splitting Strategies

As described in Section 4.5 of the manuscript, collaborative learning in video anomaly detection (VAD) may have several possible scenarios. Careful consideration of these scenarios leads to three different data splitting strategies including random, event class, and scene-based. Each of these is explained further next:

Random Split: Random Split is a baseline strategy where each participant is assigned videos randomly while ensuring a comparable number of normal and anomalous videos. Example visualizations of some videos taken from a single participant are provided in Figure 8.

Event Class Based Split: Each anomalous activity can be classified into different categories of events, e.g., road accidents, robbery, fighting, shooting, or riots. The intuition behind this split is that each collaborating participant may have a certain type of anomalous examples. A better performance of an anomaly detection network on this setting may indicate the success of collaborative learning between different organizations contributing videos containing different types of anomalous events from each other. This setting is more challenging than random distribution because each participant may have a different number of videos containing certain events. Example visualizations of some videos taken from a single participant are provided in Figure 9.

Scene Based Split: In this setting, each participant is assumed to have videos based on the scenes/locations where the videos are recorded. For example, one participant may have surveillance videos for fuel stations, another participant may have indoor videos of offices, and so on. The intuition behind this split is that similar anomalous events may occur at different locations and captured by different partic-

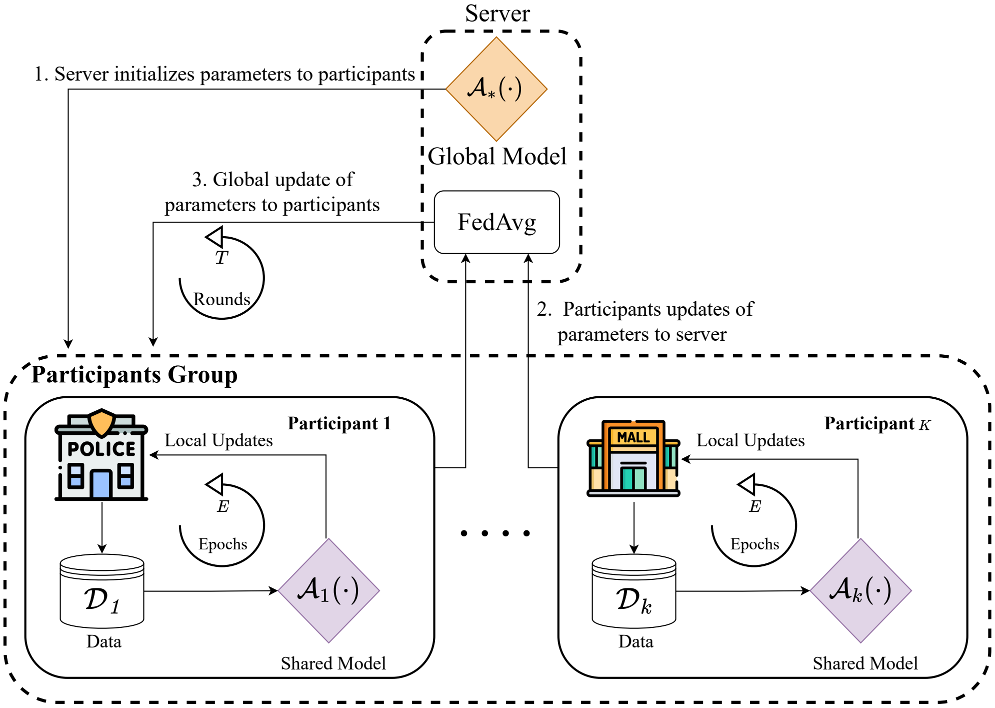


Figure 6. Abstract level Flowchart of our collaborative training scheme.

ipants. This is the most challenging setting as the dataset is not balanced either among participants or within a participant for the normal and anomalous classes. Example visualizations of some videos taken from a single participant are provided in Figure 10.

10. Architecture and Implementation Details

Our learning network, as seen in Figure 7, consists of a fully connected (FC) network and two self-attention layers. The FC network has two fully connected layers and one output layer for binary classification. A ReLU activation function and a dropout layer follow each FC layer. The FC layers have 512 and 32 neurons respectively. The self-attention layers, with dimensions matching their respective FC layers, are followed by a Softmax activation function. Unlike previous works [36, 38], we compute Softmax probabilities over the feature dimension instead of the batch size dimension. The final anomaly score prediction ranging $[0, 1]$ in our network is obtained through a Sigmoid activation function in the output layer. We use binary cross-entropy loss along with L_2 regularizer as our training loss function.

11. Datasets

Two large-scale video anomaly detection benchmark datasets are used to evaluate our approach: UCF-Crime [29] and XD-Violence [34]. These datasets are originally labeled for weakly supervised VAD tasks, where video-level labels are present for training and frame-level labels are provided only for testing. In our unsupervised VAD experiments, we completely discard the provided labels before carrying out the training.

11.1. UCF-Crime

UCF-Crime consists of 1,610 training videos and 290 testing videos covering 13 anomaly categories including Abuse, Arrest, Arson, Assault, ... etc. Some examples of these videos are shown in Figures 8, 9, & 10. These videos were gathered from actual surveillance camera feeds, amounting to a combined duration of 128 hours.

11.2. XD-Violence

XD-Violence is a multi-modal dataset sourced from various channels, including sports streaming videos, movies, and web videos. The dataset encompasses a total of 3,954 training videos and 800 testing videos. These videos collectively span approximately 217 hours.

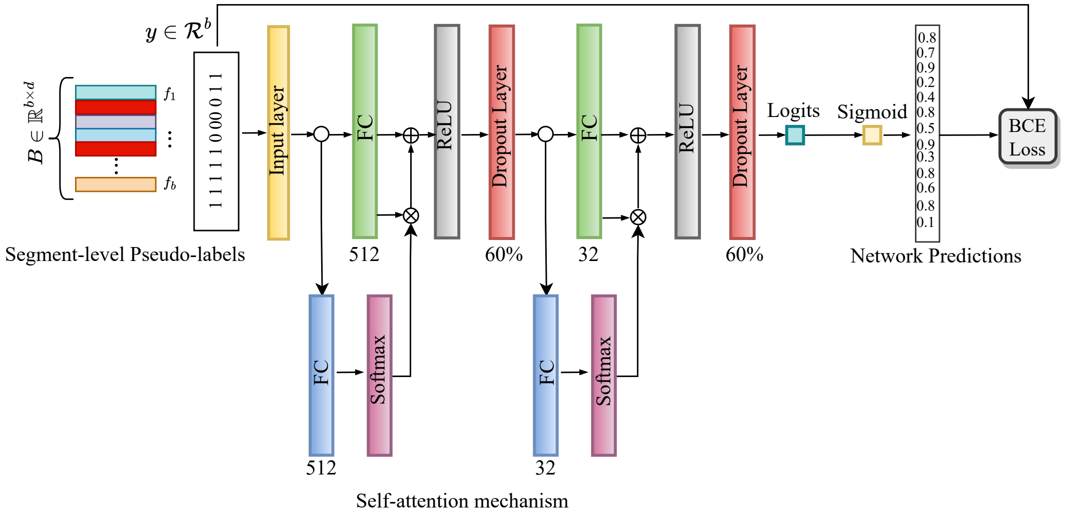


Figure 7. Learning network used in CLAP: The training batch containing pseudo-labeled feature vectors is the input to the FC backbone network (upper). In addition to the backbone network, we add two self-attention layers (lower).

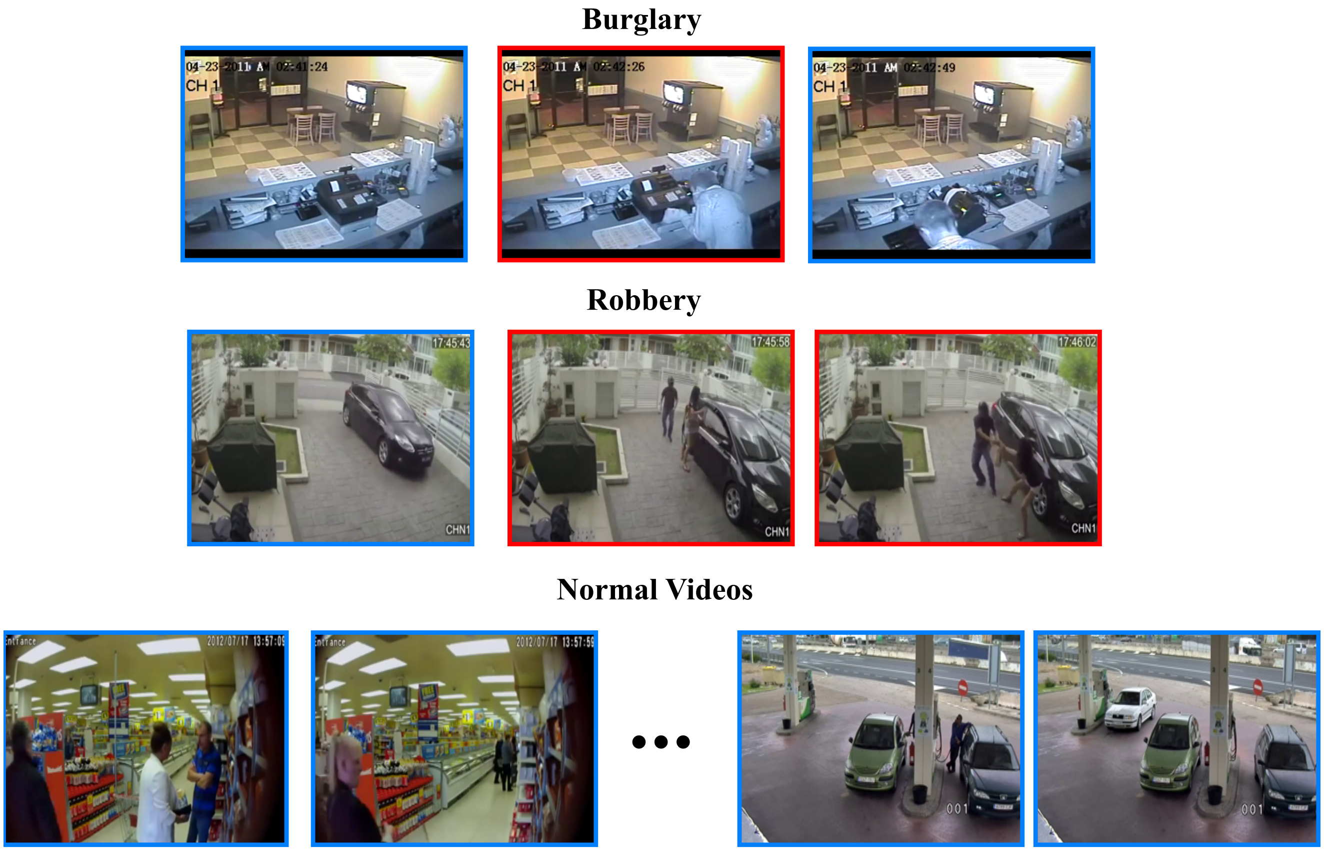


Figure 8. Example of UCF-crime videos in the random split taken from one of the participants. The blue borders represent normal events while the red borders represent anomalous events.

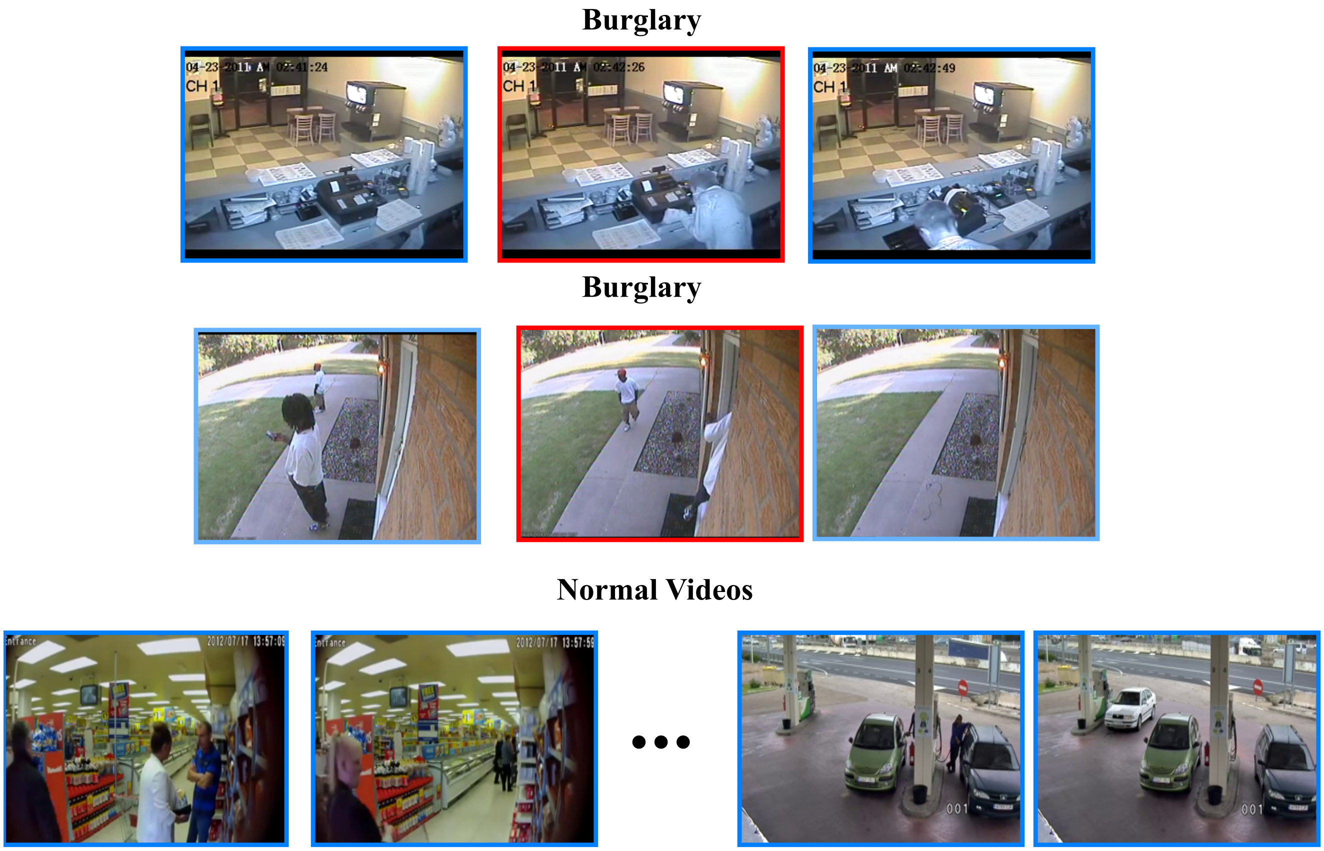
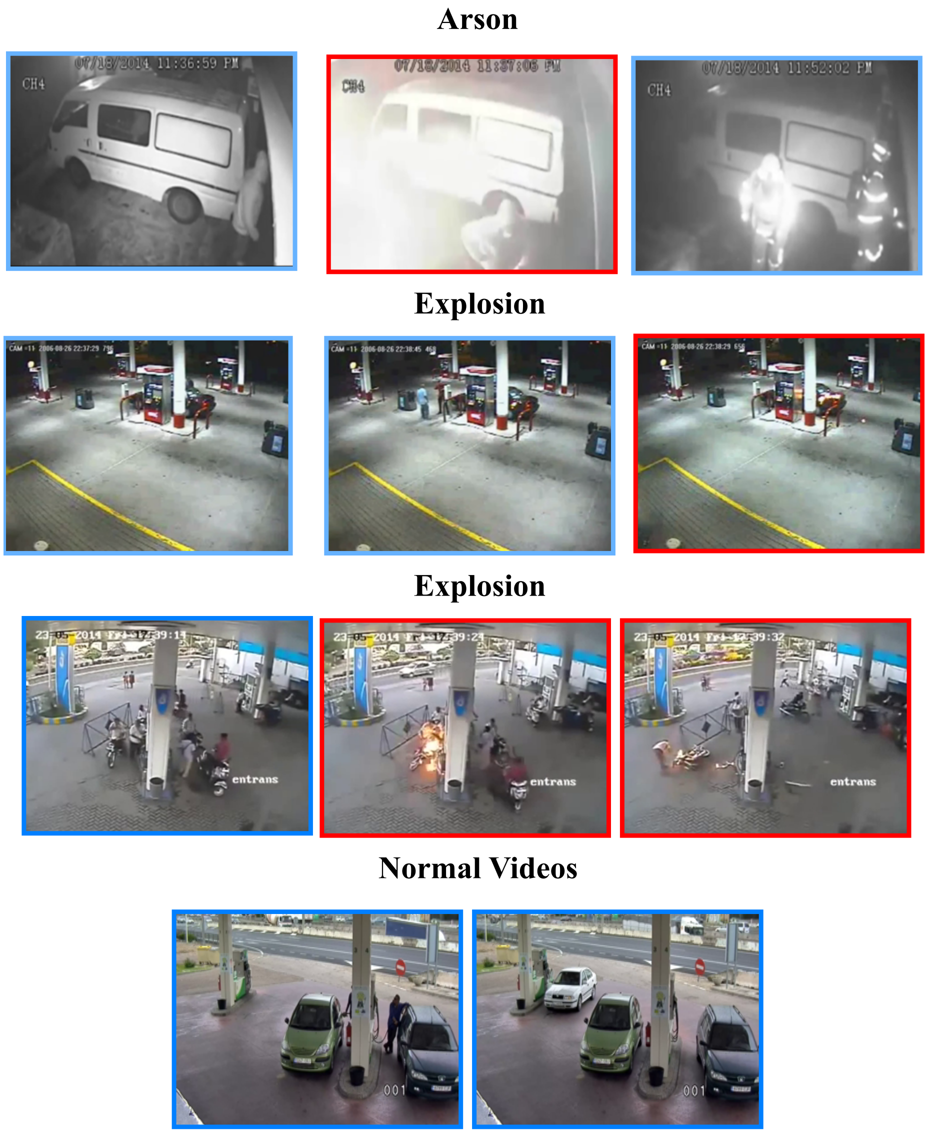


Figure 9. Example of UCF-crime videos in the event-based split taken from one of the participants. For each participant, anomalous events are the same but the background scenes can be different. The blue borders represent normal events while the red borders represent anomalous events.



Arson

Explosion

Explosion

Normal Videos

Figure 10. Example of UCF-crime videos in the scene-based split taken from a participant having videos of fuel pumps and automotive workshops). For each participant, anomalous events can be different but the overall background scenes are similar. The blue borders represent normal events while the red borders represent anomalous events.