Bayesian Nonparametric Submodular Video Partition for Robust Anomaly Detection

Hitesh Sapkota Qi Yu
Rochester Institute of Technology
{hxs1943, qi.yu}@rit.edu

Abstract

Multiple-instance learning (MIL) provides an effective way to tackle the video anomaly detection problem by modeling it as a weakly supervised problem as the labels are usually only available at the video level while missing for frames due to expensive labeling cost. We propose to conduct novel Bayesian non-parametric submodular video partition (BN-SVP) to significantly improve MIL model training that can offer a highly reliable solution for robust anomaly detection in practical settings that include outlier segments or multiple types of abnormal events. BN-SVP essentially performs dynamic non-parametric hierarchical clustering with an enhanced self-transition that groups segments in a video into temporally consistent and semantically coherent hidden states that can be naturally interpreted as scenes. Each segment is assumed to be generated through a non-parametric mixture process that allows variations of segments within the same scenes to accommodate the dynamic and noisy nature of many real-world surveillance videos. The scene and mixture component assignment of BN-SVP also induces a pairwise similarity among segments, resulting in non-parametric construction of a submodular set function. Integrating this function with an MIL loss effectively exposes the model to a diverse set of potentially positive instances to improve its training. A greedy algorithm is developed to optimize the submodular function and support efficient model training. Our theoretical analysis ensures a strong performance guarantee of the proposed algorithm. The effectiveness of the proposed approach is demonstrated over multiple real-world anomaly video datasets with robust detection performance.

1. Introduction

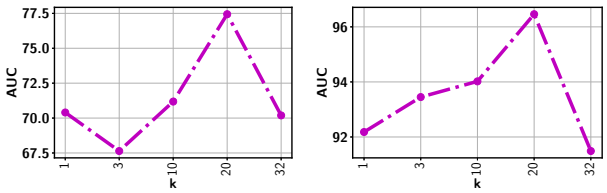
Anomaly detection from videos poses fundamental challenges as abnormal activities are usually rare, complicated, and unbounded in nature [15]. Furthermore, segment or frame labels are typically unavailable due to high labeling cost and therefore, the detection models have to rely on the weak video level labels [19]. There are two main streams of work to handle the challenging anomaly detection task. The first stream treats anomaly detection as an unsupervised learning problem [21]. It assumes that an event is considered to be abnormal if it deviates significantly from a predefined set of normal events included in the training data [2, 24, 26]. However, a model trained on limited normal data is likely to capture only specific characteristics present in the training dataset, and therefore, testing normal events deviating significantly from the training normal events will lead to a high false alarm rate [27]. The second stream of research has attempted to address the limitation by formulating the problem as multiple instance learning (MIL) that models each video as a bag and its segments (or frames) as instances within the bag [19]. The goal is to learn a model that can effectively make frame-level anomaly predictions relying on the video-level labels during the training process. One effective MIL learning objective is to maximize the gap between two instances having the respective highest anomaly scores from a pair of positive and negative bags. The maximum score based MIL (referred as MMIL) model outperformed the unsupervised approaches and achieved promising performance in real-world long surveillance videos [19].

However, there are two key limitations with the MMIL model. First, the presence of noisy outliers (different from other normal events) in both abnormal and normal videos may significantly impact the overall model performance. This is because the objective function solely focuses on the individual segments from both positive and negative bags, making the training process sensitive to noises. Figure 1 (a-b) shows the example normal frames that are significantly different from other normal ones in real-world surveillance videos. The first frame is from the burglary video that looks similar to an abnormal frame from a video with an arson event. The second frame is from the shooting video that looks similar to a fighting frame. Hence, they may serve as outliers in the corresponding videos.

Second, if multiple types of abnormal events (referred to as multi-modal anomaly) present in a single abnormal video, MMIL may only detect one type of anomaly while

(a) Burglary (Outlier 1) (b) Walking (Outlier 2) (c) Running (Modality 1) (d) Loitering (Modality 2) (e) Walking (Modality 3)

Figure 1. Examples of outlier (a-b) and multimodal frames (c-e) from the Avenue dataset



(a) Avenue (b) ShanghaiTech

Figure 2. Highly fluctuating detection performance w.r.t. k

missing other important ones due to the limitation of the objective function. Figure 1 (c)-(e) demonstrate three frames with different anomaly types from an example video in the Avenue dataset [16]. In Figure 1 (c), the person is running, which is regarded as a strange action in that context [16]. In (b), it shows a person waiting in a place holding some object in the hand, and (c) involves a person walking in the wrong direction. Therefore, the single video has multiple anomaly frames leading to a multimodal scenario.

Top-k ranking loss has been adopted in an attempt to address the issues as outlined above. It maximizes the gap between the mean score of the top-k predicted instances from a positive bag and that of a negative one [18, 22]. However, there are inherent limitations by using a top-k loss. First, it tends to be extremely sensitive to the selected k value. Figure 2 shows the highly fluctuating detection performance from two real-world surveillance video datasets. Since there is no frame (or segment) labels available during model training, setting an optimal k through cross-validation is infeasible or highly costly. Furthermore, given the diverse videos, the number of abnormal instances may vary significantly from one video to another implying we should have a different k for each video. Hence, applying the same k to all videos as in the existing approaches fails to capture the nature of the data. Another serious but more subtle issue is that all (or most of) the selected k segments may come from the same sub-sequence of the video. Using a consecutive set of visually similar segments is less effective for model training, making it more likely to suffer from outlier and multimodal scenarios. As a result, top-k approaches will fall short in providing a reliable detection performance in most practical settings as evidenced by our experiments.

To address the fundamental limitations of existing solutions, we propose novel Bayesian non-parametric construction of a submodular set function, which is integrated with multiple instance learning to deliver robust video anomaly detection performance under practical settings. Instead of choosing a set of instances with the highest prediction scores that are likely from a consecutive sub-sequence, maximizing a specially designed submodular function can involve a more diverse set of instances and expose the model to all potentially abnormal segments for more effective model training. Furthermore, the submodular set function is constructed in a non-parametric way, which induces a pairwise similarity among different segments in a video based on the diverse nature of the data. More specifically, an infinite Hidden Markov Model with a Hierarchical Dirichlet prior (HDP-HMM) [20] augmented with an enhanced self-transition is employed to partition a video through dynamic non-parametric clustering of its segments. To more effectively accommodate the dynamic and noisy nature of real-world surveillance videos, the emission process of the HMM is also governed by a non-parametric mixture model to allow segments within the same hidden state to have visual and spatial variations. This unique design is instrumental to discover temporally consistent and semantically coherent hidden states that can be naturally interpreted as scenes. Pairwise similarity among different segments is defined according to the state-component structure, which leads to the construction of a submodular set function. We then develop a novel submodularity diversified MIL loss function to ensure robust anomaly detection from real-world surveillance videos with outlier and multimodal scenarios. Our key contributions include:

- Formulation of a novel **submodularity diversified MIL loss** that simultaneously extracts a diverse set of potentially positive instances while maximizing the gap between the mean score of these instances from a positive bag and a negative one, respectively.
- **Bayesian non-parametric construction of the submodular set function** that infers the diversity from the video data to induce a pairwise similarity among different segments in a video and provide an upper bound on the size of the diverse set.
- **A greedy algorithm** that leverages the state-component hierarchical structure resulting from the non-parametric construction for submodular set function optimization and efficient model training.
- **Theoretical results** to ensure strong performance guarantee of the greedy algorithm.

The proposed approach achieves the state-of-the-art robust anomaly detection performance on real-world surveillance videos with noisy and multimodal scenarios.

2. Related Work

Encoding and sparse reconstruction-based approaches have been employed for anomaly detection, assuming that abnormal events are rare and deviate from normal patterns. They aim to capture the normal patterns using models, such as Gaussian processes (GPs) [12] and HMMs [9], to identify anomalies as outliers based on the reconstruction loss. Sparse representation-based approaches construct a dictionary for normal events and identify the events with the high reconstruction error as anomalies [16]. Recent approaches consider both abnormal and normal events in the training process. For video anomaly detection, since only video-level labels are assumed to be available during model training [6], MIL offers a natural solution by modeling each video as a bag and the associated segments (frames) as instances of the bag. Sultani et al. proposed an MIL based approach that enables to maximize the gap between highest prediction scores from a positive and negative bags, respectively [19] . However, this maximum score based MIL model (*i.e.*, MMIL) is insufficient to handle outlier and multimodal scenarios as discussed earlier.

Top-k ranking loss based MIL models have been developed to address the limitations of the MMIL model [18,22]. These models produce state-of-the-art detection performance given that a suitable k value can be assigned in advance. However, as demonstrated earlier, the detection performance of such models is highly sensitive to the chosen k value. Meanwhile, given the diverse nature of videos, applying the same k value to all the videos is sub-optimal. More importantly, since instance level labels are not available during training time, choosing a single k value through cross-validation is infeasible or incurs a high annotation cost. Distributionally Robust Optimization (DRO) has been used to convert the top-k set into an uncertainty set that allows the model to focus on instances in proportion to their prediction scores [18]. This is equivalent to assigning soft membership to involve instances into the MIL loss function. However, the size of the uncertainty set is controlled by the radius (*i.e.*, η) of the uncertainty ball, which needs to be manually set. Furthermore, the model may put more focus on a set of consecutive segments with the highest prediction scores and ignore some other potentially positive segments.

The proposed approach constructs a novel submodular set function in a non-parametric way by inferring the diversity from data automatically. By jointly optimizing the submodular function and the MIL loss, it automatically chooses a diverse set of segments and lets the model better differentiate these (potentially positive) segments from those of a negative bag to ensure good detection performance.

3. Methodology

Following the standard MIL assumption, we consider, for a positive bag, there is at least one abnormal segment whereas, for a negative bag all segments are of normal types. Table 3 in the Appendix summarizes the major symbols and their descriptions.

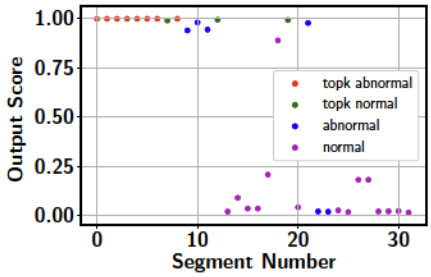


Figure 3. Output of the top-k based approach in a video from Avenue datase (missing some of the abnormal segments in top-k).

3.1. Preliminaries

Let $\mathbf{x}_i^+$ be the i^{th} segment in the positive bag $\mathcal{B}_{pos}$ and $\mathbf{x}_j^-$ indicates the j^{th} segment in the negative bag $\mathcal{B}_{neg}$. Also consider n as the total number of instances per bag. The maximum score based MIL (MMIL) model tries to maximize the gap between the maximum prediction score from positive bag and that from the negative bag [19]:

$$L(\mathcal{B}_{pos}, \mathcal{B}_{neg}) = \Big[1 - \max_{i \in \mathcal{B}_{pos}} (f(\mathbf{x}_i^+)) + \max_{j \in \mathcal{B}_{neg}} (f(\mathbf{x}_j^-))\Big]_+ \tag{1}$$

where $f(\mathbf{x}_i^+)$ (or $f(\mathbf{x}_j^-)$) is the prediction score of $\mathbf{x}_i^+$ (or $\mathbf{x}_j^-$) and $[a]_+ = \max\{0, a\}$. As mentioned earlier, MMIL is less effective to handle outlier and multimodal scenarios. The top-k ranking loss partially addresses the limitation of MMIL by maximizing the gap between an average of k highest segment predictions from the positive bag and maximum segment prediction score from a negative bag:

$$L(\mathcal{B}_{pos}, \mathcal{B}_{neg}) = \Big[1 - \frac{1}{k}\sum_{i=1}^{k} f(\mathbf{x}_{[i]}^+) + \max_{j \in \mathcal{B}_{neg}} f(\mathbf{x}_j^-)\Big]_+ \tag{2}$$

where the positive bag segment predictions are sorted in a non-decreasing order, *i.e.*, $f(\mathbf{x}_{[1]}^+) \geq ... \geq f(\mathbf{x}_{[k]}^+)$.

There are two major issues associated with the top-k ranking loss. First, choosing an optimal k value is a key challenge as the number of abnormal instances may vary significantly from one video to another implying different k value for each video. Second, all the selected top-k instances may come from same sub-sequence of the video. Including all those visually similar instances does not contribute much in the model training process. Furthermore, concentrating only on a specific sub-sequence may make the approach less effective to handle multimodal and outlier scenario. Figure 3 presents the output of the top-k based model in the Avenue dataset. It can be seen that the top-k

(a) Normal (b) Abnormal (c) Abnormal (d) Normal (e) Abnormal

Figure 4. Example frames from different scenes in an explosion video from UCF-Crime: (a-b) scene 1, (c) scene 2, (d-e) scene 3

based approach picks the consecutive video segments while missing quite a few other abnormal frames.

3.2. Bayesian Non-parametric Submodular Set Function Construction

The proposed Bayesian non-parametric submodular video partition (BN-SVP) approach offers a novel integrated solution to address the above two fundamental challenges simultaneously. In particular, since submodular set functions provide a natural measure for diversity, we design a special submodular set function that enables discovery of a representative set of segments from a video. This avoids only choosing visually similar consecutive video segments like in the top-k approach, which enhances the model's exposure to potential abnormal instances during model training. As a result, the model's capability to handle multi-modal and outlier scenarios can be effectively improved.

However, maximizing a submodular set function still requires to specify the size of the set. As mentioned above, choosing a set with an optimal size in video anomaly detection is highly challenging. To this end, we propose a novel Bayesian non-parametric construction of the submodular set function. The non-parametric construction leverages both visual features of the video segments and their temporal closeness to derive a similarity measure that allows us to define a submodular set function $F(\mathcal{C}^+)$, where $\mathcal{C}^+$ represents a subset of segments in a video. The size of $\mathcal{C}^+$ is automatically determined through Bayesian non-parametric analysis of the video. Intuitively, most videos, especially those with anomalies, usually consist of multiple scenes, where each scene is comprised of a consecutive set of visually similar segments. Figure 4 shows the example frames from three different scenes in a video that records an explosion event. Ideally, if a video could be partitioned based on these scenes, we can choose representative (and potentially positive) segments from each scene. Such information can significantly facilitate the optimization of the submodular set function. However, both the number and the types of the scenes are unavailable during model training.

The proposed BN-SVP addresses the above issue through non-parametric partition of a video. It builds upon and extends an HDP-HMM model that places a Hierarchical Dirichlet Process (HDP) prior on the state transition distribution of a Hidden Markov Model (HMM) model [20]. By using an HMM to model a video (as a sequence of segments), each discovered hidden state can be naturally interpreted as a scene in the video. The HDP prior allows us to determine the optimal number of states (*i.e.*, scenes) according to the nature of the data. However, real-world videos may be highly noisy and directly using an HDP-HMM model may extract too many scenes with less significant visual characteristics (*e.g.*, spatial changes of objects or addition/removal of a small number of objects). To address this issue, we follow the sticky HDP-HMM to encourage a stronger self-transition of a state [4]. This will result in temporal persistence of states to produce longer and semantically coherent scenes. To further accommodate spatial changes or variations in certain objects, we allow the emission distribution to follow another non-parametric DP that automatically determines the number of mixture components (*i.e.*, sub-scenes) within the same scene. For example, scene 1 in Figure 4 is comprised of two sub-scenes: the first with a clear sky and the second with smoke in the sky.

More specifically, consider a collection of hidden states (*i.e.* scenes in a video), the transition probability of state j to other states is governed by a DP:

$$\mathcal{G}_j = \sum_{l=1}^{\infty} \hat{\pi}_{jl} \delta_{\hat{\phi}_{jl}}, \ \hat{\pi}_j \sim \text{GEM}(\alpha) \tag{3}$$

where $\text{GEM}(\alpha)$ is formed through a stick breaking construction process with parameter α [20], $\hat{\phi}_{jl}$ is drawn from a base distribution $\mathcal{G}_0$, which follows another DP

$$\mathcal{G}_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}, \ \beta \sim \text{GEM}(\gamma), \ \phi_k \sim H \tag{4}$$

Because of the discrete nature of $\mathcal{G}_0$, there can be multiple $\hat{\phi}_{jl}$'s taking an identical value of ϕ_k. Considering the unique set of atoms ϕ_k, we can rewrite $\mathcal{G}_j$ as

$$\mathcal{G}_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}, \ \pi_j \sim \text{DP}(\alpha, \beta), \phi_k \sim H \tag{5}$$

Given the highly dynamic and noisy nature of many real-world surveillance videos, directly applying the standard HDP-HMM model to partition a video may result in many redundant scenes and rapidly switches among them. This is problematic in our setting in which it is critical to infer semantically coherent scenes along with a slower transition

among them. As a result, it is essential to ensure temporal persistence of the discovered scenes [4]. This can be achieved through enhanced self transitions. In particular, the transition probability of the j's state is augmented by

$$\pi_j \sim \mathrm{DP}\left(\alpha+\rho, \frac{\alpha\beta+\rho\delta_j}{\alpha+\rho}\right) \tag{6}$$

This has the effect of increasing the expected probability of staying in the same state.

$$\mathbb{E}[\pi_{jk}|\beta] = \begin{cases} \frac{\alpha\beta_j+\rho}{\alpha+\rho} & \text{if } k=j \\ \frac{\alpha\beta_k}{\alpha+\rho}, & \text{otherwise} \end{cases} \tag{7}$$

To allow certain levels of variations within the same scene and accommodate the highly dynamic nature of a video sequence, we propose to model the emission process using a mixture distribution governed by another non-parametric DP. This design offers three unique advantages. First, it further ensures the temporal persistence of a scene as for a segment with less significant visual differences, it can stay in the same scene by switching to a different mixture component instead of transitioning to another (redundant) scene. Second, it offers a fine-grained partition of the video sequence, which is instrumental to separate abnormal segments (*e.g.*, frames (b) and (e) in Figure 4) from normal ones (*e.g.*, frames (a) and (d) in Figure 4) that share a common background. Last, the number of mixture components is automatically determined by the DP (*e.g.*, scenes 1 & 3 have two mixture components while scene 2 only has one). For the k-th scene, there is an unique stick-breaking distribution $\psi_k \sim \mathrm{GEM}(\tau)$ that defines the weights of the mixture components within the scene. Then, given the scene and mixture component assignment $(z_i = k, s_i = t)$ of a segment $\mathbf{x}_i^+$ in a video, it is drawn from a specific multivariate Gaussian: $\mathcal{N}(\boldsymbol{\mu}_{k,t}, \Sigma_{k,t})$.

Posterior inference of the augmented HDP-HMM model with a DP mixture for emission can be achieved through direct assignment [20] or blocked sampling with an increasing mixing rate [3]. Hyper-parameters can also be inferred by placing a vague prior on them and conduct Gibbs sampling.

The scene and component assignments of BN-SVP induces a pairwise similarity among segments in a video:

$$\begin{cases} S_{i,j} = (\mathbf{x}_i^+)^\top \Sigma_{z_i,s_i}^{-1} \mathbf{x}_j^+ & \text{if } s_i == s_j \wedge z_i == z_j \\ S_{i,j} = 0 & \text{otherwise} \end{cases} \tag{8}$$

It is worth to note that the similarity between two segments $\mathbf{x}_i^+$ and $\mathbf{x}_j^+$ is evaluated using the learned feature representations (through a DNN) instead of the raw features. The induced similarity allows us to define a submodular set function [10, 11] summarized by the follow proposition.

Proposition 1. *Let κ denote the number of unique mixture components across all the discovered states in a bag $\mathcal{B}_{pos}$ and $\mathcal{C} \subset \mathcal{B}_{pos}$ is a subset of $\mathcal{B}_{pos}$ with size κ. Given the BN-SVP induced pairwise similarity defined in (8), the following function is a submodular set function:*

$$F(\mathcal{C}) = \sum_{i \in \mathcal{B}_{pos}} \max_{j \in \mathcal{C}} S_{i,j} \tag{9}$$

Based on the definition of $S_{i,j}$ as shown above, it is straightforward to show that $F(\mathcal{C})$ is a special instance of the location facility function [14], which is submodular. Since each mixture component captures a unique sub-scene, maximization of $F(\mathcal{C})$ can extract a diverse set of segments that best represent the all the scenes (and sub-scenes) in the entire video. By further integrating the margin loss given in (2), we achieve a submodularity diversified MIL loss:

$$\min_{\mathbf{w}, \mathcal{C}^+ \in \mathcal{B}_{pos}, |\mathcal{C}^+| \le \kappa} L(\mathcal{C}^+) - \lambda F(\mathcal{C}^+) \tag{10}$$

where the margin loss is defined over instances in a set $\mathcal{C}^+$ with size no larger than κ:

$$L(\mathcal{C}^+) = \left[1 - \frac{1}{|\mathcal{C}^+|} \sum_{i \in \mathcal{C}^+} f(\mathbf{x}_i^+) + \max_{j \in \mathcal{B}_{neg}} f(\mathbf{x}_j^-)\right]_+ \tag{11}$$

In essence, $\mathcal{C}^+$ includes the set of instances in a positive bag that participate in the model training. The constraint $|\mathcal{C}^+| \le \kappa$ has the effect of excluding some representative segments from the margin loss as these segments are less likely to be abnormal (*e.g.*, with a very low predicted score). Including these segments will increase the margin loss and coefficient λ controls the balance between the margin loss and the diversity among the chosen segments.

3.3. Greedy Submodular Function Optimization

We propose a greedy algorithm for optimizing the submodular function in (9) to ensure efficient model training. The proposed algorithm leverages the special structure of the state and mixture component space resulted from the HDP-HMM partition of the video segments. The performance guarantee of the greedy algorithm is ensured by our theoretical result presented at the end of this section.

Recall that we use s_i to denote the mixture component of segment $\mathbf{x}_i^+$. Let f_s^* denote the maximum score among all the segments assigned to the same mixture component and i_s^* be the index of the corresponding representative segment:

$$i_s^* = \arg\max_{\forall i: s_i = s} f(\mathbf{x}_i^+), \; f_s^* = f(\mathbf{x}_{i_s^*}^+) \tag{12}$$

We construct a representative set $\widehat{\mathcal{C}^+}$ as follows. Let $\widehat{\mathcal{C}^+} = \Phi$ and for each mixture component s, we set

$$\begin{cases} \widehat{\mathcal{C}^+} \leftarrow \widehat{\mathcal{C}^+} \cup \{i_s^*\}, & \text{if } f_s^* \ge \epsilon \\ \widehat{\mathcal{C}^+} \leftarrow \widehat{\mathcal{C}^+}, & \text{otherwise} \end{cases} \tag{13}$$

where ϵ is a threshold to exclude segments with a low prediction score, which plays an equivalent role as constraint $|\mathcal{C}^+| \leq \kappa$ in (10). In our experiments, we use ϵ equal to the output of the segment staying in the 35th percentile among video specific segments so as to avoid skipping any potential abnormal segments. Once a representative set $\widehat{\mathcal{C}^+}$ is constructed, model training can proceed by solving the following MIL loss:

$$\min_{\mathbf{w}} \Big[1 - \frac{1}{|\widehat{\mathcal{C}^+}|} \sum_{i \in \widehat{\mathcal{C}^+}} f(\mathbf{x}_i^+) + \max_{j \in \mathcal{B}_{neg}} f(\mathbf{x}_j^-)\Big]_+ \tag{14}$$

Given the state and mixture component assignment of each segment in a video, the representative set can be quickly constructed by sorting segments within each component according to their predicted scores and choosing the representative segment from each component by comparing its score with the threshold ϵ. Next, we provide a strong theoretical guarantee that the greedy algorithm can ensure the inclusion of a diverse set of segments for model training.

Theorem 1. *The representative set based MIL loss given in (14) is equivalent to the submodularity diversified MIL loss given in (10). Furthermore, using the proposed greedy algorithm to locate the κ representative segments essentially provides a κ-constrained greedy approximation to the maximization of the submodular set function $F(\mathcal{C})$. As a result, the obtained solution is guaranteed to be no worse $(1-e^{-1})$ of the optimal solution.*

The detailed proof is provided in the Appendix.

4. Experiments

We conduct extensive experiments to evaluate the effectiveness of the proposed BN-SVP approach. Through these experiments, we aim to demonstrate: (i) outstanding anomaly detection performance by comparing with competitive top-k, MIL, and other video anomaly detection models, (ii) robustness to outlier and multimodal scenarios, and (iii) deeper insights on the better detection performance through a qualitative study.

4.1. Datasets and Experimental Settings

Our experimentation includes three video datasets of different scales: ShanghaiTech [17], Avenue [16], and UCF-Crime [19]. Table 4 in the Appendix shows how the videos are partitioned into the training/testing sets in each dataset.

- **ShanghaiTech** consists of 437 videos (330 normal and 107 abnormal). In the original setting, all training videos are normal. To fit into our setting, we follow the data split in [27] to assign normal and abnormal videos in both training and testing sets.
- **Avenue** consists of 16 training and 21 testing videos. We perform 80:20 split separately in the abnormal and normal video sets to generate training and testing instances.
- **UCF-Crime** consists of 13 different anomalies with a total of 1900 videos, where 1610 are training videos and 290 are testing videos. In this dataset, frame labels are available only for the testing videos.

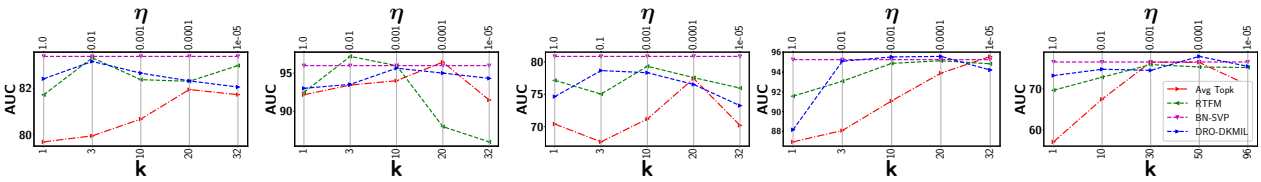
To show the robustness of the proposed approach in the multimodal and outlier scenarios, we also generate the Multimodal and Outlier datasets. Specifically, we create a multimodal scenario by extending the UCF-Crime dataset. For the outlier scenario, we deliberately impose some outliers in the ShanghaiTech dataset. More details of these two datasets are provided in Section 4.3. For evaluation, we report the frame-level receiver operating characteristics (ROC) curve along with the corresponding area under curve (AUC). The AUC score indicates the robustness of the performance at various thresholds.

For Avenue and ShanghaiTech datasets, we extract visual features from the FC7 layer of a pre-trained C3D network [23]. We re-size each video frame to 240×340 pixels and fix the frame rate to 30 fps. We compute the C3D features for every 16-frame video clip. This may yield a different number of clips (each clip having a 2048 dimensional feature vector) depending on the number of frames in each video. Thus, we fit any number of clips to the 32 segments by taking an average of clip features in a specific segment. In case of UCF-Crime, we extract the features using an I3D network [1] by using the pretrained network as described in [25]. For all datasets, we use parallel GCN networks to capture the feature similarity and temporal consistency. The outputs of the parallel branches are combined and passed through a 5-layer LSTM network where each layer has 32 hidden units followed by batch normalization. Finally, an FC layer with sigmoid activation is applied to bring the prediction score to $(0, 1)$. For model training, we use SGD with a learning rate of 0.001 and l_2 regularization with parameter $\lambda = 0.001$. Detailed information about the network architecture is provided in the Appendix.

4.2. Performance Comparison

Comparison with Top-k Models. We first compare the detection performance with two most recent top-k based models, including Robust Temporal Feature Magnitude learning (RTFM) [22] and the DRO based deep kernel MIL (DRO-DKMIL) [18]. We also compare with a standard average top-k model (Avg Topk) as the baseline. Avg Topk uses the rank loss in (2) with the same network architecture as BN-SVP. For RTFM, we get the result by re-running the original implementation for different k values. Similarly, for DRO-DKMIL, we run the original implementation for different η values that control the size of the uncertainty set. The proposed BN-SVP removes the dependency on these highly sensitive parameters through non-parametric modeling. Detailed comparison results are shown in Figure 5.

We have several important observations. First, all the



(a) UCF Crime (b) Shanghaitech (c) Avenue (d) Outlier (e) Multimodal

Figure 5. Performance comparison with top-k ranking models

top-k models are very sensitive to the selection of the k value (or η that defines a soft version of the top-k set). Both RTFM and DRO-DKMIL outperform the standard Avg Topk. DRO-DKMIL achieves relatively more stable performance across all datasets. This may attribute to its conversion of discrete optimization (*i.e.*, choose a specific k) to a continuous optimization problem (*i.e.* choosing η). However, for certain dataset (*e.g.*, Avenue), its performance still varies more than 8%. Second, while for some rare cases that RTFM or DRO-DKMIL achieves the best performance for a specific k or η, they under-perform BN-SVP in most cases. This is mainly due to that these models tend to choose a consecutive set of segments, which limits the model's exposure of other potentially positive segments. This issue has been effectively addressed by BN-SVP, which extracts a diverse set of potentially positive segments through submodular optimization.

Comparison with Other Models. We also make comparison with other existing techniques that do not depend on the k value. Specifically, our comparison study includes the maximum score based MIL model (MMIL) by Sultani et al. [19], attention based deep MIL model proposed by Ilse et al. [7], a dictionary based approach proposed by Lu et al. [16], and an MIL model for soft bags (MILS) proposed by Li & Vasconcelos [13] as common baselines for all datasets. Sultani et al. [19] used the loss function in (1) along with the temporal similarity and consistency as a regularizer. Ilse et al. used a permutation invariant aggregation function to detect the positive instances in the bag, where the function operators are learned using the attention network [7]. Li & Vasconcelos used a large-margin based latent support vector machine model with the goal to correctly classify positive and negative bags [13]. In case of the approach presented by Zhong et al. [27], we directly report the performance from the original paper for the UCF-Crime and ShanghaiTech datasets. This approach involves multiple rounds of alternative optimization between classification and cleaning and may produce unstable performance [22]. Considering its difficulty in the training and replication process, we do not include it in other datasets.

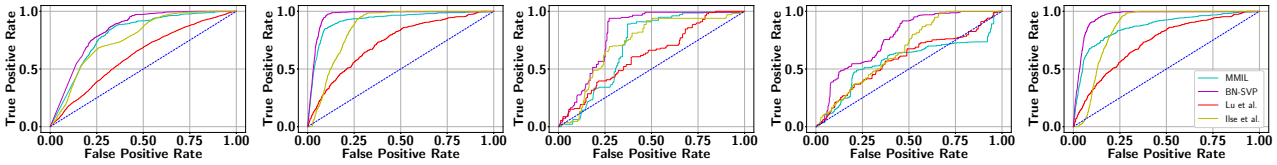
Table 1 reports the AUC scores of BN-SVP along with the results from the comparison models as described above. It can be seen that BN-SVP clearly outperforms other models in all datasets and a large margin (*i.e.*, 6-8$\%$) is achieved on both ShanghaiTech and Avenue datasets. The corresponding ROC curves are shown in Figure 6, which demonstrates a consistent trend. For example, on UCF-Crime, BN-SVP has a more than 10% better True Positive Rate (TPR) compared to MMIL at a False Positive Rate (FPR) of 0.2. Also at varying FPR, BN-SVP consistently outperforms the other competitive baselines, which justifies its outstanding detection capability.

Table 1. Comparison with Other Models

Approach	AUC (%)
UCF-CRIME	
Hasan et al. [5] (C3D)	50.60
Lu et al. [16] (C3D)	65.51
Lu et al. [16] (I3D)	61.98
MMIL [19] (C3D)	75.41
Li & Vasconcelos [13] (I3D)	77.95
Ilse et al. [7] (I3D)	76.52
Zhong et al. [27] (GCN (C3D))	81.08
Zhong et al. [27] (TSN^{RGB})	82.12
Zhong et al. [27] ($\text{TSN}^{\text{OpticalFlow}}$)	78.08
MMIL [19] (I3D)	79.68
BN-SVP (I3D)	**83.39**
SHANGHAITECH	
Lu et al. [16] (C3D)	72.90
Li & Vasconcelos [13] (C3D)	90.40
Zhong et al. [27] (GCN (C3D))	76.44
Zhong et al. [27] ($\text{GCN}(\text{TSN}^{\text{RGB}})$)	84.44
Zhong et al. [27] ($\text{GCN}(\text{TSN}^{\text{Optical Flow}})$)	84.13
Ilse et al. [7] (C3D)	85.78
MMIL [19] (C3D)	92.18
BN-SVP (C3D)	**96.00**
AVENUE	
Binary SVM (C3D)	69.11
Lu et al. [16] (C3D)	62.14
Li & Vasconcelos [13] (C3D)	72.23
Ilse et al. [7] (C3D)	72.39
MMIL [19]	70.40
BN-SVP (C3D)	**80.87**

4.3. Detecting Multimodal and Outlier Segments

Multimodal Detection. The original UCF-Crime dataset does not explicitly consider a multimodal scenario. Even though the real-world surveillance videos may indeed contain those cases (which is evidenced by the superior performance of the BN-SVP model), it is hard to identify actual videos for this specific information. In UCF-Crime dataset, different types of anomalies are present. This allows us to explicitly create multimodal scenarios by combining multiple abnormal videos from different activity types. To this

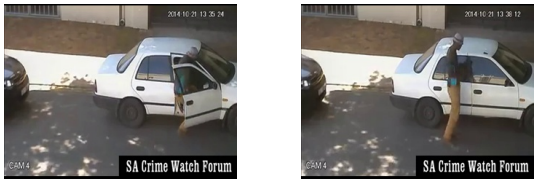


(a) UCF-Crime (b) ShanghaiTech (c) Avenue (d) Multimodal (UCF) (e) Outlier (ShanghaiTech)

Figure 6. ROC curves on three video datasets (a)-(c), multimodal (d) and outlier (e)

Table 2. AUC Scores on Multimodal and Outlier Detection

Approach	**AUC** (%)	
	Multimodal	***Outlier***
Lu et al. [16] (C3D)	58.67	72.90
Li & Vasconcelos [13] (C3D)	70.96	90.95
Ilse et al. [7] (C3D)	66.85	85.65
MMIL [19]	57.08	86.47
BN-SVP	**76.53**	**95.27**



(a) Frame 1 (b) Frame 2

Figure 7. Frames from UCF-Crime Stealing019; (a) Correct BN-SVP, Avg Topk, (b) Correct BN-SVP, Incorrect Avg Topk

end, we randomly select three activity types and form an abnormal bag by concatenating three abnormal videos, one video per activity type. The training bags are constructed using the training dataset whereas testing bags are constructed using the testing dataset. In total, we construct 50 abnormal and 50 normal training bags. In the testing set, there are 10 normal and 10 abnormal videos. Table 2 shows the AUC scores and corresponding ROC curve is shown in the Figure 6 (d). BN-SVP achieves a more superior performance compared to other baselines. Furthermore, BN-SVP stays consistently on the top in the ROC curve justifying the effectiveness of the approach toward the multimodal scenario. As an example, at FPR $= 0.1$, BN-SVP is at least 20% better than other approaches on TPR.

Outlier Detection. To assess the robustness on outlier detection, we extend the ShanghaiTech dataset with outliers. Specifically, we randomly select 120 segments from abnormal videos and replace their features with points drawn from a standard multivariate Gaussian distribution. As shown in Table 2, MMIL suffers heavily by the outliers compared to the proposed BN-SVP. This is because, it is likely to have an outlier prediction as the maximum prediction score from the abnormal video. As a result, the overall optimization process may be heavily influenced by outliers.

4.4. Qualitative Analysis

To show the effectiveness of extracting a diverse set of segments for model training, we present illustrative sample frames in a stealing video from UCF-Crime, where BN-SVP correctly identifies all abnormal frames and a top-k approach (*e.g.*, Avg Topk) misses some of them. In Figure 7, both frames are of abnormal types and but they occur in two distinct time intervals within the video. The first frame is more obvious for a stealing event. Consequently, both the proposed BN-SVP and Avg Topk are able to correctly identify it. In contrast, the second frame is less obvious for a stealing activity. Nevertheless, it is still chosen by BN-SVP due to its diverse coverage of potentially abnormal frames during the training process. On the other hand, Avg Topk only focuses on frames with high prediction scores that are usually co-located in the same time interval. This will narrow the scope of the model being exposed to other abnormal frames. Therefore, Avg Topk is not able to correctly predict the second frame and falsely classify it as normal. More qualitative analysis that demonstrates the robustness of the proposed approach on multimodal and outlier scenarios is provided in the Appendix along with an ablation study for the prediction score threshold ϵ defined in (13).

5. Conclusion

In this paper, we propose a novel Bayesian nonparametric submodularity diversified MIL model for robust video anomaly detection in practical settings that involve outlier and multimodal scenarios. By integrating submodular optimization with the minimization of an MIL loss, the proposed approach identifies a diverse set of segments to ensure comprehensive coverage of all potential positive segments for effective model training. The Bayesian nonparametric construction of the submodular set function automatically determines the upper bound on the size of the diverse set, which serves as a key constraint for minimizing the submodularity diversified MIL loss function. The resulting state-component structure also leads to a greedy submodular optimization algorithm to support efficient model training. The effectiveness of the proposed approach is demonstrated through the state-of-the-art robust anomaly detection performance on real-world surveillance videos with noisy and multimodal scenarios.

Acknowledgement

This research was supported in part by an NSF IIS award IIS-1814450 and an ONR award N00014-18-1-2875. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agency.

References

[1] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 4724–4733, 2017. 6

[2] Y. Cong, J. Yuan, and J. Liu. Sparse reconstruction cost for abnormal event detection. In *CVPR 2011*, pages 3449–3456, 2011. 1

[3] Yanbo Fan, Siwei Lyu, Yiming Ying, and Baogang Hu. Learning with average top-k loss. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *NIPS*, pages 497–505. Curran Associates, Inc., 2017. 5

[4] Emily B. Fox, Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky. A sticky hdp-hmm with application to speaker diarization. *The Annals of Applied Statistics*, 5:1020–1056, 2011. 4, 5

[5] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis. Learning temporal regularity in video sequences. In *CVPR*, pages 733–742, 2016. 7

[6] Chengkun He, Jie Shao, and Jiayu Sun. An anomaly-introduced learning method for abnormal event detection. *Multimedia Tools Appl.*, 77(22):29573–29588, Nov. 2018. 3

[7] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *ICML*, 2018. 7, 8

[8] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17, 2017. 13

[9] L. Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *CVPR*, pages 1446–1453, 2009. 3

[10] Andreas Krause. Optimizing sensing: Theory and applications. Technical report, 2008. 5

[11] Andreas Krause, H. Brendan McMahan, Carlos Guestrin, and Anupam Gupta. Robust submodular observation selection. *Journal of Machine Learning Research*, 9(93):2761–2801, 2008. 5

[12] Nannan Li, Xinyu Wu, Huiwen Guo, Dan Xu, Yongsheng Ou, and Yen-Lun Chen. Anomaly detection in video surveillance via gaussian process. *Int. J. Pattern Recognit. Artif. Intell.*, 29:1555011:1–1555011:25, 2015. 3

[13] W. Li and N. Vasconcelos. Multiple instance learning for soft bags via top instances. In *CVPR, 2015*, pages 4277–4285, 2015. 7, 8

[14] Hui Lin and Jeff Bilmes. How to select a good training-data subset for transcription: Submodular active selection for sequences. Technical report, WASHINGTON UNIV SEATTLE DEPT OF ELECTRICAL ENGINEERING, 2009. 5

[15] Wen Liu, Weixin Luo, Zhengxin Li, Peilin Zhao, and Shenghua Gao. Margin learning embedded prediction for video anomaly detection with a few anomalies. In *IJCAI*, 2019. 1

[16] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In *ICCV*, pages 2720–2727, 2013. 2, 3, 6, 7, 8

[17] W. Luo, W. Liu, and S. Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *ICCV*, pages 341–349, 2017. 6

[18] Hitesh Sapkota, Yiming Ying, Feng Chen, and Qi Yu. Distributionally robust optimization for deep kernel multiple instance learning. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2188–2196. PMLR, 13–15 Apr 2021. 2, 3, 6

[19] W. Sultani, C. Chen, and M. Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, pages 6479–6488, 2018. 1, 3, 6, 7, 8

[20] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476):1566–1581, 2006. 2, 4, 5

[21] Kai Tian, Shuigeng Zhou, Jianping Fan, and Jihong Guan. Learning competitive and discriminative reconstructions for anomaly detection. In *AAAI*, pages 5167–5174. AAAI Press, 2019. 1

[22] 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 (ICCV)*, pages 4975–4986, October 2021. 2, 3, 6, 7

[23] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, ICCV '15, page 4489–4497, USA, 2015. IEEE Computer Society. 6

[24] Hung Vu, T. Nguyen, Trung Le, Wei Luo, and Dinh Q. Phung. Robust anomaly detection in videos using multilevel representations. In *AAAI*, 2019. 1

[25] X. Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *2018*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7794–7803, 2018. 6

[26] Huan Yang, Baoyuan Wang, Stephen Lin, David Wipf, Minyi Guo, and Baining Guo. Unsupervised extraction of video highlights via robust recurrent auto-encoders. In *ICCV*, ICCV '15, page 4633–4641, USA, 2015. IEEE Computer Society. 1

[27] J. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *CVPR*, pages 1237–1246, 2019. 1, 6, 7