Appendix

Organization of Appendix. In this appendix, we first summarize the major notations used in the paper in Table 3 of Appendix A. We then provide the detailed proof of Theorem 1 in Appendix B. Next, we provide additional experimental results in Appendix C. Finally, the link to the source code is provided in Appendix D.

A. Summary of Notations

Table 3 summarizes all the major symbols along with their descriptions.

Table 3. Symbols with Descriptions

Notation	Description
$\mathcal{B}_{pos}$	Positive bag (video)
$\mathcal{B}_{neg}$	Negative bag (video)
n	Number of segments in each bag
$\mathbf{x}_i^+$	Segment in a positive bag
$\mathbf{x}_{[i]}^+$	i^{th} largest prediction segment in a positive bag
$\mathbf{x}_j^-$	Segment in a negative bag
M	Feature dimension of each video segment
$\mathbf{w}$	Network parameters
b	Network bias
k	Number of segments considered in the top-k formulation
η	Learning rate
$\mathcal{C}^+$	Set of instances from positive bag involve in model training
$\mathcal{G}_0$	Base distribution in DP
γ	Concentration parameter for the distribution $\mathcal{G}_0$
β_k	Weight associated with the k^{th} atom
ϕ_k	Atom k drawn from the distribution H
$\mathcal{G}_j$	Transition probability distribution of j^{th} state
$\hat{\pi}_{jl}$	Stick breaking weight associated with l^{th} atom in j^{th} group
α	Concentration parameter for $\hat{\pi}_j$
ϕ_{jl}	l^{th} atom corresponding j^{th} group
β_k	Stick breaking weight corresponding to atom ϕ_k
γ	Concentration parameter for β_k
ρ	Parameter defining the self transitioning
z_i	Scene assignment for the i^{th} segment in a video
s_i	Mixture component assignment for the i^{th} segment in a video
$\mathcal{N}$	Multivariate Gaussian distribution
$\boldsymbol{\mu}_{k,t}$	Mean of the k^{th} state, t^{th} mixture component
$\Sigma_{k,t}$	Covariance of k^{th} state, t^{th} mixture component
$S_{i,j}$	Pairwise similarity between i^{th} and j^{th} segments
$F(\mathcal{C})$	Submodular set function
f_s^*	Maximum output score among segments assigned to the same cluster
i_s^*	Index of the representative segment
$\widehat{\mathcal{C}^+}$	Representative set constructed using the greedy algorithm
ϵ	Threshold to exclude segments with low prediction score from the representative set
κ	Upper bound of number of representative segments

B. Proof of Theorem 1

In this section, we provide the detailed proof of Theorem 1. We first show that the representative set based MIL loss given by (14) is equivalent to the submodularity diversified MIL loss given by Equation (10) with a specific λ to balance the MIL loss and the diversity of the set. We then show that greedy algorithm to locate the κ representative segments provides

a κ-constrained greedy approximation to the maximization of the submodular set function $F(C)$ with the solution to be no worse than $(1 - e^{-1})$ of the optimal solution.

Proof of representative set based MIL loss in (14) is a special case of the submodular diversified MIL loss in (10). We first present a lemma, which is used in the proof.

Lemma 2. *Assume that $\widetilde{\mathcal{C}^+}$ with size κ is a solution that maximizes $F(\mathcal{C})$ in (9). Then, $\widetilde{\mathcal{C}^+}$ should contain one segment from each mixture component (*i.e. *sub-scene).*

Proof. The lemma can be proved by following the definition of the BN-SVP induced pairwise similarity between segments given by (8) and then use proof by contradiction. Assume that at least two segments, say $\mathbf{x}_i^{(t)}, \mathbf{x}_j^{(t)}$, are chosen from the same component t. Then, there will be at least one component, say t', where no segments are chosen by $\widetilde{\mathcal{C}^+}$. Given the definition of $F(\mathcal{C})$ in (9), for each segment in t, either $\mathbf{x}_i^{(t)}$ or $\mathbf{x}_j^{(t)}$ could be used to compute the pairwise similarity based on their closeness to that segment. Since the cohesiveness of each component is guaranteed through the BN-SVP process, both $\mathbf{x}_i^{(t)}$ and $\mathbf{x}_j^{(t)}$ should be close to the mean of their assigned Gaussian component $\mathcal{N}(\mathbf{x}_t, \Sigma_t)$ to ensure a high likelihood optimized by HDP-HMM. Due to triangle inequality, $\mathbf{x}_i^{(t)}$ and $\mathbf{x}_j^{(t)}$ should be close to each other. As a result, we can assume that $\mathbf{x}_i^{(t)}$ is always chosen to evaluate the pairwise similarity $S_{i,p}$ with each segment $\mathbf{x}_p^{(t)}$ in component t. Next, we replace $\mathbf{x}_j^{(t)}$ with another segment $\mathbf{x}_j^{(t')}$ from component t' to construct another solution set $\overline{\mathcal{C}^+}$. Since $\mathbf{x}_j^{(t')}$ has positive similarity with each segment in t' and the pairwise similarity between $\mathbf{x}_j^{(t)}$ and all segments in t' is all zero, we have $F(\overline{\mathcal{C}^+}) > F(\widetilde{\mathcal{C}^+})$, which contradicts the assumption that $\widetilde{\mathcal{C}^+}$ maximizes $F(\mathcal{C})$. □

Since the representative set $\widehat{\mathcal{C}^+}$ is constructed by choosing one segment from each mixture component, it satisfies the necessary condition to be an optimizer of $F(\mathcal{C})$ specified in the above lemma. However, choosing a set of segments with the maximum diversity is not the primary goal and the overall objective function (10) includes both the MIL loss and the diversity, which are balanced through λ. Due to the lack of instance-level labels, choosing a λ that optimally balances the MIL loss and the set diversity is challenging. We argue that construction $\widehat{\mathcal{C}^+}$ essentially offers alternative way to set a specific λ to balance these two terms. First, since the constraint $|\mathcal{C}^+| \leq \kappa$ allows the set to contain less than κ segment, $\widehat{\mathcal{C}^+}$ excludes those segments with low prediction scores. This can be viewed as setting a λ to increase $-F(\mathcal{C}^+)$ while decreasing the MIL loss $L(\mathcal{C}^+)$. Similarly, instead of choosing the instance with the largest pairwise similarity with all other segments in the same component, we choose a segment with the highest prediction score. Again, this can be viewed as further reducing the λ to give more preference to the MIL loss as such segments can further reduce the training MIL loss. Thus, instead of directly setting λ, which is highly challenging, $\widehat{\mathcal{C}^+}$ is constructed by leveraging both the mixture component assignments and the prediction scores of the segments. This is equivalent to implicitly setting a λ to balance the MIL loss and the diversity of the representative set $\widehat{\mathcal{C}^+}$, which completes the proof of the equivalence of these two objective functions.

Proof of the optimality of the greedy algorithm. We first reformulate (10) as a minimization problem $\min_{\mathbf{w}} g(\mathbf{w})$ with $g(\mathbf{w})$ defined as

$$g(\mathbf{w}) \triangleq \min_{\mathcal{C}^+ \subseteq \mathcal{B}_{pos}, |\mathcal{C}^+| \leq \kappa} L(\mathcal{B}_{pos}, \mathcal{B}_{neg}) - \lambda F(\mathcal{C}^+) \tag{15}$$

The above optimization involves finding a subset $\mathcal{C}^+ \subseteq \mathcal{B}_{pos}$ that maximizes $F(\mathcal{C}^+)$. This requires enumerating over all $\binom{n}{\kappa}$ possible subsets, which is expensive when there are large number of segments in a given video. Defining the discrete objective function $G_{\mathbf{w}}$ where

$$G_{\mathbf{w}}(\mathcal{C}^+) \triangleq L(\mathcal{B}_{pos}, \mathcal{B}_{neg}) - \lambda F(\mathcal{C}^+) \tag{16}$$

Since $-G_{\mathbf{w}}(\mathcal{C}^+)$ is monotone non-decreasing submodular, a fast greedy procedure can be used to approximately optimize $G_{\mathbf{w}}(\mathcal{C}^+)$. A typical greedy procedure involves evaluating the similarity between each pair of segments in a video and then choose the segments with the largest overall similarity with the all other segments. We make two important adjustments of this standard greedy process. First, our non-parametric HDP-HMM process follows the clustering based heuristic (Lin and Bilmes 2018) by choosing one segment from each cluster, which avoids evaluating each candidate segment in the video. Different from (Lin and Bilmes 2018), which chooses the data point that is closest to the cluster centroid, we choose the one with the highest output score. Second, our similarity evaluation takes a linear complexity with respect to the bag size by leveraging the temporal neighborhood of the segments. By leveraging the above greedy procedure, we can show that the

obtained approximate solution is guaranteed to be no worse than $(1 - e^{-1})$ of the optimal solution according to the standard result from (Nemhauser et al. 1978), which completes the proof of the second part.

Table 4. Video Level Distribution on Different Datasets

Split	ShanghaiTech		UCF-Crime		UCF-Crime Multimodal		Avenue	
	Normal	*Abnormal*	*Normal*	*Abormal*	*Normal*	*Abnormal*	*Normal*	*Abnormal*
Train	175	63	810	800	150	150	13	17
Test	155	44	150	140	30	30	3	4

C. Additional Experimental Results

In this section, we first show the detailed network architecture used in our training process. Next, we provide the ablation study demonstrating the impact of hyperparameter ϵ used in our experimentation. Finally, we provide some additional qualitative analysis justifying the effectiveness of the proposed approach. Further we also show effectiveness of the HDP-HMM technique to discover subscenes of different types in a video through qualitative analysis.

C.1. Network Architecture

First, we pass the pre-trained features through the two parallel GCN branches. The upper branch captures the feature similarity between segments and the lower one captures the temporal consistency between segments such that nearby segments will provide similar predictions. The output of the parallel branches are combined and passed through the 5 LSTM layers with 32 hidden units on each. Next, the output is passed through the BatchNorm. Finally, FCN is connected with sigmoid activation to get the final prediction score.

GCN Architecure. Next, we explain the GCN architecture in detail. Let $\mathbf{A}$ is the $n \times n$ dimensional adjacency matrix where the (i, j) entry in the matrix indicates the similarity between segment i and j. Mathematically,

$$\mathbf{A}(i, j) = k(\mathbf{x}_i, \mathbf{x}_j) \tag{17}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ be the D-dimensional representation for i^{th} and j^{th} segments respectively. It should be noted that for the feature similarity branch, we use the RBF kernel with the following form

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(\frac{|\mathbf{x}_i - \mathbf{x}_j|^2}{-2l^2}) \tag{18}$$

In case of temporal consistency branch, we use the following form between i^{th} and j^{th} segment:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-|i - j|) \tag{19}$$

This drives the temporally nearby segments to have a similar score. Based on the adjacency matrix, following Kipf and Welling [8], the graph-Laplacian with the renormalization trick can be written as

$$\hat{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}(\mathbf{A} + \mathbf{I_n})\mathbf{D}^{-\frac{1}{2}} \tag{20}$$

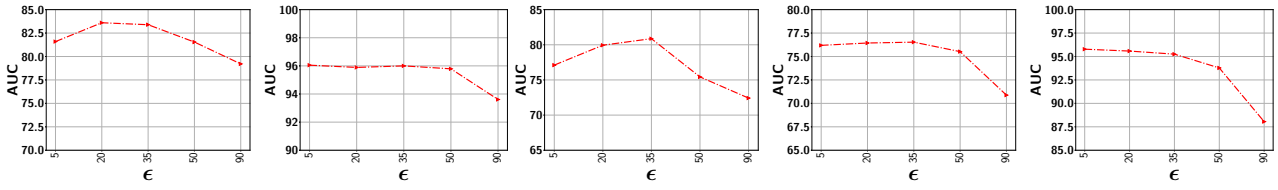
In the above equation $\mathbf{D}_{(i,i)} = \sum_j \{\mathbf{A} + \mathbf{I_n}\}_{(i,j)}$ is the corresponding degree matrix. The output of the feature similarity graph is computed as:

$$\mathbf{H} = \hat{\mathbf{A}}\mathbf{X}\mathbf{W} \tag{21}$$

where $\mathbf{W} \in \mathcal{R}^{D \times M}$ is a trainable parameter matrix and $\mathbf{X} \in \mathcal{R}^{n \times D}$ is the video specific features.

C.2. Ablation Study

Impact of ϵ. In this subsection, we show the the impact of the error threshold ϵ on the model performance. It is worth mentioning that ϵ indicates the percentile we used to determine the threshold so as to exclude the clusters with potentially all normal segments with a high probability. For example, $\epsilon = 0.1$ indicates that we first determine the output score corresponding to the segment that lies in the 10^{th} percentile based on scores of all segments sorted in the non-decreasing order. This

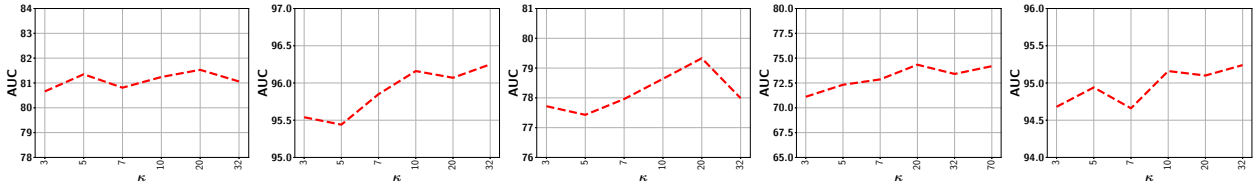


(a) UCF-Crime (b) ShanghaiTech (c) Avenue (d) Multimodal (UCF) (e) Outlier (ShanghaiTech)

Figure 8. Performance variation with respect to ϵ

selected score is used as the threshold. Next, all representative segments with a predicted score below this output threshold are discarded from the representative set $\widehat{\mathcal{C}^+}$. Figure 8 show the performance variation with respect to the different ϵ's for five different datasets. As can be seen, for a relatively lower ϵ value (*i.e.*, 20-35%), the performance is fairly stable for all datasets. This is because, with a low ϵ, the model rejects the segments from a given video with a sufficiently low output score. This way, the chance of including normal segments from abnormal videos is minimized. Further lowering ϵ may include a good number of normal segments, making the model mis-identify other similar normal segments as anomalies. On the other hand, choosing a very high ϵ results in the drop in performance. In this case, some potentially abnormal segments may be missed in the loss function and therefore, the model may have less exposure to different types of abnormal frames resulting in the degradation of performance. In sum, as long as we stay in the relatively low range when choosing ϵ (*e.g.*, 20-35% gives very stable results), we can get the stable (and nearly optimal) performance.

Impact of the constraint $|C| \leq \kappa$. It should be noted that in our approach κ only provides an upper bound on the selected segments and the actual number is determined by the non-parametric model along with the prediction threshold ϵ. This addresses the fundamental issue in the top-k models, in which a fixed k has to be set for all videos. Figures 9 shows that a stable performance can be achieved for a wide range of κ values as long as it is not set too small that may exclude some representative abnormal segments.



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

Figure 9. Performance variation with respect to κ

Impact of λ. We would like to emphasize that BN-SVP does not require to directly set λ, which is highly challenging. By leveraging the prediction score of instances and their mixture assignments, BN-SVP implicitly sets λ to balance MIL loss and the diversity of set $\widehat{\mathcal{C}^+}$. Specifically because of the constraint $|\mathcal{C}^+| \leq \kappa$, we ensure that the set contains no more than κ segments. It excludes segments with a low prediction score, which has the effect of decreasing λ to reduce the MIL loss. Similarly, instead of choosing the instance with the largest pairwise similarity with all other instances within the same mixture assignment, it chooses the instance with the highest prediction score. This can also be viewed as choosing a smaller λ to reduce the MIL loss.

Table 5. Performance (AUROC) with and without augmentation

Dataset	UCF-Crime	Avenue	Multimodal	ShanghaiTech	Outlier
w augmentation	83.39	80.87	76.53	96.00	95.27
w/o augmentation	80.56	76.71	63.23	94.99	94.52

Impact of Augmentation. We compare the performance (AUROC) with augmentation ($\rho = 1$) and without augmentation ($\rho = 0$). Table 5 shows the result for different datasets. As can be seen, BN-SVP consistently performs better on all datasets than w/o augmentation. Without augmentation, the approach transitions from one state to another state quickly for small visual changes and may not be able to keep the temporal persistence when discovering the scenes and therefore the

(a) Frame 1 (b) Frame 2

Figure 10. Example frames from UCF-Crime Stealing019; (a) Correct BN-SVP, MMIL, (b) Correct BN-SVP, incorrect MMIL

(a) Normal Frame (Component 0) (b) Abnormal Frame (Component 1) (c) Abnormal Frame (Component 2) (d) Normal Frame (Component 3)

Figure 11. Example frames from the discovered mixture components

performance is lower. We have also shown the significance of using augmentation via a qualitative analysis in Appendix C.4.

C.3. Additional Qualitative Analysis

To show the effectiveness of the proposed approach to handle multimodality, we compare BN-SVP with MMIL using some illustrative examples. Figure 10 shows two frames from the TEST06 video in Avenue with different anomaly types. In the first anomaly type, the object is thrown and in the second, a person is walking in the wrong lane. As the first anomaly is more obvious, both BN-SVP and MMIL are able to correctly predict it as abnormal. For the second one, our proposed approach correctly detects it as abnormal while MMIL fails to do that. Due to the submodular diversified loss, BN-SVP is more likely to include even less obvious frames (*e.g.*, Figure 10 (b)) during the training process and as a result the approach can make a correct prediction. On the other hand, MMIL picks the one with maximum score and therefore more likely to miss those less obvious ones during training process resulting in the mis-identification of similar frames as normal.

C.4. Effectiveness of Bayesian Non-Parametric Video Partition

In this section, we present representative frames from the mixture components (*i.e.* sub-scenes) discovered by the proposed Bayesian non-parametric video partition process. The purpose is to demonstrate that semantically coherent segments are automatically grouped into the same mixture components by the proposed BN-SVP. This significantly facilitates the optimization of the submodular function to choose a diverse set of segments and allow some of the most representative segments to participate in the MIL loss. Figure 11 shows frames randomly selected from different mixture components for video 01_0162 from the ShanghaiTech dataset. As shown in Figure 11 (a), the frame does not contain any person and its associated component (*i.e.*, Component 0) mostly consists of background segments (which are predicted as normal by the model). In Figure 11 (b), there are multiple people in the frame. Furthermore, someone is riding a bike in a wrong lane while a second person is pointing to another group of people. This frame is assigned to a newly created component (*i.e.*, Component 1) since it looks quite different from the previous frames. Also, given the abnormal behavior in the frame, the model predicts it as an anomaly. Next shown in Figure 11 (c), as the bike starts to vanish from the camera frame, it looks different from (b) and therefore the model assigns it to a new Component 2. Although (b) and (c) are both of abnormal types, the latter is much less obvious than the former. Given their distinctions, they have been assigned to different mixture components so both of them could be chosen when optimizing the submodular function to participate in model training. Finally, for Figure 11 (d), the bike completely disappears from the frame and only a group people walking normally. So, it is assigned to Component 3 and the model predicts its as normal.

D. Link to Source Code

For the source code, please click the following link: https://github.com/ritmininglab/BN-SVP.