Distributionally Robust Optimization for Deep Kernel Multiple Instance Learning

Hitesh Sapkota[1] **Yiming Ying** [2] **Feng Chen** [3] **Qi Yu** [1*]
Rochester Institute of Technology[1] University at Albany, SUNY [2] University of Texas at Dallas[3]

Abstract

Multiple Instance Learning (MIL) provides a promising solution to many real-world problems, where labels are only available at the bag level but missing for instances due to a high labeling cost. As a powerful Bayesian non-parametric model, Gaussian Processes (GP) have been extended from classical supervised learning to MIL settings, aiming to identify the most likely positive (or least negative) instance from a positive (or negative) bag using only the bag-level labels. However, solely focusing on a single instance in a bag makes the model less robust to outliers or multi-modal scenarios, where a single bag contains a diverse set of positive instances. We propose a general GP mixture framework that simultaneously considers multiple instances through a latent mixture model. By adding a top-k constraint, the framework is equivalent to choosing the top-k most positive instances, making it more robust to outliers and multimodal scenarios. We further introduce a Distributionally Robust Optimization (DRO) constraint that removes the limitation of specifying a fix k value. To ensure the prediction power over high-dimensional data (e.g., videos and images) that are common in MIL, we augment the GP kernel with fixed basis functions by using a deep neural network to learn adaptive basis functions so that the covariance structure of high-dimensional data can be accurately captured. Experiments are conducted on highly challenging real-world video anomaly detection tasks to demonstrate the effectiveness of the proposed model.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. * Corresponding author

1 Introduction

In Multiple Instance Learning (MIL), there is a collection of *positive* and *negative* bags where each bag consists of several instances. A bag is considered to be positive if at least one of the instances is positive and negative if none of the instances are positive [1]. Among many other useful applications, such as text classification and protein identification, MIL offers a particularly powerful tool to some important computer vision tasks, such as video anomaly detection, where the models have to solely rely on video level labels due to the lack of expensive frame-level labels [2, 3].

Various approaches have been developed to tackle the MIL problem by treating it as a missing-label problem [4, 5]. Those classical MIL techniques focus on the most positive instance (often referred to as the *witness*), instead of simultaneously considering multiple instances from a positive bag. In particular, the most positive instance is the one mainly responsible for determining the label of a bag [6]. For instance, in SVM based techniques [7], they maximize the margin of the instance with the most positive confidence w.r.t. the current model. Different from other works, a graph-based approach is developed to capture the interactions between instances within a bag and thereby using the information of multiple instances [8].

For many MIL tasks such as video anomaly detection, it is important to capture the interactions among frames in a video to correctly identify the abnormal frames given the temporal and spatial relationship naturally embedded in the data [6]. Further, due to the lack of instance-level labels, the model prediction may be much more uncertain and the uncertainty information is essential for many critical domains (e.g., security surveillance) [9]. Gaussian processes (GP) offer a natural way to capture the interactions among the instances through its covariance function. The Bayesian nature of GP outputs the predictive uncertainty in a principled way. In addition, as a non-parametric model, GP allows the modeling power to scale well with the increase in the dataset. By leveraging these

modeling advantages, a number of GP based MIL models have been developed, where maximum score from positive and negative bags are considered in the for model training [6, 9].

However, there are two key limitations with using a maximum score. First, the presence of noisy outliers may significantly impact the overall performance. This is because the defined objective function solely focuses on an individual instance with the highest score from positive and negative bags. Second, if a multimodal situation (e.g., multiple types of abnormal events in a single video) presents, maximum score based approaches may only detect one type of positive instances due to its inability to consider multiple instances from a single bag in the training process.

To address these limitations, we propose a general GP mixture framework that assigns a non-zero probability to each instance in a bag through a latent mixture model. By adding a top-k constraint, it is equivalent to choosing the top-k most positive instances in a bag, making it more robust to outliers and multimodal scenarios. Most importantly, we further integrate a Distributionally Robust Optimization (DRO) constraint that relaxes the limitation of specifying a fixed k value. By combining DRO with a Bayesian non-parametric GP, *this is the first work that develops a Bayesian DRO model for MIL*. To ensure the prediction power over high-dimensional data that are common in MIL problems, we augment the GP kernel with fixed basis functions by using a deep neural network to perform deep kernel (DK) learning [10]. As a result, it learns adaptive basis functions so that the covariance structure of high-dimensional input data can be accurately captured. Finally, different components of the proposed DRO-DKMIL model are jointly optimized through stochastic variational inference (SVI) that leverages local kernel interpolation and structure exploiting algebra [11] to conduct end-to-end model training, ensuring good efficiency and scalability. In summary, our key contribution is fourfold:

- a general GP mixture framework for MIL that gives flexibility for each instance to take non-zero membership probability in each bag,
- a novel Bayesian DRO MIL model that ensures the participation of multiple instances from each bag in model training, making the prediction robust to outlier and multimodal scenarios,
- the first approximate inference algorithm to train the new Bayesian DRO model in MIL setting,
- state-of-the-art prediction performance that outperforms all existing competitive MIL models.

Experiments are conducted on three challenging real-world video anomaly detection datasets with varied scales: UCF-Crime [2], ShanghaiTech [12], and Avenue [13]. Results show that DRO-DKMIL achieves best performance in all cases.

2 Related Work

In this section, we discuss existing work related to multiple instance learning, distributionally robust optimization, and deep kernel learning.

Multiple Instance Learning (MIL). SVM based approaches have been used to directly maximize the margin of the instance with the most positive confidence [7]. Similarly, a boosting based method has been proposed that treats each instance in a bag independently [14]. Later, graph based techniques have been employed to capture the relations among instances in each bag [8]. More recently, a permutation invariant aggregation function is used to detect the positive instances in the bag, where the function operators are learned using an attention network [15]. MIL has also been investigated under the Bayesian setting. For example, Gaussian Processes along with a maximum score based bag-level likelihood have been used for MIL [6]. Novel variational inference methods have also been developed for fast posterior inference in Bayesian MIL [9]. However, most of the existing MIL models focus on making bag level predictions. Therefore, the model aims to identify the most positive instance in a bag. As discussed earlier, these methods are sensitive to outliers and less effective to handle multimodal scenarios, which we aim to address in this paper.

Distributionally Robust Optimization (DRO). DRO has been employed in supervised learning to assign different weights to different losses so as to maximize the overall weighted loss over an uncertainty set for the distributional variable [16, 17]. Depending on how the uncertainty set is defined, the DRO-based loss reduces to different types of widely known loss functions. For example, by restricting the distribution of the distributional variable within a certain ball with a center given by the uniform distribution, DRO-based loss becomes variance regularized loss [18]. Similarly, by making the distributional variable take any value between 0 to 1, the corresponding DRO-based loss becomes a maximal loss and the top-k loss when further restricting the distributional variable value between 0 and $\frac{1}{k}$ [19]. By leveraging this important flexibility, we integrate DRO to constrain the parameter that governs the probability of each frame being a positive instance in the proposed GP mixture model. To our best knowledge, this is the first work that introduces DRO into MIL, leading to a DRO based GP mixture model that provides robust MIL predictions.

Deep Kernel Learning (DKL). DKL provides a powerful learning paradigm by combining the non-

parametric flexibility of kernel methods (e.g., GP) and representation learning ability of deep neural networks. State-of-the-art performance has been demonstrated over multiple supervised learning tasks [10, 11]. One of the key challenges comes from the computation bottleneck of GP which can work only for a few thousand data points [20]. Such issue has been alleviated through structure exploiting techniques [21, 22, 23], local kernel interpolation [24], and other advances in this field. Building upon these efforts, we develop a stochastic variational inference (SVI) algorithm to conduct end-to-end model training to ensure good efficiency and scalability under the MIL setting.

3 DRO Deep Kernel Multiple Instance Learning (DRO-DKMIL)

We consider that each bag has a fixed number of instances. For a positive bag, there is at least one positive instance whereas for a negative bag all instances are normal. Let $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ be a set training instances. Each $\mathbf{x}_n \in \mathcal{R}^D$ is a D-dimensional feature vector associated with bag $b \in [1, B]$ with corresponding label t_b indicating its bag type, where $+1$ denotes positive and -1 otherwise. Further, consider $\mathbf{y} = \{y_1, ..., y_B\}$ be the set of predicted labels. Table 4 in the Appendix summarizes the main symbols.

3.1 DRO-GP MIL

For most MIL problems, it is essential to capture the interactions among instances in the same bag (e.g. frames in a video) as the spatially and/or temporally close instances usually belong to the same event, which should be assigned the same labels. Further, capturing the uncertainty associated with each instance is crucial in MIL tasks such as anomaly detection from surveillance videos. Gaussian Processes (GP) naturally capture the interactions among instances through its covariance function and its non-parametric flexibility allows the modeling power to scale well with the increase of data. Being a Bayesian model, GP also directly outputs the predictive distribution that quantifies the prediction uncertainty in a principled way.

We propose a GP based mixture framework to address the limitation of existing models as discussed earlier. By integrating GP with a latent mixture model, the proposed framework assigns a non-zero membership probability for each instance present in a bag resulting in robustness to the outlier and multimodal scenarios.

We start by defining the bag level likelihood:

$$p(y_b|\mathbf{f}_b, \mathbf{z}_b) = \prod_{i=1}^{n} \left\{ \frac{1}{1+\exp(-t_b f_{bi})} \right\}^{z_{bi}} \tag{1}$$

$$p(\mathbf{z}_b|\boldsymbol{\pi}_b) = \prod_{i=1}^{n} \pi_{bi}^{z_{bi}}, \pi_{bi} \geq 0, \sum_{i}^{n} \pi_{bi} = 1 \tag{2}$$

where $\mathbf{z}_b$ is an indicator variable drawn from a multinomial distribution parameterized by $\boldsymbol{\pi}_b, \forall b \in [1, B]$. For a negative bag with $t_b = -1$, the model is expected to output a small *score* f_{bi} (which can be negative) to maximize the bag level likelihood. In contrast, f_{bi} will be high for a positive bag with $t_b = 1$. Since $\pi_{bi} \geq 0$, each instance has a chance to be predicted as positive.

We denote $\mathcal{P}_{\boldsymbol{\pi}_b,n}$ as an uncertainty set, defining the constraints over the mixing coefficient $\boldsymbol{\pi}_b$. Without adding any additional constraints other than being non-negative and summing to one, we have $\mathcal{P}^{max}_{\boldsymbol{\pi}_b,n} := \{\boldsymbol{\pi}_b \in R^n : \boldsymbol{\pi}_b^T \mathbb{1} = 1, 0 \leq \boldsymbol{\pi}_b\}$. It turns out performing multiple instance learning under the GP mixture framework with constraints $\mathcal{P}^{max}_{\boldsymbol{\pi}_b,n}$ is equivalent to a maximum score based model.

Lemma 1. *With $\mathcal{P}^{max}$ as constraints, MIL under the GP mixture framework only considers the most positive instance (equivalent to maximum score MIL).*

Proof. Marginalizing over $\mathbf{z}_b$ leads to the marginal likelihood of the bag-level label:

$$p(y_b|\mathbf{f}_b, \boldsymbol{\pi}_b) = \sum_{i=1}^{n} \pi_{bi} \frac{1}{1+\exp(-t_b f_{bi})} \tag{3}$$

Denote $p(f_{bi}) = \frac{1}{1+\exp(-t_b f_{bi})}$ and maximizing (3) over $\boldsymbol{\pi}_b$ leads to

$$\pi_{bi} = \begin{cases} 1, & \text{if } p(f_{bi}) = \max_{i \in b} p(f_{bi}) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Thus, the bag level likelihood is given by:

$$p(y_b|\mathbf{f}_b) = \max_{i \in b} p(f_{bi}) \tag{5}$$

which only relies on the most positive instance [6]. □

To more effectively involve multiple instances, we can instead consider a top-k constraint, given by

$$\mathcal{P}^{top-k}_{\boldsymbol{\pi}_b,n} := \{\boldsymbol{\pi}_b \in R^n : \boldsymbol{\pi}_b^T \mathbb{1} = 1, 0 \leq \pi_{bi} \leq \frac{1}{k}\} \tag{6}$$

where k indicates the number of instances being potentially positive.

Lemma 2. *With $\mathcal{P}^{top-k}$ as constraints, MIL under the GP mixture framework considers the top-k most positive instances (equivalent to average top-k MIL).*

Proof. Maximizing (3) under $\mathcal{P}^{top-k}_{\boldsymbol{\pi}_b,n}$ constraints gives

$$\pi_{bi} = \begin{cases} \frac{1}{k}, & \text{if } p(f_{bi}) \geq p(f_{b[k]}) \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where $p(f_{b[k]})$ indicates the instance with the k^{th} highest value. Thus, the bag level likelihood is

$$p(y_b|\mathbf{f}_b) = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{1+\exp(-t_b f_{b[i]})} \tag{8}$$

which is collectively determined by the top-k instances with largest scores. □

Lemma 2 shows that by leveraging the top-k constraint, the GP mixture framework involves the top-k most positive instances that can effectively overcome outlier and multimodal situations. However, a remaining issue is how to set a suitable k, which can be quite challenging in practice. More importantly, since k takes discrete values, the prediction performance may fluctuate significantly when k changes. To address this fundamental challenge, we propose to integrate a DRO constraint into the GP mixture framework, which can essentially function as a soft version of the top-k constraint, thus removing the need to specify a fixed k value while ensuring a more stable (and robust) prediction. More specifically, the DRO constraint restricts $\boldsymbol{\pi}_b$ within a certain ball with a center given by the uniform distribution [25]:

$$\mathcal{P}^{DRO}_{\boldsymbol{\pi}_b,n} := \{\boldsymbol{\pi}_b \in R^n : \boldsymbol{\pi}_b^T \mathbb{1} = 1, \boldsymbol{\pi}_b \geq 0, D_f(\boldsymbol{\pi}_b || \frac{\mathbb{1}}{n}) \leq \eta\} \tag{9}$$

where η controls the radius of a ball and D_f is the f-divergence. A large η gives more flexibility on $\boldsymbol{\pi}_b$, which allows it to deviate significantly from the uniform distribution so that one single instance may play a dominant role in the bag level likelihood (equivalent to maximum score MIL when $\eta \to \infty$); a small η leads to near equal probability for each instance (equivalent to averaging overall all instances in a bag when $\eta \to 0$).

3.2 Deep Kernel MIL

While a GP has the non-parametric flexibility along with its Bayesian nature to capture model uncertainty, it is restricted by the kernels with fixed basis functions that are less effective when applied to high dimensional data. To address this issue, one viable solution is to integrate a deep neural network (DNN), which uses adaptive basis functions to learn the rich representations from high dimensional input data.

In terms of network architecture, the proposed deep kernel multiple instance learning consists of three main components: (1) deep neural network, (2) additive Gaussian Processes, and (3) mixing model. For each instance ($\mathbf{x}_i \in \mathcal{R}^D$) present in a bag b, we perform non-linear transformation using a mapping function $\mathbf{h}(\mathbf{x}, \mathbf{w})$ parameterised by neural network weights $\mathbf{w}$ to generate Q-dimensional features at the final layer L, i.e., $h_i^{L1}, ..., h_i^{LQ}$. Next, we use J Gaussian Processes with corresponding base kernels $k_1, ..., k_J$ applied to the subset of those extracted features constituting an additive GP model [20]. As the base kernels act on low dimensional inputs, local kernel interpolation (for scalability) become more natural. The resulting GP functional values from J-GPS ($f_i^1, ..., f_i^J$) are linearly mixed by a training matrix $A \in R^{J \times 1}$ to produce a single functional value f_i. Finally collecting the functional values for all instances present in a bag b, we arrive at the bag-level likelihood in (1).

For the j^{th} Gaussian process in the additive GP layer, let $\mathbf{f}^j = \{f_i^j\}_{i=1}^N$ be the latent functions on the input data features for all the instances in a bag. By introducing a set of latent inducing variables $\mathbf{u}^j$ indexed by m inducing inputs [26] (denoted as R), we have

$$\begin{aligned} p(\mathbf{f}^j|\mathbf{u}^j) &= \mathcal{N}(\mathbf{f}^j|\mathbf{K}^j_{X,R}{\mathbf{K}^j}^{-1}_{R,R}\mathbf{u}^j, \hat{\mathbf{K}}^j), \\ \hat{\mathbf{K}} &= \mathbf{K}_{X,X} - \mathbf{K}_{X,R}\mathbf{K}^{-1}_{R,R}\mathbf{K}_{R,X} \end{aligned} \tag{10}$$

where $X \in R^{N \times Q}$ is the feature representation learned from N training instances through DNN. Performing the local interpolation approximation (similar to [20]) $\mathbf{K}_{X,X} \approx \mathbf{M}\mathbf{K}_{R,R}\mathbf{M}^T$, $\hat{\mathbf{K}}^j$ becomes zero, yielding $\mathbf{f}^j = \mathbf{K}_{X,R}\mathbf{K}^{-1}_{R,R}\mathbf{u} = \mathbf{M}\mathbf{u}$, where $\mathbf{M}$ is $N \times m$ matrix of interpolation weights that can be extremely sparse with the relationship $\mathbf{K}_{X,R} \approx M\mathbf{K}_{R,R}$. This means with the help of local interpolation along with inducing points, we can obtain a deterministic relationship between $\mathbf{f}$ and $\mathbf{u}$ governed by the sparse matrix $\mathbf{M}$.

Denote $\mathbf{U} = \{\mathbf{u}^j\}_{j=1}^J$ as the collection of inducing variables for J additive GPs along with the posterior distribution as $q(\mathbf{U}) = \prod_{j=1}^J \mathcal{N}(\mathbf{u}^j|\boldsymbol{\mu}_j, \mathbf{S}_j)$. Further, let $q(\mathbf{z}_b|\mathbf{r}_b) = \prod_{i=1}^n r_{bi}^{z_{bi}}$ be the posterior distribution for a multinomial variable corresponding to a bag b paramerized by $\mathbf{r}_b$. To update: (1) variational parameters ($\{\boldsymbol{\mu}_j, \mathbf{S}_j\}_{j=1}^J$, $\{r_{bi}\}_{i=1}^n; \forall b \in [1, B]$) (2) GP kernel hyper-parameters, (3) $\{\pi_{bi}\}_{i=1}^n; \forall b \in [1, B]$, (4) mixing coefficients A, and (5) neural network parameters $\mathbf{w}$, we optimize a lower bound of the marginal likelihood using an efficient stochastic variational procedure.

3.3 Stochastic Variational Inference

Exact inference and parameter learning with a non-Gaussian bag level likelihood is intractable. We develop the first stochastic variational inference method that combines a fast sampling scheme to work on a mini-batch setting to ensure efficient and scalable end-to-end training of the new DRO-DK-MIL model.

We start by defining the log marginal bag-level likelihood and applying Jensen's inequality

$$\begin{aligned} \log p(\mathbf{y}) &= \log \int p(\mathbf{y}, \mathbf{Z}, \mathbf{F}, \mathbf{U}) d\mathbf{Z} d\mathbf{F} d\mathbf{U} \\ &\geq \mathbb{E}_q[\log p(\mathbf{y}, \mathbf{Z}, \mathbf{F}, \mathbf{U})] - \mathbb{E}_q[\log q(\mathbf{Z}, \mathbf{F}, \mathbf{U})] \end{aligned}$$

We formally define the lower bound as

$$\begin{aligned} L(q) &\stackrel{\partial}{=} \mathbb{E}_q[\log p(\mathbf{y}, \mathbf{Z}, \mathbf{F}, \mathbf{U})] - \mathbb{E}_q[\log q(\mathbf{Z}, \mathbf{F}, \mathbf{U})] \\ &= \mathbb{E}_q[\log p(\mathbf{y}|\mathbf{Z}, \mathbf{F})] - KL(q(\mathbf{Z})||p(\mathbf{Z})) \\ &\quad - KL[q(\mathbf{U})||p(\mathbf{U})] \end{aligned} \tag{11}$$

where $KL(P||Q)$ is the KL divergence between two distributions P and Q.

Since the likelihood function presented in (11) factorizes over each bag, i.e., $p(\mathbf{y}|\mathbf{Z},\mathbf{F}) = \prod_{b=1}^{B} p(y_b|\mathbf{f}_b,\mathbf{z}_b)$, we can optimize the lower bound in a minibatch setting. The variational parameters corresponding to $q(\mathbf{U})$, kernel hyper-parameter parameters, mixing coefficients A, and neural network parameters are updated using SGD through the noisy approximation of the gradient of the lower bound on mini-batches, as detailed below.

Update $q(\mathbf{Z})$. To update $q(\mathbf{Z})$, we further simplify (11) by absorbing terms that do not depend on $\mathbf{Z}$ to a constant term,

$$L(q(\mathbf{Z})) \overset{\partial}{=} \mathbb{E}_{q(\mathbf{U})q(\mathbf{Z})}[\log p(\mathbf{y}|\mathbf{F},\mathbf{Z})] - \mathbb{E}_{q(\mathbf{Z})}[\log(q(\mathbf{Z}))] + \mathbb{E}_{q(\mathbf{Z})}[\log(p(\mathbf{Z}))] + \text{const}$$

By taking derivative with respect to r_{bi}, we have

$$r_{bi} = \pi_{bi} \exp(\mathbb{E}_{q(\mathbf{U})}[\log(p(t_b f_{bi}))]), \quad \forall i \in [1,n] \quad (12)$$

As long as $\pi_{bi} \geq 0$, we have $r_{bi} \geq 0$. To satisfy the second constraint $\sum_{i=1}^{n} r_{bi} = 1$, we normalize it as

$$r_{bi} = r_{bi} / \sum_{j=1}^{n} r_{bj} \quad (13)$$

Update $\boldsymbol{\pi}$. To update $\boldsymbol{\pi}_b$, we focus on $\mathbb{E}_{q(\mathbf{Z})}[\log(p(\mathbf{Z}))]$, which is the only term as a function of $\boldsymbol{\pi}_b$ and proceed as

$$\max_{\boldsymbol{\pi}_b} \mathbb{E}_{q(\mathbf{z})} \sum_{i=1}^{n} z_{bi} \log(\pi_{bi}) = \max_{\boldsymbol{\pi}_b} \sum_{i=1}^{n} r_{bi} \log(\pi_{bi}) \quad (14)$$

where $\mathbb{E}_{q(\mathbf{Z})}[z_{bi}] = r_{bi}, \forall i \in [1,n]$. It should be noted that maximization of the above objective function is performed under the DRO constraints in (9).

Update $q(\mathbf{U})$. Due to the non-Gaussian bag-level likelihood function in (11), expectation cannot be evaluated analytically. Therefore, we use a sampling method, which is proven to be highly efficient with structured reparametrization, local kernel interpolation, and structure exploiting algebra [20, 11]. Using the local kernel interpolation, the latent function $\mathbf{f}$ is expressed as a deterministic local interpolation of the inducing variables $\mathbf{u}$ and therefore, allowing us to make the difficult posterior approximation over $\mathbf{f}$ easier. As such, we can perform direct reparameterization over $q(\mathbf{U})$ and compute $\mathbf{f}$ directly through interpolation $\mathbf{f}^t = \mathbf{M}\mathbf{u}^t$ (for notation simplicity, we have omitted the index j corresponding to j^{th} GP). Usinig Cholesky decomposition for the covariance matrix of $q(\mathbf{U})$: $\mathbf{S} = \mathbf{L}^T\mathbf{L}$, we have the sampling procedure:

$$\mathbf{u}^t = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{\epsilon}^t, \quad \boldsymbol{\epsilon}^t \sim \mathcal{N}(\mathbf{0},\mathbf{I}) \quad (15)$$

where each step of the above standard sampler has a complexity of $O(m^2)$ with m inducing points.

As the above sampling procedure requires a matrix-vector product, it may become expensive with many inducing points which are required for large datasets with a high dimensional input [20]. To further scale up the sampling procedure, we can take the advantage of both Toeplitz and circulant structure along with the Kronecker decomposition on $\mathbf{L} = \bigotimes_{d=1}^{D} \mathbf{L}_d$ with D being the input dimension of the base kernel.

As both KL divergence terms have a closed form, only the bag-level likelihood function requires sampling for the expectation computation. With T samples of $\mathbf{u}$ and mini-batch of bag size P, we can estimate the marginal likelihood lower bound as:

$$L(q) \approx \frac{1}{TP} \sum_{t=1}^{T} \sum_{b=1}^{P} \sum_{i=1}^{n} z_{bi}^t \log\left(\frac{1}{1+\exp(-t_b f_{bn}^t)}\right) - KL[q(\mathbf{U})||p(\mathbf{U})] - KL[q(\mathbf{Z})||p(\mathbf{Z})] \quad (16)$$

where we can efficiently compute the $KL(q(\mathbf{U})||p(\mathbf{U}))$ term and its gradient with the Kronecker method (details are provided in the Appendix).

Update other parameters. Update of other parameters, including the mixing coefficients A, kernel hyper-parameters, variational parameters $\{\boldsymbol{\mu}, \{\mathbf{L}_d\}_{d=1}^{D}\}$, and neural network parameters, can be achieved through gradient decent as detailed in the Appendix.

4 Experiments

We conduct extensive experiments to evaluate the proposed DRO-DKMIL model. We first introduce three real-world video datasets for anomaly detection. Anomaly detection is regarded as one of the most challenging computer vision tasks under the MIL setting. Next, we demonstrate the overall performance of DRO-DKMIL and compare it with existing state-of-the-art video anomaly detection models. Further, we assess the effectiveness of our proposed model in multimodal and outlier scenarios. We also provide a qualitative analysis to justify the superior performance of our model. Finally, we investigate the impact of the key parameters to the model performance. The GitHub repository that includes the source code and detailed documentation can be accessed via this link.

4.1 Datasets and Experimental Settings

Datasets. Our experiments involve three anomaly detection video datasets of different scales: ShanghaiTech [27], Avenue [13], and UCF-Crime [2]. On those videos, the assumption is that in the training set, frame level annotation is missing and only video level information (indicating whether the video is of abnormal type or normal type) is available.

Table 1: Video Level Distribution on Different Datasets

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

- **ShanghaiTech** consists of 437 videos with 330 normal and 107 abnormal videos. In the original setting, all training videos are normal. To fit into our scenario, we follow the data split in [28] 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: 1610 for training and 290 for testing. In this dataset, frame labels are available only for the testing videos.

Table 1 shows how the videos are partitioned into the training and testing sets in each dataset.

Evaluation metric and model training. For evaluation, we report the frame-level receiver operating characteristics (ROC) curve along with the corresponding AUC score, which captures the robustness of the prediction performance at varying thresholds. For the Avenue and ShanghaiTech datasets, we extract the visual features from FC7 layer of a pre-trained C3D network [29]. To extract the features, we first re-size each video frame to 240×340 pixels and fix the frame rate to 30fps. Next, we use a pre-trained C3D model to compute the C3D features for every 16-frame video clip. This may yield a different number of clips (each clip having 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 terms of the DNN architecture, we follow the 2-dimensional neural network followed by the GP base kernels. The first FC layer has 32 units followed by 16 units. We adopt a 60% dropout regularization between FC layers along with the ReLU activation. For the UCF-Crime dataset, we extract features using I3D network [30]. We uniformly sample 1512 frames and pass an 8-frame video clip into the network. This yields 189 segment clips each with 1024 dimensional feature vector. For this dataset, we use a 5-layer LSTM network, where each layer has 189 hidden units followed by a batch normalization layer and FC layer of 16 nodes. Finally, base GP kernels are applied to the DNN output features. In the uncertainty set of parameter $\boldsymbol{\pi}$, we define the f-divergence as a Kullback-Leibler (KL)-divergence. For hyper-parameter η, we conduct a grid search in a range from 10^{-9} to 1.0 and find the one with best validation AUC score as the optimal η value. The details about η value selection and its impact are provided in the Appendix. For DNN training, we use SGD with a learning rate of 0.001 and l_2 regularization with parameter $\lambda = 0.001$ whereas, for variational parameters, mixing coefficient (A), and hyper-parameters, we use a learning rate of 0.1.

4.2 Performance Comparison

In our comparison study, we include baselines that are used in the video anomaly detection tasks. We also compare with the maximum score based GP model [6] but augment it with deep kernel learning to properly handle high-dimension data (referred as DK-MMIL). We further implement the variational inference algorithm developed in [9] and refer to this model as VGP-MIL. We also compare with the average top-k constraint as introduced in Lemma 2 (refer to as DK-TKMIL) with a pairwise hinge loss (details are provided in the Appendix). In addition, for each dataset, we also include other competitive models that have been applied to that dataset.

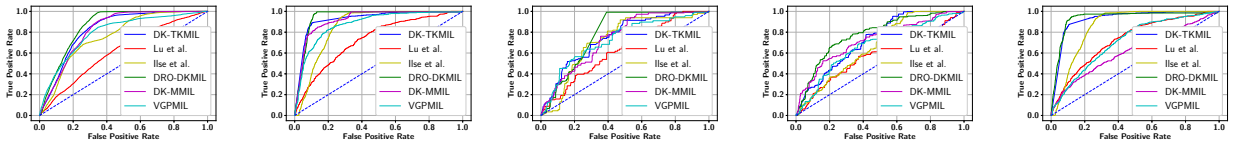
UCF-Crime. Table 2 shows the AUC scores of all competitive techniques. As can be seen, DRO-DKMIL has superior performance compared to other existing techniques. The corresponding ROC performance is shown in Figure 1 (a). As shown, DRO-DKMIL has higher TPR for all FPR below 0.5, which demonstrates the robustness of the approach.

ShanghaiTech. Besides the common baselines, we also compare our method with the recent GCN based model using three feature extractors ($C3D$, TSN^{RGB}, $TSN^{OptimalFlow}$) [28]. The result is reported in Table 2. The corresponding ROC curves are shown in Figure 1 (b). The result shows that DRO-DKMIL significantly outperforms other competitive methods.

Avenue. Table 2 summarizes the AUC scores on Avenue of the proposed approach along with other techniques. The result confirms that DRO-DKMIL outperforms all existing techniques. The corresponding ROC performance is shown in Figure 1 (c). Similarly, the proposed approach achieves higher recall compared to other approaches.

4.3 Multimodal and Outlier Detection

In this section, we assess the effectiveness of the proposed DRO-DKMIL in outlier and multimodal set-



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

Figure 1: ROC Performance on Three Video Datasets (a)-(c); Multimodal (d) and Outlier Prediction (e)

Table 2: Comparison of AUC Scores

Approach	AUC (%)
UCF-Crime	
Hasan et al. [31] (C3D)	50.60
Lu et al. [13] (C3D)	65.51
Lu et al. [13] (I3D)	61.98
Sultani et al. [2] (C3D)	75.41
Ilse et al. [15] (I3D)	76.52
Zhong et al. [28] (GCN (C3D))	81.08
Zhong et al. [28] (TSN^{RGB})	82.12
Zhong et al. [28] ($TSN^{OpticalFlow}$)	78.08
Haußmann et al. [9] VGPMIL (I3D)	79.56
DK-MMIL (I3D)	82.32
DK-TKMIL (I3D)	82.66
DRO-DKMIL (I3D)	**85.93**
ShanghaiTech	
Lu et al. [13] (C3D)	72.90
Zhong et al. [28] (GCN (C3D))	76.44
Zhong et al. [28] (TSN^{RGB})	84.44
Zhong et al. [28] ($TSN^{OpticalFlow}$)	84.13
Ilse et al. [15] (C3D)	85.78
Haußmann et al. [9] VGPMIL (C3D)	87.78
DK-MMIL (C3D)	92.00
DK-TKMIL (C3D)	92.30
DRO-DKMIL (C3D)	**94.39**
Avenue	
Lu et al. [13] (C3D)	62.14
Ilse et al. [15] (C3D)	72.39
Haußmann et al. [9] VGPMIL (C3D)	72.84
DK-MMIL (C3D)	73.93
DK-TKMIL (C3D)	75.12
DRO-DKMIL (C3D)	**78.66**

tings. For this, we create a multimodal scenario by extending the UCF-Crime dataset. For the outlier scenario, we deliberately impose some outliers in the ShanghaiTech dataset and evaluate the performance.

Multimodal Detection. The original UCF-Crime dataset does not explicitly consider the multimodal scenario. Although it is natural to have multimodal scenario in the real-world videos (as evidenced by the superior performance of the proposed model), it is hard to identify the actual videos for this specific evaluation. In case of UCF-Crime, we have abnormal videos categorized into different activity types. Therefore, we create a multimodal scenario by combining multiple abnormal videos from different anomaly types. To create a multimodal scenario, we randomly select three activity types. Then, we form a positive (abnormal) bag by concatenating three abnormal videos, one video per activity type. To construct a normal bag, we randomly pick three normal videos and concatenate them. In the process, the training bags are constructed using training videos only and testing bags are constructed using testing videos only. The corresponding video statistics is shown in the Table 1. Each bag is a concatenation of three videos yielding total 50 abnormal and 50 normal bags in the training set. The testing set consists of 10 normal and 10 normal videos. Table 3 reports the AUC scores and the ROC plot is shown in Figure 1 (d). We can observe that the ROC curve from DRO-DKMIL clearly outperforms all baselines. This means, compared to the baselines, our approach is more robust to the multimodal scenario at various thresholds.

Table 3: AUC on Multimodal and Outlier Detection

Approach	AUC (%)	
	Multimodal	***Outlier***
Lu et al. [13] (C3D)	58.67	72.90
Ilse et al. [15] (C3D)	66.85	85.65
Haußmann et al. [9] VGPMIL (C3D)	67.16	71.31
DK-MMIL (C3D)	72.44	62.89
DK-TKMIL (C3D)	72.75	92.61
DRO-DKMIL (C3D)	**77.89**	**93.49**

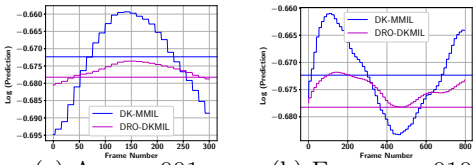
Outlier Detection. To test the robustness of the proposed approach with outliers, we extend the ShanghaiTech dataset by explicitly including outliers. Specifically, we randomly select 120 segments from abnormal videos and replace their features with points drawn from a standard multivariate Gaussian distribution. As shown in Table 3, DK-MMIL suffers heavily by the outliers compared to the proposed DRO-DKMIL. This is because, it is likely to have an outlier predicted as the maximum prediction score from an abnormal video. As a result, the overall training process may be heavily influenced by outliers. However, as our approach gives chance to other actual abnormal segments as well in the training process, it makes the model robust to the outliers. It is also noted that DK-TKMIL performs very well with outliers, which benefits from the top-k constraints. However, setting a proper k value is highly challenging in practice and the prediction performance fluctuates significantly with the change of k (see Appendix for details).

4.4 Qualitative Analysis

To get deeper insight regarding the effectiveness of our approach, we analyze videos where the proposed DRO-

(a) ARREST001 (b) EXPLOSION010

Figure 2: Abnormal Frames Identified by DRO-DKMIL but not DK-MMIL



(a) ARREST001 (b) EXPLOSION010

Figure 3: Abnormal Frame Prediction

DKMIL and maximum score-based DK-MMIL provide different predictions. Figure 2 shows abnormal frames from two videos in which DRO-DKMIL correctly predicts as abnormal whereas DK-MMIL fails. The resulting prediction scores for all abnormal frames for the videos ARREST001 and EXPLOSION010 are shown in Figure 3. The prediction threshold (shown as a horizontal line) in each approach is determined such that FPR is maintained at 0.3. As shown in the video ARREST001, DK-MMIL fails to detect the abnormal frames near the transition phase. Since transitioning frames may be far from the abnormal frame with the maximum prediction score, DK-MMIL does not consider those types of abnormal frames during model training. However, for the proposed DRO-DKMIL, it is more likely to involve these transitioning abnormal frames in the training process. Thus, it can correctly identify similar frames during testing.

In the video EXPLOSION010, DK-MMIL fails to correctly identify the abnormal frames that are in the middle. This may be because more extreme abnormal frames of the explosion type may only participate in the training process. As a result, the maximum score based approach may not consider the frame as shown in Figure 2(b). However, the proposed approach may be more likely to involve this type of abnormal frames as it allows the participation of multiple abnormal frames from each abnormal video.

4.5 Uncertainty Analysis

Being a Bayesian model, the proposed DRO-DKMIL is able to accurately capture the prediction uncertainty, which provides important complementary information for video anomaly detection. The uncertainty score can guide a human decision maker to not only focus on the predicted positive frames to examine the abnormality but also pay attention to the highly uncertain

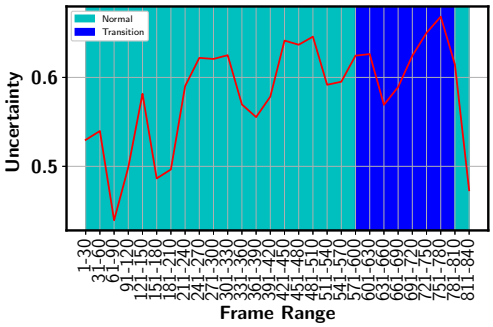


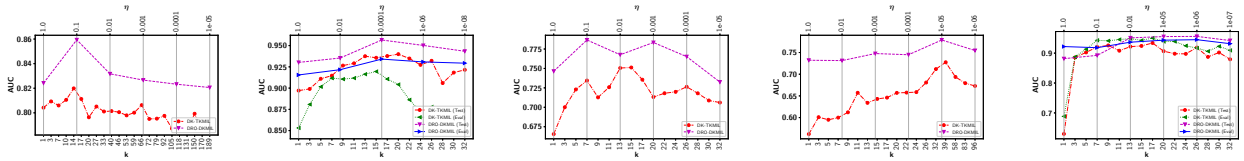
Figure 4: Uncertainty of Different Frames

areas in the videos that may also include important information to support decision-making. To show this, we use the Avenue dataset and report the standard deviation (SD) output by DRO-DKMIL for each testing frame. We maintain FPR = 0.3 and identify all correct and incorrect instances. By setting a threshold as 0.67, we identify 70 incorrect and 42 correct frames with a SD above the threshold. This means a larger uncertainty score (i.e., sd) indicates a higher chance of prediction errors, which is a desirable property.

Figure 4 shows the uncertainty associated with different frames in Avenue video TEST-10. In the video, the first 569 are normal frames, where the model has a relatively high confidence. After that until frame 817, the transition occurs nine times between abnormal and normal frames. Therefore, we observe much higher uncertainty. As the transition is rapid, the consecutive frames may look very similar to each other, which may confuse the model, leading to a (correctly) predicted high uncertainty score for those frames.

4.6 Impact of Key Model Parameters

Impact of η. We analyze the impact of the hyperparameter η in the AUC score for all datasets (UCF-Crime, ShanghaiTech, Avenue, UCF Crime Multimodal, and ShanghaiTech outlier). For the ShanghaiTech and ShanghaiTech outlier, we use 20% of the original testing set to construct a validation set and use the rest to report the model performance. The propose of constructing the validation set is to determine the optimal η value. To get robustness in the performance, we randomly choose the validation set 20 times producing 20 pairs of the validation-testing split. Figures 5 (b) and (e) show the validation and testing AUC change for the randomly selected test-validation pair from ShanghaiTech and ShanghaiTech Outlier datasets, respectively. For a lower η value, the model allows the participation of most of the frames. As a result, the model tries to make a prediction score of most of the frames from an abnormal video to be higher than from a normal video, resulting in the misclassification of many normal frames from abnormal videos. Therefore, we observe lower performance for a



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

Figure 5: AUC Performance vs. η and Comparison with Average Top-k

lower η value. As we increase η, the model limits the participation of the frames from both abnormal and normal videos. This increases the chances of including only abnormal frames while leaving out normal frames from abnormal videos in the optimization process. As a result, the model learns to have a higher score for the abnormal frames compared to all the normal frames, resulting in improvement in the performance. However, a very high η value allows the participation of a very limited number of abnormal as well as normal frames during the training process. As a result, the model may be highly influenced by outliers and multimodal scenarios. Therefore, we can see the degradation in the performance for a high η value.

For the UCF-Crime, Avenue, and Multimodal datasets, we directly report the performance in the testing dataset, instead of using a separate validation set. For UCF-Crime, the use of a separate validation set may not be effective because of the limited testing videos of a given type. Therefore, similar to [2], we evaluate the testing performance with respect to η and report the one with the best performance as the best η value. For the Avenue dataset, there are very limited testing videos, so determining the η value using a separate validation set may not be feasible. As can be seen in the Figures 5 (a), (c), and (d), the trend is similar to what we have observed in the ShanghaiTech dataset for the same reason explained above.

Comparison with Average top-k. As proved in Lemma 2, by using a top-k constraint, the proposed framework is equivalent to perform average top-k MIL. As the top-k most positive frames are simultaneously considered by the training process, it can potentially handle the outlier and multimodal scenarios as well. In this set of experiments, we further compare the proposed DRO-DKMIL with the average top-k model (the deep kernel version is referred to as DK-TKMIL). Figure 5 compares the AUC scores between DRO-DKMIL and DK-TKMIL while varying η and k. We have three key observations: (i) With a properly chosen k, DK-TKMIL achieves a decent prediction performance, especially when dealing with outliers, as shown in 5(e). (ii) DRO-DKMIL achieves even better prediction performance. In all datasets, the test AUC curve of DRO-DKMIL stays on top of DK-TKMIL for almost all different η and k values. (iii) In almost all datasets, the AUC score of DK-TKMIL changes more significantly when compared with the AUC score change of DRO-DKMIL with η. In addition, η varies in a much wider range (i.e., 10^{-8} to 1) than the k values. This clearly confirms the advantage of DRO-DKMIL over an average top-k model as setting a proper k may be highly challenging. In addition, due to the discrete nature of k, the prediction performance may fluctuate significantly when k changes. The DRO based constraint essentially offers a soft version of the top-k constraint, which effectively addresses the limitation of an average top-k model.

5 Conclusions

We present a general GP mixture framework for multiple instance learning under noisy and multimodal settings. The proposed framework can flexibly incorporate multiple instances into the bag-level likelihood so that the model can most effectively learn from these potentially positive instances to make more robust predictions with the presence of outliers and different event types in the same bag. A key modeling ingredient is a DRO constraint applied to the mixture model parameters that acts as a soft top-k constraint to identify the subset of most positive instances in a bag. We further augment the GP kernel by using a deep neural network that uses adaptive basis functions to learn the rich representations from high dimensional input data. A stochastic variational inference method combines a fast sampling scheme to work on a mini-batch setting that ensures efficient and scalable end-to-end model training. Experiments on three challenging real-world video anomaly detection datasets clearly demonstrate the effectiveness of the proposed model.

Acknowledgement

Hitesh Sapkota and Qi Yu are supported in part by an ONR award N00014-18-1-2875 and an NSF IIS award IIS-1814450. Yiming Ying is supported by NSF grants IIS-1816227 and IIS-2008532. Feng Chen is supported by the NSF under Grant No #1815696 and #1750911. 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] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1–2, p. 31–71, Jan. 1997.

[2] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6479–6488.

[3] J.-X. 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," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1237–1246, 2019.

[4] Z.-H. Zhou and M.-L. Zhang, "Neural networks for multi-instance learning," *Proceedings of the International Conference on Intelligent Information Technology*, 11 2002.

[5] X. Wei, J. Wu, and Z. Zhou, "Scalable multi-instance learning," in *2014 IEEE International Conference on Data Mining (ICDM)*, 2014, pp. 1037–1042.

[6] M. Kim and F. Torre, "Gaussian processes multiple instance learning," in *ICML*, 2010.

[7] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *NIPS*, 2002.

[8] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li, "Multi-instance learning by treating instances as non-i.i.d. samples," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 1249–1256.

[9] M. Haußmann, F. A. Hamprecht, and M. Kandemir, "Variational bayesian multiple instance learning with gaussian processes," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 810–819.

[10] A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing, "Deep kernel learning," in *Artificial intelligence and statistics*, 2016, pp. 370–378.

[11] A. Wilson, Z. Hu, R. Salakhutdinov, and E. Xing, "Stochastic variational deep kernel learning," *ArXiv*, vol. abs/1611.00336, 2016.

[12] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection - a new baseline," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6536–6545.

[13] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in matlab," in *2013 IEEE International Conference on Computer Vision*, 2013, pp. 2720–2727.

[14] X. Xu and E. Frank, "Logistic regression and boosting for labeled bags of instances," vol. 3056, 08 2004.

[15] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *ICML*, 2018.

[16] H. Namkoong and J. C. Duchi, "Stochastic gradient methods for distributionally robust optimization with f-divergences," in *NIPS*, 2016.

[17] Q. Qi, Y. Yan, Z. Wu, X. Wang, and T. Yang, "A simple and effective framework for pairwise deep metric learning," *ArXiv*, vol. abs/1912.11194, 2019.

[18] H. Namkoong and J. C. Duchi, "Variance-based regularization with convex objectives," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 2975–2984.

[19] Y. Fan, S. Lyu, Y. Ying, and B. Hu, "Learning with average top-k loss," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 497–505.

[20] A. G. Wilson and H. Nickisch, "Kernel interpolation for scalable structured gaussian processes (kiss-gp)," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML'15. JMLR.org, 2015, p. 1775–1784.

[21] A. Dezfouli and E. V. Bonilla, "Scalable inference for gaussian process models with black-box likelihoods," in *NIPS*, 2015.

[22] R. Turner, "Statistical models for natural sounds," 2010.

[23] J. Cunningham, K. Shenoy, and M. Sahani, "Fast gaussian process methods for point process intensity estimation," in *ICML '08*, 2008.

[24] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 6, pp. 1153–1160, 1981.

[25] H. Namkoong and J. C. Duchi, "Stochastic gradient methods for distributionally robust optimization with f-divergences," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 2216–2224.

[26] J. Quiñonero Candela and C. E. Rasmussen, "A unifying view of sparse approximate gaussian process regression," *J. Mach. Learn. Res.*, vol. 6, p. 1939–1959, Dec. 2005.

[27] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked rnn framework," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 341–349.

[28] 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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1237–1246.

[29] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ser. ICCV '15. USA: IEEE Computer Society, 2015, p. 4489–4497.

[30] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4724–4733.

[31] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 733–742.