Appendix: Distributionally Robust Optimization for Deep Kernel Multiple Instance Learning

In this appendix, we first present a table summarizing the major notations used by the main paper in Appendix A. We then present in Appendix B the detailed proof of Lemma 1 and Lemma 2 presented in Section 3. Next, we present the update rules of key parameters in the model. We then discuss details regarding the impact of η on the performance and compare it with the average top-k model. Finally, we perform an in-depth qualitative analysis that shows key evidence that the proposed DKL-MIL approach is robust to the outliers and multimodal scenarios. The link for the source code is provided at the end of the Appendix.

A Summary of Notation

Table 4 summarizes all the major symbols along with their descriptions.

Table 4: Symbols with Descriptions

Notation	Description
$\mathbf{X}$	Set of training bag instances
B	Total number of bags in a training set
$\mathbf{y}$	Set of predicted probabilities of B bags
t_b	Binary label for bag b
$\mathbf{f}_b$	Set of functional values of instances in a bag b
$\mathbf{z}_b$	Indicator variable drawn from a multinomial distribution
n	Total number of instances in each bag
Q	DNN final layer (L) feature representation of each bag instance
$\mathbf{w}$	DNN parameters
u^j	Inducing variables for j^{th} GP
$\boldsymbol{\mu}_j$	Posterior distribution mean for u^j
$\mathbf{S}_j$	Posterior distribution co-variance matrix for u^j
A	Mixing parameter to combine J-GPS functional values
$\mathbf{U}$	Set of inducing variables for J GPS
$\mathbf{Z}$	Set of multinomial variables for bags B
$\mathbf{F}$	Set of functional values for bags B
M	Sparse interpolation matrix
$\mathbf{r}_b$	Posterior distribution parameter for $\mathbf{z}_b$
η	Hyper-parameter used to define the ball radius in DRO framework
T	Total number of likelihood samples used
P	Mini-batch of bag size
$\mathbf{L}$	Lower diagonal matrix with real and positive entries
$\boldsymbol{\pi}_b$	Prior distribution parameter for $\mathbf{z}_b$
N	Total number of segments in a training set
D	Feature dimension of each bag instance
$\mathcal{N}$	Gaussian distribution
$\mathbb{E}_q$	Expectation with respect to the distribution q
$\mathbf{K}_{A,B}$	Kernel matrix computed between A and B
R	Total number of inducing inputs considered in each GP
$L(q)$	Marginal likelihood lower bound with variational distribution q
$\bigotimes$	Kronecker decomposition operator

B Detailed Proofs

In this section, we show the detailed steps of the proofs for Lemma 1 and Lemma 2.

Proof of Lemma 1: We have the bag level likelihood from Eqs. (1) and (2) as follows:

$$p(y_b|\mathbf{f}_b, \mathbf{z}_b) = \prod_{i=1}^{n} \left\{ \frac{1}{1+\exp(-t_b f_{bi})} \right\}^{z_{bi}} \tag{17}$$

$$p(\mathbf{z}_b|\boldsymbol{\pi}_b) = \prod_{i=1}^{n} \pi_{bi}^{z_{bi}}, \pi_{bi} \geq 0, \sum_{i}^{n} \pi_{bi} = 1 \tag{18}$$

Marginalizing over $\mathbf{z}_b$, we have the following expression:

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

Let's denote $p(f_{bi}) = \frac{1}{1+\exp(-t_b f_{bi})}$, which yields,

$$p(y_b|\mathbf{f}_b, \pi_b) = \sum_{i=1}^{n} \pi_{bi} p(f_{bi})$$

In Lemma 1 , we maximize the above likelihood over $\boldsymbol{\pi}_b$ with respect to the following uncertainty set:

$$\mathcal{P}_{\boldsymbol{\pi}_b,n}^{max} := \{\boldsymbol{\pi}_b \in R^n : \boldsymbol{\pi}_b^T \mathbb{1} = 1, 0 \leq \boldsymbol{\pi}_b\}$$

The resulting optimization becomes:

$$\max_{\boldsymbol{\pi}_b} \sum_{i=1}^{n} \pi_{bi} p(f_{bi}) \quad \text{s.t.} \sum_{i=1}^{n} \pi_{bi} = 1, \pi_{bi} \geq 0, \forall i \in [1, n]$$

Adding the Lagrange multipliers $u_i \geq 0$ and λ, we get:

$$L(\boldsymbol{\pi}_b, \mathbf{u}, \lambda) = \sum_{i=1}^{n} [\pi_{bi} p(f_{bi}) + u_i \pi_{bi}] + \lambda \left[\sum_{i=1}^{n} \pi_{bi} - 1 \right]$$

Taking derivative with respect to π_{bi} and setting to zero yields

$$p(f_{bi}) + u_i + \lambda = 0 \tag{19}$$

The corresponding KKT conditions are:

$$u_i \geq 0, \quad u_i \pi_{bi} = 0, \quad \forall i \in [1, n] \tag{20}$$

Considering $j = \arg\max_{i \in b} p(f_{bi})$, we have the following condition

$$p(f_{bj}) + u_j + \lambda = 0 \tag{21}$$

Combining Eqs. (19) and (21) results in:

$$p(f_{bj}) + u_j = p(f_{bi}) + u_i, \quad \text{s.t. } j = \arg\max_{i} p(f_{bi}), \forall i \in [1, n], i \neq j \tag{22}$$

Since $p(f_{bj}) > p(f_{bi})$, we have $u_j < u_i$. As $u_i \geq 0, \forall i \in [1, n]$, we have $u_i \neq 0, \forall i \neq j$. By leveraging the complementary slackness condition, we have

$$u_i \neq 0, \quad \pi_{bi} = 0, \quad \forall i \in [i, n], i \neq j \tag{23}$$

Further using summation constraint, i.e., $\sum_{i=1}^{n} \pi_{bi} = 1$, we have the following

$$\pi_{bi} = \begin{cases} 1, & \text{if } p(f_{bi}) = \max_i p(f_{bi}) \\ 0, & \text{otherwise} \end{cases} \tag{24}$$

In case of equality condition with $p(f_{bj}) = p(f_{bi})$ with $i \neq j$, we randomly select one and assign $\pi_{bi} = 1$ for one instance whereas 0 for others.

Proof of Lemma 2: We have the following marginalized likelihood function (from Lemma 1 proof):

$$p(y_b|\mathbf{f}_b\pi_b) = \sum_{i=1}^{n} \pi_{bi} p(f_{bi}) \; \forall i \in [1, n]$$

In Lemma 2 , we maximize the above likelihood function with respect to the following uncertainty set:

$$\mathcal{P}^{top-k}_{\boldsymbol{\pi}_b,n} := \{\boldsymbol{\pi}_b \in R^n : \boldsymbol{\pi}_b^T \mathbb{1} = 1, 0 \le \pi_{bi} \le \frac{1}{k}\} \tag{25}$$

The resulting optimization becomes

$$\max_{\boldsymbol{\pi}_b} \sum_{i=1}^{n} \pi_{bi} p(f_{bi}), \quad \text{s.t.} \; \sum_{i=1}^{n} \pi_{bi} = 1, \quad 0 \le \pi_{bi} \le \frac{1}{k}, \quad \forall i \in [1, n]$$

Adding the Lagrange multipliers $u_i \ge 0$, $v_i \ge 0$, and λ we get:

$$L(\boldsymbol{\pi}_b, \mathbf{u}, \mathbf{v}, \lambda) = \sum_{i=1}^{n} \left[\pi_{bi} p(f_{bi}) + u_{bi}\pi_{bi} + v_{bi}(\frac{1}{k} - \pi_{bi}) \right] + \lambda \left(\sum_{i=1}^{n} \pi_{bi} - 1 \right)$$

Taking the derivative with respect to π_{bi} yields the following:

$$p(f_{bi}) + u_{bi} - v_{bi} + \lambda = 0$$

Considering $p(f_{b[1]}) > p(f_{b[2]}), ..., > p(f_{b[n]})$ with $p(f_{b[i]})$ be the i^{th} highest probability score, we have the following conditions:

$$\begin{aligned} &p(f_{b[1]}) + u_{b[1]} - v_{b[1]} + \lambda \\ =&p(f_{b[2]}) + u_{b[2]} - v_{b[2]} + \lambda \\ &... \\ =&p(f_{b[n]}) + u_{b[n]} - v_{b[n]} + \lambda \end{aligned}$$

Removing the $p(f_{b[i]})$ and λ terms, we get the following inequalities:

$$u_{b[1]} - v_{b[1]} < u_{b[2]} - v_{b[2]} < .... < u_{b[k]} - v_{b[k]} < .... < u_{b[n]} - v_{b[n]} \tag{26}$$

Consider the following KKT conditions $\forall i \in [1, n]$

$$\sum_{i=1}^{n} \pi_{bi} = 1 \tag{27}$$

$$u_{b[i]}\pi_{bi} = 0 \tag{28}$$

$$v_{b[i]} \left(\frac{1}{k} - \pi_{bi} \right) = 0 \tag{29}$$

$$u_{b[i]} \ge 0 \tag{30}$$

$$v_{b[i]} \ge 0 \tag{31}$$

Case 1: Assume $\pi_{b[1]} = 0$, so $v_{b[1]} = 0$ according to (29). This implies $u_{b[1]} < u_{b[2]} - v_{b[2]}$. Using KKT condition (30), we can write the following:

$$u_{b[2]} - v_{b[2]} > 0 \Rightarrow u_{b[2]} > v_{b[2]} \Rightarrow u_{b[2]} > 0 \text{ (according to KKT condition (31))}$$

This means, to satisfy the constraint $u_{b[2]}\pi_{b[2]} = 0$ we need to have the following:

$$\pi_{b[2]} = 0$$

Again $\pi_{b[2]}$=0 makes $\pi_{b[3]} = 0$ and so on. As $\pi_{b[i]} = 0 \; \forall i \in [1, n]$, and therefore violating the summation constraint $\sum_{i=1}^{n} \pi_{bi} = 1$. **Therefore, $\pi_{b[1]}$ can not be 0.**

Case 2: Assume $0 < \pi_{b[1]} < \frac{1}{k}$, then $v_{b[1]} = 0$, $u_{b[1]} = 0$, We have the following expression:

$$u_{b[2]} > v_{b[2]}$$

Using KKT condition $v_{b[2]} \geq 0$ we can write:

$$u_{b[2]} > 0$$

Now again using KKT condition $u_{b[2]}\pi_{b[2]} = 0$, we write:

$$\pi_{b[2]} = 0$$

Using Case 1, once $\pi_{b[2]}$ becomes 0 all of the proceeding values also become zero:

$$\pi_{b[3]} = 0, ..., \pi_{b[n]} = 0$$

This again violates the summation constraint constraint $\sum_{i=1}^{n} \pi_{bi} = 1$. **Therefore, $\pi_{b[1]}$ can not be less than** $\frac{1}{k}$. This leads to the following conclusion:

$$\pi_{b[1]} = \frac{1}{k}$$

The process of having $\pi_{b[i]} = \frac{1}{k}$ continues until $\pi_{b[k]} = \frac{1}{k}$. As long as we reach to the k^{th} highest element, $\sum_{i=1}^{n} \pi_{bi} = 1$. This implies the following:

$$\pi_{b[i]} = 0, \quad \forall i > k$$

In conclusion, we can write the following:

$$\pi_{bi} = \begin{cases} \frac{1}{k}, & \text{if } p(f_{bi}) \geq p(f_{b[k]}) \\ 0, & \text{otherwise} \end{cases} \tag{32}$$

which proves Lemma 2.

C Parameter Update

To update the parameters, we take the derivative of the lower bound $L(q)$ with respect to the parameter we want to update. We then, use SGD to update the parameter with a given learning rate. Therefore, in this section, we show the computation of derivative of $L(q)$, w.r.t. each parameter.

Derivative of $L(q)$ w.r.t. Base Kernel Hyperparameters We have following expression for a lower bound

$$L(q) \approx \mathbb{E}_{q(\mathbf{U})q(\mathbf{Z})}[\log p(\mathbf{y}|\mathbf{F}, \mathbf{Z})] - KL[q(\mathbf{U})||p(\mathbf{U})] - KL[q(\mathbf{Z})||p(\mathbf{Z})]$$

In the above equation, the base kernel hyperparameters θ are only involved in the second term i.e., $KL[q(\mathbf{U})||p(\mathbf{U})]$. Before taking its derivative let us simplify it further,

$$KL[q(\mathbf{U})||p(\mathbf{U})] = \frac{1}{2}\left\{\log|K| - \log|\mathbf{S}| - D + \text{tr}(K^{-1}\mathbf{S}) + \mu^T K^{-1}\mu\right\}$$

Taking derivative w.r.t. θ, it gives:

$$\frac{\partial L(q)}{\partial \theta} = -\frac{\partial KL(q(\mathbf{U})||p(\mathbf{U})}{\partial \theta} = -\frac{1}{2}\left\{\text{tr}(K^{-1}\frac{\partial K}{\partial \theta}) - \text{tr}(K^{-1}\frac{\partial K}{\partial \theta}K^{-1}S) - \boldsymbol{\mu}^T K^{-1}\frac{\partial K}{\partial \theta}K^{-1}\boldsymbol{\mu}\right\}$$

Depending on type of a given kernel, we can find $\frac{\partial K}{\partial \theta}$ in the above equation. The corresponding matrix inversions and traces can be computed efficiently by using Kronecker product decomposition [11].

Derivative of $L(q)$ w.r.t. variational parameters of distribution $q(\mathbf{U})$ Both variational parameters $\boldsymbol{\mu}$ and $\mathbf{L}$ depend on the first bag-level likelihood term and $KL[q(\mathbf{U})||p(\mathbf{U})]$. The first term in $L(q)$ is given as:

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

Taking the derivative of i^{th} instance with respect to (p,q)-th element of $L_d^{(j)}$, i.e. λ, where $j \in [1, J]$ indicates the j^{th} base GP:

$$\nabla_\lambda \log p(y_b|\mathbf{f}_b, \mathbf{z}_b) = z_{bi} \frac{\partial \log p(f_{bi})}{\partial f_{bi}} \frac{\partial f_{bi}}{\partial \lambda}$$

In the above equation f_{bi} is defined as:

$$f_{bi} = \sum_{j=1}^{J} A_j f_{bi}^j \tag{33}$$

As we are taking with respect to the j^{th} GP, it means:

$$\frac{\partial f_{bi}}{\partial \lambda} = A_j \frac{\partial f_{bi}^j}{\partial \lambda}$$

As we know we have the following relationship $f = M(\boldsymbol{\mu} + \mathbf{L}\boldsymbol{\epsilon})$. Using this relationship we get:

$$\frac{\partial f_{bi}^j}{\partial \lambda} = A_j M_i^j \nabla_\lambda L^{(j)} \boldsymbol{\epsilon}$$

where M_i^j is the i^{th} row of the j^{th} GP. Now the first term can be computed as:

$$\frac{\partial \log p(f_{bi})}{\partial f_{bi}} = \frac{t_b}{1+\exp(t_b f_{bi})}$$

Combining both and considering all elements present in a bag b, we have the following update rule:

$$\nabla_\lambda \log p(y_b|\mathbf{f}_b, \mathbf{z}_b) = \mathbb{E}_{p(\boldsymbol{\epsilon})q(\mathbf{Z})} \left[z_{bi} \frac{t_b A_j M_i^j \nabla_\lambda L^{(j)} \boldsymbol{\epsilon}}{1+\exp(t_b f_{bi})} \right]$$

We can write down the derivatives w.r.t. the whole matric $\mathbf{L}^{(j)}$ which is efficient for computing:

$$\nabla_{L^{(j)}} \log p(y_b|f_b, z_b) = \mathbb{E}_{p(\boldsymbol{\epsilon})q(\mathbf{Z})} \left[z_{bi} \frac{t_b A_j \left(\boldsymbol{\epsilon} M_i^j\right)^T}{1+\exp(t_b f_{bi})} \right]$$

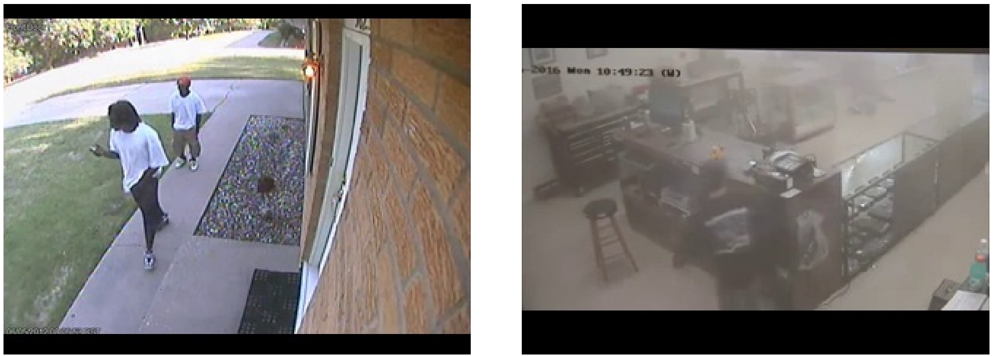
Now the derivative of $KL[q(\mathbf{U})||p(\mathbf{U})]$ can be written as:

$$\nabla_\lambda KL[q(\mathbf{U})||p(\mathbf{U})] = -\frac{1}{2} \frac{\partial[-\log|S| + \mathrm{tr}(K^{-1}S)]}{\partial \lambda}$$

We can efficiently compute the above by using Kronecker decomposition matrix. The derivatives w.r.t. the variational mean μ can be computed similarly as that of $\mathbf{L}$.

Derivative w.r.t. other parameters The mixing weight only depends on the likelihood function. Therefore, it can be easily computed as:

$$\nabla_{A_j} \log p(y_b|\mathbf{f}_b, \mathbf{z}_b) = \mathbb{E}_{p(\boldsymbol{\epsilon})q(\mathbf{Z})} \left[\frac{z_{bi} t_b f_{bi}^j}{1+\exp(t_b f_{bi})} \right] \tag{34}$$



(a) Burglary005 (b) Shooting028

Figure 6: Normal Frames where DRO-DKMIL Correctly Predicts whereas DK-MMIL fails

To update the neural network parameters $\mathbf{w}$, we take the derivative of $L(q)$ with respect to $\mathbf{w}$. As the network parameters are related to $L(q)$ through the kernel matrix K, our update procedure is given as:

$$\frac{\partial L(q)}{\partial \mathbf{w}} = \frac{\partial L(q)}{\partial K} \frac{\partial K}{\partial h(\mathbf{x}, \mathbf{w})} \frac{\partial h(\mathbf{x}, \mathbf{w})}{\partial \mathbf{w}}$$

In the above expression, the term $\frac{\partial K}{\partial h(\mathbf{x}, \mathbf{w})}$ is the implicit derivative of the deep kernel with respect to $h(\mathbf{x}, \mathbf{w})$, holding base kernel hyperparameters θ fixed. The derivatives with respect to network weight variables $\frac{\partial h(\mathbf{x}, \mathbf{w})}{\partial \mathbf{w}}$ are computed using the standard back-propagation techniques.

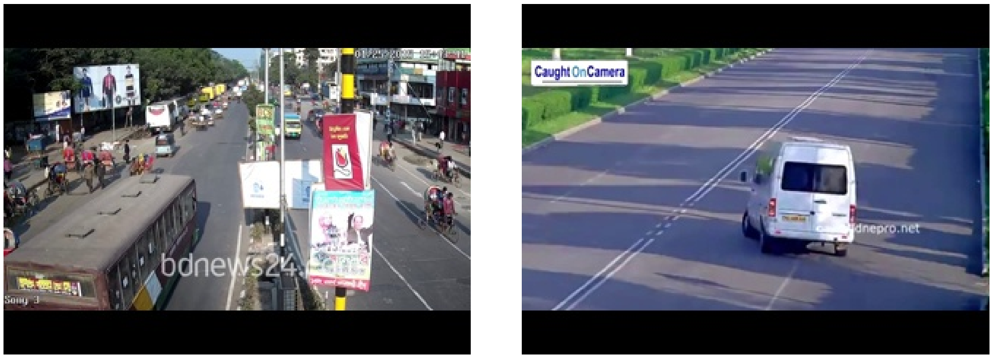
D Additional Experimental Results

D.1 Qualitative Analysis

We perform in-depth analysis in the UCF-Crime dataset considering the rich set of anomalies and complex nature of the videos. In this dataset, the proposed DRO-DKMIL approach is able to correctly predict around 10% more abnormal frames than the DK-MMIL based approach while maintaining a similar FPR (0.3). The higher performance is because of the ability of the DRO-DKMIL approach to facilitate the participation of multiple types of anomaly frames in the training process along with its robustness toward the outlier frames.

The presence of the outlier frames in the abnormal video may result in the degradation in the performance of the maximum score based approach. This may be because during the training process, the outlier may take the maximum prediction score and therefore, the training may be influenced heavily by the outlier. Figure 6 demonstrates two normal frames from abnormal videos Burglary005 and Shooting028, respectively. DK-MMIL incorrectly predicts both of them as abnormal with a high confidence. As the frame in (a) is very similar to abnormal frames from the arson category, during training, this type of frames is more likely to be involved in the process. For instance, suppose that the video Burglary005 is in the training set. It is more likely that a normal frame in (a) would have participated in the training process compared to the other actual abnormal frames of the Burglary type. So, this may cause the failure of the actual abnormal frames from Burglary005 to participate in the optimization process. This eventually may lead to the misclassification of: (1) many normal frames that look similar to the arson type, and (2) many abnormal frames from Burglary005 video. Similarly, the frame in (b) looks more similar to the explosion type of frames because of the foggy environment. Due to the similar reason as in frame (a), it may cause misclassification of many normal frames that look similar to the explosion type. However, using DRO-DKMIL approach, we can correctly classify both frames (a) and (b). This is because, in our approach, actual abnormal frames are also more likely to be involved in the optimization process (in addition to the outliers) and the model may be influenced more by actual abnormal frames instead of the outliers especially when we have a small number of outliers. In the UCF-Crime dataset, outlier frames from abnormal videos are those normal frames that resemble to abnormal frames from other activity types (as demonstrated in Figure 6 (a) and (b))

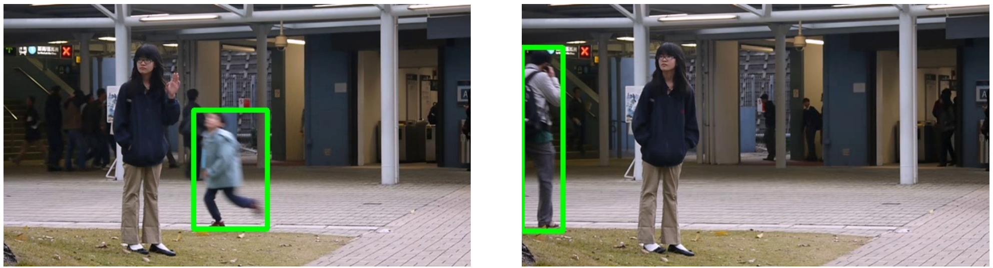
Figure 7 shows abnormal frames where both DRO-DKMIL as well as DK-MMIL fail to predict correctly. Both frames are labeled as abnormal and are from the transition phase, where anomaly is about to happen. It is still



(a) EXPLOSION011 (b) ROADACCIDENT133

Figure 7: Abnormal Frames where both DRO-DKMIL as well as DK-MMIL fails.

meaningful that the predictions made by DRO-DKMIL and DK-MMIL as those frames look more normal. The wrong prediction is caused by the inconsistency with the frame labels assigned by humans that may not truly reflect the actual situation. Therefore, these examples also reveal some inherent challenges of video anomaly detection even for humans.



(a) Anomaly One (b) Anomaly Two

Figure 8: Abnormal Frames from Avenue Test09: (a) Correct-DRO-DKMIL, DKMMIL; (b) Correct-DRO-DKMIL, Incorrect-DK-MMIL

We show the robustness of the proposed DRO-DKMIL approach on a multimodal scenario in a real-world dataset. As multi-modality exists explicitly (with available ground truth) in the Avenue dataset, we perform a qualitative analysis on this. Specifically, we set the FPR = 23% and look at the FNR in both DRO-DKMIL as well as DK-MMIL. We particularly look for the abnormal frames that are classified as normal because of the multimodality. Figure 8 shows two frames taken from video Avenue Test09. As multimodality occurs in the same video, both approaches correctly predict the first type of anomaly present in Figure 8(a) but not the one shown in 8 (b). Only DRO-DKMIL can correctly predict the abnormal frame shown in Figure 8 (b). DK-MMIL may fail to detect the multiple types of abnormal frames in a multimodal scenario during the training process. As a result, it may only be able to detect a similar abnormality type presented in Figure 8 (a) but not the one shown in Figure 8(b) resulting in misclassification. In contrast, as DRO-DKMIL allows the participation of multiple anomaly types during training, it can detect similar frames as that of Figure 8 (a) and (b).

D.2 Computational Complexity Analysis

The potential high computational cost of the GP component has been significantly reduced through the inducing points and key optimization tricks, such as Kronecker decomposition. In particular, the update of $q(\mathbf{U})$ involves sampling of $\mathbf{u}$, which can be efficiently performed through fast Kronecker matrix-vector product with a cost of $O(m^{(1+\frac{1}{D})})$, where m is the number of inducing points and D is the dimension of input to the GP. In addition, it also involves computing the gradient of $KL(q(\mathbf{U}||p(\mathbf{U}))$ with respect to $\mathbf{L}$ and $\boldsymbol{\mu}$ with a total cost of $O(Dm^{3/D})$ through Kronecker decomposition. Update of $q(\mathbf{Z})$ using Eqs. 12 and 13 has a cost of $O(n)$, where n is the number of instances in a bag. Update of $\boldsymbol{\pi}$ involves solving a constrained convex optimization problem with a cost of $O(n^3)$. Finally, update of the mixing coefficients A using the sampled $\mathbf{u}$ and $\mathbf{z}$ has a cost of $O(J)$, where J is the number of GPs. So, the overall complexity of updating the GP related parameters per iteration is $O(m^{(1+\frac{1}{D})} + Dm^{3/D} + n^3)$. It is also worth to note that both D and n are typically much smaller than m. The parameters of the deep neural network (DNN) are updated using standard backpropagation as described at the end of Section C in the Appendix so its cost is similar for training a DNN.

D.3 Link to Source Code

For the source code of this paper, please click here.