RobustA: robust anomaly detection in multimodal data

RobustA: Robust Anomaly Detection in Multimodal Data

Salem AlMarri[1†], Muhammad Irzam Liaqat[2†], Muhammad Zaigham Zaheer[1], Shah Nawaz[3], Karthik Nandakumar[1], Markus Schedl[3,4]

[1]Mohamed Bin Zayed University of Artificial Intelligence, [2]IMT School for Advanced Studies, [3]Johannes Kepler University Linz, [4]Human-centered AI Group, AI Lab, Linz Institute of Technology

***Abstract*— In recent years, multimodal anomaly detection methods have demonstrated remarkable performance improvements over video-only models. However, real-world multimodal data is often corrupted due to unforeseen environmental distortions. In this paper, we present the first-of-its-kind work that comprehensively investigates the adverse effects of corrupted modalities on multimodal anomaly detection task. To streamline this work, we propose `RobustA`, a carefully curated evaluation dataset to systematically observe the impacts of audio and visual corruptions on the overall effectiveness of anomaly detection systems. Furthermore, we propose a multimodal anomaly detection method, which shows notable resilience against corrupted modalities. The proposed method learns a shared representation space for different modalities and employs a dynamic weighting scheme during inference based on the estimated level of corruption. Our work represents a significant step forward in enabling the real-world application of multimodal anomaly detection, addressing situations where the likely events of modality corruptions occur. The proposed evaluation dataset with corrupted modalities and respective extracted features will be made publicly available.**

***Index Terms*—Anomaly detection, Audio-visual modalities, Corrupted and missing modalities**

I. Introduction

VIDEO anomaly detection (VAD) is a challenging computer vision task with various real-world applications including surveillance, autonomous navigation, packaging, and biomedical imaging [1]. Generally, VAD methods aim to predict high anomaly scores for the frames in a video that deviate significantly from the norm, where the application context determines the norm. For example, events such as shoplifting or violence can be considered anomalies in the CCTV surveillance context [2]. Since it is laborious to obtain fine-grained (pixel-level or frame-level) labels of anomalies in videos, a weakly-supervised VAD (WS-VAD) setting that learns to detect anomalous frames using only video-level binary labels is typically used. In WS-VAD, a training video is labeled as normal if no anomalous event is present, whereas it is labeled as an anomaly if any anomalous event is present [3]–[5]. In recent years, WS-VAD has gained considerable popularity and became a mainstream VAD paradigm [3], [6], [7]. More recently, WS-VAD has been transformed into a multimodal learning task by including other informative modalities such as audio to discriminate and locate events better [2], [8]. For instance, it is challenging to rely only on visual signals in shaky videos covering an explosion event. However, if accompanied by the audio modality, the event can be easily classified. Prior works have successfully shown the effectiveness of using video and audio pairs to improve the overall performance of WS-VAD [2], [9], [10].

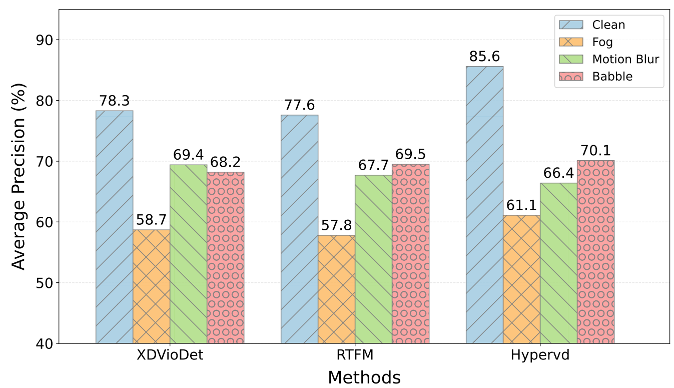


Fig. 1. Existing multimodal anomaly detection approaches [2], [11], [12] are not robust when subjected to corruptions in either modality, including vision (e.g., fog and motion blur) or audio (e.g., babble).

Despite their many advantages, deep neural network models are vulnerable to typical corruptions in the input data (e.g., degradations, distortions, and disturbances caused by weather changes, system errors, etc.) [13]. Though multimodal learning systems achieve better performance on multimodal data [14] compared to their unimodal counterparts, they are also susceptible to such vulnerabilities. Multimodal VAD systems, in particular, are prone to such issues (see Fig. 1), as they are often subjected to extreme environmental conditions [15]. For this reason, it is essential to design methods that can withstand corrupted data during deployment. While it may be possible to achieve this goal by including corrupted data during the training process, *this work aims to evaluate the robustness of multimodal VAD systems trained only on clean data, when encountering corrupted data during deployment*. Existing literature lacks a detailed study and appropriate dataset that investigates the impact of corrupted modalities on WS-VAD and multimodal VAD systems.

In this work, we empirically investigate this problem with a proposed dataset, `RobustA`, specifically to study this issue. We observe that existing multimodal anomaly detection models deteriorate dramatically under corrupted data. For example, in Fig. 1, the performance of XDVioDet [2] drops from 78.36% to 58.70% when exposed to fog in visual modality during testing. To alleviate this issue, we propose a robust anomaly detection method that is resilient against corrupted modalities. The proposed method independently

† Equal contribution

maps multiple modalities into a shared representation space. This shared space can be particularly beneficial in scenarios where one modality is compromised. For example, if one modality becomes distorted, inaccurate, noisy, or is entirely missing, the shared space can leverage information from the remaining modalities to compensate for the loss. Results reported in Section V-B suggest that our approach not only yields better multimodal performance but also demonstrates resilience against corruptions.

The key contributions of our work are as follows:

- We present `RobustA`, a first-of-its-kind, carefully curated dataset comprising 8 visual corruptions and 8 audio corruptions, enabling evaluation of the multimodal anomaly detection methods against various types and levels of modality corruptions.
- We propose an anomaly detection approach that independently maps multiple modalities to learn representations in a shared space while dynamically adjusting the weights of individual modalities to reduce the impact of the compromised modality towards multimodal anomaly detection.
- We study the plug-and-play nature of our proposed approach, demonstrating its resilience against compromised modalities when used in conjunction with existing approaches.

II. Related Work

A. Weakly Supervised Video Anomaly Detection and Multimodality

Weakly Supervised Video Anomaly Detection (WS-VAD) refers to the process of identifying abnormal events in a given video while the training is carried out using only video-level binary labels. The problem was first introduced by Sultani et al. [3] where the authors utilized multi-instance learning based ranking to carry out the overall training. The research was advanced by several researchers [16]–[18]. However, most of these studies are focused on unimodal (video-only) training and testing.

More recently, anomaly detection task is transformed into multimodal learning where more than one modalities (usually vision and audio) are present during training and testing. For example, XDVioDet [2] utilized a multi-branch neural network comprising holistic, localized, and score branches to effectively leverage multimodal feature representations. Similarly, HyperVD [11] proposed a method using two graph-based hyperbolic convolutional networks for spatio-temporal feature extraction from multimodal input. While the multimodal anomaly detection task has gained popularity due to superior performance compared to the unimodal counterpart approaches, existing literature lacks rigorous studies understanding the adverse effects of compromised data on multimodal learning systems. As, CCTV surveillance cameras equipped with anomaly detection systems are highly prone to environmental corruptions, it is pertinent to explore this novel research direction.

B. Addressing Compromised Modalities

Corrupted Modalities. Data collected from diverse multiple sources may contain corrupted samples due to various factors, such as sensor failure, obstructions in the video stream, noise in audio signals, data storage errors, and many more. Recent years have seen an increased interest in investigating the vulnerability of deep models against modality corruptions [19]–[23]. For example, Hendrycks et al. [24] built benchmarks to evaluate the performance of various Convolutional Network Networks on the image classification task. Similarly, Yi et al. [15] curated robust video classification benchmark to evaluate state-of-the-art Convolutional Networks and Transformers against video corruptions. The growing interest has also broadened to multimodal learning where handling compromised modalities is critical to improve performance and robustness. For example, Beemelmanns et al. [20] introduces MultiCorrupt framework to evaluate the robustness of multimodal 3D object detectors against various corruption categories. Similarly, Hong et al. [25] proposed a multimodal input corruption modeling to develop robust audio-visual speech recognition models. These studies demonstrate that the performance of unimodal or multimodal methods deteriorate dramatically when a modality is corrupted.

Missing Modalities Multimodal learning has shown remarkable performance improvements over unimodal methods. However, such methods exhibit deteriorated performances if one or more modalities are missing [26]–[29]. Considering the significance of multimodal learning, recent years have seen an increased interest in studies addressing missing modalities for various tasks including classification, such as [30]–[32].

While these methods have notably enhanced model robustness in addressing compromised modalities across various application areas, to the best of our knowledge, there remains a gap in the literature for the multimodal anomaly detection task. In this work, we investigated the robustness of multimodal anomaly detection under a carefully crafted dataset of compromised modalities. Moreover, we propose a robust approach that maintains superior performance when faced with extreme missing and corrupted modalities.

III. Methodology

A. Background and Overview

Existing methods generally concatenate audio and visual embeddings to learn fused representations for the multimodal anomaly detection task [2], [9], [11], [33]. While effective under modality-complete settings, these methods rely heavily on the complete availability of modalities and suffer dramatic performance deterioration when a modality is corrupted or missing. Since surveillance data may involve corruptions due to weather conditions, connection latency, broken/tempered equipment, etc., a lack of robustness against such scenarios may deter the deployment of these methods in real-world applications. In this work, we develop an approach to *learn multimodal representations by mapping audio and visual modalities independently into a common feature space*. This enables the model to maintain *robustness when a modality is*

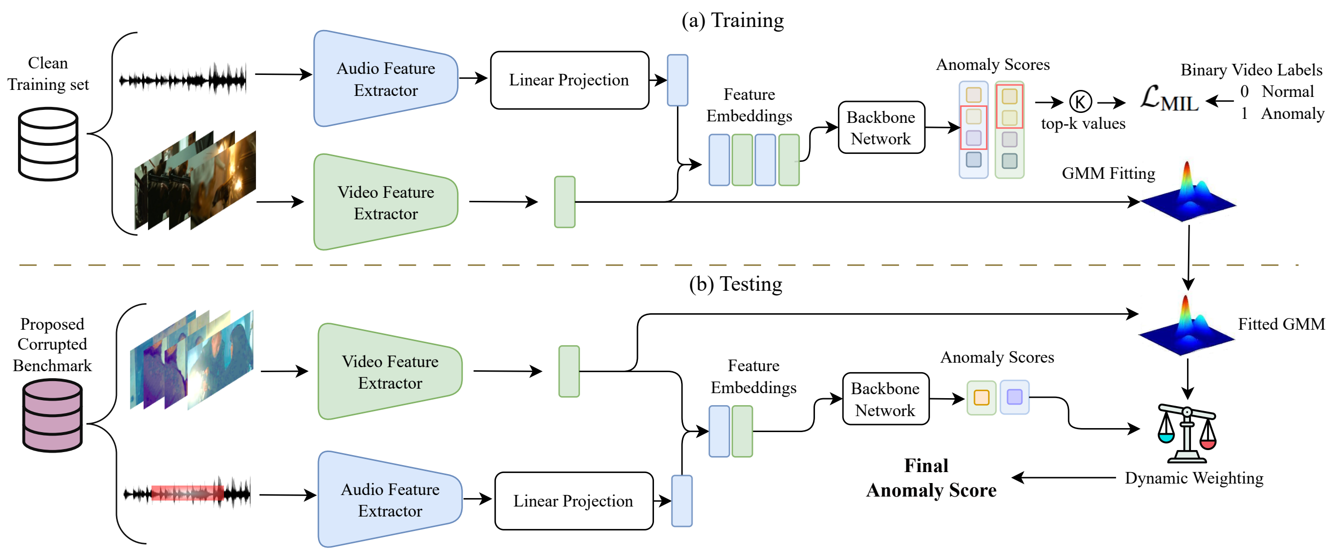


Fig. 2. Architecture of our approach. Modality-specific features are extracted using pre-trained audio and visual encoders. A linear projection is used to match the feature dimensions. Modality embeddings are independently mapped to learn representations in a shared space. During inference, the shared learning space helps mitigate the adverse effects of modality corruptions. Moreover, a dynamic weighting scheme is utilized to adjust the weights of the corrupted modality for better anomaly detection.

compromised by leveraging a shared representation of cross-modal interactions, effectively compensating for missing or corrupted modality data. The proposed approach is illustrated in Fig. 2 and the details are discussed next.

B. Preliminaries

Given a dataset of n videos, each video V is preprocessed into m non-overlapping audio and visual segments and passed to a pre-trained feature extractor. The audio and video feature extractors generate the feature embeddings $E^A \in \mathbb{R}^{m\times d_A}$ and $E^V \in \mathbb{R}^{m\times d_V}$ for each video, respectively with the feature dimensions are denoted as d_A and d_V. The class label $y \in \{1, 0\}$ indicates the presence (1) or absence (0) of anomalous events in the video. The goal is to learn an anomaly detector $\mathcal{A}_\theta(E^*)$ that takes a feature vector (audio or visual) as input and predicts an anomaly score in the range of 0 and 1.

C. Learning a Shared Representation Space

Prior multimodal anomaly detection methods rely on concatenation of audio and visual representations and learning the anomaly detector based on the concatenated representation. Given the audio and visual feature embeddings E^A and E^V for n training videos in a dataset, Eq. 1 outlines the traditional fusion-based learning approach utilizing Multi-Instance Learning (MIL) loss [2]:

$$\mathcal{L}_{\text{total}} = \sum_{i=1}^{n} \mathcal{L}_{\text{MIL}}(\mathcal{A}_\theta(E^A||E^V), y_i), \tag{1}$$

where $E^A||E^V$ is the fused multimodal representation obtained by concatenation of E^V and E^A, allowing the architecture to capture the inter-modality relationship. Consequently, the success of these methods is highly dependent on the availability of both modalities during inference. To alleviate this dependence on modality completeness, we propose a method that learns multimodal representations by mapping each modality independently into a shared space.

Specifically, we map individual modality independently into a common representation space, enabling the model to capture inter- and intra-modality relationships through cross-modal data interactions. As a result, when a modality is compromised, the model compensates by leveraging the shared representation space. Thus, Eq. 1 needs to be modified as follows:

$$\mathcal{L}_{\text{total}} = \sum_{i=1}^{2n} \mathcal{L}_{\text{MIL}}(\mathcal{A}_\theta(E_i^*), y_i) \tag{2}$$

where E^* is randomly sampled from either E^V or E^A. However, in existing methods [2], depending on the choice of feature extractor, the output dimensions of E^V and E^A are usually inconsistent. To address this issue, we introduce a linear projection module $\mathcal{P}_\phi$ that transforms E^A to a higher dimensional space, matching the dimension of E^V. Thus, E^* in Eq. (2) will be randomly sampled from either E^V or $\mathcal{P}_\phi(E^A)$. The parameters θ of the anomaly detector and ϕ of the projection module are learned by minimizing the total loss $\mathcal{L}_{\text{total}}$. Intuitively, the proposed approach attempts to learn a single modality-agnostic anomaly detector, which implicitly forces the two modalities to share a common representation space.

D. Dynamic Weighting During Inference

As the anomaly detector is agnostic to the input modality, it is possible to obtain an anomaly score using both audio and visual features during inference. Hence, a strategy to optimally combine individual anomaly scores is needed to achieve good multimodal performance. A straightforward approach would be to compute the weighted average of the anomaly scores produced by the anomaly detector $\mathcal{A}$ for each modality in a late fusion manner. For instance, the final anomaly score S for a given video segment based on audio-visual features can be obtained as:

$$S = \lambda_A \mathcal{A}(\overline{E}^A) + \lambda_V \mathcal{A}(\overline{E}^V), \tag{3}$$

where $\overline{E}^A$ and $\overline{E}^V$ are the audio and visual feature embeddings of the test sample, respectively, which may also contain a compromised modality. Here, λ_A and λ_V are the weights assigned to the audio and visual modalities, respectively, and these weights are linearly normalized such that they add up to 1. In a naive approach, λ_A and λ_V can be set to 0.5 giving equal weightage to both modalities. While, as shown in our experiments, this approach works reasonably well due to the shared representation space mitigating the impacts of an individual compromised modality, a more sophisticated approach may be required to handle extreme corruption in one of the modalities.

Since the training data does not have prior information about the compromised modality (since training is performed only on clean data), we devise a dynamic weighting scheme based on the clean data distribution. Specifically, we fit a Gaussian Mixture Model (GMM) to the clean (non-corrupted) training visual features and evaluate if the test data matches with this clean data distribution. Let $\mathcal{G}^{\mathcal{V}}$ denote a K-component GMM with parameters $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$ representing the clean distribution of visual features, where π_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$ denote the mixture probability, mean, and covariance matrix, respectively, of the k-th Gaussian component.

The negative log-likelihood ℓ of the test features ($\overline{E}^V$) is then computed using the fitted GMM as follows:

$$\ell(\overline{E}^V) \;=\; -\log\left[\sum_{k=1}^{k} \pi_k \, \mathcal{N}\Big(\overline{E}^V \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\Big)\right] \tag{4}$$

where $\mathcal{N}$ denotes the likelihood of observing the feature vector $\overline{E}^V$ under the k^{th} Gaussian component of the mixture model. A higher ℓ value indicates that the sample is more likely to be compromised, while a lower value of ℓ suggests the sample is clean. Finally, λ_V is computed using a sigmoid function as:

$$\lambda_V \;=\; \frac{0.5}{1 \;+\; \exp\Big(c\big[\,\ell(\overline{E}^V) + x_0\big]\Big)}, \tag{5}$$

where c and x_0 are the scale and shift hyperparameters that shape the sigmoid function, respectively. The above equation maps the negative log-likelihood $\ell\left(\bar{E}^V\right)$ to a value of λ_V between 0 and 0.5. Similarly, the value of λ_A can also be computed and the two weights can be linearly normalized so that they sum to 1. Extensive ablation studies on the proposed dynamic weighting scheme are provided in Section VI-A, which demonstrate the effectiveness of our approach. When one of the modalities is completely missing, its corresponding weight is set to zero because the anomaly score for that modality cannot be computed.

IV. Modality Corruptions

In this work, we propose a carefully curated dataset, `RobustA`, to evaluate the critical impacts of audio/visual corruptions on the overall performance of multimodal anomaly detection systems. To the best of our knowledge, `RobustA` is the first rigorous attempt to study anomaly detection in such scenarios. We leverage XD-Violence multimodal anomaly detection dataset as the base to create several corruption scenarios, by varying the type and severity level, for both audio and visual modalities. We choose these modalities because of their globally standard availability in the existing real-world CCTV systems, while other modalities, such as text, lack practical application. Examples of some visual corruptions and audio corruption mel spectograms [34], [35] are provided in Figure 3.

Types of Corruptions in Visual Modality. We consider the following corruptions in the visual modality: Bit error, brightness, contrast, fog, rain, motion blur, saturation, and shot noise. Details about each corruption and its implementation are provided in the Supplementary.

Types of Corruptions in Audio Modality. We consider the following corruptions in the audio modality: Overlay, bit error, pitch shift, random dropout, and reverb. Details about each corruption and the implementation details are provided in the Supplementary.

Number of Corrupted Samples. We consider varying numbers of corrupted modality samples. In particular, we consider 0%, 10%, 30%, 50%, 70%, 90%, and 100% of samples in each modality as corrupted.

V. Experiments

A. Implementation Details

Feature Extraction: For feature extraction of visual modality, we employ ResNet-I3D model pretrained on Kinetics-400 dataset. Following prior work [12], we perform 10-crop augmentation by using a 16 frame sliding window with a sample rate of 24 FPS. The audio embeddings are extracted by using the VGGish network [36] pre-trained on large-scale YouTube dataset. Moreover, the extraction is performed by pre-processing the audio signal into 960 ms of overlapping segments and setting 96×64 bin size for the mel-spectrogram.

Training Details: We trained the proposed architecture on Nvidia RTX $A100$ for 50 epochs using Adam optimizer with initial learning rate of 10^{-5} and dynamic adjustment using multi-step cosine scheduler. We use a batch size of 640 and weight decay of 0.00001.

Backbone Network: We adopt XDVioDet [2] as the primary backbone network in our experiments (also referred to as baseline hereafter) and provide extensive comparisons with it. In addition, to explore the general applicability of our proposed approach, we provide additional experiments on two other models as backbone networks, including RTFM [12] and HyperVD [11]. Nevertheless, unless specified otherwise, the results are reported with XDVioDet as the primary backbone network.

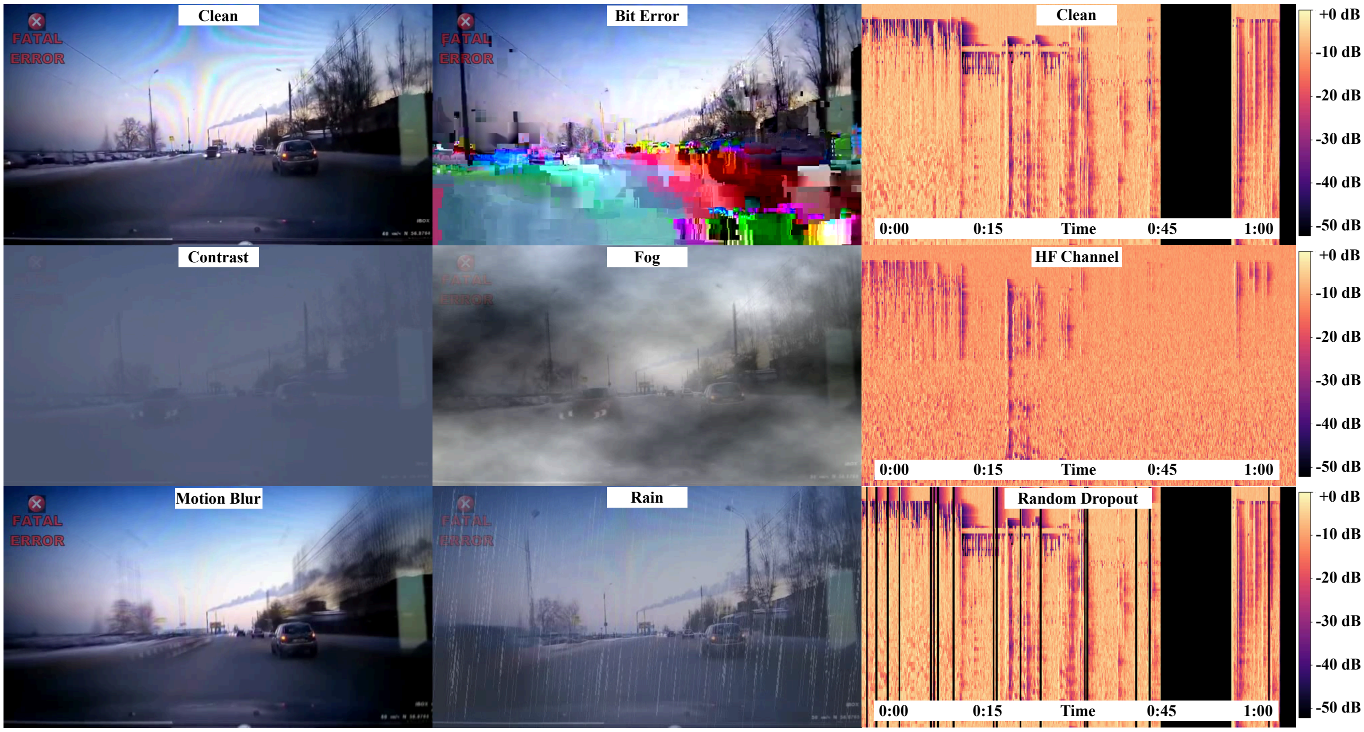


Fig. 3. Examples of a few visual and audio corruptions demonstrating the challenging scenarios presented in our proposed benchmark.

Dataset: We use XD-Violence multimodal anomaly detection dataset [2] as a stepping stone for our dataset. More specifically, we apply corruptions on the test set of the XD-Violence dataset based on the details provided in Section IV to create `RobustA`. Similar to the original dataset, the train and test splits include $3,954$ and 800 anomalous and normal videos, with varying levels of corruptions based on the experiment. Following prior works [2], [11], we use Average Precision (AP) as the evaluation metric.

B. Experimental Results

1) Robustness Against Compromised Modalities: Table I provides extensive performance evaluation of our method and the baseline [2] under various types of audio and visual corruptions. It may be noted that the training set remains clean, as our goal is to achieve robustness against corruptions when trained with clean data.

Visual Corruptions: During testing, various types of corruptions, such as bit error, fog, rain etc., are present in the visual modality (as outlined in Section IV). We evaluate the performance of our method against the baseline on varying levels of corruption, which reflect the proportion of corrupted video in the test set. A corruption level of 0% means no corrupted video, while a level of 100% indicates that all test videos are corrupted. The results (Table I) demonstrate that our method outperformed the baseline against various types and levels of corruptions. Notably, for extreme cases of corruptions such as contrast, our method achieves performance scores of 80.00% and 69.21%, whereas the baseline attains scores of 78.36% and 52.47% under 0% and 100% of corrupted videos. Although the performance of our method and the baseline method is comparable under 0% corruption, our method notably outperforms it under 100% corruption, highlighting its robustness against visual modality corruption.

Audio Corruptions. During testing, various types of corruptions, such as babble, bitrate, and hfchannel, are present in the audio modality (as described in Section IV). Consistent with the trends observed in the visual corruption section, we find similar patterns for the audio modality. The results highlight that our method is resilient against various levels of audio corruptions.

Missing Modalities. While corruptions interrupt the overall quality of a modality, some information may still be retained, which might be useful for the network to perform the predictions. To take the challenge up a notch, in this section, we consider the missing modality setting in which a modality is entirely removed. This may be considered as an extreme case where no information regarding the corrupted modality is available to the network. Although the performances of our method (80.00%) and the baseline (78.30%) are comparable when modalities are completely available, our method notably outperforms the baseline method under missing modalities. Notably, our method achieves performance of 67.15% and 72.55%, whereas the baseline method attains scores of 49.50% and 69.10% when removing visual and audio modality, respectively, highlighting the robustness of our approach.

Mixed Corruptions. As real-world environments may have several distortions present, we also explore the effects of the modalities compromised by mixed corruptions. To carry out this experiment, for each level of corruption, it is ensured that the test set contains all corruption types accounting for the total percentage of the corruption level. Performance trends similar

TABLE I
COMPARISON OF OUR APPROACH AND BASELINE UNDER VARIOUS CORRUPTION TYPES AND CORRUPTION LEVELS (PERCENTAGES). AP(%) IS USED AS THE EVALUATION METRIC (HIGHER IS BETTER). BOLD REPRESENTS BEST PERFORMANCE IN EACH COMPARISON.

Corruption Levels		0%		10%		30%		50%		70%		90%		100%	
Corruption	**Modality**	[2]	Ours	[2]	Ours	[2]	Ours	[2]	Ours	[2]	Ours	[2]	Ours	[2]	Ours
Bit Error	**Visual**	78.36	**80.00**	77.41	**79.19**	72.57	**76.67**	72.48	**76.35**	68.21	**74.70**	66.61	**74.78**	65.55	**74.45**
Brightness	**Visual**	78.36	**80.00**	77.39	**79.30**	76.35	**77.74**	75.54	**77.02**	75.06	**76.23**	74.34	**75.42**	73.76	**74.72**
Contrast	**Visual**	78.36	**80.00**	77.08	**78.67**	69.84	**75.75**	66.82	**74.44**	59.85	**71.86**	54.54	**70.35**	52.47	**69.21**
Fog	**Visual**	78.36	**80.00**	76.45	**78.05**	69.21	**73.47**	68.87	**72.87**	62.32	**69.68**	60.48	**69.26**	58.70	**67.97**
Rain	**Visual**	78.36	**80.00**	77.73	**79.52**	74.27	**77.66**	72.88	**77.16**	70.94	**75.54**	69.43	**74.56**	68.69	**73.89**
Motion Blur	**Visual**	78.36	**80.00**	77.22	**78.73**	74.45	**77.02**	74.03	**76.46**	71.43	**74.55**	70.28	**73.80**	69.46	**73.08**
Saturate	**Visual**	78.36	**80.00**	77.02	**78.84**	73.64	**77.14**	69.94	**76.10**	68.20	**75.19**	65.57	**74.26**	65.23	**73.86**
Shot Noise	**Visual**	78.36	**80.00**	77.69	**79.10**	74.65	**76.08**	74.60	**75.62**	72.35	**73.62**	71.40	**72.72**	70.99	**71.68**
Average	**Visual**	78.36	**80.00**	77.25	**78.93**	73.12	**76.44**	71.90	**75.75**	68.54	**73.92**	66.58	**73.14**	65.61	**72.36**
Babble	**Audio**	78.36	**80.00**	77.16	**79.06**	75.18	**77.64**	72.65	**74.36**	70.94	**73.29**	68.57	**70.96**	68.21	**70.45**
Bitrate	**Audio**	78.36	**80.00**	77.03	**78.45**	75.60	**76.66**	**72.76**	72.37	**71.12**	70.21	**68.67**	65.48	**68.31**	65.11
HF Channel	**Audio**	78.36	**80.00**	76.37	**79.01**	73.73	**77.77**	69.36	**73.24**	67.59	**72.09**	65.35	**69.78**	65.56	**69.47**
Pink	**Audio**	78.36	**80.00**	76.32	**79.02**	73.45	**77.94**	68.97	**73.63**	67.30	**72.67**	64.99	**70.85**	65.36	**70.60**
Pitch Shift	**Audio**	78.36	**80.00**	76.24	**78.87**	73.49	**77.62**	69.53	**72.83**	67.01	**71.23**	63.66	**66.91**	63.07	**66.39**
Random Dropout	**Audio**	78.36	**80.00**	76.45	**78.93**	74.22	**77.70**	69.80	**73.05**	67.79	**71.81**	64.01	**66.92**	63.53	**66.48**
Reverb	**Audio**	78.36	**80.00**	76.65	**79.01**	75.02	**77.78**	71.99	**73.23**	70.01	**71.90**	67.01	**70.02**	66.84	**69.73**
White	**Audio**	78.36	**80.00**	76.48	**78.95**	73.69	**77.81**	69.24	**73.48**	67.88	**72.36**	66.02	**70.43**	66.60	**70.21**
Average	**Audio**	78.36	**80.00**	76.59	**78.91**	74.30	**77.74**	70.41	**73.27**	68.71	**71.94**	66.04	**68.92**	65.94	**68.56**
Missing Visual	**Visual**	78.30	**80.00**	75.10	**76.94**	69.50	**71.87**	63.10	**71.18**	58.20	**68.61**	52.10	**68.00**	49.50	**67.15**
Missing Audio	**Audio**	78.36	**80.00**	77.30	**78.86**	75.20	**77.96**	73.90	**74.75**	72.50	**73.91**	70.00	**72.74**	69.10	**72.55**
Mixed Corruptions	**Visual**	78.36	**80.00**	77.70	**79.54**	73.93	**78.15**	73.75	**78.65**	72.75	**77.35**	72.09	**77.15**	71.45	**76.27**
Mixed Corruptions	**Audio**	78.36	**80.00**	76.94	**79.09**	74.77	**77.62**	72.40	**74.13**	70.48	**72.99**	67.71	**70.50**	67.14	**69.97**
Motion Blur+Babble	**Audio+Visual**	78.36	**80.00**	75.71	**77.59**	70.91	**73.54**	67.19	**69.03**	61.94	**64.22**	56.82	**60.13**	54.66	**58.04**
Brightness+Rand. Dropout	**Audio+Visual**	78.36	**80.00**	75.42	**78.08**	71.66	**74.49**	65.20	**67.83**	61.62	**65.11**	56.05	**56.94**	54.19	**55.12**
Bit Error+HF-Channel Noise	**Audio+Visual**	78.36	**80.00**	74.56	**78.07**	69.43	**74.29**	62.89	**70.27**	55.53	**65.87**	47.73	**60.60**	45.50	**59.47**

TABLE II
COMPARISON OF BASELINE VS. PROPOSED WHEN ONE MODALITY IS COMPLETELY MISSING AND THE OTHER IS CORRUPTED. HIGHER ACCURACY IS BETTER.

Corruption Levels		0%		30%		70%		100%	
Corruption	**Mod**	[2]	**Ours**	[2]	**Ours**	[2]	**Ours**	[2]	**Ours**
bit_error	V	69.17	**73.82**	63.76	**67.31**	61.57	**64.70**	59.23	**63.42**
brightness	V	69.17	**73.82**	65.64	**70.27**	63.82	**68.76**	60.84	**66.45**
contrast	V	69.17	**73.82**	59.51	**65.79**	48.12	**55.83**	38.65	**49.07**
fog	V	69.17	**73.82**	56.97	**60.76**	52.48	**54.49**	50.25	**50.41**
rain	V	69.17	**73.82**	63.89	**69.15**	59.60	**64.49**	58.11	**60.33**
motion_blur	V	69.17	**73.82**	63.21	**68.83**	58.21	**64.79**	54.61	**62.29**
saturate	V	69.17	**73.82**	63.27	**68.94**	57.61	**64.61**	54.16	**62.22**
shot_noise	V	69.17	**73.82**	65.11	**67.97**	**64.65**	63.23	**65.44**	59.28
average	V	69.17	**73.82**	62.67	**67.38**	58.26	**62.61**	55.16	**59.18**
babble	A	49.50	**67.65**	41.42	**59.27**	32.94	**42.12**	28.47	**31.47**
bitrate	A	49.50	**67.65**	43.10	**56.95**	34.36	**40.57**	28.74	**32.11**
hfchannel	A	49.50	**67.65**	40.92	**58.91**	29.05	**36.21**	**24.55**	24.01
pink	A	49.50	**67.65**	41.92	**59.12**	28.86	**36.80**	**25.93**	24.57
pitch_shift	A	49.50	**67.65**	36.68	**57.96**	26.11	**34.90**	22.23	**23.27**
dropout	A	49.50	**67.65**	38.74	**58.46**	27.53	**37.90**	23.62	**27.20**
reverb	A	49.50	**67.65**	40.07	**58.47**	30.83	**37.22**	26.09	**26.39**
white	A	49.50	**67.65**	42.38	**58.90**	29.77	**35.85**	**24.40**	23.40
average	A	49.50	**67.65**	40.65	**58.51**	29.93	**37.70**	25.50	**26.55**

TABLE III
PLUG-AND-PLAY PERFORMANCE OF OUR APPROACH FOR MIXED AUDIO AND VISUAL DISTORTIONS AT 0% AND 100% LEVELS OF CORRUPTION ON MULTIPLE BASELINES INCLUDING XDVIODET, HYPERVD, AND RTFM. BOLD: BEST PERFORMANCE.

Corrupted Modality	**Visual**		**Audio**	
	0%	**100%**	**0%**	**100%**
XDVioDet (Baseline)	78.36	71.45	78.36	67.14
XDVioDet (Ours)	**80.00**	**76.27**	**80.00**	**69.97**
HyperVD (Baseline)	85.60	73.28	85.60	76.90
HyperVD (Ours)	**85.90**	**77.29**	**85.90**	**78.18**
RTFM (Baseline)	**77.66**	70.43	**77.66**	**77.44**
RTFM (Ours)	77.30	**72.67**	77.30	74.21

to the previously discussed experiments are observed, where our approach outperforms the baseline in all compared cases, signifying the importance of our method against modality corruptions in real-world settings.

Audio and Visual Corruptions. Since real-world environments often contain multiple distortions at the same time, we investigate their impact on both modalities. Consistent with prior experiments, the corruptions cause significant drops in the performance of the baseline, whereas our approach demonstrates better resilience.

2) Unimodal Testing with Corruptions: In a series of experiments, we explore a use case on our method in which one modality is completely missing while the other modality is corrupted. The intuition behind this experiment is to observe the importance of each modality for training and testing a robust multimodal anomaly detection system. Table II summarizes the results obtained during these experiments. It may be seen that the video modality is notably dominant than the audio modality. In the case of 0% corruption cases, the presence of video modality outperforms the presence of audio modality (AP of 73.82% vs. 67.65%). When the noise is added to the visual modality, a notable performance drop is observed in each corruption case (an average drop in AP from 73.80% to 59.18%). However, the performance drop increases dramatically when the system is exposed to 100% missing visual modality while adding noise to the audio modality (an average drop in AP from 67.65% to 26.55%). This series of experiments highlights that while audio modality plays an important role in the overall performance, the visual modality is generally more informative, and the performance deteriorates dramatically if the visual modality is compromised.

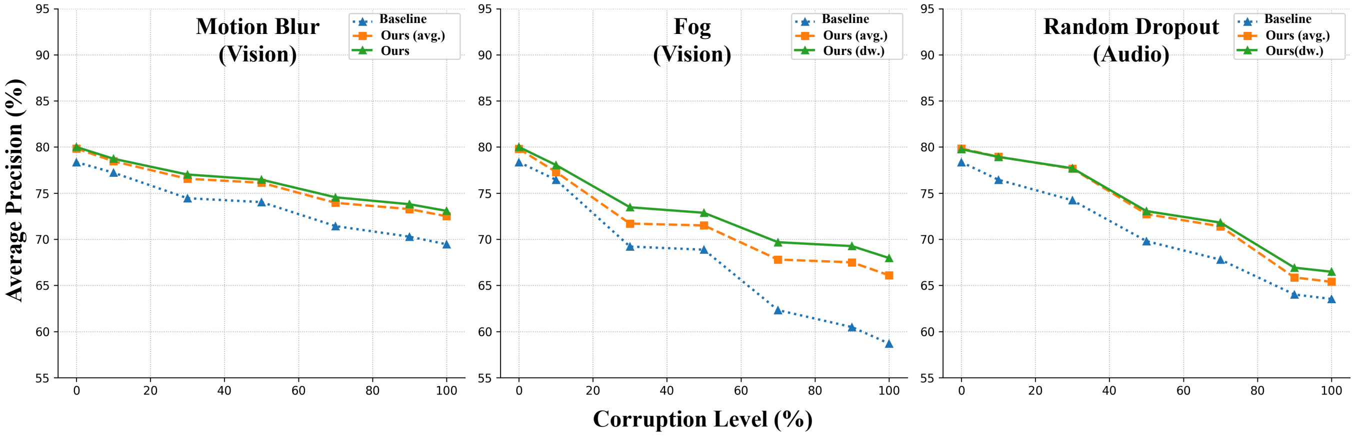


Fig. 4. Comparison of weighting schemes (average and dynamic) with baseline concatenation approach [2] on two visual and one audio corruption.

VI. Analysis and Discussion

A. Importance of Dynamic Weighting

During training, our method processes each modality independently, while at test time, we employ a dynamic weighting strategy (Section III-D) to effectively combine the predictions of individual modalities for multimodal performance. In this section, we compare the empirical performance of three approaches: the naive averaging approach (as chalked in Eq. 3), dynamic weighting (as chalked in Eq. 5), and the concatenation-based inference (as proposed by the baseline XDVioDet [2]). Results on two visual corruptions (motion blur and Fog) and one audio Corruption (Random Dropout) provided in Fig. 4 show that the naive averaging approach with our method outperforms the baseline in all cases, demonstrated the importance of shared space representation learning. Furthermore, the best results are achieved with the proposed dynamic weighting approach.

B. Is Our Approach Plug-and-Play?

The components of our approach, such as projection layer, shared representation learning, and dynamic weighting, can be plugged into existing multimodal anomaly detection methods, including XDVioDet [2], HyperVD [11], and RTFM [12]. These methods usually incorporate multimodal information by concatenating audio-visual modalities at feature-level input, making them prone to corrupted modalities during testing. As seen in Table III, when the components of our approach are added to XDVioDet and HyperVD, it outperforms the baseline in both cases of mixed corrupted modalities by a notable margin. In the case of RTFM as baseline, our method outperforms in the case of visual modality. However, in the case of corrupted audio modality, RTFM baseline performs better. This may be attributed to the projection layer used in our approach that upscale the audio features to match the dimensions of visual features. In contrast, RTFM by default uses the visual feature dimension of 2048, whereas the dimension of added audio features is 128.

C. Does Multimodal Robustness Translate to Better Zero-shot Performance

While the proposed shared space learning for multimodal data improves robustness against compromised modality, we explore its effectiveness in zero-shot setting. To this end, we utilize the models trained on XD-Violence dataset and evaluate on the test set of a benchmark anomaly detection dataset, UCF-crime [3]. As UCF-Crime is a visual-only dataset with no audio, this experiment poses two challenges: 1) Generalizability across datasets. 2) Missing audio data during testing. The results are summarized in Table IV. As seen, our method outperformed RTFM and XDVioDet baselines as well as existing unsupervised video anomaly detection methods [18], [33], [37].

D. Is Linear Projection Necessary?

In this section, we explore the importance of the linear projection layer used in our approach. To this end, we replace it with zero-padding to match the embedding dimension of visual and audio modalities. The results are reported in Table V. As seen, the padding demonstrates lower performance than even the baseline, whereas our design choice of using projection layer notably helps achieve better performance. This demonstrates that projection layer complements the proposed shared space representation learning for robust anomaly detection under multimodal setting, as proposed in our approach.

E. Qualitative Results

Fig. 5 shows anomaly score plots of `RobustA` and the baseline on different videos highlighting cases with and without corruption. As seen, both `RobustA` and the baseline demonstrate reasonable anomaly scores when neither of the modalities is compromised. The performance of the baseline drops notably when modalities are corrupted, while `RobustA` retains the performance.

VII. Conclusion

In this paper, we presented a comprehensive study to investigate the adverse effects of corrupted modalities on multimodal

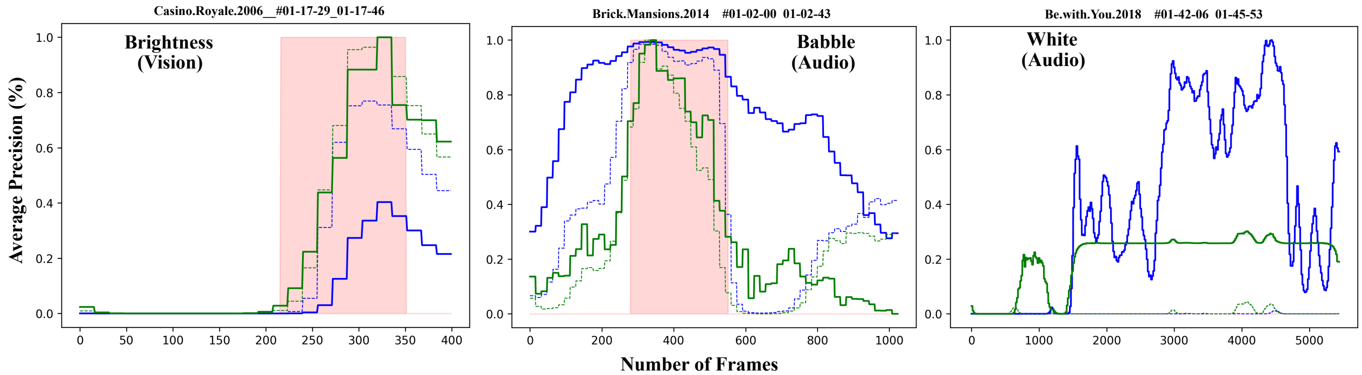


Fig. 5. Qualitative results of `RobustA` and the baseline on three videos taken from XD-Violence dataset. Blue represents baseline anomaly scores whereas green represents `RobustA` results. Dotter represents clean test samples whereas solid highlights corruption cases. As seen, our approach generally outputs comparable anomaly scores for clean and corruption cases. The baseline, while generating competitive anomaly scores for clean samples, demonstrates deteriorated performance when the input is corrupted. Red shaded area represents anomaly ground truth.

TABLE IV
AUC COMPARISON ON THE TEST SET OF UCF-CRIME DATASET UNDER ZERO-SHOT SETTING WHERE THE MODELS ARE TRAINED ON XD-VOILENCE DATASET. OUR APPROACH DEMONSTRATES BETTER ZERO-SHOT GENERALIZATION, HIGHLIGHTING THE IMPORTANCE OF OUR PROPOSED SHARED SPACE REPRESENTATION LEARNING. AS SEEN, OUR APPROACH ALSO OUTPERFORMS THE EXISTING UNSUPERVISED APPROACHES TRAINED AND TESTED ON UCF-CRIME [3].

	Method	**UCF-Crime**
Unsupervised	Kim et al. [38]	52.00
	GCL [18]	65.32
	C2FPL [37]	65.85
	CLAP [33]	67.74
Zero-shot	RTFM [12] (Baseline)	59.6
	RTFM (Ours)	**68.9**
	XDVioDet [2] (Baseline)	63.3
	XDVioDet (Ours)	**68.4**

TABLE V
IMPORTANCE OF LINEAR PROJECTION COMPARED TO PADDING AND CONCATENATION (BASELINE) UNDER DIFFERENT LEVELS (0% TO 100%) OF AN ARBITRARILY SELECTED CORRUPTION CASE (PITCH SHIFT) FOR CONCISENESS.

Method	**0%**	**10%**	**30%**	**50%**	**70%**	**90%**	**100%**
Baseline	78.36	76.24	73.49	69.53	67.01	63.66	63.07
Padding	76.21	73.46	70.52	64.30	60.15	56.83	55.86
Linear Proj.	80.00	78.87	77.62	72.83	71.23	66.91	66.39

anomaly detection task. To address this, we introduced a novel dataset `RobustA`, to systematically evaluate the impact of audio and visual corruptions on anomaly detection performance. Moreover, we proposed a robust multimodal anomaly method that demonstrate resilience against corrupted and missing data. Extensive experiments on the XD-Violence dataset, across various corruption types and levels, demonstrated the effectiveness and robustness of our proposed method. `RobustA` would be instrumental in evaluating anomaly detection methods under the settings more closer to the real-world scenarios.

REFERENCES

[1] Yang Liu, Dingkang Yang, Yan Wang, Jing Liu, Jun Liu, Azzedine Boukerche, Peng Sun, and Liang Song, "Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models," *ACM Computing Surveys*, 2023.

[2] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang, "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. Springer, 2020, pp. 322–339.

[3] Waqas Sultani, Chen Chen, and Mubarak Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6479–6488.

[4] MyeongAh Cho, Minjung Kim, Sangwon Hwang, Chaewon Park, Kyungjae Lee, and Sangyoun Lee, "Look around for anomalies: Weakly-supervised anomaly detection via context-motion relational learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 12137–12146.

[5] Muhammad Zaigham Zaheer, Jin-Ha Lee, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee, "Stabilizing adversarially learned one-class novelty detection using pseudo anomalies," *IEEE Transactions on Image Processing*, vol. 31, pp. 5963–5975, 2022.

[6] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang, "Autoloc: Weakly-supervised temporal action localization in untrimmed videos," in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 154–171.

[7] Sanath Narayan, Hisham Cholakkal, Fahad Shahbaz Khan, and Ling Shao, "3c-net: Category count and center loss for weakly-supervised action localization," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8679–8687.

[8] Jiashuo Yu, Jinyu Liu, Ying Cheng, Rui Feng, and Yuejie Zhang, "Modality-aware contrastive instance learning with self-distillation for weakly-supervised audio-visual violence detection," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 6278–6287.

[9] Peng Wu, Xiaotao Liu, and Jing Liu, "Weakly supervised audio-visual violence detection," *IEEE Transactions on Multimedia*, 2022.

[10] Ayush Ghadiya, Purbayan Kar, Vishal Chudasama, and Pankaj Wasnik, "Cross-modal fusion and attention mechanism for weakly supervised video anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1965–1974.

[11] Xiao Zhou, Xiaogang Peng, Hao Wen, Yikai Luo, Keyang Yu, Ping Yang, and Zizhao Wu, "Learning weakly supervised audio-visual violence detection in hyperbolic space," *Image and Vision Computing*, vol. 151, pp. 105286, 2024.

[12] 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*, 2021, pp. 4975–4986.

[13] Arthur Josi, Mahdi Alehdaghi, Rafael M. O. Cruz, and Eric Granger, "Multimodal data augmentation for visual-infrared person reid with corrupted data," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, January 2023, pp. 32–41.

[14] Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro, "Deep

multimodal learning with missing modality: A survey," *arXiv preprint arXiv:2409.07825*, 2024.

[15] Chenyu Yi, Siyuan Yang, Haoliang Li, Yap-Peng Tan, and Alex C. Kot, "Benchmarking the robustness of spatial-temporal models against corruptions," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, Joaquin Vanschoren and Sai-Kit Yeung, Eds., 2021.

[16] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee, "Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*. Springer, 2020, pp. 358–376.

[17] Muhammad Zaigham Zaheer, Arif Mahmood, Hochul Shin, and Seung-Ik Lee, "A self-reasoning framework for anomaly detection using video-level labels," *IEEE Signal Processing Letters*, vol. 27, pp. 1705–1709, 2020.

[18] M Zaigham Zaheer, Arif Mahmood, M Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee, "Generative cooperative learning for unsupervised video anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 14744–14754.

[19] Vikramjit Mitra, Horacio Franco, Richard M Stern, Julien Van Hout, Luciana Ferrer, Martin Graciarena, Wen Wang, Dimitra Vergyri, Abeer Alwan, and John HL Hansen, "Robust features in deep-learning-based speech recognition," *New Era for Robust Speech Recognition: Exploiting Deep Learning*, pp. 187–217, 2017.

[20] Till Beemelmanns, Quan Zhang, Christian Geller, and Lutz Eckstein, "Multicorrupt: A multi-modal robustness dataset and benchmark of lidar-camera fusion for 3d object detection," in *2024 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2024, pp. 3255–3261.

[21] Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu, "Robo3d: Towards robust and reliable 3d perception against corruptions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19994–20006.

[22] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu, "Benchmarking bird's eye view detection robustness to real-world corruptions," in *International Conference on Learning Representations 2023 Workshop on Scene Representations for Autonomous Driving*, 2023.

[23] Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V Le, et al., "Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 17182–17191.

[24] Dan Hendrycks and Thomas Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *arXiv preprint arXiv:1903.12261*, 2019.

[25] Joanna Hong, Minsu Kim, Jeongsoo Choi, and Yong Man Ro, "Watch or listen: Robust audio-visual speech recognition with visual corruption modeling and reliability scoring," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18783–18794.

[26] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng, "Smil: Multimodal learning with severely missing modality," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 2302–2310.

[27] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng, "Are multimodal transformers robust to missing modality?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18177–18186.

[28] Christian Ganhör, Marta Moscati, Anna Hausberger, Shah Nawaz, and Markus Schedl, "A multimodal single-branch embedding network for recommendation in cold-start and missing modality scenarios," in *Proceedings of the 18th ACM Conference on Recommender Systems*, 2024, pp. 380–390.

[29] Muhammad Irzam Liaqat, Shah Nawaz, Muhammad Zaigham Zaheer, Muhammad Saad Saeed, Hassan Sajjad, Tom De Schepper, Karthik Nandakumar, Muhammad Haris Khan, Ignazio Gallo, and Markus Schedl, "Chameleon: A multimodal learning framework robust to missing modalities," *International Journal of Multimedia Information Retrieval*, vol. 14, no. 2, pp. 21, 2025.

[30] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee, "Multimodal prompting with missing modalities for visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14943–14952.

[31] Vijay John and Yasutomo Kawanishi, "Multimodal cascaded framework with multimodal latent loss functions robust to missing modalities," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2025.

[32] Ronghao Lin and Haifeng Hu, "Missmodal: Increasing robustness to missing modality in multimodal sentiment analysis," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1686–1702, 2023.

[33] Anas Al-lahham, Muhammad Zaigham Zaheer, Nurbek Tastan, and Karthik Nandakumar, "Collaborative learning of anomalies with privacy (clap) for unsupervised video anomaly detection: A new baseline," *arXiv preprint arXiv:2404.00847*, 2024.

[34] Vincent Lostanlen, Justin Salamon, Mark Cartwright, Brian McFee, Andrew Farnsworth, Steve Kelling, and Juan Pablo Bello, "Per-channel energy normalization: Why and how," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 39–43, 2018.

[35] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in python.," *SciPy*, vol. 2015, pp. 18–24, 2015.

[36] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

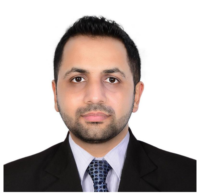
[37] Anas Al-Lahham, Nurbek Tastan, Muhammad Zaigham Zaheer, and Karthik Nandakumar, "A coarse-to-fine pseudo-labeling (c2fpl) framework for unsupervised video anomaly detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 6793–6802.

[38] Jin-Hwa Kim, Do-Hyeong Kim, Saehoon Yi, and Taehoon Lee, "Semi-orthogonal embedding for efficient unsupervised anomaly segmentation," *arXiv preprint arXiv:2105.14737*, 2021.

Salem AlMarri is a Ph.D. graduate from Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi. He received his Masters degree in Electrical Engineering from Rochester's Institute of Technology. His thesis was on Utilizing Auxiliary Information for Weakly-Supervised Video Anomaly Detection.

Muhammad Irzam Liaqat is a Ph.D. candidate in System Science at IMT School for Advanced Studies, Lucca, Italy. He received his master's degree in computer science from the University of Engineering and Technology, Lahore, Pakistan. He held visiting researcher positions at Johannes Kepler University and at Mohamed bin Zayed University of Artificial Intelligence. His research interests include machine learning, deep learning, multimodal learning, computer vision, and medical imaging.

Shah Nawaz is an assistant professor at Johannes Kepler University Linz, Austria. He received bachelor degree in computer engineering from University of Engineering & Technology, Taxila Pakistan and master degree in embedding systems from Technical University of Eindhoven, Netherlands, and PhD degree in computer science from University of Insubria, Italy. His research interests are focused on multimodal representation learning.



Muhammad Zaigham Zaheer received his PhD degree from the University of Science and Technology in 2022. He is currently associated with Mohamed bin Zayed University of Artificial Intelligence as a Research Scientist. Previously, he worked at the Electronics and Telecommunications Research Institute (ETRI) in South Korea, as a postdoctoral researcher. His current research interests include vision language models and their applications, self-supervised learning, and unsupervised learning.

Markus Schedl is a full professor at the Johannes Kepler University Linz / Institute of Computational Perception, leading the Multimedia Mining and Search group. In addition, he is head of the Human-centered AI group at the Linz Institute of Technology (LIT) AI Lab. He graduated in Computer Science from the Vienna University of Technology and earned his Ph.D. from the Johannes Kepler University Linz. Markus further studied International Business Administration at the Vienna University of Economics and Business Administration as well as at the Handelshögskolan of the University of Gothenburg, which led to a Master's degree. His main research interests include recommender systems, information retrieval, natural language processing, multimedia, machine learning, and web mining. He (co-)authored more than 250 refereed articles in journals and conference proceedings as well as several book chapters.

Karthik Nandakumar is an associate professor of computer vision at Mohamed bin Zayed University of Artificial Intelligence. His primary research interests include computer vision, machine learning, biometric recognition, applied cryptography, and blockchain. Prior to joining Mohamed bin Zayed University of Artificial Intelligence, Nandakumar was a research staff member at IBM Research – Singapore from 2014 to 2020 and a scientist at the Institute for Infocomm Research, A*STAR, Singapore from 2008 to 2014..