Towards Efficient Cross-Modal Anomaly Detection Using Triple-Adaptive Network and Bi-Quintuple Contrastive Learning

Shu-Juan Peng, Ye Fan, Yiu-ming Cheung, *Fellow, IEEE*, Xin Liu, *Senior Member, IEEE*, Zhen Cui, *Member, IEEE*, and Taihao Li

***Abstract*—Cross-modal anomaly detection is a relatively new and challenging research topic in machine learning field, which aims at identifying the anomalies whose patterns are disparate across different modalities. As far as we know, this topic has yet to be well studied, and existing works often suffer from the incomplete anomalous data detection and low data utilization problems. To alleviate these limitations, this paper proposes an efficient deep cross-modal anomaly detection approach via Triple-adaptive Network and Bi-quintuple Contrastive Learning (TN-BCL), which lies among the earliest attempt to detect various cross-modal anomalies within the heterogeneous multi-modal data. To be specific, a triple-adaptive network is explicitly designed to identify various anomalies, whose patterns are disparate in both single-modal scenario and cross-modal scenario. On the one hand, the top branch network is utilized to adaptively detect the attribute anomalies and part of mixed anomalies in multi-modal data samples. On the other hand, the bottom two-branch network, with shared residual blocks, is leveraged to learn the discriminative cross-modal embeddings. At the same time, an efficient bi-quintuple contrastive learning method is designed to enhance the feature correlation between the same attribute data, while maximally enlarging the feature difference between different attribute data. Besides that, the bidirectional learning scheme is employed to significantly improve the data utilization. Through the joint exploitation of the above, different kinds of anomalous samples can be well detected across different modalities. Extensive experiments show that the proposed framework outperforms the state-of-the-art competing methods, with a large improvement margin.**

***Index Terms*—Cross-modal anomaly detection, triple-adaptive network, bi-quintuple contrastive learning, shared residual block.**

Manuscript received 21 September 2022; revised 3 January 2023; accepted 3 March 2023. Date of publication 24 March 2023; date of current version 23 January 2024. This work was supported in part by the Open Project of Zhejiang Lab under Grant 2021KH0AB01, in part by the National Science Foundation of China under Grants 61673185 and 61672444, in part by RGC General Research Fund under Grants 12201321 and 12202622, in part by the NSFC/RGC Joint Research Scheme under Grant N_HKBU214/21, in part by the Innovation and Technology Fund of Innovation and Technology Commission of the Government of the Hong Kong SAR under Grant ITS/339/18, and in part by the National Science Foundation of Fujian Province under Grants 2020J01083 and 2020J01084. *(Corresponding authors: Yiu-ming Cheung; Xin Liu.)*

Shu-Juan Peng is with the Department of Artificial Intelligence, Huaqiao University, Xiamen 361021, China, also with the Artificial Intelligence Research Institute, Zhejiang Lab, Hangzhou 311121, China, and also with the Key Laboratory of Computer Vision and Machine Learning (Huaqiao University), Fujian Province University, Xiamen 361021, China (e-mail: pshujuan@163.com).

Ye Fan is with the Xiamen Key Laboratory of Computer Vision, and Pattern Recognition, Xiamen 361021, China, and also with the Fujian Key Laboratory of Big Data Intelligence and Security, Huaqiao University, Xiamen 361021, China (e-mail: yfan@hqu.edu.cn).

Yiu-ming Cheung is with the Department of Computer Science, Hong Kong Baptist University Hong Kong SAR, China (e-mail: ymc@comp.hkbu.edu.hk).

Xin Liu is with the Department of Computer Science, Huaqiao University, Xiamen 361021, China, also with the Artificial Intelligence Research Institute, Zhejiang Lab, Hangzhou 311121, China, and also with the Department of Computer Science, Hong Kong Baptist University Hong Kong SAR, China (e-mail: starxliu@163.com).

Zhen Cui is with the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: zhen.cui@njust.edu.cn).

Taihao Li is with the Artificial Intelligence Research Institute, Zhejiang Lab, Hangzhou 311121, China (e-mail: lith@zhejianglab.com).

Digital Object Identifier 10.1109/TETCI.2023.3256466

I. Introduction

ANOMALY detection aims to identify the abnormal samples that are significantly different from the other normal instances, which is an important data analysis technique due to the fact that anomalies often provide significant and critical information [1]. For instance, anomaly detection in driving behaviours is of paramount importance to reduce millions of vehicle accidents occurring in worldwide [2], while anomalous pattern identification in medical imaging system plays a critical role in diagnosing a certain disease [3]. Note that, these anomaly detectors predominately focus on examining the data from a single source, i.e., single-view data [4]. As shown in Fig. 1, data samples are practically acquired from different sources, and the ability to detect anomalies in multi-modal data is highly desirable in many applications, such as micro-expression detection [5], purchase behavior analysis [6] and malicious intruder detection [7]. As such, traditional single view detectors cannot discover multi-model anomalies, and existing multi-modal anomaly detectors are mainly designed to identify possible anomalies in case where the instances in one modalities are temporarily not available or include noise [8].

In this paper, we focus on a relatively new topic in the anomaly detection field, i.e., cross-modal anomaly detection. It aims at identifying the anomalies whose patterns are disparate across different modalities. That is, some anomalous data instances from multi-modal data are often not anomalous when they are viewed separately in each individual modality, but which may contain inconsistent semantic patterns when these multi-modal instances are considered jointly. In practice, anomaly detection across multi-modal data often benefits lots of valuable applications, such as transaction record detection [9] and mobile

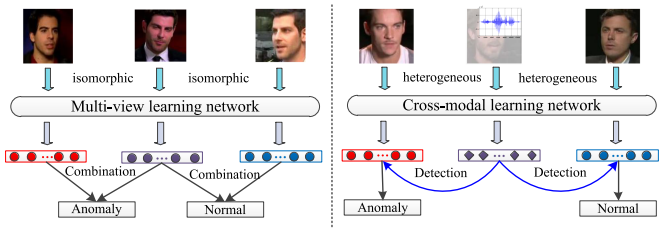

Fig. 1. Illustration of difference between the multi-view and cross-modal anomaly detection mechanisms.

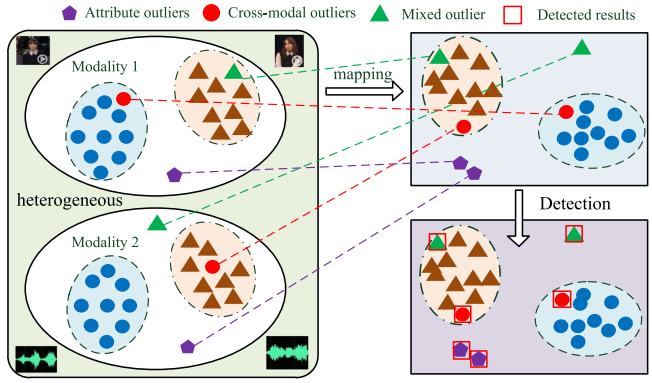

Fig. 2. Illustration of three types of outliers and the process of detection task across heterogeneous modalities.

robot navigation [10]. For instance, there are increasingly more people consuming news through social media, and the news with images, video and text information can provide a better storytelling and attract more attention from readers. In this case, if the image profile is not consistent with other sources of the same news, e.g., the image behaving inconsistently with its textual description, might cause a fake news to mislead readers and may therefore bring negative effects to the public events.

From the research perspective, a straightforward approach for multi-modal anomaly detection is to concatenate all modalities together and convert it to single-view anomaly detection task [11], [12]. However, this concatenation often neglects the inconsistent information across multiple modalities, which often fails to detect the anomalous samples that behave inconsistently across different modalities. Recently, a number of multi-view anomaly detection methods have been proposed to detect the outliers that have abnormal behaviors in each view or have inconsistent behaviors across different views. Note that, most existing approaches rely on the assumption that multi-view data of a normal instance share consistent clustering structures, while the anomalous samples tend to fall into different clusters or consistently deviate from all clusters. Nevertheless, if there are no clusters in data, it is difficult for these approaches to detect the anomalous samples. Besides, these anomaly detectors mainly focus on splitting the object feature representation into different subsets and consider each subset as one particular view of the data, but very few works pay attention to completely heterogeneous multi-modal data acquire from different modalities.

Linking the heterogeneous, not directly comparable sources of multi-modal data, cross-modal anomaly detection remains a challenging task. As shown in Fig. 2, there are three types of outliers possibly existing in multi-modal anomalies. For terminological convenience, they are 1) **Attribute outliers**: these samples have abnormal behaviors in each modality, which will be considered as outliers in each modality individually. 2) **Cross-modal outliers**: this kind of samples may not be identified as outliers when they are viewed separately, but which will be identified as anomalies when their mutual behaviors do not behave consistently across different modalities. 3) **Mixed outliers**: these samples look like an attribute outlier in some modality and exhibit cross-modal outlier property in another modality. Note that, the outliers resulted from the missing modality cannot support the cross-modal anomaly detection task, and this kind of outlier is not considered. Recently, the work [9] specifies the concept of cross-modal anomaly detection, but which is only designed to detect the cross-modal outliers. To the best of our knowledge, there is still a lack of approaches that can simultaneously detect all kinds of outliers across heterogeneous modalities. Although multi-modal data instances belonging to the same category have high semantic relevance, it is still a non-trivial task to perform efficient cross-modal anomaly detection, mainly due to the complex integration of heterogeneous data distributions, modality heterogeneity and semantic gap problem.

In this paper, we propose an efficient deep cross-modal anomaly detection approach via Triple-adaptive Network and Bi-quintuple Contrastive Learning (TN-BCL), which lies among the early attempts to detect various kinds of cross-modal anomalies within the heterogeneous multi-modal data. To be specific, the proposed framework seamlessly unifies the unimodal anomaly detection and cross-modal anomaly detection together, in which the attribute outliers are isolated from the neighborhood structures of one modality data while cross-modal outliers and mixed outliers are detected via the inconsistent behaviors across different modalities. The main contributions are summarized as follows:

- A triple-adaptive network is designed to explicitly identify various kinds of anomalies whose patterns are disparate in both single-modal scenario and cross-modal scenario. To the best of our knowledge, this work is the first attempt to detect three types of anomalies simultaneously across heterogeneous modalities.
- An efficient bi-quintuple contrastive learning method is proposed to guide the cross-modal embedding learning, which significantly enhances the feature correlation among the similar attribute data and maximally enlarges the feature difference between different attribute data.
- The bidirectional learning scheme is developed to significantly improve the data utilization, which can well promote the outlier detection performance.
- Extensive experiments verify the advantages of the proposed approach under various abnormal scenarios and show its superiority over state-of-the-arts.

The remaining part of this paper is organized as follows: Section II makes an overview of the existing multi-view and cross-modal anomaly detection works, and Section III elaborates

the proposed deep cross-modal anomaly detection framework in detail. The experimental results are provided in Section IV. Finally, we draw a conclusion in Section V.

II. Related Work

Anomaly detection across different modalities is a relatively new topic, and this section mainly surveys the most relevant multi-view or multi-modal anomaly detection works.

A. Multi-View Anomaly Detection

Traditional anomaly detection methods mainly focus on detecting the unusual samples from a single view. With the popularity of multi-view learning, a number of multi-view anomaly detectors have been developed, and the earlier efforts mainly attempt to find the samples that have inconsistent cross-view cluster memberships. For instance, Gao et al. [13] present a horizontal anomaly detection (HOAD) approach to identify objects that exhibit inconsistent characteristics across different views. This approach constructs a combined similarity graph in multiple views and calculates the anomalous score of each sample with the cosine distance. Alvarez et al. [14] propose an affinity propagation (AP) algorithm to detect multi-view anomalies by analyzing the affinity vectors of each sample in different views. Note that, these approaches perform the clustering in different views, which may fail to detect the possible anomalies when there are no clusters in data.

In general, a normal sample usually serves as a good contributor in representing the other normal samples while the outlier fails. Accordingly, Zhao et al. [11], [15] utilize the dual-regularization and consensus regularization on the latent representations to achieve multi-view outlier detection. Specifically, these two methods first characterize the outlier by the latent coefficients or intrinsic cluster assignments, and then quantify the inconsistency by a well-designed outlier criterion. Similarly, Li et al. [16] represent the multi-view data by a global low-rank representation shared by all views and define an outlier score function for anomalous sample detection. Although these approaches are able to detect data-anomalies that have inconsistent features across multiple views, they still cannot identify the mixed outliers and often suffer from low data utilization problem. To detect different outliers, Sheng et al. [17] first build a nearest neighbor-based anomaly criterion and then exploit the nearest neighbor based MUlti-View Anomaly Detection (MUVAD) approach to identify various kinds of outliers. Wang et al. [18] first build a hierarchical Bayesian model to represent the multi-view data, and then employ variational inference to evaluate anomalous scores of multi-view instances. In recent years, non-linear mapping has been widely employed in representation learning for complex data structures. Along this line, Ji et al. [19] perform multi-view outlier detection in deep intact space (MODDIS), and define an outlier score to detect different outliers. Note that, these anomaly detectors mainly focus on splitting the feature representation into different subsets, and consider each subset as one particular view of the data. Evidently, the different feature subsets often share the similar distributions, which make these detectors unsuitable to handle completely heterogeneous data acquire from different modalities.

B. Multi-Modal Anomaly Detection

Multi-modal anomaly detectors mainly aim to identify possible anomalies from completely heterogeneous multi-modal data. Deep neural networks are capable of learning nonlinear mappings by extracting high-level abstractions from the input raw features, which have dramatically improved the multi-modal representation performance [20]. Along this line, Wang et al. [21] investigate two-branch neural networks (TBNN) to learn an explicit shared latent embedding, and predict a similarity score between image and text data. Nevertheless, this approach is incapable of discovering the inherent anomalous samples across different modalities. Until very recently, Li et al. [9] specify the concept of cross-modal anomaly detection (CMAD), which aims to identify the anomalies whose patterns are disparate across different modalities. This approach leverages a series of nonlinear mapping functions to map the heterogeneous information of each modality into a comparable consensus feature space, whereby the cross-modal anomalies can be identified by measuring the similarity with a pre-defined threshold. Noted that, this method is only designed to detect the abnormal samples with inconsistent behaviors across different modalities, which cannot identify the attribute or mixed outliers. To the best of our knowledge, there is still a lack of efficient models to adaptively detect different kinds of outliers across different modalities.

III. Deep Cross-Modal Anomaly Detection

Cross-modal anomaly detection across heterogeneous modalities is a relatively new topic in data mining field. Without loss of generality, this section mainly focuses on cross-modal anomaly detection with only two modalities, particularly for visual and textual modalities. Note that, the proposed learning framework can be easily extended to other different modalities, e.g., visual and audio data. First, this section clarifies the notations and formal definitions of multi-modal anomalies. Then, the proposed triple-adaptive network architecture and bi-quintuple contrastive learning method are introduced in tandem. Finally, the detection of anomalous data and its optimization process are explicitly provided.

A. Notation and Problem Formulation

Suppose that we have training data $\mathbf{X} = \{\mathbf{X}^A, \mathbf{X}^B\}$ with two modalities $\mathbf{X}^A \in \mathbb{R}^{n \times d_1}$ and $\mathbf{X}^B \in \mathbb{R}^{n \times d_2}$, with n being the numbers of data samples and d_1, d_2 (in general $d_1 \neq d_2$) the dimensions of these two modalities, the i-th input multi-modal data $\mathbf{X}_i = \{\mathbf{X}_i^A, \mathbf{X}_i^B\}$ is regarded as an anomaly, if one of the following cases is appeared:

Case 1:: If $\mathbf{X}_i^A$ and $\mathbf{X}_i^B$ are both the abnormal samples in each modality individually, $\mathbf{X}_i$ is regarded as attribute outlier.

Case 2: Suppose $\mathbf{X}_i^A$ and $\mathbf{X}_i^B$ both behave normally in each modality, $\mathbf{X}_i$ is regarded as the cross-modal outlier if $\mathbf{X}_i^A$ and $\mathbf{X}_i^B$ have inconsistent behaviors across different modalities.

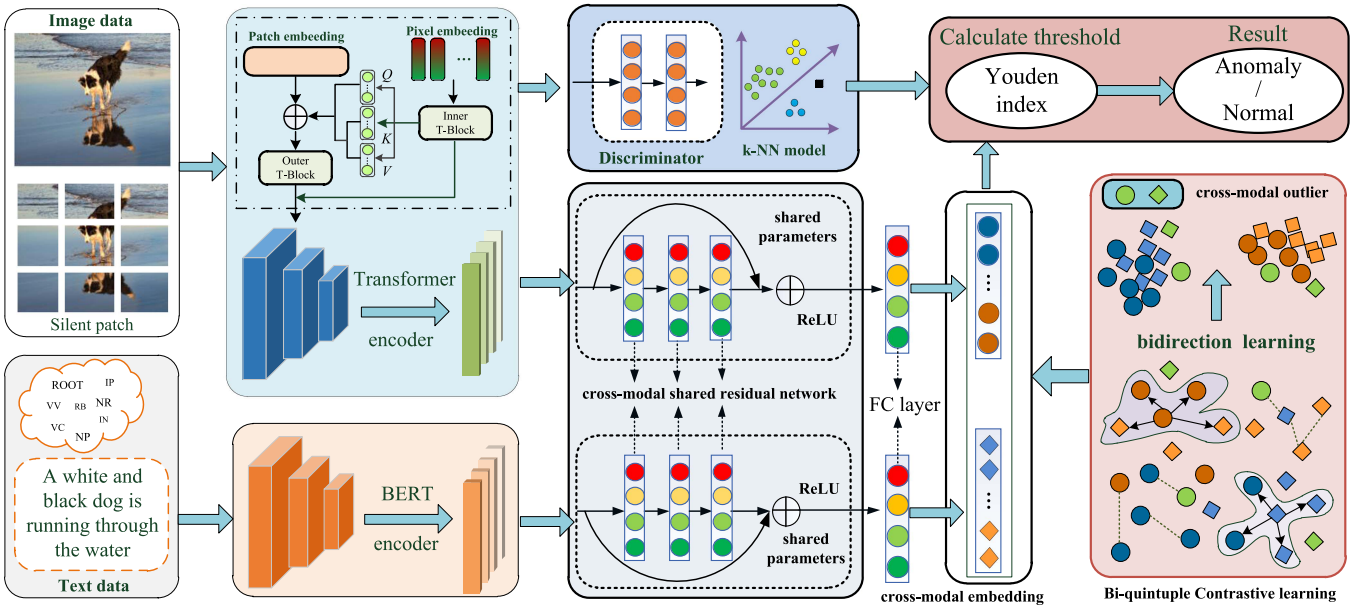

Fig. 3. The schematic architecture of the proposed deep cross-modal anomaly detection approach (TN-BCL).

Case 3: Suppose $\mathbf{X}_i^A$ (resp. $\mathbf{X}_i^B$) is an abnormal sample in modality A (resp. B), and $\mathbf{X}_i^B$ (resp. $\mathbf{X}_i^A$) is a normal sample in modality B (resp. A), $\mathbf{X}_i$ is regarded as the mixed outlier if the mutual behaviors between $\mathbf{X}_i^A$ and $\mathbf{X}_i^B$ do not behave consistently across different modalities.

After an in-depth study of these cases, it can be easily found that if $\mathbf{X}_i^A$ is an abnormal sample in modality A, $\mathbf{X}_i$ can be regarded as the anomalous data, regardless of whether $\mathbf{X}_i^B$ is an abnormal sample or not. To be specific, if $\mathbf{X}_i^B$ is an abnormal sample, $\mathbf{X}_i$ falls into the first case. If $\mathbf{X}_i^B$ is a normal sample, and the mutual behaviors between $\mathbf{X}_i^A$ and $\mathbf{X}_i^B$ are abnormal, $\mathbf{X}_i$ corresponds to the third case. To find these potential outliers, we propose an efficient deep cross-modal anomaly detection approach via triple-adaptive network and bi-quintuple contrastive learning (TN-BCL), which lies among the early attempts to detect various kinds of cross-modal anomalies within the heterogeneous multi-modal data.

B. Triple-Adaptive Network Architecture

The goal of the proposed framework is to align the high-level representations of all semantically relevant samples from heterogeneous modalities, while enlarging the distance between semantically irrelevant ones. Multi-modal deep neural networks have been successfully utilized to learn the compatible features among different types of data, including text, image and audio data. For multi-modal data from heterogeneous sources, we mainly focus on cross-modal anomaly detection with visual and other modalities, e.g., image and text, face and voice. Without loss of generality, we select the image and text data to introduce the proposed learning framework. Let $\mathbf{X}^v = \{\mathbf{x}_i^v\}_{i=1}^N$ represent the visual image data, $\mathbf{X}^t = \{\mathbf{x}_i^t\}_{i=1}^N$ denote the text data, where $\{\mathbf{x}_i^v, \mathbf{x}_i^t\}$ represents the i-th input image-text data pair with its semantic label $\mathbf{y}_i$, and N represents the total number of multi-modal samples. Often, there are three types of outliers possibly existing in multi-modal anomalies. To tackle this problem, as shown in Fig. 3, a triple-adaptive network, is explicitly designed to identify different kinds of anomalies whose patterns are disparate in both single-modal scenario and cross-modal scenario. On the one hand, the top branch network is utilized to adaptively detect the attribute anomaly and part of class-attribute anomaly in multi-modal data samples. On the other hand, the bottom two-branch network associated with shared residual blocks, is leveraged to detect the cross-modal outlier and other outliers.

As discussed in Section III-A, if an abnormal sample $\mathbf{x}_i^v$ (resp. $\mathbf{x}_i^t$) is appeared in one modality, this multi-modal data $\{\mathbf{x}_i^v, \mathbf{x}_i^t\}$ can be regarded as the anomalous data, regardless of whether its corresponding data $\mathbf{x}_i^t$ (resp. $\mathbf{x}_i^v$) is an abnormal sample or not. Specifically, if a normal sample $\mathbf{x}_i^v$ (resp. $\mathbf{x}_i^t$) is appeared in one modality, while its corresponding data $\mathbf{x}_i^t$ (resp. $\mathbf{x}_i^v$) is an abnormal sample in another modality, the input data pair $\{\mathbf{x}_i^v, \mathbf{x}_i^t\}$ can be identified via the cross-modal outlier detection network if $\mathbf{x}_i^v$ and $\mathbf{x}_i^t$ do not behave consistently across different modalities. Since the image data is the most popular data in various applications, we therefore select image data as a typical example to perform anomaly discriminator in the top branch network. First, we utilize transformer network to encode the visual features, and employ the BERT encoder to extract the text features. Then, these heterogeneous features are fed into the cross-modal shared residual network (CM-SRN) to learn the compatible cross-modal representations. At the meantime, the network outputs are regularized by the bi-quintuple contrastive loss, which can well enlarge the feature difference between different attribute data, while enhancing the feature correlation between the similar attribute data.

Visual Transformer and Anomaly Discriminator: Given an image, it is well accepted that not all information is equally

important for image representation. By considering this nature of the visual data representation problem, transformer is a new kind of neural architecture which can encode the input data into discriminative features via the attention mechanism. For the input visual data, we utilize transformer network to encode the visual feature. Formally, the transformer utilizes the linear projections to compute a set of queries ($\mathbf{K}$), keys ($\mathbf{Q}$) and values ($\mathbf{V}$), and calculates the scaled dot products to obtain the attention weights, followed by the value aggregation for each query:

$$\texttt{Att}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \texttt{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\mathbf{V} \tag{1}$$

where d_k is the dimension of $\mathbf{K}$. The transformer applies the attention mechanism multiple times throughout the architecture and results in a multi-head self-attention model. Accordingly, image transformer allows the model to jointly attend to information from different representations at different positions, which has h parallel attention 'heads' to generate several $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ values, and their values are concatenated to aggregate the attended representations:

$$\texttt{head}_i = \texttt{Att}(\mathbf{Q}\mathbf{W}_i^{\mathbf{Q}}, \mathbf{K}\mathbf{W}_i^{\mathbf{K}}, \mathbf{V}\mathbf{W}_i^{\mathbf{V}}) \tag{2}$$

$$\texttt{MultiAtt}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \texttt{Concat}(\texttt{head}_1,\ldots,\texttt{head}_h)\mathbf{W} \tag{3}$$

where $\mathbf{W}_i^{\mathbf{Q}}$, $\mathbf{W}_i^{\mathbf{K}}$, $\mathbf{W}_i^{\mathbf{V}}$ and $\mathbf{W}$ are the projection matrices. Accordingly, each image is divided into several image patches, and the position of each patch is encoded by the transformer representation. Specifically, the number of transformer block is varied for different datasets, which is generally set at a small value if the dataset is simple, and set at a large value if the dataset is complex. The head number is fixed at 6 and MLP layer is set at 4. Finally, the transformer utilizes the non-linear transformation to calculate the output features. Remarkably, ReLU is a non-linear function that will output the input directly if it is positive, otherwise, it will output zero. In practice, ReLU is favored for its simple computation, fast implementation and efficiency, and it can also avoid vanishing gradient problem in the network learning process. Besides, training a deep network with ReLU tends to converge much more quickly and reliably than training a deep network with sigmoid activation. Therefore, we utilize the ReLU as the nonlinear transformations to train the learning model, and the high-level image semantic feature vector is obtained by:

$$\mathbf{v}_i^{c,1} = \texttt{ReLU}(\texttt{TNT}(\mathbf{x}_i^v)) \tag{4}$$

where $\mathbf{v}_i^{c,1}$ represents the high-level image feature vector derived from the top network branch. Within the visual data, the distribution of normal data points is relatively dense, while the distribution of outliers is relatively scattered. Accordingly, the distance of each data point with its local neighborhood can reflect the states that the data point is normal or abnormal.

Specifically, if the distance between one data point and its local neighborhood is small, it means that the neighbors of the data point are distributed around it. In contrast to this, if the distance between the data point and its local neighborhood is large, it means that the neighbors of the data point are distributed far away from the data point. Alternatively, the k-NN model is a classic data-driven method which is relatively effective yet simple to exploit neighborhood relationship among the data. Inspired by this distribution property, we utilize ball-tree based k-NN to search the local geometric structures in the dataset:

$$\left\{\mathbf{v}_{i,r}^{c,1}\right\}_{r=1}^{k} = \underset{1,\ldots,k}{\texttt{Rank}}\left\{\texttt{min}\,\mathbf{d}(\mathbf{v}_i^{c,1},\mathbf{v}_j^{c,1})_{j\neq i, j\in[1,N]}\right\} \tag{5}$$

where $\mathbf{v}_{i,r}^{c,1}$ represents the r-th nearest neighbor for the i-th instance, $\mathbf{d}(\mathbf{v}_i^{c,1},\mathbf{v}_j^{c,1})$ denotes the Euler distance between feature vectors $\mathbf{v}_i^{c,1}$ and $\mathbf{v}_j^{c,1}$. Then, we compute the averaged distance between each point and its k-nearest neighbor data, and search the results recursively. More specifically, we first utilize all the normal data in the verification set to construct the ball-tree, and set the contamination at a small value (e.g., 0.001) to calculate the threshold. If the distance is less than the threshold, the neighboring data points around the current data point are densely distributed, and this data is not anomalous. On the contrary, if the distance is greater than the threshold, the data distribution around the data point is sparser and this data is anomalous. Under such circumstances, the proposed learning framework shall perform the subsequent cross-modal anomaly detection.

Cross-modal Network Encoder: Cross-modal anomalies often present inconsistent behaviors across different modalities. If the multi-modal instances are learned separately, such inconsistent behaviors cannot be well detected. In order to explore the semantic associations between different modalities, we utilize the cross-modal shared residual network (CM-SRN) with four layers to learn the compatible cross-modal representations. In particular, a series of nonlinear mapping functions are employed to map the data points of different modalities into a consensus feature space, in which the instance pairs with consistent patterns are pulled together while the data pairs with inconsistent cross-modal patterns are pushed away.

For the text data, we utilize the BERT encoder to extract the textual features $\mathbf{t}_i = \texttt{BERT}(\mathbf{x}_i^t)$, and then feed this feature into the CR-SRN module. Note that, the networks with the shared weight parameters can well reduce the semantic gap between the heterogeneous modalities, while enhancing their the semantic associations:

$$\mathbf{v}_i^{c,2} = \sigma\left(\texttt{FC}\left(\mathbf{v}_i^{c,1} + \texttt{CM-SRN}\left(\mathbf{v}_i^{c,1}\right)\right)\right) \tag{6}$$

$$\mathbf{t}_i^{c,2} = \sigma\left(\texttt{FC}\left(\mathbf{t}_i + \texttt{CM-SRN}\left(\mathbf{t}_i\right)\right)\right) \tag{7}$$

where $\mathbf{v}_i^{c,2}$ and $\mathbf{t}_i^{c,2}$ are respectively the feature output of image and text data, σ is the non-linear action function such as ReLU or Tanh, and FC is a fully connected layer. Note that, the nonlinear mapping functions could help to fully capture the nonlinear correlations among different modalities, and we utilize ReLU to train the network model.

C. Bi-Quintuple Contrastive Learning

The purpose of cross-modal anomaly detection is to find the anomalous instances whose patterns are inconsistent across different modalities. Towards this end, an efficient cross-modal embedding learning method should pull together the instances

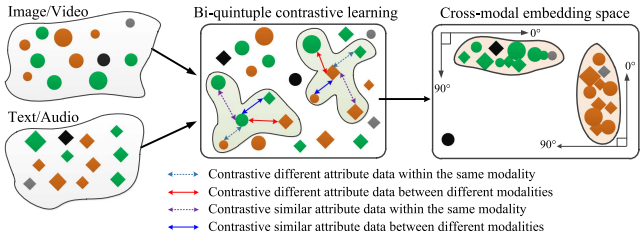

Fig. 4. Illustration of the proposed bi-quintuple contrastive learning.

with consistent pattern, while pushing away the instances with inconsistent patterns. It has been theoretically and practically validated that the contrastive learning can be well utilized for better cross-modal representation learning, which allows the model to flexibly define powerful losses by contrasting the positive pairs from sets of negative samples. Nevertheless, most previous cross-modal contrastive learning methods mainly map the features of heterogeneous modalities into a common embedding space, but which do not simultaneously consider the intra-modal similarity and inter-modal similarity to learn efficient embeddings. To tackle this problem, as shown in Fig. 4, we design an efficient bi-quintuple contrastive learning method to maximally enlarge the feature distance between different attribute data, while enhancing the feature correlation between the similar attribute data. More specifically, the proposed learning method consists of multiple triple losses, which yield a flexible principle: pull an anchor and a positive sample in the embedding space together, and push apart the anchor from many negative samples. Let $\mathbf{o}$ represent an anchor sample, $\mathbf{p}$ denotes a positive sample that belongs to the same attribute with $\mathbf{o}$, $\mathbf{n}$ represents a negative sample that belong to different attributes from $\mathbf{o}$, the formal definition of the triple loss is formulated as follows:

$$\mathcal{L}_{\text{tri}} = \sum_{\mathbf{o},\mathbf{p},\mathbf{n}} \left(\gamma_0 - \mathcal{S}(\mathbf{o},\mathbf{p})\right) + \max\left(0, \mathcal{S}(\mathbf{o},\mathbf{n}) - \gamma\right) \tag{8}$$

where $\mathcal{S}(\cdot,\cdot)$ measures the similarity between a pair of observations, γ_0 and γ are the regularization parameters. Evidently, the sample $\mathbf{o}$ is orthogonal to the negative sample $\mathbf{n}$ in case where $\mathcal{S}(\mathbf{o},\mathbf{n}) = 0$. In practice, γ is utilized to accelerate the convergence. For each triplet $< \mathbf{o}, \mathbf{p}, \mathbf{n} >$, the optimization of the triplet loss is to make $\mathcal{S}(\mathbf{o},\mathbf{p})$ as large as possible, and force $\mathcal{S}(\mathbf{o},\mathbf{n})$ as small as possible.

For contrastive learning across heterogeneous modalities, we design multiple triples to discriminatively regularize the cross-modal representations. Without loss of generality, taking any image data $\mathbf{v}_{anc}^{c,2}$ as an anchor point in visual modality, we construct image-to-text similarity triple $(\mathbf{v}_{anc}^{c,2}, \mathbf{t}_{pos}^{c,2}, \mathbf{t}_{neg}^{c,2})$, where text data $\mathbf{t}_{pos}^{c,2}$ shares the consistent semantic with the image example $\mathbf{v}_{anc}^{c,2}$, and $\mathbf{t}_{neg}^{c,2}$ behaves inconsistent behavior with $\mathbf{v}_{anc}^{c,2}$. To be specific, when the attributes $\mathbf{v}_{anc}^{c,2}$ and $\mathbf{t}_{j}^{c,2}$ exhibit consistent behaviors, their labels are expressed as $\mathbf{y}_j^t = \mathbf{y}_{anc}^v$. On the contrary, if the attributes $\mathbf{p}_{anc}^{c,2}$ and $\mathbf{t}_{j}^{c,2}$ do not behave consistently across different modalities, their semantic labels are expressed as $\mathbf{y}_j^t \neq \mathbf{y}_{anc}^v$. Accordingly, $\mathbf{t}_{pos}^{c,2}$ and $\mathbf{t}_{neg}^{c,2}$ can be derived by:

$$\mathbf{t}_{pos}^{c,2} = \underset{\mathbf{y}_j^t = \mathbf{y}_{anc}^v}{\texttt{argmax}}\ \mathcal{S}\left(\mathbf{v}_{anc}^{c,2}, \mathbf{t}_j^{c,2}\right)|_{j\in[1,N]} \tag{9}$$

$$\mathbf{t}_{neg}^{c,2} = \underset{\mathbf{y}_j^t \neq \mathbf{y}_{anc}^v}{\texttt{argmin}}\ \mathcal{S}\left(\mathbf{v}_{anc}^{c,2}, \mathbf{t}_j^{c,2}\right)|_{j\in[1,N]} \tag{10}$$

$$\mathcal{S}\left(\mathbf{v}_{anc}^{c,2}, \mathbf{t}_j^{c,2}\right) = \frac{\mathbf{v}_{anc}^{c,2} \cdot \mathbf{t}_j^{c,2}}{||\mathbf{v}_{anc}^{c,2}|| \cdot ||\mathbf{t}_j^{c,2}||}. \tag{11}$$

Accordingly, the image-to-text contrastive loss is defined as:

$$\mathcal{L}_{\mathbf{vt}}(\mathbf{v}_{anc}^{c,2}, \mathbf{t}_{pos}^{c,2}, \mathbf{t}_{neg}^{c,2}) = \left(\gamma_0 - \mathcal{S}\left(\mathbf{v}_{anc}^{c,2}, \mathbf{t}_{pos}^{c,2}\right)\right) + \max\left(0, \mathcal{S}(\mathbf{v}_{anc}^{c,2}, \mathbf{t}_{neg}^{c,2}) - \gamma\right). \tag{12}$$

Similarly, we also construct the text-to-image similarity triple $(\mathbf{t}_{anc}^{c,2}, \mathbf{v}_{pos}^{c,2}, \mathbf{v}_{neg}^{c,2})$ to perform corresponding constraint, and its corresponding contrastive loss is defined as follows:

$$\mathbf{v}_{pos}^{c,2} = \underset{\mathbf{y}_j^t = \mathbf{y}_{anc}^v}{\texttt{argmax}}\ \mathcal{S}\left(\mathbf{t}_{anc}^{c,2}, \mathbf{v}_j^{c,2}\right)|_{j\in[1,N]} \tag{13}$$

$$\mathbf{v}_{neg}^{c,2} = \underset{\mathbf{y}_j^t \neq \mathbf{y}_{anc}^v}{\texttt{argmin}}\ \mathcal{S}\left(\mathbf{t}_{anc}^{c,2}, \mathbf{v}_j^{c,2}\right)|_{j\in[1,N]} \tag{14}$$

$$\mathcal{S}\left(\mathbf{t}_{anc}^{c,2}, \mathbf{v}_j^{c,2}\right) = \frac{\mathbf{t}_{anc}^{c,2} \cdot \mathbf{v}_j^{c,2}}{||\mathbf{t}_{anc}^{c,2}|| \cdot ||\mathbf{v}_j^{c,2}||} \tag{15}$$

$$\mathcal{L}_{\mathbf{tv}}(\mathbf{t}_{2,anc}^{c,2}, \mathbf{v}_{pos}^{c,2}, \mathbf{v}_{neg}^{c,2}) = \left(\gamma_0 - \mathcal{S}\left(\mathbf{t}_{anc}^{c,2}, \mathbf{v}_{pos}^{c,2}\right)\right) + \max\left(0, \mathcal{S}(\mathbf{t}_{anc}^{c,2}, \mathbf{v}_{neg}^{c,2}) - \gamma\right). \tag{16}$$

Accordingly, the inter-modal contrastive loss across different modalities can be expressed as:

$$\mathcal{L}_{\text{inter}} = \frac{1}{N}\sum_{anc=1}^{N} \mathcal{L}_{\mathbf{vt}}\left(\mathbf{v}_{anc}^{c,2}, \mathbf{t}_{pos}^{c,2}, \mathbf{t}_{neg}^{c,2}\right) + \frac{1}{N}\sum_{anc=1}^{N} \mathcal{L}_{\mathbf{tv}}\left(\mathbf{t}_{anc}^{c,2}, \mathbf{v}_{pos}^{c,2}, \mathbf{v}_{neg}^{c,2}\right). \tag{17}$$

Recent studies have demonstrated that the intrinsic manifold structure residing in individual modalities is also able to promote the cross-modal learning. Therefore, the proposed learning framework also exploit the intra-modal contrastive learning to promote the representation learning. Given any text data $\mathbf{t}_{2,anc}^{c,2}$ as an anchor sample, we construct text-to-text similarity triple $(\mathbf{t}_{anc}^{c,2}, \mathbf{t}_{pos}^{w,2}, \mathbf{t}_{neg}^{w,2})$, where $\mathbf{t}_{pos}^{w,2}$ and $\mathbf{t}_{neg}^{w,2}$ respectively denote the positive sample and negative sample within the text data. If the attributes of $\mathbf{t}_{anc}^{w,2}$ and $\mathbf{t}_{j}^{c,2}$ are the similar, it is expressed as $\mathbf{y}_j^t = \mathbf{y}_{anc}^t$. On the contrary, if the attributes of $\mathbf{t}_{anc}^{w,2}$ and $\mathbf{t}_{j}^{c,2}$ are different, it is expressed as $\mathbf{y}_j^t \neq \mathbf{y}_{anc}^t$. Accordingly, $\mathbf{t}_{pos}^{wc,2}$ and $\mathbf{t}_{neg}^{w,2}$ can be obtained by:

$$\mathbf{t}_{pos}^{w,2} = \underset{\mathbf{y}_j^t = \mathbf{y}_{anc}^t}{\texttt{argmax}}\ \mathcal{S}\left(\mathbf{t}_{anc}^{c,2}, \mathbf{t}_j^{c,2}\right)|_{j\in[1,N]} \tag{18}$$

$$\mathbf{t}_{neg}^{w,2} = \underset{\mathbf{y}_j^t \neq \mathbf{y}_{anc}^t}{\texttt{argmin}}\ \mathcal{S}\left(\mathbf{t}_{anc}^{c,2}, \mathbf{t}_j^{c,2}\right)|_{j\in[1,N]} \tag{19}$$

$$\mathcal{S}\left(\mathbf{t}_{anc}^{c,2}, \mathbf{t}_j^{c,2}\right) = \frac{\mathbf{t}_{anc}^{c,2} \cdot \mathbf{t}_j^{c,2}}{||\mathbf{t}_{anc}^{c,2}|| \cdot ||\mathbf{t}_j^{c,2}||}. \tag{20}$$

Accordingly, the text-to-text contrastive loss is defined as:

$$\mathcal{L}_{\mathbf{tt}}(\mathbf{t}_{anc}^{c,2}, \mathbf{t}_{pos}^{w,2}, \mathbf{t}_{neg}^{w,2}) = \left(\gamma_0 - \mathcal{S}(\mathbf{t}_{anc}^{c,2}, \mathbf{t}_{pos}^{w,2})\right) + \texttt{max}\left(0, \mathcal{S}\left(\mathbf{t}_{anc}^{c,2}, \mathbf{t}_{neg}^{w,2}\right) - \gamma\right). \tag{21}$$

Similarly, we also construct the image-to-image similarity triple $(\mathbf{v}_{anc}^{c,2}, \mathbf{v}_{pos}^{w,2}, \mathbf{v}_{neg}^{w,2})$ to perform intra-modal constraint on image data, where $\mathbf{v}_{pos}^{w,2}$ and $\mathbf{v}_{neg}^{w,2}$ respectively denote the positive sample and negative sample within the image data:

$$\mathbf{v}_{pos}^{w,2} = \underset{\mathbf{y}_j^v = \mathbf{y}_{anc}^v}{\texttt{argmax}}\ \mathcal{S}\left(\mathbf{v}_{anc}^{c,2}, \mathbf{v}_j^{c,2}\right)|_{j\in[1,N]} \tag{22}$$

$$\mathbf{v}_{neg}^{w,2} = \underset{\mathbf{y}_j^v \neq \mathbf{y}_{anc}^v}{\texttt{argmin}}\ \mathcal{S}\left(\mathbf{v}_{anc}^{c,2}, \mathbf{v}_j^{c,2}\right)|_{j\in[1,N]} \tag{23}$$

$$\mathcal{S}\left(\mathbf{v}_{anc}^{c,2}, \mathbf{v}_j^{c,2}\right) = \frac{\mathbf{v}_{anc}^{c,2} \cdot \mathbf{v}_j^{c,2}}{||\mathbf{v}_{anc}^{c,2}|| \cdot ||\mathbf{v}_j^{c,2}||} \tag{24}$$

Similarly, the image-to-image contrastive loss is defined as:

$$\mathcal{L}_{\mathbf{vv}}\left(\mathbf{v}_{anc}^{c,2}, \mathbf{v}_{pos}^{w,2}, \mathbf{v}_{neg}^{w,2}\right) = \left(\gamma_0 - \mathcal{S}(\mathbf{v}_{anc}^{c,2}, \mathbf{v}_{pos}^{w,2})\right) + \texttt{max}\left(0, \mathcal{S}(\mathbf{v}_{anc}^{c,2}, \mathbf{v}_{neg}^{w,2}) - \gamma\right). \tag{25}$$

Consequently, the intra-modal contrastive learning and cross-modal contrastive learning not only increases the data usage, but also enhances discrimination power of the entire learning framework. Further, we seamlessly combine these intra-modal triples and inter-modal triples to form a quintuple representation. Let $(\mathbf{v}_{anc}^{c,2}, \mathbf{v}_{pos}^{w,2}, \mathbf{v}_{neg}^{w,2}, \mathbf{t}_{pos}^{c,2}, \mathbf{t}_{neg}^{c,2})$ and $(\mathbf{t}_{anc}^{c,2}, \mathbf{t}_{pos}^{w,2}, \mathbf{t}_{neg}^{w,2}, \mathbf{v}_{pos}^{c,2}, \mathbf{v}_{neg}^{c,2})$, respectively, denote the quintuple combination for each image and text sample, their quintuple contrastive loss formulations are defined as:

$$\mathcal{L}_{\mathbf{v}}(\mathbf{v}_{anc}^{c,2}, \mathbf{v}_{pos}^{w,2}, \mathbf{v}_{neg}^{w,2}, \mathbf{t}_{pos}^{c,2}, \mathbf{t}_{neg}^{c,2}) = \mathcal{L}_{\mathbf{vt}}\left(\mathbf{v}_{anc}^{c,2}, \mathbf{t}_{pos}^{c,2}, \mathbf{t}_{neg}^{c,2}\right) + \mathcal{L}_{\mathbf{vv}}\left(\mathbf{v}_{anc}^{c,2}, \mathbf{v}_{pos}^{w,2}, \mathbf{v}_{neg}^{w,2}\right). \tag{26}$$

$$\mathcal{L}_{\mathbf{t}}(\mathbf{t}_{anc}^{c,2}, \mathbf{t}_{pos}^{w,2}, \mathbf{t}_{neg}^{w,2}, \mathbf{v}_{pos}^{c,2}, \mathbf{v}_{neg}^{c,2}) = \mathcal{L}_{\mathbf{tv}}\left(\mathbf{t}_{anc}^{c,2}, \mathbf{v}_{pos}^{c,2}, \mathbf{v}_{neg}^{c,2}\right) + \mathcal{L}_{\mathbf{tt}}\left(\mathbf{t}_{anc}^{c,2}, \mathbf{t}_{pos}^{w,2}, \mathbf{t}_{neg}^{w,2}\right). \tag{27}$$

By integrating the (26) and (27), the entire bi-quintuple contrastive loss is defined as follows:

$$\mathcal{L}_{\text{quintuple}} = \frac{1}{N}\sum_{anc=1}^{N} \mathcal{L}_{\mathbf{v}}\left(\mathbf{v}_{anc}^{c,2}, \mathbf{v}_{pos}^{w,2}, \mathbf{v}_{neg}^{w,2}, \mathbf{t}_{pos}^{c,2}, \mathbf{t}_{neg}^{c,2}\right) + \frac{1}{N}\sum_{anc=1}^{N} \mathcal{L}_{\mathbf{t}}\left(\mathbf{t}_{anc}^{c,2}, \mathbf{t}_{pos}^{w,2}, \mathbf{t}_{neg}^{w,2}, \mathbf{v}_{pos}^{c,2}, \mathbf{v}_{neg}^{c,2}\right). \tag{28}$$

Through the joint exploitation of the bi-quintuple loss, the proposed TN-BCL framework can well push the representations of normal instance pairs closer while pulling those representations of abnormal instance pairs away.

D. Optimization and Anomaly Detection

The objective of the proposed TN-BCL framework is to minimize (28). Similar to work [9], the proposed model is optimized by Adam optimizer [22], which is an adaptive stochastic gradient descent method and its optimization process can be iteratively solved until the convergence is reached. The whole network is trained in an end-to-end manner and network parameters are updated by the backpropagation. On the one hand, the instances with consistent patterns will be pulled together within a small distance, and their cross-modal similarity in the transformed feature space should be very high. On the other hand, the instances with inconsistent patterns will be pushed away from each other, and their cross-modal similarity in the transformed feature space should be very small. Since the Youden index [23] provides the best tradeoff between sensitivity and specificity, we utilize the validation set to obtain the ROC curve, and select the Youden index [23] to optimize the threshold value. For the testing, we calculate the similarity between heterogeneous data pairs, and utilize the derived threshold to detect various cross-modal anomalies.

IV. Experiment

This section conducts a series of quantitative experiments to investigate the effectiveness and robustness of the proposed deep cross-modal anomaly detection method. The experiments and analysis will be detailed in the following subsections.

A. Datasets and Implementations

The popular MNIST, FashionMNIST, CIFAR10 [24] and Voxceleb [25] datasets are selected for evaluation. The main description of each dataset is briefly described as follows:

1) MNIST dataset: It consists of 70,000 original digital images to represent 10 different numbers of pixels. The entire MNIST dataset is divided into a training set of 60,000 images and a test set of 10,000 instances. We randomly selects 5000 pieces of data from the training set as the validation set.

2) FashionMNIST dataset: This dataset replaces the MNIST handwritten digit set with fashion products derived from 10 categories, which covers a total of 70,000 product images. The size of the training set, testing set and validation set are set as the same as the MNIST dataset.

3) CIFAR10 dataset: It is a computer vision dataset that contains 60,000 RGB color images from 10 categories. The entire dataset is divided into a training set of 50,000 images and a test set of 10,000 instances. We also randomly selects 5000 pieces of data from the training set as the validation set.

4) Voxceleb dataset: It is a popular face-voice dataset collected from 1251 celebrities. We utilize MTCNN [26] to crop RGB faces with a size of $224 \times 224 \times 3$ from the video frames, and employ the voice activity detector [27] to eliminate the voice

segment with 64-dimensional log melspectrograms. We select 901 celebrities as the training set, 100 people as the validation set, and 250 people as the test set. The identities between these splits are fully disjoint.

For MNIST, FashionMNIST and CIFAR10 datasets, the text tag is added to each image example by GloVe word embedding [28], and the BERT encoder [29] is utilized to map the tag information into a 100-dimensional vector. Accordingly, the image-text dataset can be synthetically generated for multi-modal data analysis. Note that, the paired instances derived from these datasets exactly share the same semantic information between each other, which can be well utilized to detect the anomalous samples that behave inconsistently across different modalities. Specifically, some other multi-label multi-modal datasets may not have exactly the same semantic information for the paired instances, which are therefore unsuitable for evaluating the cross-modal anomaly detection tasks. Since there is no ground truth of anomalies in these datasets, we refer to work [19] and select the popular injection method to create a number of inconsistent multi-modal data pairs. For the image-text data, we modify a certain proportion of one modality data and randomly inject other data to generate attribute anomalies and mixed anomalies. Meanwhile, we randomly scramble a certain proportion of image-text pairs to generate cross-modal anomalies, in which the scrambled image-text pairs do not match semantically. For the Voxceleb datasets, we replace a certain proportion of faces with different celebrities and randomly generate the disguised voice sequence of the same dimensionality to generate attribute anomaly and mixed anomaly data. Meanwhile, we shuffle a certain proportion of face-voice pairs to generate cross-modal anomaly data. In the experiments, the number of transformer blocks is set at 3 for MNIST and FashionMNIST datasets, set at 5 for CIFAR10 dataset, and set at 7 for Voxceleb dataset.

B. Baseline Methods and Evaluation Metrics

To the best of our knowledge, there exist limited cross-modal anomaly detection methods, except for CMAD [9]. For meaningful comparison, we select three multi-view outlier detection algorithms that can be utilized for cross-modal anomaly detection, i.e., MUVAD [17], MODDIS [19] and TBNN [21]. Besides, cross-modality metric learning (CMML) [30] is also selected to learn the relative distance between the positive and negative pairs, which can be utilized to detect the cross-modal outliers. To be concept, MUVAD exploits the nearest neighbor based multi-view anomaly detection to identify multi-view outliers, while MODDIS utilizes the neural networks to integrate multi-view data into a latent intact space and defines an outlier score measurement to detect different outliers. For these two methods, we consider each modality as one particular view of data to process the multi-modal data, and attempt to detect various anomalous samples in the datasets. TBNN [21] constructs two network structures to learn the shared latent embedding space and measures the similarity between image and text. CMML [30] applies the bi-directional triplet variants to optimize the relative distance between the cross-modal positive and negative pairs.

TABLE I
THE AUC, FPR, TPR AND ACCURACY RESULTS EVALUATED ON MNIST DATASET

Method	AUC	FPR	TPR	Accuracy
MUVAD [17]	0.7592	0.3689	0.8872	0.7577
MODDIS [19]	0.8567	0.0386	0.7519	0.8723
TBNN [21]	0.9574	0.0258	0.9407	0.9592
CMAD [9]	0.9871	0.0072	0.9814	0.9869
CMML [30]	0.9761	0.0195	0.9705	0.9725
Ours($\mathcal{L}_{\text{inter}}$)	**0.9974**	**0.0051**	**0.9934**	**0.9942**
Ours($\mathcal{L}_{\text{quintuple}}$)	0.9971	0.0062	0.9922	0.9932

TABLE II
THE AUC, FPR, TPR AND ACCURACY RESULTS EVALUATED ON FASHIONMNIST DATASET

Method	AUC	FPR	TPR	Accuracy
MUVAD [17]	0.8077	0.1109	0.7263	0.8067
MODDIS [19]	0.8367	0.1477	0.8211	0.8448
TBNN [21]	0.8879	0.1439	0.9196	0.8837
CMAD [9]	0.9025	**0.0362**	0.8412	0.9069
CMML [30]	0.9315	0.0651	0.9227	0.9232
Ours($\mathcal{L}_{\text{inter}}$)	0.9859	0.0435	**0.9617**	0.9591
Ours($\mathcal{L}_{\text{quintuple}}$)	**0.9878**	0.0369	0.9611	**0.9621**

For these methods, we utilize the propposed threshold optimization scheme to detect the possible anomalous samples across different modalities.

In the experiment, the hyperparameter γ is set at 0.4 for MNIST, FashionMNIST, and CIFAR10 dataset, and set at 0.2 for Voxceleb dataset. The quantitative performance is evaluated by the popular true positive rate (TPR) and false positive rate (FPR):

$$\texttt{TPR} = \frac{\texttt{TP}}{\texttt{TP + FN}}, \texttt{FPR} = \frac{\texttt{FP}}{\texttt{TN + FP}} \tag{29}$$

where TP, FN, TN, and FP, respectively, represent the number of true positives, false negatives, true negatives, and false positives. In general, the larger TPR values generally indicate the better detection performance, while the smaller FPR values show the better detection results. Consequently, the accuracy is also utilized to evaluate the detection performances:

$$\texttt{Accuracy} = \frac{\texttt{TP + TN}}{\texttt{TP + TN + FP + FN}} \tag{30}$$

In addition, we also utilize the ROC curves and AUC values to evaluate the detection performances [31]. Specifically, AUC is the area of the ROC curve, with larger values indicating the better outlier detection performance.

C. Performance Comparison and Analysis

1) Results of Detection Performances: The anomaly detection results tested on different datasets are shown in Tables I, II, III, and IV, respectively. It can be found that the competing multi-view anomaly detection methods have delivered relatively lower AUC, TPR and Accuracy values, while generating a large bit FPR values. Specifically, MUVAD [17] exploits an anomaly measurement criterion to estimate the set of normal instances, while MODDIS [19] construct a multi-view latent intact space to encode outlier information. For the FashionMNIST dataset, the AUC score and accuracy value obtained by MUVAD approach

TABLE III
THE AUC, FPR, TPR AND ACCURACY RESULTS EVALUATED ON CIFAR10 DATASET

Method	AUC	FPR	TPR	Accuracy
MODDIS [19]	0.6681	0.2199	0.5561	0.6783
TBNN [21]	0.7432	0.2467	0.7331	0.7428
CMAD [9]	0.7621	**0.1579**	0.6821	0.7651
CMML [30]	0.8267	0.1826	0.7782	0.7941
Ours($\mathcal{L}_{\text{inter}}$)	0.8803	0.2014	**0.8595**	0.8268
Ours($\mathcal{L}_{\text{quintuple}}$)	**0.8845**	0.1698	0.8476	**0.8389**

TABLE IV
THE AUC, FPR, TPR AND ACCURACY RESULTS EVALUATED ON VOXCELEB DATASET

Method	AUC	FPR	TPR	Accuracy
TBNN [21]	0.7471	0.4129	0.8229	0.7273
CMAD [9]	0.7106	0.4776	0.8207	0.6991
CMML [30]	0.7597	0.3940	0.8095	0.7317
Ours($\mathcal{L}_{\text{inter}}$)	0.7543	0.3793	0.818	0.7378
Ours($\mathcal{L}_{\text{quintuple}}$)	**0.7737**	**0.3718**	**0.8261**	**0.7459**

are respectively equal to 0.8077 and 0.8067, while the AUC score and accuracy value obtained by MODDIS method are respectively equal to 0.8367 and 0.8448. These two multi-view outlier detection methods are able to detect some easy-to-identify outliers, but which often fail to detect some inconsistent semantic patterns. As a result, their detection performances are uncompetitive when processing the heterogeneous modalities.

Specifically, TBNN [21] encodes both bidirectional ranking constraints and neighborhood-preserving constraints to regularize the correspondence between different modalities, which can learn their semantic similarity to differentiate the possible anomalous samples. Accordingly, the accuracy scores obtained by this approach are only equal to 0.8837 and 0.7428, respectively, tested on FashionMNIST and CIFAR10 datasets. Note that, this approach ignores the intra-modal structure embedded in real-world data and therefore results in a lower detection performance. CMAD [9] exploits a deep structured framework to characterize the feature representations between heterogeneous data samples, which can identify the anomalies whose patterns are disparate across different modalities. It can be found that CMAD has yielded the better anomalous sample detection performance than the results obtained by the competitive multi-view outlier detection methods. Nevertheless, CMAD has delivered very poor detection performance on the inconsistent behavior matching across face and voice modalities, and the detection accuracy is only equal to 0.6991. The main reason lies that CMAD is only designed to detect the cross-modal outliers, which cannot differentiate the attribute outliers. Besides, CMAD just utilizes the cross-modal negative samples to penalize the instance pairs with inconsistent patterns, whereby some confused outliers cannot be well detected. CMML [30] applies the bi-directional triplet variants to optimize the relative distance between the cross-modal positive and negative pairs, which can promote the outlier detection performances. Nevertheless, this approach ignores the intra-modal negative samples to enlarge the feature distance between different attribute data, and its detection performances need further improvement.

Comparatively speaking, the proposed TN-BCL method always yields the better cross-modal outlier detection performance than the competing multi-view outlier detection methods. To be specific, the proposed TN-BCL approach with $\mathcal{L}_{\text{quintuple}}$ loss always delivers the better AUC scores in most tested datasets, and yields the highest accuracies in most cases. For instance, the AUC values obtained by TN-BCL ($\mathcal{L}_{\text{quintuple}}$) reach up to 0.9878 and 0.8845, respectively, evaluated on the FashionMNIST and CIFAR10 datasets. Comparing with the FashionMNIST dataset, CIFAR10 dataset has more complex visual appearances, and such complexity makes it difficult to identify the anomalies whose patterns are disparate across different modalities. Noted that, the CMAD and TN-BCL methods are both designed to identify the abnormal samples across different modalities. It can be found that the AUC, TPR and accuracy values obtained by the proposed TN-BCL approach are all higher than that produced by the CMAD method, which indicates that the proposed TN-BCL approach is able to identify more complex abnormal samples across heterogeneous modalities. As shown in Fig. 5, the similar detection performance can be also evaluated using the ROC curve, which graphically demonstrates the changes of true positive rate with respect to the changes of false positive rate in the detection. It can be observed that the proposed approach has achieved the best detection performances and improved the state-of-the-art results significantly. That is, the cross-modal embedding derived from the proposed TN-BCL framework are more discriminative and semantically meaningful, which can well guarantee the semantic consistency between the similar heterogeneous samples and inconsistency between dissimilar heterogeneous samples.

2) Results of Different Anomaly Ratios: We also sample different ratio of anomaly samples in the dataset to evaluate the effectiveness of the proposed TN-BCL framework. For simplicity, the ratio is defined as the sample proportion between normal instance and abnormal instance in the dataset. As shown in Table V, it can be observed that the accuracy value changes under different anomaly ratios. Remarkably, the CMAD method [9] is specifically designed to identify cross-modal anomalies whose patterns are disparate across different modalities, which can achieve very competitive performances than the competing multi-view outlier detection methods, i.e., MUVAD and MODDIS, especially when there exist large anomaly examples. For instance, the accuracies obtained by MUVAD and MODDIS, are respectively equal to 0.5969 and 0.7322, when tested on Cifar10 dataset with 1:4 ratio. The main reason lies that multi-view anomaly detectors mainly consider each feature subset as one particular view of the data, but which often degrade their performance on completely heterogeneous data acquired from different modalities. TBNN [21] constructs two network structures to measure the similarity between different modalities, which can well differentiate the anomaly samples within the multi-modal datasets. Nevertheless, this approach is very sensitive to the anomaly ratio values, and the detection accuracies vary significantly under different ratio values. Although CMAD approach [9] and CMML [30] are able to identify the cross-modal anomalies, such approach ignores the

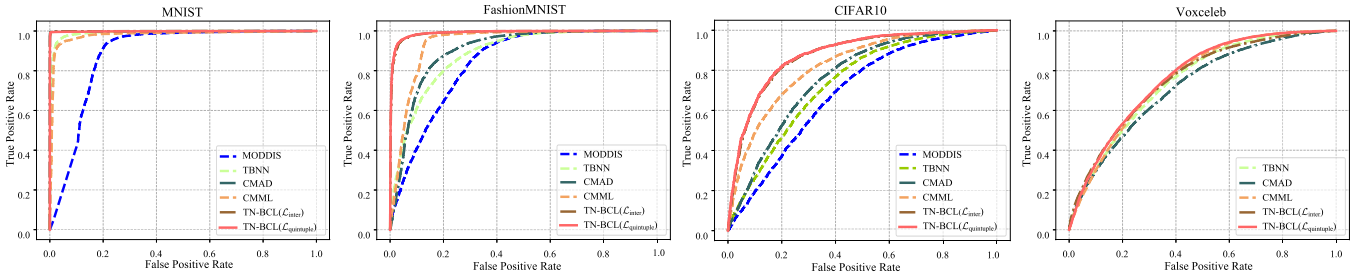

Fig. 5. The ROC curves obtained by different approaches and tested on four datasets.

TABLE V
ANOMALY DETECTION ACCURACIES OBTAINED BY DIFFERENT APPROACHES AND DIVERSE ABNORMAL RATIOS

Dataset	Method	Different Anomaly Ratios (Normal/Abnormal)								
		19:1	9:1	4:1	3:1	2:1	1:1	1:2	1:3	1:4
MNIST	MUVAD [17]	0.6543	0.6815	0.692	0.7112	0.7433	0.7515	0.838	0.8651	0.8821
	MODDIS [19]	0.9320	0.9287	0.9188	0.9072	0.8896	0.8585	0.8201	0.8029	0.7936
	TBNN [21]	0.9730	0.9716	0.967	0.9656	0.9618	0.9536	0.948	0.9466	0.9458
	CMAD [9]	0.9922	0.9920	0.9915	0.9908	0.9883	0.9865	0.9821	0.9802	0.9790
	CMML [30]	0.9703	0.9719	0.9723	0.9729	0.9738	0.9755	0.9758	0.9762	0.9771
	TN-BCL ($\mathcal{L}_{\text{inter}}$)	**0.9935**	**0.9941**	**0.9937**	**0.9938**	**0.9939**	**0.9942**	**0.9941**	**0.9945**	**0.9946**
	TN-BCL ($\mathcal{L}_{\text{quintuple}}$)	0.9923	0.9924	0.9926	0.9929	0.9928	0.9931	0.9933	0.9931	0.9935
FashionMNIST	MUVAD [17]	0.8784	0.8692	0.855	0.8395	0.8183	0.7865	0.7543	0.7393	0.7194
	MODDIS [19]	0.8568	0.8626	0.8516	0.8483	0.8446	0.8460	0.8331	0.8317	0.8284
	TBNN [21]	0.8597	0.8626	0.8704	0.8722	0.8768	0.8980	0.8987	0.9051	0.9095
	CMAD [9]	0.9335	0.9301	0.9254	0.9188	0.9127	0.9083	0.8877	0.8763	0.8696
	CMML [30]	0.9233	0.9236	0.9245	0.9252	0.9266	0.9279	0.9308	0.9318	0.9320
	TN-BCL ($\mathcal{L}_{\text{inter}}$)	**0.9614**	0.9612	0.9608	0.9606	0.9601	0.9591	0.9582	0.9571	0.9575
	TN-BCL ($\mathcal{L}_{\text{quintuple}}$)	0.9612	**0.9613**	**0.9615**	**0.9616**	**0.9614**	**0.9621**	**0.9625**	**0.9626**	**0.9625**
Cifar10	MODDIS [19]	0.7404	0.7307	0.7113	0.7016	0.6954	0.6731	0.6408	0.6146	0.5969
	TBNN [21]	0.7520	0.7507	0.7481	0.7467	0.7445	0.7411	0.7358	0.7335	0.7322
	CMAD [9]	**0.8581**	0.8461	0.8331	0.8133	0.7971	0.7346	0.6821	0.6659	0.6561
	CMML [30]	0.7801	0.7821	0.7849	0.7880	0.7902	0.7978	0.8013	0.8056	0.8075
	TN-BCL ($\mathcal{L}_{\text{inter}}$)	0.8565	**0.8534**	**0.8473**	**0.8457**	0.8392	0.8291	0.8189	0.8136	0.8108
	TN-BCL ($\mathcal{L}_{\text{quintuple}}$)	0.8467	0.8459	0.8445	0.8434	**0.8419**	**0.8387**	**0.8358**	**0.8346**	**0.8338**
Voxceleb	TBNN [21]	0.8111	0.7993	0.7757	0.7636	0.7443	0.7052	0.6657	0.6463	0.6341
	CMAD [9]	0.8058	0.7854	0.7614	0.7461	0.7213	0.6716	0.6218	0.5970	0.5821
	CMML [30]	0.7993	0.7891	0.7697	0.7586	0.7436	0.7125	0.6738	0.6588	0.6436
	TN-BCL ($\mathcal{L}_{\text{inter}}$)	0.8081	0.7981	0.7783	0.7687	0.7523	0.7183	0.6865	0.6697	0.6601
	TN-BCL ($\mathcal{L}_{\text{quintuple}}$)	**0.8162**	**0.8063**	**0.7865**	**0.7766**	**0.7611**	**0.7273**	**0.6938**	**0.6781**	**0.6675**

intra-modal negative samples to enlarge the feature distance between different attribute data and therefore may result missing outlier detections.

In contrast to this, the proposed TN-BCL approach always delivers the better detection performances under different anomaly ratios, which again demonstrates the effectiveness of the proposed model. For the MNIST dataset, the proposed TN-BCL approach with $\mathcal{L}_{\text{inter}}$ loss delivers the best detection performance with different anomaly ratios. For the Voxceleb dataset, the proposed TN-BCL approach with $\mathcal{L}_{\text{quintuple}}$ loss yields the best detection accuracy under different anomaly ratios. Further, we also sample different types of anomaly instances separately. Specifically, we utilize the learnt representations of CMAD associated with optimized threshold to detect the attribute outliers and mixed outliers. The similar results tested on FashionMNIST and CIFAR10 datasets can be also found in Fig. 6, it can be found that MODDIS, TBNN and CMAD approaches have induced an obvious fluctuations on the detection accuracies. For instance, if the anomaly ratios are different, CMAD has resulted different accuracies when detecting the cross-modal outliers and mixed outliers. Comparatively speaking, our proposed TN-BCL approach has achieved very stable anomaly detection performance and the detection accuracies are always higher than the results obtained by the competing baselines. That is, our proposed TN-BCL approach not only can identify different kinds of anomalous samples, but also could produce relatively stable detection performance on different anomaly detection tasks.

3) Analysis on Training Time: Further, we record the execution times on each training epoch and 1000 testing examples to show the time complexity of the different framework. The proposed model and the competing baselines are trained on GPU NVIDIA RTX 2080Ti. Since the proposed model aggregates more modules to discriminatively learn the cross-modal representations, the execution time of training time or testing time could be much higher than that obtained by the competing methods. Fortunately, as illustrated in Table VI, the proposed TN-BCL framework does not significantly increase the training time and testing time to a large extent, while achieving the best outlier detection performances. From a practical viewpoint, the proposed TN-BCL method achieves a good balance between the

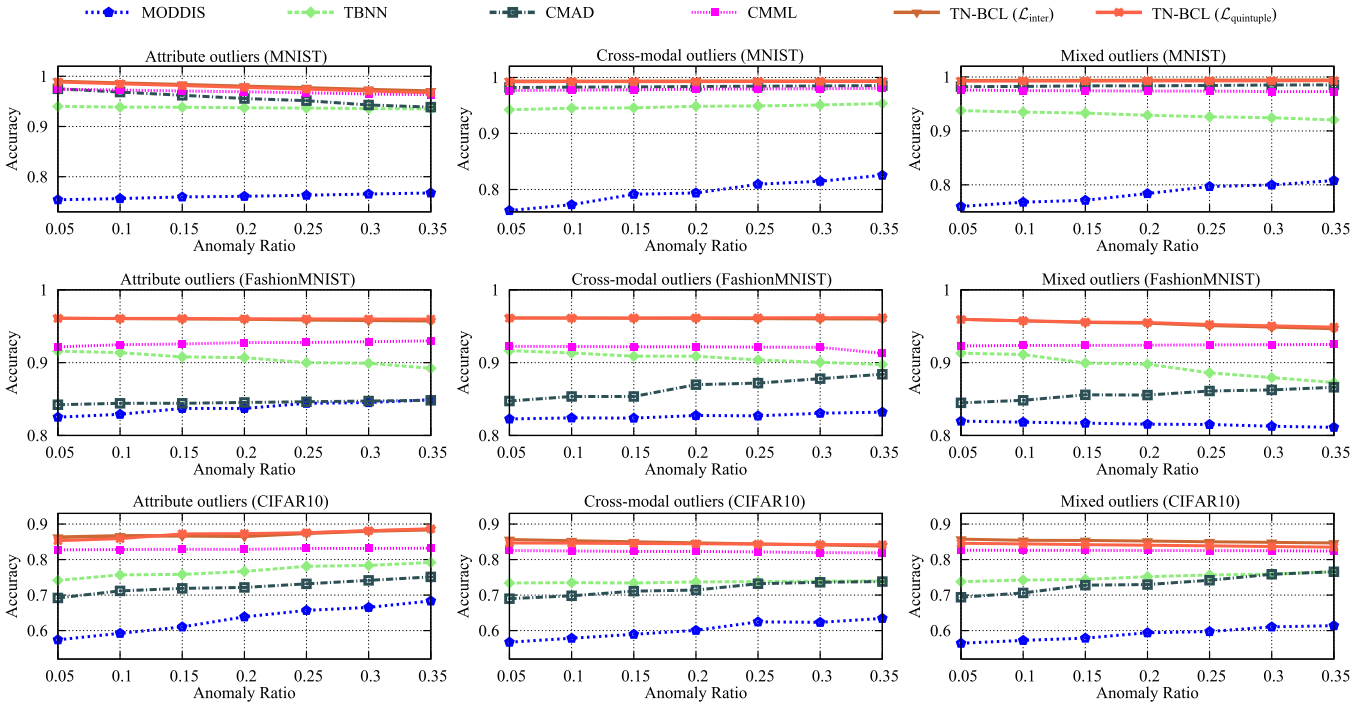

Fig. 6. Anomaly detection results obtained by different approaches and tested on diverse anomaly types.

TABLE VI
EVALUATION OF EXECUTION TIMES ON FASHIONMNIST DATASET

Method	Training (epoch/h)	Testing (1000/s)	Acc (%)
MODDIS	0.18	4.3	0.8448
TBNN	0.16	2.5	0.8837
CMAD	0.18	3.2	0.9069
CMML	0.20	3.2	0.9232
TN-BCL ($\mathcal{L}_{\text{inter}}$)	0.28	6.6	0.9591
TN-BCL ($\mathcal{L}_{\text{quintuple}}$)	0.29	6.6	0.9621

time cost and outlier detection performance, which is suitable for identifying the anomalies whose patterns are disparate across different modalities.

4) Visualization and Analysis: To visually verify the superiority of the proposed TN-BCL model, we show some representative cross-modal anomaly detection examples obtained by the proposed TN-BCL framework. As shown in Fig. 7, the right parts show the visual detection results specified by text or voice query. On the one hand, it can be found that the proposed TN-BCL approach is able to identify the semantically similar examples from one modality to another modalities. For instance, given a 'ship' query in the text modality, the proposed TN-BCL method is able to match the relevant visual examples. This indicates that the proposed framework aggregates more semantic relationships within semantically relevant multi-modal data, which can explicitly learn the semantic correspondence to correlate heterogeneous data samples. On the other hand, the proposed TN-BCL framework is capable of detecting the abnormal examples that exhibit inconsistent behaviors across different modalities. For instance, given an 'Horse' query in the text modality, the proposed TN-BCL framework is able to identify the semantically irrelevant examples (i.e., cross-modal outliers) that show inconsistent meanings in the visual modality, e.g., 'camel' and 'cattle'. It indicates that the proposed network structure exhibits high discriminability to learn the semantically differentiable embeddings, which can well push representations of the semantically relevant examples closer while pulling those of semantically irrelevant instances away. The experiments constantly show its outstanding performance.

D. Further Discussion

Cross-modal anomaly detection across heterogeneous modalities is a very challenging topic in the anomaly detection field. The extensive experiments show that the proposed TN-BCL framework is able to detect all kinds of outliers from the heterogeneous multi-modal data. The main advantages contributed to these very competitive performances are three-fold: 1) The designed triple-adaptive network is able to identify different kinds of anomalies whose patterns are disparate in both single-modal scenario and cross-modal scenario. Accordingly, the proposed network structure can well promote the outlier detection performance. 2) The proposed bi-quintuple contrastive learning is capable of enhancing the feature correlation between the same attribute data, while maximally enlarging the feature distance between different attribute data. As a result, the instance pairs with consistent patterns across different modalities are pulled together, while the data pairs with inconsistent cross-modal patterns are pushed away. 3) The designed bidirectional learning scheme is able to significantly improve the data utilization, which can well benefit the network to learn the discriminative

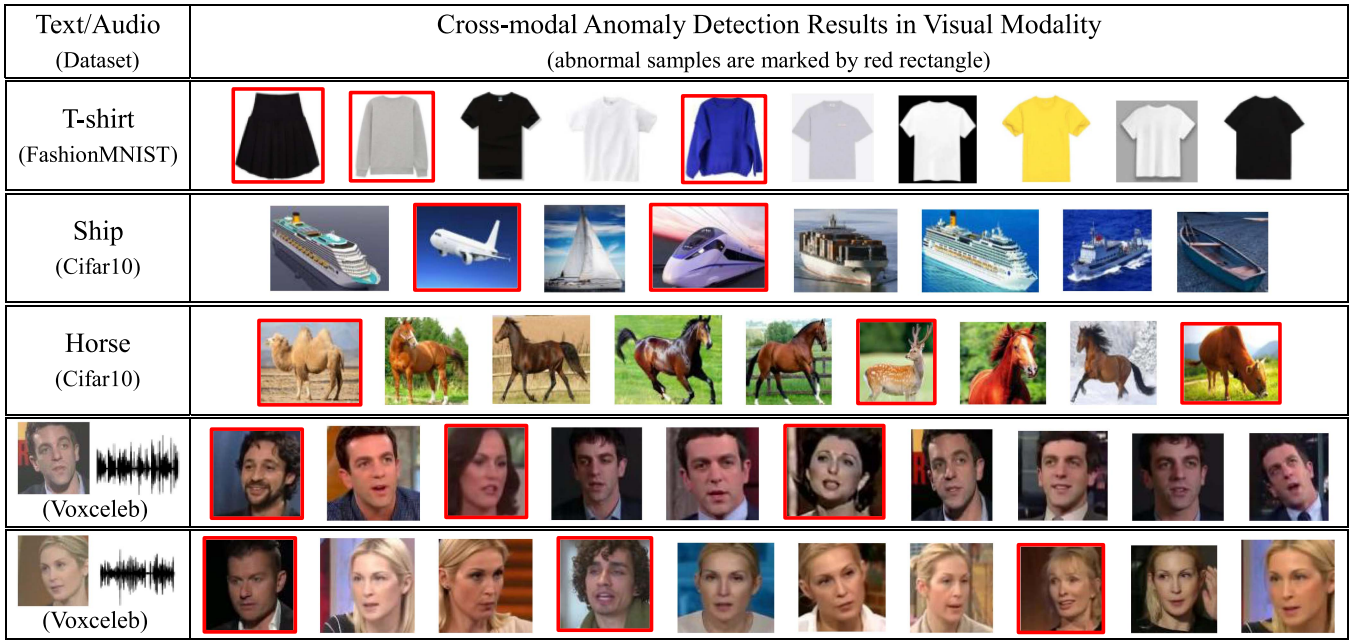

Fig. 7. Visualization of representative cross-modal anomaly detection examples on different datasets. For each text query or voice query, the detected anomalous samples are marked with red rectangle.

cross-modal embeddings and therefore promote the outlier detection performance. The experimental results consistently validate the advantage and effectiveness of the proposed TN-BCL framework in detecting different kinds of anomaly examples.

V. Conclusion

In this paper, we have proposed an efficient deep cross-modal anomaly detection approach via triple-adaptive network and bi-quintuple contrastive learning. Within the proposed framework, a triple-adaptive network is explicitly designed to identify different kinds of anomalies whose patterns are disparate in both single-modal scenario and cross-modal scenario. Meanwhile, an efficient bi-quintuple contrastive learning method is discriminatively designed to guide the cross-modal embedding learning process, which can maximally enlarge the feature distance between different attribute data and enhance the feature correlation between the same attribute data. As a result, the multi-modal data pairs with consistent patterns are pulled together, while the data pairs with inconsistent patterns are pushed away. In addition, the bidirectional learning scheme is able to improve the data utilization significantly and therefore benefit the abnormal sample detection in a more interpretable and plausible way. Extensive experiments conducted on various kinds of cross-modal anomaly detection tasks have shown its promising performance.

Along the line of the present work, several open problems also deserve our further research. For example, the missing modality problem is another challenging topic in cross-modal outlier detection field, and the adaptive cross-modal anomaly detection method should also consider this practical problem. Also, the current learning framework aggregates multiple modules to promote the outlier detection performance. If the training dataset is very large, the updating of model parameters will need more computational load. Besides, if the multi-modal datasets have more than three modalities or incorporate the imbalanced multi-modal data collections, the current work will need more designs to tackle these problems. These investigations will be studied in our future works.

References

[1] P. Jain, S. Jain, O. R. Zaiane, and A. Srivastava, "Anomaly detection in resource constrained environments with streaming data," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 3, pp. 649–659, Jun. 2022.

[2] C. Ryan, F. Murphy, and M. Mullins, "End-to-end autonomous driving risk analysis: A behavioural anomaly detection approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1650–1662, Mar. 2021.

[3] H. D. P. d. Santos, A. H. D. P. S. Ulbrich, V. Woloszyn, and R. Vieira, "DDC-outlier: Preventing medication errors using unsupervised learning," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 874–881, Mar. 2019.

[4] W. Lin, J. Gao, Q. Wang, and X. Li, "Learning to detect anomaly events in crowd scenes from synthetic data," *Neurocomputing*, vol. 436, pp. 248–259, 2021.

[5] Y. Zong, W. Zheng, X. Huang, J. Shi, Z. Cui, and G. Zhao, "Domain regeneration for cross-database micro-expression recognition," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2484–2498, May 2018.

[6] Q. Li, M. Gu, K. Zhou, and X. Sun, "Multi-classes feature engineering with sliding window for purchase prediction in mobile commerce," in *Proc. IEEE Int. Conf. Data Mining Workshop*, 2015, pp. 1048–1054.

[7] N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, "A deep learning approach to network intrusion detection," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 1, pp. 41–50, Feb. 2018.

[8] A.-U. Rehman, H. S. Ullah, H. Farooq, M. S. Khan, T. Mahmood, and H. O. A. Khan, "Multi-modal anomaly detection by using audio and visual cues," *IEEE Access*, vol. 9, pp. 30587–30603, 2021.

[9] Y. Li, N. Liu, J. Li, M. Du, and X. Hu, "Deep structured cross-modal anomaly detection," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2019, pp. 1–8.

[10] J. Simanek, V. Kubelka, and M. Reinstein, "Improving multi-modal data fusion by anomaly detection," *Auton. Robots*, vol. 39, no. 2, pp. 139–154, 2015.

[11] H. Zhao and Y. Fu, "Dual-regularized multi-view outlier detection," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, 2015, pp. 4077–4083.
[12] X. Li, H. Zhang, R. Wang, and F. Nie, "Multiview clustering: A scalable and parameter-free bipartite graph fusion method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 330–344, Jan. 2022.
[13] J. Gao, W. Fan, D. Turaga, S. Parthasarathy, and J. Han, "A spectral framework for detecting inconsistency across multi-source object relationships," in *Proc. IEEE 11th Int. Conf. Data Mining*, 2011, pp. 1050–1055.
[14] A. M. Alvarez, M. Yamada, A. Kimura, and T. Iwata, "Clustering-based anomaly detection in multi-view data," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2013, pp. 1545–1548.
[15] H. Zhao, H. Liu, Z. Ding, and Y. Fu, "Consensus regularized multi-view outlier detection," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 236–248, Jan. 2018.
[16] K. Li, S. Li, Z. Ding, W. Zhang, and Y. Fu, "Latent discriminant subspace representations for multi-view outlier detection," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 3522–3529.
[17] X. R. Sheng, D.-C. Zhan, S. Lu, and Y. Jiang, "Multi-view anomaly detection: Neighborhood in locality matters," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 4894–4901.
[18] W. Zhen and L. Chao, "Towards a hierarchical Bayesian model of multi-view anomaly detection," in *Proc. Int. Joint Conf. Artif. Intell.*, 2020, pp. 2420–2426.
[19] Y. X. Ji et al., "Multi-view outlier detection in deep intact space," in *Proc. IEEE Int. Conf. Data Mining*, 2019, pp. 1132–1137.
[20] X. Liu, Y.-m. Cheung, Z. Hu, Y. He, and B. Zhong, "Adversarial tri-fusion hashing network for imbalanced cross-modal retrieval," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 5, no. 4, pp. 607–619, Aug. 2021.
[21] L. Wang, Y. Li, J. Huang, and S. Lazebnik, "Learning two-branch neural networks for image-text matching tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 394–407, Feb. 2019.
[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–15.
[23] J. Yin and L. Tian, "Joint inference about sensitivity and specificity at the optimal cut-off point associated with Youden index," *Comput. Statist. Data Anal.*, vol. 77, pp. 1–13, 2014.
[24] E. F. Carvalho and P. M. Engel, "Convolutional sparse feature descriptor for object recognition in CIFAR-10," in *Proc. IEEE Braz. Conf. Intell. Syst.*, 2013, pp. 131–135.
[25] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Comput. Speech Lang.*, vol. 60, 2020, Art. no. 101027.
[26] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[27] X. L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 4, pp. 697–710, Apr. 2013.
[28] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. Int. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.
[29] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 4171–4186.
[30] M. Ye, W. Ruan, B. Du, and M. Z. Shou, "Channel augmented joint learning for visible-infrared recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 13547–13556.
[31] X. Liu, Z. Hu, H. Ling, and Y.-M. Cheung, "MTFH: A matrix tri-factorization hashing framework for efficient cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 964–981, Mar. 2021.

Shu-Juan Peng received the Ph.D. degree in computer science from Wuhan University, Wuhan, China, in 2009. She is currently an Associate Professor with the Department of Artificial Intelligence, Huaqiao University, Xiamen, China, and also a Research Fellow with the Key Laboratory of Pattern Recognition and Computer Vision, Xiamen, and also a Research Fellow with the Key Laboratory of Computer Vision and Machine Learning (Huaqiao University), Fujian Province University, Xiamen. Her research interests include multimedia data analysis, pattern recognition, and computer animation.

Ye Fan received the M.S. degree in computer science from Huaqiao University, Xiamen, China, in 2023. He is currently a Research Fellow with the Xiamen Key Laboratory of Computer Vision and Pattern Recognition, Xiamen, and also a Research Fellow with the Fujian Key Laboratory of Big Data Intelligence and Security, Xiamen. His research interests include multimedia content analysis, pattern recognition, and deep learning.

Yiu-ming Cheung (Fellow, IEEE) received the Ph.D. degree from the Department of Computer Science and Engineering, the Chinese University of Hong Kong, Hong Kong. He is currently a Chair Professor (Artificial Intelligence) with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. His research interests include machine learning, pattern recognition and visual computing. He is the Editor-in-Chief (since 2023) of IEEE Transactions on Emerging Topics in Computational Intelligence, and is an Associate Editor for IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS, *Pattern Recognition*, *Knowledge and Information Systems*, to name a few. He is an IET Fellow, AAAS Fellow and BCS Fellow. For details, please visit: https://www.comp.hkbu.edu.hk/ymc.

Xin Liu (Senior Member, IEEE) received the Ph.D. degree in computer science from Hong Kong Baptist University, Hong Kong, in 2013. He was a Visiting Scholar with Computer & Information Sciences Department, Temple University, Philadelphia, PA, USA, from 2017 to 2018. Currently, he is a Professor with the Department of Computer Science and Technology, Huaqiao University, Xiamen, China, and also a Research Fellow with Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China. His present research interests include multimedia data analysis, anomaly detection, pattern recognition and deep learning based data mining. He is a senior member of the IEEE.

Zhen Cui (Member, IEEE) received the B.S. degree from Shandong Normal University, Jinan, China, in 2004, the M.S. degree from Sun Yat-sen University, Guangzhou, China, in 2006, and the Ph.D. degree from the Institute of Computing Technology (ICT), Chinese Academy of Sciences, Beijing, China, in 2014. He also spent half a year as a Research Assistant on Nanyang Technological University (NTU), Singapore, from June 2012 to December 2012. He was a Research Fellow with the Department of Electrical and Computer Engineering, National University of Singapore (NUS), Singapore, from 2014 to 2015. He is currently a Professor with the Nanjing University of Science and Technology, Nanjing, China. His research interests include deep learning, computer vision, and pattern recognition.

Taihao Li received the Ph.D. degree in information science and systems engineering from National University of Tokushima, Tokushima, Japan, in 2006. He was a Researcher with Harvard University, Cambridge, MA, USA, from 2006 to 2011. He was a Principle Scientist with Flatley Discovery Lab from 2011 to 2019. He is currently the Deputy Director of Cross-Media Intelligence Research Center in, Zhejiang Lab, Hangzhou, China. He has published authored or coauthored more than 30 related papers in well-known journals and conferences around topics like affective computing, image processing, and multi-modal information fusion, etc. He also hosted or participated in 18 projects in the United States, Japan and China and has applied more than 30 patents for multi-modal emotion recognition.