# Multiview Deep Anomaly Detection: A Systematic Exploration

Siqi Wang, Jiyuan Liu, Guang Yu, Xinwang Liu, *Senior Member, IEEE*, Sihang Zhou, En Zhu, Yuexiang Yang, Jianping Yin, and Wenjing Yang

***Abstract*— Anomaly detection (AD), which models a given normal class and distinguishes it from the rest of abnormal classes, has been a long-standing topic with ubiquitous applications. As modern scenarios often deal with massive high-dimensional complex data spawned by multiple sources, it is natural to consider AD from the perspective of multiview deep learning. However, it has not been formally discussed by the literature and remains underexplored. Motivated by this blank, this article makes fourfold contributions: First, to the best of our knowledge, this is the first work that formally identifies and formulates the multiview deep AD problem. Second, we take recent advances in relevant areas into account and systematically devise various baseline solutions, which lays the foundation for multiview deep AD research. Third, to remedy the problem that limited benchmark datasets are available for multiview deep AD, we extensively collect the existing public data and process them into more than 30 multiview benchmark datasets via multiple means, so as to provide a better evaluation platform for multiview deep AD. Finally, by comprehensively evaluating the devised solutions on different types of multiview deep AD benchmark datasets, we conduct a thorough analysis on the effectiveness of the designed baselines and hopefully provide other researchers with beneficial guidance and insight into the new multiview deep AD topic.**

***Index Terms*— Deep anomaly detection (AD), multiview deep AD, multiview deep learning.**

## I. Introduction

ANOMALY detection (AD) [1] is a classic task in machine learning. At the training stage of AD, only data from one single class (the normal class) are provided to train an AD model, while no data from other classes (abnormal classes) are available. For inference, the trained AD model is expected to classify whether the incoming data belong to the normal or abnormal class. AD catches the eyes of researchers from both academia and industry for its pervasive applications in practice. For instance, a public video surveillance system usually has easy access to massive data of normal daily events, while abnormal events, such as robbery or vehicle intrusion, are rare and extremely hard to encounter. Therefore, it is often unrealistic to collect sufficient anomalies, which constitutes to one typical application scenario for AD. Besides, AD techniques are also applicable to various realms, such as information retrieval [2], fault detection [3], authorship verification [4], enhanced multiclass classification [5], and so on. In the literature, AD is sometimes referred as outlier detection (OD) [6], one-class learning [7], novelty detection [8] or out-of-distribution detection [9], and so on. In particular, we strictly define AD in this context to be the *semi-supervised* task that labels pure normal data for training. By contrast, we follow the taxonomy in [8] and refer the *unsupervised* task that directly detects peculiar data from contaminated unlabeled dataset to be OD [10]. We distinguish OD from AD, because their terms are often interchangeably used in the literature and cause confusion.

Compared with fully supervised binary/multiclass classification, AD remains a special and challenging problem. This is mainly due to the absence of anomalous data in training, which makes it impossible to train a classifier directly by discriminating the normal and abnormal class. Meanwhile, AD is also different from fully unsupervised tasks, such as OD or clustering, since training data in AD share a common positive label and provide partial supervision information. So far, various solutions have been proposed [1], [11] to tackle AD and will be reviewed in Section II and Supplementary Material.

Nevertheless, as the modern society has witnessed an explosive development in data acquisition capabilities, people find it increasingly difficult to leverage classic models for modern data in many learning tasks, which include but are not confined to AD. In this article, we will focus on two of the most important challenges: First, unlike traditional data, modern data, such as images, are often endowed with high-dimensional and complex latent structures. Classic methods usually fail to exploit such latent information embedded in data, due to their shallow model architectures and limited representation power. Second, with significantly enriched sources to acquire data, one object is often described from multiple viewpoints, such as different modalities, sensors, or angles, which gives

Manuscript received 7 June 2021; revised 23 March 2022; accepted 11 June 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62006236, Grant 61922088, and Grant 61906020; in part by the National University of Defense Technology Research Project under Grant ZK20-10; in part by the State Key Laboratory of High Performance Computing (HPCL) Autonomous Project under Grant 202101-15; in part by the National Key Research and Development Program of China under Grant 2020AAA0107100; and in part by the Education Ministry-China Mobile Research Funding under Grant MCM20170404. *(Siqi Wang and Jiyuan Liu contributed equally to this work.) (Corresponding author: Xinwang Liu.)*

Siqi Wang, Jiyuan Liu, Guang Yu, Xinwang Liu, Sihang Zhou, En Zhu, Yuexiang Yang, and Wenjing Yang are with the College of Computer, National University of Defense Technology (NUDT), Changsha 410073, China (e-mail: wangsiqi10c@nudt.edu.cn; liujiyuan13@nudt.edu.cn; yuguangnudt@gmail.com; xinwangliu@nudt.edu.cn; sihangjoe@gmail.com; enzhu@nudt.edu.cn; yyx@nudt.edu.cn; wenjing.yang@nudt.edu.cn).

Jianping Yin is with the School of Cyberspace Security, Dongguan University of Technology, Dongguan 523808, China (e-mail: jpyin@dgut.edu.cn).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TNNLS.2022.3184723.

Digital Object Identifier 10.1109/TNNLS.2022.3184723

birth to a large amount of multiview data. However, classic methods are usually designed for single-view data, and they lack the ability to exploit complementary information and cross-view correlation embedded in multiview data. To handle the abovementioned two challenges posed by modern data, an emerging realm named *multiview deep learning* has attracted surging attention from researchers. Specifically, multiview deep learning resorts to artificial neural networks with deep architecture to conduct layer-wise data abstraction and representation learning [12], which is proven highly effective in vast applications. The remarkable success of deep learning has made it a standard tool to handle massive complex data. Meanwhile, multiview deep learning methods usually leverage multiview fusion or multiview alignment techniques [13] to exploit inter-view information embedded in multiview data. Multiview deep learning has already been successfully applied to many tasks [13]–[15]. Therefore, it is quite natural for us to consider the intersection of AD and multiview deep learning, i.e., *multiview deep AD*. Multiview deep AD has a large potential to many practical problems, and a straightforward example is the aforementioned video surveillance system that aims to detect anomalies: The collected normal events can be described by both RGB and optical flow data, which are both high-dimensional data with rich underlying semantics, while the AD model needs to be trained with such data to build a normality model and discriminate the anomalies.

Although both AD and multiview deep learning methods have been thoroughly studied in the literature, the problem of multiview deep AD has not been formally defined and systematically explored to our best knowledge. Such a blank constitutes to the biggest motivation of this article. There are three major obstacles when looking into multiview deep AD: First, *above all, the lack of formal formulation of the problem.* Despite its huge application potential in various real-world scenarios, multiview deep AD has not been formally identified and formulated, which prevents researchers from giving sufficient attention to this novel but challenging problem. Second, *the lack of baseline methods*. Although many attempts have been made to approach multiview deep learning, they are typically designed for other tasks and, therefore, not explored for AD. In the meantime, the existing AD approaches are merely applicable to the single-view case. Third, *the lack of proper benchmark datasets for evaluation*. Previous researches usually evaluate AD models by the "one *versus* all" protocol [16]. For any binary/multiclass benchmark datasets, it assumes data from a certain class to be normal, while data from the rest of classes to be abnormal. Besides, datasets that are specifically designed for AD are also proposed recently [17]. However, frequently used benchmark datasets in AD are basically single view. In the meantime, the existing multiview benchmark datasets are often limited in size, and none of them are specifically designed for the background of AD. As a result, the effort to build eligible benchmark datasets for multiview deep AD is still insufficient.

To bridge the abovementioned gaps, this article, for the first time, formally identifies and formulates the problem of multiview deep AD, and carries out a systematic study on this new area. Our contributions can be summarized as follows.

1) To the best of our knowledge, this is the first work that formally identifies and formulates the multiview deep AD problem, which points out a brand new realm for both AD and multiview deep learning research.
2) Inspired by recent progress of AD and multiview deep learning, we systematically design 11 multiview deep AD solutions as baselines, which are ground-breaking efforts in this new realm and lay its research foundation.
3) To facilitate the evaluation of the new multiview deep AD problem, we extensively collect the existing public data and process them into more than 30 multiview benchmark datasets via various means.
4) We comprehensively evaluate the proposed multiview deep AD baselines on both constructed and existing multiview datasets and conduct in-depth analysis on their performances. It sheds the first light on multiview deep AD research and hopefully provides informative guidance and insights into future research.

## II. Related Work

In this section, we will focus on reviewing deep AD, multiview OD, and multiview deep learning, which are the most relevant areas to multiview deep AD. In Sections I and II, Supplementary Material, we also briefly review classic methods for AD and multiview learning due to the page limit.

### A. Deep AD

There is a surging interest in AD to leverage deep neural networks (DNNs) to handle high-dimensional complex data [18]. Since only data of a single class are available, the most frequently used models for deep AD are generative DNNs. A simple but effective way is to extend the shallow autoencoder (AE) into a deep one. For example, stacked denoising AE (SDAE) [19] and deep convolutional AE (DCAE) [20] have been leveraged to perform AD with raw video data. Meanwhile, many attempts are also made to improve AE's AD performance, such as using the ensemble technique [21] and combining AE with an energy-based model [22]. In addition to AE-based methods, other popular generative neural networks, such as generative adversarial networks (GANs) [23], [24] and U-Net [25], [26], are also actively explored to perform deep AD. Such generative DNNs typically perform AD by measuring the reconstruction error of the generated target data, while other methods (e.g., the discriminator outputs and latent representations of GANs) are also explored. Apart from generative deep models, several representative discriminative approaches are also proposed recently. Ruff *et al.* [27] extend the classic support vector data description (SVDD) into deep SVDD (DSVDD), which learns to map latent representations of positive data into a hypersphere with minimal radius. Golan and El-Yaniv [16], for the first time, leverage self-supervised learning for image AD. They impose multiple geometric transformations to create pseudo-classes, which are classified by a discriminative DNN to enable highly effective representation learning. Statistics of the discriminative DNN outputs are then used to score each image. Bergman and Hoshen [28] further extend self-supervised learning-based deep AD to generic tabular data by introducing random projection for creating pseudo-classes. Goyal *et al.* [29] assume a low-dimensional manifold in given positive data, which can be utilized to sample accurate pseudo-outliers to train a discriminative component. The detailed review can be found in [30]. Despite that great progress has been made in deep AD, current discussion is typically limited to the single-view setting.

### *B. Multiview Outlier (Anomaly) Detection*

Multiview OD is a relevant but essentially different area from multiview deep AD in this article. As a comparison, multiview OD is an unsupervised task that aims to detect either intra-view outliers ("attribute outlier") or outliers with cross-view inconsistency ("class outlier") from contaminated unlabeled data [31]. In particular, it should be noted that multiview OD is often termed "multiview AD" in some prior works, such as [32]–[34], but their setup is evidently different from AD or multiview deep AD in this context (see Section I). The pioneer work of multiview OD is proposed by Gao *et al.* [31], while a series of improved solutions are developed [10], [32]–[35]. Most multiview OD methods spot outliers by the cluster structure of given unlabeled data, which are obtained by classic techniques, such as spectral clustering [32] or outlierness estimation [34], while only very recent works [36], [37] begin to explore DNNs to perform multiview deep OD (MDOD). We notice that the latest work [37], for the first time, leverages AE-based reconstruction paradigm, and our work differs from [37] in terms of two aspects: First, two works target at essentially different problems with different setups: AEs for multiview OD [37] are fed with contaminated unlabeled data, while pure data from a single class are used to train AEs in this work. Second, our work places more emphasize on designing a generic framework rather than a specific solution, such as [37]. For example, we explore three different ways to realize latent representation alignment, while [37] only uses the simplest distance-based alignment. Besides, it is also noted that [38] leverages a hierarchical Bayesian model to address "semi-supervised multiview AD," which is a multiview AD task by our definition. Nevertheless, their method cannot perform DNN-like representation learning. Meanwhile, it is only tested on classic benchmarks and suffers from poor scalability to large-scale data. Thus, there is still a gap between their work and the multiview deep AD in this article.

### *C. Multiview Deep Learning*

As classic multiview learning does not involve representation learning and lacks the ability to handle with complex data, multiview deep learning has rapidly become an emerging topic. Current multiview deep learning methods are usually categorized into two groups, i.e., *multiview fusion* and *multiview alignment*-based methods. Multiview fusion-based methods fuse the learned representations from different views into a joint representation, which can be realized by either simple operations, such as max/sum/concatenation [13], or sophisticated means, such as a neural network. Specifically, the pioneer work of Ngiam *et al.* [39] proposes a multimodal deep AE for multiview deep fusion, while Srivastava and Salakhutdinov [40] perform the fusion by a multimodal deep Boltzmann machine (DBM). Such neural network-based multiview fusion can also be conducted on modern neural network architecture, such as convolutional neural networks (CNNs) [41] and recurrent neural networks (RNNs) [42]. Latest work from Sun *et al.* [43] employs a multiview deep Gaussian process to obtain the joint representation and perform classification. Apart from the prevalent neural network-based fusion, Zadeh *et al.* [44] propose a novel tensor-based fusion scheme, while Liu *et al.* [45] extend it to the generic multiview case by low-rank decomposition. Unlike multiview fusion, multiview alignment intends to align the learned representations from each view, so as to exploit the common information among different views. The most popular and representative multiview alignment method is canonical correlation analysis (CCA) [46] and its deep variant deep CCA (DCCA) [47], which seeks to maximize the correlation of two views. Wang *et al.* [48] later develop a variant named deep canonically correlated AEs (DCCAEs), which is regularized by the reconstruction objective, and Benton *et al.* [49] propose deep generalized CCA (DGCCA) to handle with the case of more than two views. In addition to correlation, deep multiview alignment also leverages other metrics. For example, Frome *et al.* [50] maximize the dot-product similarity by a hinge rank loss, while Feng *et al.* [51] minimize the $l_2$-norm distance between the learned representations of two views. Besides, inspired by GANs, adversarial training is also borrowed to improve multiview representation learning by learning modality-invariant representations [52] or cross-view transformation [53]. Consequently, many solutions have been proposed for multiview deep learning, and they are widely adopted to serve many tasks, such as action recognition, sentiment analysis, and image captioning. However, none of those works has considered the marriage of AD and multiview deep learning, which motivates this article.

## III. Problem Formulation

To tackle the first obstacle mentioned in Section I, we will provide a formal problem formulation of multiview deep AD in the first place. Given the normal class $\mathcal{C}_n$, the multiview data $\{\mathbf{x}_{\text{train}}^{(v)}\}_{v=1}^{V}$ are sampled from $\mathcal{C}_n$ for training, where $V \geq 2$ is the number of views and $\mathbf{x}_{\text{train}}^{(v)} \in \mathbb{R}^{d_1^{(v)} \times d_2^{(v)} \times \cdots d_{M_v}^{(v)}}$ is an $M_v$-dimensional tensor with the shape $d_1^{(v)} \times d_2^{(v)} \times \cdots d_{M_v}^{(v)}$. To be more specific, $\mathbf{x}_{\text{train}}^{(v)}$ denotes the observation from the $v$th view, while $M_v = 1$ and $M_v > 1$ correspond to tabular data and complex data (e.g., images or videos), respectively. Note that the observations from different views can be heterogeneous. With the training data $\{\mathbf{x}_{\text{train}}^{(v)}\}_{v=1}^{V}$, the goal of multiview deep AD is to obtain a DNN model

$$\mathcal{M}_{\boldsymbol{\Theta}} : \mathbb{R}^{\prod_{i=1}^{M_1} d_i^{(1)}} \times \mathbb{R}^{\prod_{i=1}^{M_2} d_i^{(2)}} \cdots \times \mathbb{R}^{\prod_{i=1}^{M_V} d_i^{(V)}} \mapsto \{0, 1\} \tag{1}$$

where $\boldsymbol{\Theta}$ represents the set of all learnable parameters for the model $\mathcal{M}_{\boldsymbol{\Theta}}$. In the inference phase, $\mathcal{M}_{\boldsymbol{\Theta}}$ aims to classify whether the incoming multiview testing data $\{\mathbf{x}_{\text{test}}^{(v)}\}_{v=1}^{V}$ belongs to the normal class $\mathcal{C}_n$ or not, where $\mathbf{x}_{\text{test}}^{(v)} \in \mathbb{R}^{d_1^{(v)} \times d_2^{(v)} \times \cdots d_{M_v}^{(v)}}$ denotes the data from the $v$th view, i.e.,

$$\mathcal{M}_{\boldsymbol{\Theta}}\left(\left\{\mathbf{x}_{\text{test}}^{(v)}\right\}_{v=1}^{V}\right) = \begin{cases} 1, & \text{if } \left\{\mathbf{x}_{\text{test}}^{(v)}\right\}_{v=1}^{V} \in \mathcal{C}_n \\ 0, & \text{if } \left\{\mathbf{x}_{\text{test}}^{(v)}\right\}_{v=1}^{V} \notin \mathcal{C}_n. \end{cases} \tag{2}$$

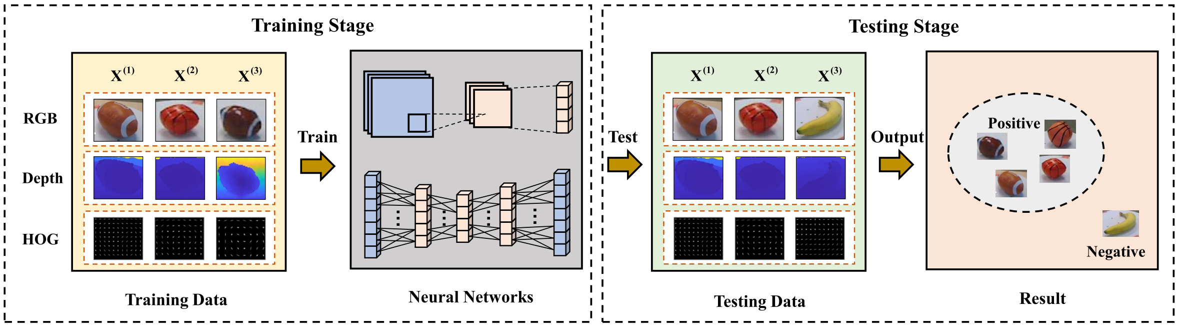


Fig. 1. Overview of multiview deep AD. The example shows soccer balls described by data from three views (RGB, depth, and HOG), and the goal is to determine whether incoming multiview data are from the soccer ball class or not.

In practice, $\mathcal{M}_{\mathbf{\Theta}}$ is usually supposed to obtain a score $\mathcal{S}(\{\mathbf{x}_{\text{test}}^{(v)}\}_{v=1}^{V})$, which indicates the likelihood that $\{\mathbf{x}_{\text{test}}^{(v)}\}_{v=1}^{V}$ belongs to the normal class. A threshold can be then chosen to binarize the score into the final decision of $\mathcal{M}_{\mathbf{\Theta}}$. It should be noted that the DNN-based model $\mathcal{M}_{\mathbf{\Theta}}$ can be constructed by either pure DNNs or a mixture of DNNs and classic AD models. Since a mixture of DNNs and classic AD models often suffers from some issues (e.g., the decoupling of representation learning and classification), we will focus on discussing the model that consists of pure DNNs. In other words, we discuss the case where $\mathcal{M}_{\mathbf{\Theta}}$ is able to perform end-to-end AD. An overview of multiview deep AD task is presented in Fig. 1.

## IV. Proposed Baselines

Having provided a formal problem formulation, we will address the second issue in Section I by designing baseline solutions to multiview deep AD, so as to provide the first sense to approach this topic. In this section, we systematically design four types of baseline solutions: fusion-based solutions, alignment-based solutions, tailored deep AD methods, and self-supervision-based solutions.

### A. Fusion-Based Solutions

A core issue for multiview learning is how to maximally exploit the information embedded in different views to perform downstream tasks. To this end, the most straightforward idea is to fuse data from multiple views into a joint embedding. Therefore, it is natural for us to propose fusion-based multiview deep AD solutions, which fuse the data embeddings learned from different views into a joint embedding to conduct AD. We will discuss its framework and specific implementations of each component in the following.

*1) Framework:* Given a multiview data $\{\mathbf{x}_{\text{train}}^{(v)}\}_{v=1}^{V}$ with $V$ views, fusion-based solutions first introduce a set of DNN-based encoders to encode the input observation of each view into their latent embeddings. For the $v$th view, an encoder $\text{Enc}^{(v)}:\mathbb{R}^{d_1^{(v)}\times d_2^{(v)}\times\cdots d_{M_v}^{(v)}}\mapsto\mathbb{R}^{d_l^{(v)}}$ encodes $\mathbf{x}_{\text{train}}^{(v)}$ into a latent embedding $\mathbf{h}^{(v)}$ with a dimension $d_l^{(v)}$

$$\mathbf{h}^{(v)}=\text{Enc}^{(v)}\left(\mathbf{x}_{\text{train}}^{(v)}\right),\quad v=1,2,\ldots,V \tag{3}$$

where $\mathbf{h}^{(v)}$ is a $d_l^{(v)}$-dimensional column vector. In this way, embeddings from different views can be collected into a set $\{\mathbf{h}^{(v)}\}_{v=1}^{V}$. Subsequently, fusion-based methods select a fusion function $F_f:\mathbb{R}^{d_l^{(1)}\times d_l^{(2)}\times\cdots d_l^{(V)}}\mapsto\mathbb{R}^{D}$ to fuse the embeddings of different views into a $D$-dimensional vector $\mathbf{h}$ as the joint embedding of the multiview data

$$\mathbf{h}=F_f\left(\{\mathbf{h}^{(v)}\}_{v=1}^{V}\right). \tag{4}$$

Since only very weak supervision is available (i.e., all training data share a common positive label), discriminative information is unavailable for guiding the representation learning of encoders in multiview deep AD. Therefore, as a baseline, we propose to leverage the frequently used reconstruction paradigm to guide the model training. To this end, a set of DNN-based decoders are introduced to decode the input data of each view from the joint embedding $\mathbf{h}$: For the $v$th view, a decoder $\text{Dec}^{(v)}:\mathbb{R}^{D}\mapsto\mathbb{R}^{d_1^{(v)}\times d_2^{(v)}\times\cdots d_{M_v}^{(v)}}$ intends to map $\mathbf{h}$ back to the $v$th view's original input $\mathbf{x}_{\text{train}}^{(v)}$

$$\hat{\mathbf{x}}_{\text{train}}^{(v)}=\text{Dec}^{(v)}(\mathbf{h}) \tag{5}$$

where $\hat{\mathbf{x}}_{\text{train}}^{(v)}$ is the reconstructed input of $v$th view. To train the DNN-based model, one can simply minimize the differences between original inputs and reconstructed inputs

$$\mathcal{L}_r=\sum_{v=1}^{V}\ell\left(\mathbf{x}_{\text{train}}^{(v)},\hat{\mathbf{x}}_{\text{train}}^{(v)}\right)=\sum_{v=1}^{V}\left\|\mathbf{x}_{\text{train}}^{(v)}-\hat{\mathbf{x}}_{\text{train}}^{(v)}\right\|_2^2. \tag{6}$$

In addition to the mean square errors (MSEs) mentioned earlier, other types of reconstruction loss $\ell(\cdot)$ are also applicable, such as $L_1$-norm reconstruction loss. During testing, the incoming multiview data $\{\mathbf{x}_{\text{test}}^{(v)}\}_{v=1}^{V}$ are fed into the network to obtain the reconstructed data $\{\hat{\mathbf{x}}_{\text{test}}^{(v)}\}_{v=1}^{V}$ by (3)–(5). As the DNN-based model is trained with only data from the normal class $\mathcal{C}_n$, one can follow the standard practice in AD to assume that a lower reconstruction error indicates a higher likelihood that the testing data belong to $\mathcal{C}_n$. In other words, a baseline score for the $v$th view can be directly obtained by $\mathcal{S}^{(v)}(\mathbf{x}_{\text{test}}^{(v)})=-\ell(\mathbf{x}_{\text{train}}^{(v)},\hat{\mathbf{x}}_{\text{train}}^{(v)})$. Finally, we can obtain a score function by the reconstruction errors of all views

$$\mathcal{S}\left(\left\{\mathbf{x}_{\text{test}}^{(v)}\right\}_{v=1}^{V}\right)=F_l\left(\mathcal{S}^{(1)}\left(\mathbf{x}_{\text{test}}^{(1)}\right),\ldots,\mathcal{S}^{(V)}\left(\mathbf{x}_{\text{test}}^{(V)}\right)\right) \tag{7}$$

where $F_l(\cdot)$ is a late fusion function that combines scores of different views into a final score, which is discussed later. An intuitive illustration of the framework is given in Fig. 2.

*2) Implementations:* The key to a fusion-based multiview deep AD method is the implementation of fusion function $F_f(\cdot)$. Thus, we design four specific ways to realize $F_f(\cdot)$.

1) *Summation-Based Fusion (SUM):* Summation-based fusion combines latent embeddings from different views by

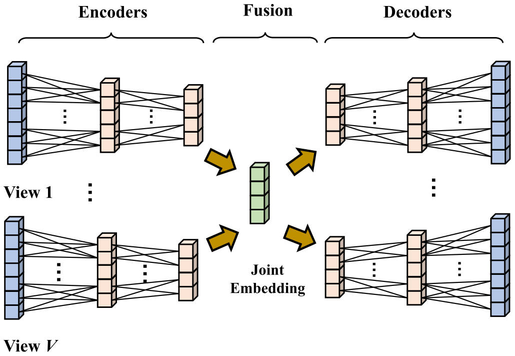


Fig. 2. Fusion-based solutions for multiview deep AD.

summing them up. Specifically, it assumes that embeddings of all views share the same dimension $d_l^{(v)} = D$, and a joint embedding $\mathbf{h}$ can be yielded by

$$F_f\left(\{\mathbf{h}^{(v)}\}_{v=1}^V\right) = \frac{1}{V}\sum_{v=1}^{V}\mathbf{h}^{(v)}. \tag{8}$$

When the embedding dimensions are different, we can introduce a linear mapping parameterized by a learnable matrix $\mathbf{P}^{(v)}$ to map the $v$th embedding to the shared dimension $D$: $\hat{\mathbf{h}}^{(v)} = \mathbf{P}^{(v)}\cdot\mathbf{h}^{(v)}$. Since neural networks can flexibly map a data into an embedding with any dimension with a linear mapping layer, we simply assume that all embeddings $\mathbf{h}^{(v)}$ share the same dimension $D$ here to facilitate analysis in the rest parts of this article.

2) *Max-Based Fusion (MAX):* Similar to summation-based fusion, max-based fusion also assumes a shared dimension $d_l^{(v)} = D$ and use the maximum of embeddings of different views as the joint embedding $\mathbf{h}$

$$F_f\left(\{\mathbf{h}^{(v)}\}_{v=1}^V\right) = \max\left(\{\mathbf{h}^{(v)}\}_{v=1}^V\right). \tag{9}$$

3) *Network-Based Fusion (NN):* It is easy to notice that both summation-based fusion and max-based fusion assume a shared embedding dimension across different views. To make the fusion more flexible, it is also natural to map all latent embeddings into the joint embedding $\mathbf{h}$ by a fully connected neural network with learnable parameters

$$F_f\left(\{\mathbf{h}^{(v)}\}_{v=1}^V\right) = \sigma\left(\mathbf{W}\cdot Cat(\{\mathbf{h}^{(v)}\}_{v=1}^V) + \mathbf{b}\right) \tag{10}$$

where $\mathbf{W}$ and $\mathbf{b}$ are learnable weights and biases of corresponding neurons, and $Cat(\cdot)$ and $\sigma(\cdot)$ denote the concatenation operation and the activation function, respectively. Note that we can also leverage a multilayer fully connected network to perform DNN-based fusion.

4) *Tensor-Based Fusion (TF):* Tensor-based fusion [44] is an emerging method in multiview deep learning. The core idea of tensor-based fusion is to combine the embeddings of different views by the tensor outer product $\mathcal{Z} = \bigotimes_{v=1}^V \mathbf{h}^{(v)}$, where $\mathcal{Z}$ is a $d_l^{(1)} \times d_l^{(2)} \times \cdots d_l^{(V)}$ tensor. Afterward, $\mathcal{Z}$ is fed into a linear layer with weight tensor $\mathcal{W} \in \mathbb{R}^{D\times d_l^{(1)}\times d_l^{(2)}\times\cdots d_l^{(V)}}$ and bias vector $b \in \mathbb{R}^D$ to obtain the unified representation $\mathbf{h}$

$$F_f\left(\{\mathbf{h}^{(v)}\}_{v=1}^V\right) = \mathcal{W}\cdot\mathcal{Z} + b. \tag{11}$$

Note here we slightly abuse the notation of matrix-vector multiplication by considering $\mathcal{W}$ and $\mathcal{Z}$ as $D \times K$ matrix and $K$-dimensional vector, where $K = \prod_{v=1}^V d_l^{(v)}$. However, a severe practical problem is that tensor-based fusion requires computing the tensor $\mathcal{Z}$ and recording $\mathcal{W}$, which incurs exponential computational cost. To address this problem, we leverage the low-rank approximation technique in [45] by considering the calculation of the unified representation $\mathbf{h}$'s $k$th element, $\mathbf{h}(k)$. Suppose that the weight $\mathcal{W}$ is yielded by stacking $D$ tensors $\mathcal{W} = [\mathcal{W}_1; \mathcal{W}_2 \cdots ; \mathcal{W}_D]$, where $\mathcal{W}_k \in \mathbb{R}^{1\times d_l^{(1)}\times d_l^{(2)}\times\cdots d_l^{(V)}}$ and $k = 1, \ldots, D$. Thus, we have

$$\mathbf{h}(k) = \mathcal{W}_k \cdot \mathcal{Z} + b(k) \tag{12}$$

where $b(k)$ is the $k$th element of $b$. Then, $\mathcal{W}_k$ can be approximated by a set of learnable vectors as follows:

$$\mathcal{W}_k = \sum_{r=1}^{R}\bigotimes_{v=1}^{V}\mathbf{w}_{r,k}^{(v)} \tag{13}$$

where $\mathbf{w}_{r,k}^{(v)} \in \mathbb{R}^{d_l^{(v)}}$, and $R$ is the rank of low-rank approximation. Since $\mathcal{Z} = \bigotimes_{v=1}^V \mathbf{h}^{(v)}$, tensor-based fusion can be computed in a highly efficient manner by rearranging the order of inner product and outer product [45], which enables tensor-based fusion to be computationally tractable.

### *B. Alignment-Based Solutions*

Compared with multiview fusion, multiview alignment is another popular category of methods in multiview deep learning. It does not require to obtain a joint embedding. Instead, they attempt to align the representations learned by different views, so as to make those representations share some common characteristics. Likewise, we also present the overall framework and specific implementations of alignment-based multiview deep AD solutions in the following.

1) *Framework:* In a training batch with $N$ multiview data, we denote the embeddings of the $n$th multiview data $\{\mathbf{x}_n^{(v)}\}_{v=1}^V$ by $\{\mathbf{h}_n^{(v)}\}_{v=1}^V$, which are learned by a set of encoder networks $\{\text{Enc}^{(v)}\}_{v=1}^V$. Then, an alignment function $F_a$ is defined to compute a quantitative measure of alignment across learned embeddings of different views

$$\mathcal{A} = F_a\left(\left\{\{\mathbf{h}_n^{(v)}\}_{v=1}^V\right\}_{n=1}^N\right), \quad F_a \in \mathcal{F}_a \tag{14}$$

where $\mathcal{F}_a$ is the set of available alignment functions. As shown in (14), a key difference between alignment-based solutions and fusion-based solutions is that fusion usually occurs within one multiview data, while the alignment of two views can involve multiple multiview data. To maximize the alignment across different views, we can equivalently minimize the alignment loss $\mathcal{L}_a = -\mathcal{A}$. Similar to fusion-based solutions, we also resort to the reconstruction paradigm and a set of decoder networks $\{\text{Dec}^{(v)}\}_{v=1}^V$ to guide the training of DNNs. As a result, alignment-based solutions minimize the following loss function:

$$\mathcal{L} = \mathcal{L}_r + \alpha\,\mathcal{L}_a \tag{15}$$

where $\mathcal{L}_r$ is the reconstruction loss defined in (6), and $\alpha$ is the weight of alignment loss. Given the testing data $\{\mathbf{x}_{\text{test}}^{(v)}\}_{v=1}^V$, we also leverage the reconstruction errors as baseline scores, which is the same as (7). An intuitive illustration of the framework is given in Fig. 3.

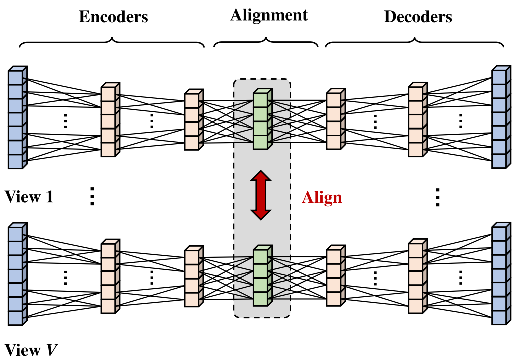


Fig. 3. Alignment-based solutions for multiview deep AD.

*2) Implementations:* The core issue of alignment-based solutions is the design of alignment function $F_a$. Inspired by the literature of multiview deep learning, we propose to implement the alignment function by the following ways.

1) *Distance-Based Alignment (DIS):* A commonly seen technique to align two embeddings is to minimize their distance, as a smaller distance usually indicates better alignment. Therefore, we propose to adopt the widely used pair-wise $L_p$-norm distance of all embeddings to measure alignment

$$F_a\left(\left\{\left\{\mathbf{h}_n^{(v)}\right\}_{v=1}^{V}\right\}_{n=1}^{N}\right) = -\sum_{n=1}^{N}\sum_{i=1}^{V-1}\sum_{j=i+1}^{V}\left|\left|\mathbf{h}_n^{(i)} - \mathbf{h}_n^{(j)}\right|\right|_p^p \tag{16}$$

where $||\cdot||$ denotes the $L_p$-norm, and $p$ is a non-negative integer. As can be seen from (16), it requires the embeddings of different views to share a common dimension. Note that the alignment function in (16) is equivalent to the correspondent AE proposed in [51], which leverages multiview alignment for cross-modal retrieval. A drawback of such an alignment function is that it only performs the view alignment within one multiview data.

2) *Similarity-Based Alignment (SIM):* In addition to distance, similarity is another intuitive way to measure the degree of alignment. Given a similarity function $s(\cdot)$, we can similarly define an alignment function, such as (16). Nevertheless, such an alignment function only considers the view similarity within one multiview data. To consider the view similarity across different data, we are inspired by [50] and propose to adopt a more sophisticated similarity measure between different views: For the $i$th and $j$th view, a similarity loss $Sim(i, j)$ is computed based on $s(\cdot)$ and a hinge loss

$$\mathrm{Sim}(i, j) = \sum_{a\neq b}\max\left\{0, m - s\left(\mathbf{h}_a^{(i)}, \mathbf{h}_a^{(j)}\right) + s\left(\mathbf{h}_a^{(i)}, \mathbf{h}_b^{(j)}\right)\right\} \tag{17}$$

where $m$ is a margin. The abovementioned similarity loss encourages the embeddings from the same multiview data to be similar, while embeddings from two different multiview data to be dissimilar. The similarity function $s(\cdot)$ can be realized by multiple forms, such as inner product and cosine similarity. Then, the final alignment function can be calculated by

$$F_a\left(\left\{\left\{\mathbf{h}_n^{(v)}\right\}_{v=1}^{V}\right\}_{n=1}^{N}\right) = -\sum_{i=1}^{V-1}\sum_{j=i+1}^{V}\mathrm{Sim}(i, j). \tag{18}$$

3) *Correlation-Based Alignment (DCCA):* CCA is a classic statistical technique for finding the maximally correlated linear projections of two vectors. Thus, a natural way for us to align different views in deep learning is the CCA's deep variant, DCCA [47]. To conduct correlation-based alignment, we intend to maximize the correlation between two views. Specifically, we stack the $v$th view's embeddings of $N$ multiview data in a training batch into a $d_l^{(v)} \times N$ embedding matrix: $\mathbf{H}^{(v)} = [\mathbf{h}_1^{(v)}, \ldots, \mathbf{h}_N^{(v)}]$, while $\mathbf{H}^{(v)}$ can be centered by $\bar{\mathbf{H}}^{(v)} = \mathbf{H}^{(v)} - {}^1\!/\!{}_N\mathbf{H}^{(v)} \cdot \mathbf{1}$, where $\mathbf{1}$ is an $N \times N$ all-1 matrix. With the embedding matrix $\mathbf{H}^{(i)}$ and $\mathbf{H}^{(j)}$ for the $i$th and $j$th view, we first estimate the covariance matrices $\sum_{ii} = (1/(N-1))\bar{\mathbf{H}}^{(i)} \cdot \bar{\mathbf{H}}^{(i)\top} + r\mathbf{I}$, $\sum_{ij} = (1/(N-1))\bar{\mathbf{H}}^{(i)} \cdot \bar{\mathbf{H}}^{(j)\top}$, and $\sum_{jj} = (1/(N-1))\bar{\mathbf{H}}^{(j)} \cdot \bar{\mathbf{H}}^{(j)\top} + r\mathbf{I}$, where $r$ is the coefficient for regularization, and $\mathbf{I}$ is an identity matrix. With estimated covariance matrices, we compute an intermediate matrix $T_{ij} = \sum_{ii}^{-1/2} \cdot \sum_{ij} \cdot \sum_{jj}^{-1/2}$. It can be proved that the correlation of view $i$ and $j$ is the matrix trace norm of $T_{ij}$ [47]

$$\mathrm{Corr}(i, j) = ||T_{ij}||_{tr} = tr\left(T_{ij}^{\top} \cdot T_{ij}\right)^{\frac{1}{2}}. \tag{19}$$

The final alignment function can be calculated by

$$F_a\left(\left\{\left\{\mathbf{h}_n^{(v)}\right\}_{v=1}^{V}\right\}_{n=1}^{N}\right) = \sum_{i=1}^{V-1}\sum_{j=i+1}^{V}\mathrm{Corr}(i, j). \tag{20}$$

### *C. Deep AD Tailored Solutions*

Apart from baselines based on multiview deep learning, we design the third type of baseline solutions by tailoring the existing deep AD solutions. The basic idea is to train a deep AD model for data of each view. During inference, the AD results of each view are fused to yield the final results. The framework and specific implementations of deep AD tailored solutions are presented in the following.

*1) Framework:* Suppose that the deep AD model $\mathcal{M}^{(v)}$ is trained with data from the $v$th view. Given the newly incoming multiview data $\{\mathbf{x}_{\text{test}}^{(v)}\}_{v=1}^{V}$, the AD result for the $v$th view is given by

$$\mathcal{S}^{(v)} = \mathcal{M}^{(v)}\left(\mathbf{x}_{\text{test}}^{(v)}\right), \quad v = 1, \ldots, V. \tag{21}$$

The final score for the multiview data $\{\mathbf{x}_{\text{test}}^{(v)}\}_{v=1}^{V}$ is computed by a late fusion function $F_l(\cdot)$

$$\mathcal{S}\left(\left\{\mathbf{x}_{\text{test}}^{(v)}\right\}_{v=1}^{V}\right) = F_l\left(\mathcal{S}^{(1)}\left(\mathbf{x}_{\text{test}}^{(1)}\right), \mathcal{S}^{(2)}\left(\mathbf{x}_{\text{test}}^{(2)}\right), \ldots, \mathcal{S}^{(V)}\left(\mathbf{x}_{\text{test}}^{(V)}\right)\right). \tag{22}$$

An intuitive illustration of the framework is given in Fig. 4.

*2) Implementations:* The choice of deep AD model plays a center role in designing tailored deep AD solutions. This article introduces two representative deep AD methods in the literature to construct baseline models for multiview deep AD: standard deep AEs (DAEs) and the recent DSVDD [27].

1) *DAE-Based Solution (DAE):* DAE leverages DNNs as the encoder $\mathrm{Enc}^{(v)}$ and the decoder $\mathrm{Dec}^{(v)}$ to reconstruct input data from a low-dimensional embedding. Formally, given $N$

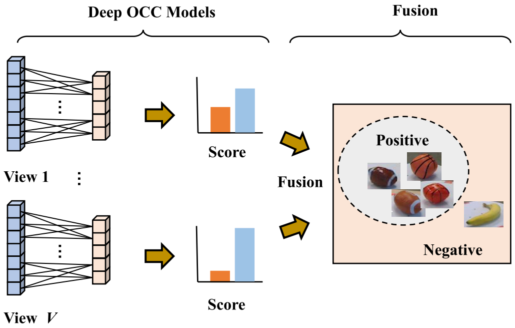


Fig. 4. Framework of deep AD tailored solutions.

training data $\{\mathbf{x}_n^{(v)}\}_{n=1}^{N}$ from the $v$th view, DAE requires to minimize the reconstruction loss

$$\min_{\boldsymbol{\theta}_E^{(v)},\boldsymbol{\theta}_D^{(v)}} \frac{1}{N}\sum_{n=1}^{N}\left\|\mathrm{Dec}^{(v)}\left(\mathrm{Enc}^{(v)}\left(\mathbf{x}_n^{(v)}\right)\right)-\mathbf{x}_n^{(v)}\right\|_2^2 + \frac{\lambda}{2}\left(\left\|\boldsymbol{\theta}_E^{(v)}\right\|_2^2+\left\|\boldsymbol{\theta}_D^{(v)}\right\|_2^2\right) \tag{23}$$

where $\boldsymbol{\theta}_E^{(v)}$ and $\boldsymbol{\theta}_D^{(v)}$ represent the learnable parameters of the encoder network $\mathrm{Enc}^{(v)}$ and decoder network $\mathrm{Dec}^{(v)}$ respectively, and $\lambda$ is the weight for the $L_2$-norm regularization term. For inference, the reconstruction errors are often directly used as scores

$$\mathcal{S}^{(v)} = -\left\|\mathrm{Dec}^{(v)}\left(\mathrm{Enc}^{(v)}\left(\mathbf{x}_{\text{test}}^{(v)}\right)\right)-\mathbf{x}_{\text{test}}^{(v)}\right\|_2^2. \tag{24}$$

DAE-based baseline is also viewed as the most fundamental baseline for multiview deep AD.

2) *DSVDD-Based Solution (DSV):* In general, DSVDD intends to map embeddings of data from the normal class $\{\mathbf{h}_n^{(v)}\}_{n=1}^{N}$ to a hypersphere with minimal radius. To be more specific, DSVDD can be implemented by a simplified version or a soft-boundary version [27]. Since the simplified version enjoys less hyperparameters and better performance in practice, we choose it to perform multiview deep AD. Specifically, simplified DSVDD encourages embeddings of all data to be as close to a center $\mathbf{c}^{(v)}$ as possible. Formally, simplified DSVDD requires to solve the following optimization problem:

$$\min_{\boldsymbol{\theta}_E^{(v)}} \frac{1}{N}\sum_{n=1}^{N}\left\|\mathrm{Enc}^{(v)}\left(\mathbf{x}_n^{(v)}\right)-\mathbf{c}^{(v)}\right\|_2^2+\frac{\lambda}{2}\left\|\boldsymbol{\theta}_E^{(v)}\right\|_2^2 \tag{25}$$

where the values of $\boldsymbol{\theta}^{(v)}$ are the learnable parameters of DSVDD, and $\lambda$ is the weight of the $L_2$-norm regularization term. The encoder $\mathrm{Enc}^{(v)}$ can be pre-trained in a DAE fashion. The non-zero center $\mathbf{c}^{(v)}$ is initialized before training and can be adjusted during training. The abovementioned optimization problem can be efficiently solved by gradient descent. During inference, one can score the test data $\mathbf{x}_{\text{test}}^{(v)}$ by calculating the distance between its embedding $\mathbf{h}_{\text{test}}^{(v)}$ and the center

$$\mathcal{S}^{(v)} = -\left\|\mathrm{Enc}^{(v)}\left(\mathbf{x}_{\text{test}}^{(v)}\right)-\mathbf{c}^{(v)}\right\|_2^2. \tag{26}$$

### *D. Self-Supervision-Based Solutions*

Self-supervised learning is a hot topic in recent research, and it has been demonstrated as a highly effective way to conduct unsupervised representation learning [54]. Specifically, self-supervised learning introduces a certain pretext task to provide additional supervision signal and enable better representation learning. Due to the lack of supervision in multiview deep AD, creating self-supervision can be an appealing solution. Multiview data intrinsically contain richer information than single-view data, which makes it possible to design pretext tasks in a more flexible way. In this section, we mainly focus on designing generative pretext tasks to realize self-supervised multiview deep AD. We also explore discriminative pretext tasks in the Supplementary Material.

1) *Framework:* The intuition for generative pretext tasks is to generate data from some views based on other views. Formally, given the multiview data $\{\mathbf{x}^{(v)}\}_{v=1}^{V}$, we partition the view indices into two subsets $\mathcal{P}$ and $\mathcal{Q}$, which satisfy

$$\mathcal{P} \neq \mathcal{Q}, \quad \mathcal{P} \cup \mathcal{Q} = \{1, 2, \ldots, V\}. \tag{27}$$

Note that the intersection of $\mathcal{P}$ and $\mathcal{Q}$ may not be empty. By $\mathcal{P}$ and $\mathcal{Q}$, we can partition the multiview data into two sets of data, $\{\mathbf{x}^{(i)}\}_{i\in\mathcal{P}}$ and $\{\mathbf{x}^{(j)}\}_{j\in\mathcal{Q}}$. The goal of generative pretext tasks is to generate $\{\mathbf{x}^{(j)}\}_{j\in\mathcal{Q}}$ by taking $\{\mathbf{x}^{(i)}\}_{i\in\mathcal{P}}$ as an input. To fulfill this task, we propose to introduce $|\mathcal{P}|$ encoder networks and $|\mathcal{Q}|$ decoder networks, where $|\cdot|$ denotes the number of elements in the set. $\{\mathbf{x}^{(i)}\}_{i\in\mathcal{P}}$ is first mapped to the embedding set $\{\mathbf{h}^{(i)}\}_{i\in\mathcal{P}}$ by encoders, and a joint embedding is then obtained by

$$\mathbf{h}^{\mathcal{P}} = F_f(\{\mathbf{h}^{(i)}\}_{i\in\mathcal{P}}) \tag{28}$$

where $F_f(\cdot)$ can be any fusion function defined in Section IV-A2. The decoder networks use $\mathbf{h}^{\mathcal{P}}$ as the input to infer the data $\{\mathbf{x}^{(j)}\}_{j\in\mathcal{Q}}$, which aims to learn

$$\min_{\boldsymbol{\theta}_D^{(j)}} \sum_{j\in\mathcal{Q}} ||\mathrm{Dec}^{(j)}(\mathbf{h}^{\mathcal{P}})-\mathbf{x}^{(j)}||_2^2 \tag{29}$$

where $\boldsymbol{\theta}_D^{(j)}$ is the set of learnable weights for decoder $\mathrm{Dec}^{(j)}$. In this way, $\{\mathbf{x}^{(j)}\}_{j\in\mathcal{Q}}$ is used as supervision to guide the training of encoders/decoders. Similarly, the generation errors $\ell(\mathrm{Dec}^{(j)}(\mathbf{h}^{\mathcal{P}}), \mathbf{x}^{(j)})$ can be used for scoring during inference.

2) *Implementations:* There are many ways to divide $\mathcal{P}$ and $\mathcal{Q}$, and we select two of them to build our baseline solutions here.

1) *Plain Prediction (PPRD):* Here, $\mathcal{Q} = \{v\}$ and $\mathcal{P} = \{1, 2, \ldots, V\} - \mathcal{Q}$. It means that we predict data from the $v$th view by data from the rest of views. Due to the lack of standard to select a specific $v$, we vary $v$ from 1 to $V$ and alternatively use each view as learning target, which results in multiple rounds of prediction. To avoid excessive computational cost, we introduce $V$ encoders and $V$ decoders in total, and the $v$th encoder and decoder are specifically responsible for data of the $v$th view in each round of prediction. The final score is yielded by averaging the results of all rounds of prediction.

2) *Split Prediction (SPRD):* Here, $\mathcal{P} = \{v\}$ and $\mathcal{Q} = \{1, 2, \ldots, V\}$. It means that we predict data of all views by a data from the $v$th view. Likewise, we also alternatively use data of each view to predict data of all views and introduce $V$ encoders and $V$ decoders that are shared in different rounds of prediction. As shown earlier, generative pretext tasks aim to maximally capture the inter-view correspondence during

representation learning, which cannot be realized by previous baseline solutions.

### E. Additional Remarks

*1) Late Fusion:* Except for the self-supervision-based solution that uses discriminative pretext tasks, all other baselines require to fuse the results yielded by different views via a late fusion function $F_l$. For traditional tasks, such as classification and clustering [55], [56], numerous strategies have been proposed to carry out late fusion. However, since AD lacks discriminative supervision information and trains the model with only data from the normal class, it is not straightforward to exploit prior knowledge or propose an assumption on different views to perform late fusion. Thus, considering that the average strategy is usually viewed as a non-trivial baseline in traditional multiview learning [57], [58], we also adopt the simple averaging strategy for late fusion in our baseline solutions mentioned earlier, namely

$$F_l(\mathcal{S}^{(1)}, \mathcal{S}^{(2)}, \ldots, \mathcal{S}^{(V)}) = \frac{1}{V}\sum_{v=1}^{V} \mathcal{S}^{(v)}. \tag{30}$$

Apart from the simple averaging, one can certainly adopt more sophisticated late fusion strategy, such as the covariance-based late fusion strategy proposed in [19]. However, our later empirical evaluations show that late fusion-based averaging is a fairly strong baseline, which often prevails in both effectiveness and efficiency.

*2) Other Potential Baselines:* Actually, we have explored more ways to design baseline solutions for multiview deep AD. Eleven baseline solutions presented earlier are the most representative ones that enjoy easier implementation, satisfactory performance, and sound extendibility. Due to the limit of pages, we introduce other potential baseline solutions in the Supplementary Material. Besides, we also introduce two latest methods that are designed for unsupervised multiview/multimodal deep AD [37], which is essentially multimodal deep OD/MDOD by our definition in Section II-B. We customize them to learn from pure normal training data and design two additional solutions: multiview OD based on deep intact space (MODDIS) [36] and cross-aligned AE (CAAE) [37]. Their performances are also reported in experiments as a reference.

## V. Benchmark Datasets

### A. Limitations of Existing Datasets

Public benchmark datasets play a pivotal role in prompting the development of machine learning algorithms. However, as we briefed in Section I (the third issue), the existing datasets basically suffer from some important limitations when they are used for evaluating multiview deep AD: First, *the existing multiview datasets are not adequate for multiview deep AD*. To be more specific, frequently used multiview benchmark datasets (e.g., *Flower17/Flower102*[1]) are originally designed to evaluate traditional multiview learning algorithms, and the number of samples is too small to train DNNs (the average sample number of a class is often less than 100). As a consequence, very few existing multiview datasets can be directly adopted for multiview deep AD. Second, *popular benchmark datasets for deep learning are typically single view*. Recent years have witnessed a surging interest in deep learning, which gives rise to a rapid growth of available benchmark datasets. By contrast, multiview deep learning is still a relatively new area with much less applicable benchmark datasets. Third, *most importantly, very few benchmark dataset is specifically designed for the background of multiview deep AD*. As we introduced in Section I, multiview deep AD actually enjoys broad applications in many realms, such as vision-based AD and fault detection, but benchmark datasets in such background are quite rare. In the literature, many works adopt the "one *versus* all" protocol to convert a binary/multiclass dataset into an AD dataset, which is non-comprehensive for the evaluation of multiview deep AD.

To this end, we need to build more benchmark datasets for multiview deep AD. However, collecting multiview data from scratch can be expensive and time-consuming, and it takes a long time to obtain sufficient and diverse benchmark datasets in this way. Therefore, our strategy is to extensively collect existing public datasets that come from mature public benchmark datasets, and process them via various means into proper multiview datasets, so as to construct abundant benchmark datasets in a highly efficient manner. Collected data and processing techniques will be elaborated in the following.

### B. More Multiview Benchmark Datasets

We intend to build our new multiview benchmark datasets based on vision data, which is due to the fact that computer vision is the earliest realm where deep learning is thoroughly studied and successfully applied. Hence, abundant accessible public vision data can be exploited in the realm. Specifically, we process existing data into *image-based multiview datasets* and *video-based multiview datasets*.

*1) Image-Based Multiview Datasets:* Image data are the most fundamental data type in deep learning. We process image data into multiview data by the two means: First, *multiple image descriptors*: many image descriptors have been proposed to depict different attributes of images, such as texture, color, and gradient. Therefore, it is natural to convert a single image into a multiview data by describing it with different image descriptors. In this article, we choose several popular image benchmark datasets with comparatively small images (e.g., 32 × 32 images), which are less complex for image descriptors to depict: *MNIST*,[2] *FashionMNIST*,[3] *CIFAR10*,[4] *SVHN*,[5] *CIFAR100*[4], and nine image datasets from the *MedMNIST* dataset collection.[6] To obtain multiview data, we extract six types of features, i.e., color histogram, GIST, Histogram of Oriented Gradient (HOG)2 × 2, HOG3 × 3, Local Binary Pattern (LBP), and Scale-Invariant Feature Transform (SIFT), which are implemented by a feature extraction toolbox.[7] Second, *multiple pre-trained DNN mod-*

[1] https://www.robots.ox.ac.uk/vgg/data/flowers/17/

[2] http://yann.lecun.com/exdb/mnist/

[3] https://github.com/zalandoresearch/fashion-mnist/

[4] https://www.cs.toronto.edu/~kriz/cifar.html

[5] http://ufldl.stanford.edu/housenumbers/

[6] https://medmnist.github.io/

[7] https://github.com/adikhosla/feature-extraction

*els*: classic image descriptors often find it hard to describe high-resolution images effectively. To convert high-resolution images to multiview data, we propose to describe them by multiple pre-trained DNN models with different network architectures. Those DNN models are usually pretrained on a large-scale generic image dataset, such as ImageNet [59], while different network architectures enable them to acquire image knowledge from different views. We extract the outputs from the penultimate layer of each pretrained DNN model as the representations of the image. For high-resolution image data, we collect image data from the *Cat_vs_Dog*[8] dataset and *MvTecAD*[9] dataset collection that contains 15 datasets. As for DNN architectures, we select VGGNet [60], Inceptionv3 [61], ResNet34 [62], and DenseNet121 [63] pretrained on ImageNet. It is worth mention that the *MvTecAD* dataset collection is specifically designed for evaluating AD models, which makes it even more favorable for multiview deep AD.

*2) Video-Based Multiview Datasets:* Compared with image data, video data contain both spatial and temporal information, so it is even more natural to transform them into multiview representation. Since AD is a representative application of AD, we simply collect video data from benchmark datasets that are designed for the video AD (VAD) task [64]. To yield video data from a different view, we calculate the optical flow map of each video frame by a pretrained FlowNetv2 model [65]. In this way, each video frame is represented from the view of both RGB and optical flow, which depicts videos by both appearance and motion. Afterward, we leverage the joint foreground localization strategy from [66], so as to localize both daily and novel video foreground objects by bounding boxes. Based on those bounding boxes, we can extract both corresponding RGB and optical flow patches from the original video frame and optical flow map, respectively, which serve as a two-view representation of each foreground object in videos. Extracted patches are then normalized into the same size (32 × 32 patches). As for VAD datasets, we select *UCSDped1/UCSDped2*,[10] *Avenue*,[11] *UMN*[12], and *ShanghaiTech*.[13] For VAD datasets that provide pixel-level ground-truth mask for abnormal video foreground (*UCSD ped1/ped2*, *Avenue*, and *ShanghaiTech*), those patches that are overlapped with any anomaly mask are labeled as 0, and other patches are labeled as 1. Although *UMN* dataset does not provide pixel-level mask, its anomalies happen at a certain stage, and all foreground objects exhibit abnormal behavior at that stage. Therefore, we simply label each foreground patch in that stage as 0, otherwise labeled as 1. In this way, we can yield video-based multiview datasets, which are readily applicable to evaluate multiview deep AD with real-world application background.

### C. Existing Multiview/Multimodal Datasets

Besides, we collect some existing multiview/multimodal datasets for more comprehensive evaluation. We collect 13 multiview/multimodal datasets: *Citeseer*,[14] *Cora*[14], *Reuters*[14], *BBC*,[15] *Wiki*,[16] *BDGP*,[17] *Caltech20*,[18] *AwA*,[19] *NUS-Wide*,[20] *SunRGBD*,[21] *YoutubeFace*[22] (shorted as *YtFace*), *CMU-MOSEI*[23], and *DriverAD*,[24] which cover a wide range of scales and data types. For those multiview/multimodal datasets that are not specifically designed for AD, we adopt the "one *versus* all" protocol to evaluate multiview deep AD methods on them: At each round, a certain class of the dataset is viewed as the normal class, while all of other classes are viewed as the negative class. The final AD performance can be obtained by averaging the performance of all rounds. The selection criterion is that at least one class in the multiview dataset can provide more than 300 data for training. As *DriverAD* dataset is designed for AD [67], we use the normal videos in the training set as training data, while the evaluation is performed on the test set that contains both normal and abnormal videos. A summary of all multiview/multimodal benchmark datasets used in this article is given in the Supplementary Material.

## VI. Empirical Evaluations

Having established formulation, baselines, and benchmark datasets for multiview deep AD, we perform empirical evaluations to give the first glimpse into this new topic. In addition to head-to-head performance comparison between different baselines, we also conduct an in-depth analysis on the characteristics of each model.

### A. Experimental Setup

For multiview datasets that are specifically designed for AD (*MvTecAD*, video-based multiview datasets, and *DriverAD*), we directly use the given normal class to train the AD model, and data from the abnormal class are used to evaluate AD performance. For other binary or multiclass multiview datasets, we apply the "one *versus* all" protocol (detailed in Section V-C) for training and evaluating the AD performance. For multiclass datasets that possess more than ten classes, we select the first ten qualified classes ($\geq$ 300 training data) for experiments. For those multiview datasets that have already provided the train/test split, we simply use the data of normal class in the training set to train the AD model, and the test set is used to evaluate the AD performance. As for those datasets that do not provide train/test split, we randomly sample 70% data of the current normal class as the training set, while the rest of normal class data are mixed with data of the negative class to serve as the testing set. The sampling process is repeated for ten times, and the average performance is reported. Before training, training data from each view are

[8] www.diffen.com/difference/Cat_vs_Dog/

[9] https://www.mvtec.com/company/research/datasets/mvtec-ad/

[10] http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm

[11] http://www.cse.cuhk.edu.hk/leojia/projects/detectabnormal/dataset.html

[12] http://mha.cs.umn.edu/proj_events.shtml#crowd

[13] https://svip-lab.github.io/dataset/campusdataset.html

[14] http://lig-membres.imag.fr/grimal/data.html

[15] http://mlg.ucd.ie/datasets/segment.html

[16] http://www.svcl.ucsd.edu/projects/crossmodal/

[17] http://ranger.uta.edu/~heng/Drosophila/

[18] http://www.vision.caltech.edu/Image_Datasets/Caltech101/

[19] https://cvml.ist.ac.at/AwA/

[20] https://lms.comp.nus.edu.sg/wp-content/uploads/2019/research/nuswide/NUS-WIDE.html

[21] http://rgbd.cs.princeton.edu/

[22] http://archive.ics.uci.edu/ml/datasets/YouTube+Multiview+Video+Games+Dataset

[23] https://github.com/A2Zadeh/CMU-MultimodalSDK

[24] https://github.com/okankop/Driver-Anomaly-Detection

TABLE I

AUROC (%) OF DIFFERENT BASELINES ON IMAGE-BASED MULTIVIEW DATASETS (BEST PERFORMER IN BOLDFACE)

| Type | | MNIST | FashionMNIST | Cifar10 | Cifar100 | SVHN | Cat_vs_Dog | MedMNIST | MvTecAD |
|---|---|---|---|---|---|---|---|---|---|
| Fusion | SUM | 97.57 | 92.76 | 81.10 | **76.04** | 77.89 | 97.88 | 78.63 | **89.83** |
| | MAX | 97.53 | 92.60 | 80.24 | 75.41 | 77.37 | 97.89 | 78.62 | 89.62 |
| | NN | 97.61 | 92.87 | 79.61 | 75.42 | 77.93 | 97.99 | 78.68 | 89.22 |
| | TF | 97.58 | 92.54 | 80.60 | 75.47 | 78.40 | 97.74 | 78.42 | 89.06 |
| Alignment | DIS | 97.52 | 92.69 | 80.77 | 75.72 | 78.80 | 97.96 | 78.86 | 89.28 |
| | SIM | 97.57 | 92.67 | 81.22 | 76.00 | 79.03 | 98.01 | 78.71 | 89.58 |
| | DCCA | 97.59 | 91.97 | 76.53 | 73.58 | **79.06** | 95.11 | 78.78 | 89.81 |
| Tailored | DAE | 97.57 | 92.64 | 81.05 | 75.70 | **79.06** | **98.02** | 78.52 | 89.44 |
| | DSV | 97.20 | 91.89 | 75.15 | 70.70 | 70.48 | 86.07 | 77.57 | 80.43 |
| Self-supervision | PPRD | 97.51 | 92.76 | 80.31 | 74.89 | 75.93 | 97.72 | 79.35 | 89.53 |
| | SPRD | **97.63** | 92.98 | **81.30** | 75.52 | 77.29 | 97.90 | **79.59** | 89.50 |
| MDOD | MODDIS | 93.86 | 86.54 | 64.40 | 63.67 | 57.40 | 32.87 | 75.49 | 75.02 |
| | CAAE | 97.52 | **93.00** | 74.10 | 70.53 | 70.58 | 75.40 | 78.02 | 83.87 |

TABLE II

AUROC (%) OF DIFFERENT BASELINES ON VIDEO-BASED MULTIVIEW DATASETS (BEST PERFORMER IN BOLDFACE)

| Type | | UCSDped1 | UCSDped2 | UMN_scene1 | UMN_scene2 | UMN_scene3 | Avenue | ShanghaiTech |
|---|---|---|---|---|---|---|---|---|
| Fusion | SUM | **83.26** | 86.66 | 97.99 | 88.08 | 90.59 | 84.26 | 67.41 |
| | MAX | 81.48 | 83.85 | 97.70 | 87.08 | 89.73 | 83.54 | 64.75 |
| | NN | 82.19 | 83.81 | 98.15 | 87.35 | 90.55 | 83.92 | 67.13 |
| | TF | 82.95 | 86.17 | 98.12 | 87.78 | 90.82 | 84.28 | 66.86 |
| Alignment | DIS | 81.08 | 82.18 | 97.28 | 86.89 | 90.35 | 82.53 | 64.35 |
| | SIM | 82.92 | 84.12 | 97.32 | 86.74 | 89.35 | 79.08 | 65.53 |
| | DCCA | 77.61 | 84.62 | 96.79 | 86.18 | 89.68 | 82.56 | 65.77 |
| Tailored | DAE | 80.51 | 81.86 | 97.01 | 87.47 | 88.91 | 83.35 | 64.93 |
| | DSV | 66.30 | 87.02 | 98.47 | 83.49 | 94.03 | 83.35 | 59.12 |
| Self-supervision | PPRD | 78.40 | **89.62** | 98.45 | 86.31 | **94.76** | 80.73 | 49.89 |
| | SPRD | 79.14 | 88.49 | 98.51 | **88.41** | 93.12 | 82.11 | 56.09 |
| MDOD | MODDIS | 78.07 | 83.28 | 98.64 | 85.48 | 93.88 | **84.61** | 55.54 |
| | CAAE | 76.68 | 86.66 | **98.60** | 85.17 | 93.95 | **84.61** | **70.24** |

normalized into the interval $[-1, 1]$, while the testing set is similarly normalized by the statistics (i.e., min–max value) of the training set. For inference, the reconstruction error-based scores of each view are further normalized by the input data dimension, which aims to make scores from different views share the same scale, so they can be comparable and applicable to averaging-based late fusion. To quantify the AD performance, we follow the deep AD literature and utilize three commonly used threshold-independent metrics: area under the receiver operation characteristic curve (AUROC), area under the precision–recall curve (AUPR), and true negative rate at 95% true positive rate (TNR@95%TPR). We also provide more implementation details in the Supplementary Material.

### B. Head-to-Head Comparison of Baselines

We test the designed 11 multiview deep AD solutions on both our new multiview datasets and selected existing multiview datasets. Due to the page limit, we report the most frequently used AUROC of each baseline for the head-to-head comparison, while the results under other metrics are provided in the supplementary material. Since other metrics actually exhibit a similar trend to AUROC, we will focus on discussing the AUROC performance in this section. The experimental results on image-based multiview datasets, video-based multiview datasets, and selected existing multiview datasets are given in Tables I–IV. Note that the performance of *MedMNIST* and *MvTecAD* is given by averaging the performance of each datasets in a collection (detailed results of each dataset in those dataset collections are reported in the Supplementary Material). From those results, we can draw the following observations.

1) *In some cases, most of the baseline solutions actually achieve fairly close performance*, despite of their differences in type and implementation. Concretely, as shown in Table III, baseline solutions attain almost identical performance on several existing multiview datasets that are widely used in the literature, e.g., *BBC*, *Caltech20*, *Cora*, and *Reuters*. On many image-based multiview datasets, we also note that the best performer usually leads other counterparts by a less than 1% AUROC. However, baselines could also obtain evidently different performance on other multiview datasets, such as some video-based and image-based datasets. This also justifies the necessity for a comprehensive evaluation.
2) *There does not exist a single baseline that can consistently outperform other baselines*. For example, we notice that self-supervision-based baselines (PPRD and SPRD) attain the optimal or near-optimal performance (i.e., not significantly different from the best performer) on 16 out of the total 28 datasets (*MedMNIST* and *MvTecAD* are viewed as two datasets here). However, self-supervision-based baselines also suffer from evidently inferior performance to other baselines on some datasets, such as *UCSDped1* and *ShanghaiTech*.
3) *Simple fusion functions (SUM and MAX) can readily compete with comparatively complex fusion functions (NN and TF)*. In fact, all fusion-based baselines yield fairly comparable performance on most datasets. To our

TABLE III

AUROC (%) OF DIFFERENT BASELINES ON THE EXISTING MULTIVIEW/MULTIMODAL DATASETS WITH RANDOM TRAIN/TEST SET SPLIT. THE VALUE IN THE BRACKET IS THE $p$-VALUE OF STUDENT-$t$ TEST ($p < 0.05$ INDICATES A SIGNIFICANT DIFFERENCE FROM THE BEST PERFORMER)

| | BBC | BDGP | Caltech20 | Citeseer | Cora | Reuters | Wiki | AwA | NUS-Wide | SunRGBD |
|---|---|---|---|---|---|---|---|---|---|---|
| SUM | 94.35$_{\pm 0.54}$<br>(1.00) | 81.27$_{\pm 0.77}$<br>(0.06) | 99.76$_{\pm 0.11}$<br>(0.23) | 83.85$_{\pm 0.33}$<br>(0.89) | 87.79$_{\pm 0.53}$<br>(0.98) | 65.05$_{\pm 0.43}$<br>(0.88) | 88.84$_{\pm 0.80}$<br>(0.00) | 63.15$_{\pm 0.72}$<br>(0.32) | 67.94$_{\pm 0.54}$<br>(0.02) | 84.81$_{\pm 0.45}$<br>(1.00) |
| MAX | 94.35$_{\pm 0.54}$<br>(1.00) | 81.36$_{\pm 0.84}$<br>(0.10) | 99.69$_{\pm 0.17}$<br>(0.05) | 83.86$_{\pm 0.33}$<br>(0.96) | 87.79$_{\pm 0.51}$<br>(0.98) | 65.04$_{\pm 0.42}$<br>(0.85) | 88.93$_{\pm 0.64}$<br>(0.00) | 63.34$_{\pm 0.77}$<br>(0.64) | 68.25$_{\pm 0.49}$<br>(0.15) | 84.63$_{\pm 0.51}$<br>(0.60) |
| NN | 94.35$_{\pm 0.54}$<br>(1.00) | 80.98$_{\pm 0.84}$<br>(0.01) | 99.77$_{\pm 0.11}$<br>(0.27) | 83.87$_{\pm 0.34}$<br>(0.99) | 87.78$_{\pm 0.52}$<br>(0.97) | 65.03$_{\pm 0.42}$<br>(0.81) | 88.85$_{\pm 0.51}$<br>(0.00) | 63.27$_{\pm 0.71}$<br>(0.50) | 68.56$_{\pm 0.47}$<br>(0.81) | 84.55$_{\pm 0.42}$<br>(0.42) |
| TF | 94.35$_{\pm 0.54}$<br>(1.00) | 81.03$_{\pm 0.86}$<br>(0.02) | 97.79$_{\pm 1.30}$<br>(0.00) | 83.87$_{\pm 0.32}$<br>(0.98) | 87.78$_{\pm 0.52}$<br>(0.96) | 65.05$_{\pm 0.42}$<br>(0.88) | 89.24$_{\pm 1.03}$<br>(0.00) | 62.76$_{\pm 0.72}$<br>(0.05) | 67.24$_{\pm 0.56}$<br>(0.00) | 84.37$_{\pm 0.56}$<br>(0.25) |
| DIS | 94.35$_{\pm 0.54}$<br>(1.00) | 82.03$_{\pm 0.80}$<br>(1.00) | 99.82$_{\pm 0.08}$<br>(1.00) | 83.86$_{\pm 0.34}$<br>(0.94) | 87.79$_{\pm 0.53}$<br>(0.96) | 65.05$_{\pm 0.42}$<br>(0.90) | 86.58$_{\pm 0.77}$<br>(0.00) | 62.96$_{\pm 0.67}$<br>(0.13) | 66.91$_{\pm 0.64}$<br>(0.00) | 84.16$_{\pm 0.43}$<br>(0.07) |
| SIM | 94.35$_{\pm 0.54}$<br>(1.00) | 81.85$_{\pm 0.80}$<br>(0.64) | 99.77$_{\pm 0.12}$<br>(0.27) | 83.87$_{\pm 0.32}$<br>(0.98) | 87.78$_{\pm 0.53}$<br>(0.97) | 65.08$_{\pm 0.42}$<br>(1.00) | 86.11$_{\pm 0.77}$<br>(0.00) | 62.67$_{\pm 0.65}$<br>(0.02) | 67.02$_{\pm 0.42}$<br>(0.00) | 84.27$_{\pm 0.59}$<br>(0.11) |
| DCCA | 94.35$_{\pm 0.54}$<br>(1.00) | 81.74$_{\pm 0.84}$<br>(0.47) | 99.74$_{\pm 0.14}$<br>(0.15) | 83.86$_{\pm 0.34}$<br>(0.97) | 87.78$_{\pm 0.52}$<br>(0.97) | 65.08$_{\pm 0.42}$<br>(1.00) | 87.49$_{\pm 0.84}$<br>(0.00) | 62.76$_{\pm 0.67}$<br>(0.04) | 66.89$_{\pm 0.48}$<br>(0.00) | 84.00$_{\pm 0.49}$<br>0.04 |
| DAE | 94.35$_{\pm 0.54}$<br>(1.00) | 81.99$_{\pm 0.79}$<br>(0.93) | 99.80$_{\pm 0.11}$<br>(0.60) | 83.86$_{\pm 0.32}$<br>(0.95) | 87.79$_{\pm 0.52}$<br>(1.00) | 65.05$_{\pm 0.42}$<br>(0.87) | 85.87$_{\pm 0.56}$<br>(0.00) | 62.84$_{\pm 0.70}$<br>(0.07) | 66.59$_{\pm 0.57}$<br>(0.00) | 84.18$_{\pm 0.43}$<br>(0.08) |
| DSV | 93.64$_{\pm 0.59}$<br>(0.02) | 76.09$_{\pm 1.51}$<br>(0.00) | 98.11$_{\pm 0.24}$<br>(0.00) | 72.86$_{\pm 0.65}$<br>(0.00) | 82.66$_{\pm 1.08}$<br>(0.01) | 64.53$_{\pm 0.40}$<br>(0.00) | 84.81$_{\pm 0.54}$<br>(0.00) | 61.96$_{\pm 0.47}$<br>(0.00) | 66.33$_{\pm 0.78}$<br>(0.00) | 68.33$_{\pm 1.22}$<br>(0.00) |
| PPRD | 94.35$_{\pm 0.54}$<br>(1.00) | 81.13$_{\pm 1.00}$<br>(0.05) | 99.55$_{\pm 0.20}$<br>(0.00) | 83.86$_{\pm 0.33}$<br>(0.94) | 87.78$_{\pm 0.51}$<br>(0.96) | 65.03$_{\pm 0.42}$<br>(0.79) | 90.93$_{\pm 0.53}$<br>(1.00) | 63.51$_{\pm 0.79}$<br>(1.00) | 67.71$_{\pm 0.40}$<br>(0.00) | 83.39$_{\pm 0.40}$<br>(0.00) |
| SPRD | 94.35$_{\pm 0.54}$<br>(1.00) | 79.50$_{\pm 0.93}$<br>(0.00) | 99.61$_{\pm 0.19}$<br>(0.01) | 83.87$_{\pm 0.33}$<br>(1.00) | 87.78$_{\pm 0.52}$<br>(0.95) | 65.01$_{\pm 0.42}$<br>(0.75) | 90.82$_{\pm 0.63}$<br>(0.70) | 63.50$_{\pm 0.64}$<br>(0.98) | 68.62$_{\pm 0.55}$<br>(1.00) | 84.81$_{\pm 0.44}$<br>(1.00) |
| MODDIS | 93.80$_{\pm 0.49}$<br>(0.04) | 59.00$_{\pm 1.21}$<br>(0.00) | 78.77$_{\pm 1.73}$<br>(0.00) | 78.37$_{\pm 0.52}$<br>(0.00) | 86.71$_{\pm 0.40}$<br>(0.00) | 64.38$_{\pm 0.42}$<br>(0.00) | 86.40$_{\pm 1.21}$<br>(0.00) | 59.42$_{\pm 0.65}$<br>(0.00) | 63.45$_{\pm 0.53}$<br>(0.00) | 46.79$_{\pm 1.16}$<br>(0.00) |
| CAAE | 93.07$_{\pm 0.50}$<br>(0.00) | 76.00$_{\pm 1.34}$<br>(0.00) | 99.29$_{\pm 0.16}$<br>(0.00) | 74.95$_{\pm 0.45}$<br>(0.00) | 84.45$_{\pm 0.57}$<br>(0.00) | 64.52$_{\pm 0.59}$<br>(0.03) | 87.47$_{\pm 0.56}$<br>(0.00) | 62.24$_{\pm 0.63}$<br>(0.00) | 67.78$_{\pm 0.69}$<br>(0.01) | 73.46$_{\pm 0.86}$<br>(0.00) |

surprise, summation turns out to be the most effective way to conduct fusion in our evaluation.

4) *Correlation-based alignment undergoes more fluctuations than other ways of alignment*. It can be observed that DCCA-based alignment sometimes performs evidently worse than its two alignment-based counterparts, e.g., on *Cifar10/Cifar100*, *Cat_vs_Dog*, and *YtFace*. By contrast, distance-based alignment maintains the most stable performance in the evaluation.
5) *DAE proves to be a strong baseline, while the performance of DSV is typically unsatisfactory in most cases*. Although DAE is a simple extension from the single-view DAE, it is able to produce acceptable or even superior performance to other baselines that are more sophisticated. However, DSVDD-based baseline often achieves lower AUROC than other baselines, although it is the best performer on the recent *YtFace* dataset.
6) *The performance of customized MDOD methods is unstable in multiview deep AD*. Interestingly, we notice that customized MDOD methods (MODDIS and CAAE) work effectively in certain cases, e.g., several video-based datasets (see Table II). However, they may also suffer from significantly worse performance than designed baselines on some datasets (e.g., many multiview/multimodal datasets in Table III). Meanwhile, CAAE is generally better than MODDIS, which validates the use of AE in multiview deep AD.

In the Supplementary Material, we also show the performance of baselines under other metrics (AUPR and TNR@95%TPR), as well as the performance of four miscellaneous baselines. We believe those results lay a firm foundation for future research on multiview deep AD.

TABLE IV

AUROC (%) OF DIFFERENT BASELINES ON THE EXISTING MULTIVIEW/MULTIMODAL DATASETS WITH GIVEN TRAIN/TEST SET SPLIT

| Type | | YtFace | CMU-MOSEI | DirverAD |
|---|---|---|---|---|
| Fusion | SUM | 88.12 | 56.96 | 96.00 |
| | MAX | 87.86 | 56.84 | 95.31 |
| | NN | 87.98 | 56.99 | 91.97 |
| | TF | 86.41 | 57.03 | 95.96 |
| Alignment | DIS | 88.31 | 52.89 | 96.33 |
| | SIM | 88.42 | 56.99 | 96.46 |
| | DCCA | 87.64 | 56.94 | 96.47 |
| Tailored | DAE | 88.33 | 57.30 | 94.80 |
| | DSV | **90.04** | 48.14 | 83.53 |
| Self-supervision | PPRD | 88.25 | 56.95 | 91.09 |
| | SPRD | 87.67 | 56.87 | 93.99 |
| MDOD | MODDIS | 83.14 | 56.47 | **97.01** |
| | CAAE | 89.05 | **63.12** | 91.45 |

### C. Further Analysis

*1) Comparison With Single-View Performance:* To enable a better insight into devised multiview deep AD baselines, we conduct an experiment to compare best baselines' performance and the best single-view performance on each benchmark dataset in terms of AUROC. The best single-view performance is obtained by training a DAE with data from one single views and selecting the best performer among the obtained DAEs. In particular, it should be noted that the best single-view performance is actually hindsight; i.e., it is usually not practically accessible due to the absence of the negative class in multiview deep AD. Therefore, it is merely used as a reference to reflect how existing baselines exploit multiview information. The results are shown in Fig. 5, and we can come to an interesting conclusion: *Despite of*

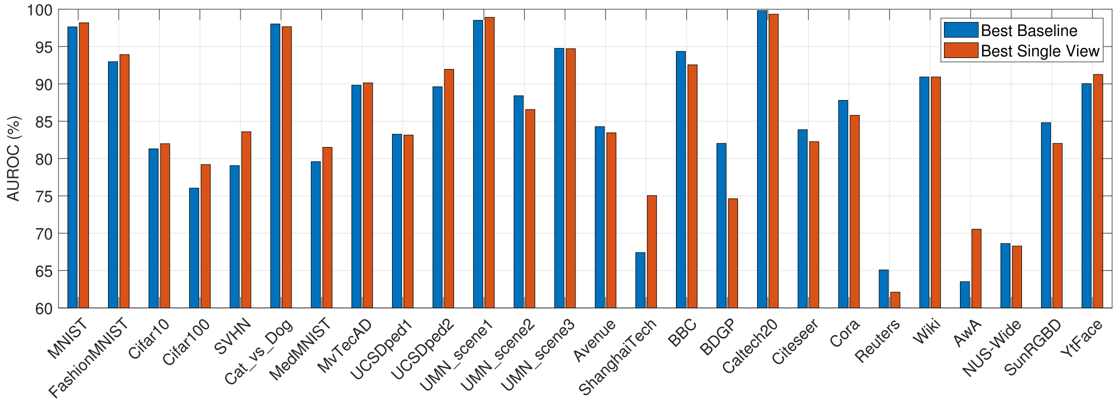


Fig. 5. AUROC (%) between the best baseline and best single view.

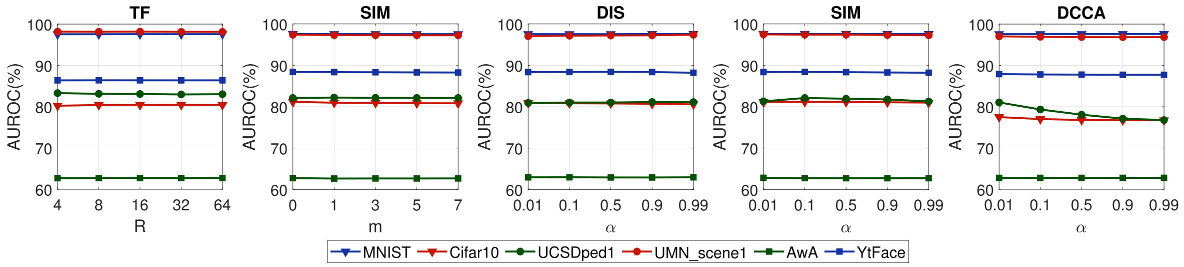


Fig. 6. Sensitivity analysis of typical hyperparameters in multiview deep AD.

TABLE V

AUROC (%) of Different Late Fusion Strategies on Selected Existing Multiview Datasets

| | BBC | BDGP | Caltech20 | Citeseer | Cora | Reuters | Wiki | AwA | NUS-Wide | SunRGBD |
|---|---|---|---|---|---|---|---|---|---|---|
| LF-AVG | **94.35**$\pm$**0.54** | 81.99$\pm$0.79 | **99.80**$\pm$**0.11** | **83.86**$\pm$0.32 | **87.79**$\pm$**0.52** | **65.05**$\pm$**0.42** | 85.87$\pm$0.56 | **62.84**$\pm$**0.70** | **66.59**$\pm$**0.57** | **84.18**$\pm$**0.43** |
| LF-MIN | 93.24$\pm$0.64 | **82.70**$\pm$**0.87** | 99.44$\pm$0.20 | 81.99$\pm$0.34 | 82.35$\pm$0.74 | 63.36$\pm$0.39 | 79.94$\pm$0.72 | 61.80$\pm$0.70 | 63.88$\pm$0.47 | 81.48$\pm$0.67 |
| LF-MAX | 94.11$\pm$0.41 | 51.64$\pm$1.24 | 91.91$\pm$0.99 | 53.11$\pm$0.76 | 55.88$\pm$0.51 | 59.74$\pm$0.22 | **90.49**$\pm$**0.84** | 54.93$\pm$0.59 | 64.44$\pm$0.50 | 82.16$\pm$0.36 |

*a systematic exploration, current baselines still suffer from insufficient capability to exploit multiview information for multiview deep AD.* Specifically, on 12 out of the total 26 datasets, the performance of best baseline is still inferior to the best single-view performance. However, an ideal multiview learning model is supposed to be superior or comparable to the best single-view performance. Such results imply two facts: First, it is discovered that the existing baselines are still unable to find a perfect way to exploit the contributing information embedded in each view. Second, redundant information in multiview data could be detrimental to the multiview deep AD performance. As a consequence, there is a large room for developing improved multiview deep AD solutions.

*2) Sensitivity Analysis:* In this section, we will discuss the impact of typical hyperparameters for the devised baselines: the rank number $R$ for tensor-based fusion, the margin $m$ for similarity-based fusion, and the weight of alignment loss $\alpha$ for alignment-based baselines (DIS, SIM, and DCCA). We choose the $R$, $m$, and $\alpha$ values from $\{4, 8, 16, 32, 64\}$, $\{0, 1, 3, 5, 7\}$, and $\{0.01, 0.1, 0.5, 0.9, 0.99\}$, respectively, and show the corresponding performance on representative datasets in Fig. 6. Surprisingly, we notice that the performance under different hyperparameter settings remains stable in the majority of cases. The performance fluctuations are usually within the range of 1%, except for the case of DCCA on UCSDped1. Consequently, we can speculate that a breakthrough of performance requires progress on model design, and tuning hyperparameter may not produce a performance leap.

*3) Influence of Late Fusion:* As a common component for almost all baselines, late fusion has a major influence on the performance multiview deep AD. Since we assume that no data from negative classes are available for validation, it is hard to apply many existing late fusion solutions here. As a preliminary effort, we take DAE for an example and explore three simple strategies for late fusion: averaging strategy (LF-AVG, used by default), max-value strategy (LF-MAX), and min-value strategy (LF-MIN), which compute the final score by the mean, maximum, and minimum of all views' scores. For simplicity, we test them on the existing multiview datasets and show the results in Table V. As shown in Table V, the averaging strategy almost constantly outperforms max-value and min-value strategies (except for *BDGP* and *Wiki*). The min-value strategy also achieves acceptable results in most cases, which is consistent with our intuition that any abnormal view should signify the abnormal data. However, it is noted that the max-value strategy can produce very poor fusion results, e.g., on *Citeseer* and *Cora* datasets. Therefore, the averaging strategy could still be an informative baseline late fusion strategy for multiview deep AD, which is somewhat similar to the case of multiview learning.

## VII. Discussion

Based on the results of previous experiments, we would like to make the following remarks on the multiview deep AD, which may inspire further research on this new topic.

1) *A non-trivial "killer" approach to multiview deep AD still requires exploration.* As we have shown in Section VI-B, there is not a single baseline that can consistently outperforms its counterparts. In the meantime, the performance gap between different baselines can be very small in many cases. Thus, it will be very attractive to explore the possibility to design a new multiview deep AD solution. In particular, we believe that self-supervised learning can be a promising direction to find such a solution, considering its comparatively better performance among baselines and the remarkable progress achieved by the self-supervised learning community.
2) *It will be interesting to assess the quality or contribution of each view to multiview deep AD.* Since prior knowledge on negative classes is not given, it will be natural to describe a sample by as many views as possible. However, as it is shown in Section VI-C1, it may degrade the performance when data of multiple views are blindly fused or aligned. Therefore, it is of high value to develop a strategy to perform knowledgeable multiview fusion or alignment. This is also applicable to the late fusion stage.
3) *The revolution of the learning paradigm may breed a breakthrough.* The generative learning paradigm (i.e., generation or prediction) has been a standard practice in deep AD, which is followed in this article when designing most baselines. However, other learning paradigms, such as the discriminative learning [16] and contrastive learning [68] paradigm, have been proven more effective than generative learning paradigm in realms, such as unsupervised representation learning. Naturally, a brand-new learning paradigm may be a good remedy to multiview deep AD.
4) *Newly emerging DNN models can be explored for enhancing multiview deep AD.* In this article, most baselines are developed based on the classic encoder–decoder, such as DNN models. This is due to the fact that DAE and its variants are the most commonly used tool for deep AD, and they can be good reference to understand multiview deep AD. However, the deep AD realm also witnesses the emergence of many emerging DNN models, such as GANs [69] and transformers [70]. Such new techniques pave the way for better multiview deep AD. For example, it will be interesting to leverage the self-attention mechanism of transformers to capture the inter-view correspondence within multiview data.
5) *Multiview deep AD is a relevant but different topic from other realms, such as MDOD.* The effectiveness of two customized MDOD solutions on several datasets (e.g., *ShanghaiTech* and *DriverAD*) suggests that multiview deep AD also benefits from the progress in other related realms, such as MDOD. However, the severe performance degradation of those MDOD solutions in some other cases (e.g., BDGP and SunRGBD) also shows that they are not universally applicable solutions to multiview deep AD. Thus, mutliview deep AD cannot be simply equalized to MDOD.

## VIII. Conclusion

This article investigates a pervasive but unexplored problem: multiview deep AD. Within the scope of our best knowledge, we are the first to formally identify and formulate multiview deep AD. To overcome the practical difficulties to look into this problem, we systematically design baseline solutions by extensively reviewing relevant areas in the literature, and we also construct abundant new multiview datasets by processing public data via various means. Together with some existing multiview datasets, a comprehensive evaluation of designed baselines is carried out to provide the first glimpse to this new topic. Hopefully, our baseline solutions and experimental results can facilitate later research on this topic.

## References

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, p. 15, Jul. 2009.

[2] L. M. Manevitz and M. Yousef, "One-class SVMs for document classification," *J. Mach. Learn. Res.*, vol. 2, pp. 139–154, Dec. 2001.

[3] H. J. Shin, D.-H. Eom, and S.-S. Kim, "One-class support vector machines—An application in machine fault detection and classification," *Comput. Ind. Eng.*, vol. 48, no. 2, pp. 395–408, 2005.

[4] M. Koppel and J. Schler, "Authorship verification as a one-class classification problem," in *Proc. 21st Int. Conf. Mach. Learn. (ICML)*, 2004, p. 62.

[5] B. Krawczyk, M. Woźniak, and F. Herrera, "On the usefulness of one-class classifier ensembles for decomposition of multi-class problems," *Pattern Recognit.*, vol. 48, no. 12, pp. 3969–3982, 2015.

[6] V. Chandola and V. Kumar, "Outlier detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–83, 2007.

[7] D. M. J. Tax, "One-class classification," Ph.D. thesis, Delft Univ. Technol., Delft, Netherlands, Jun. 2001.

[8] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[9] A. Sedlmeier, R. Müller, S. Illium, and C. Linnhoff-Popien, "Policy entropy for out-of-distribution classification," in *Proc. Int. Conf. Artif. Neural Netw.* Piscataway, NJ, USA: Institute of Electrical and Electronics Engineers, 2020, pp. 420–431.

[10] H. Zhao and Y. Fu, "Dual-regularized multi-view outlier detection," in *Proc. 24th Int. Joint Conf. Artif. Intell. (IJCAI)*. Buenos Aires, Argentina: AAAI Press, Jul. 2015, pp. 4083–4077.

[11] S. S. Khan and M. G. Madden, "One-class classification: Taxonomy of study and review of techniques," *Knowl. Eng. Rev.*, vol. 29, no. 3, pp. 345–374, Jan. 2014.

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[13] Y. Li, M. Yang, and Z. Zhang, "A survey of multi-view representation learning," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 10, pp. 1863–1883, Oct. 2018.

[14] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[15] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63373–63394, 2019.

[16] I. Golan and R. El-Yaniv, "Deep anomaly detection using geometric transformations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 9758–9769.

[17] S. Rayana. (2016). *ODDS Library*. [Online]. Available: http://odds.cs.stonybrook.edu

[18] G. Pang, C. Shen, L. Cao, and A. van den Hengel, "Deep learning for anomaly detection: A review," 2020, *arXiv:2007.02500*.

[19] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion," *Comput. Vis. Image Understand.*, vol. 156, pp. 117–127, Mar. 2016.

[20] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 733–742.

[21] J. Chen, S. Sathe, C. Aggarwal, and D. Turaga, "Outlier detection with autoencoder ensembles," in *Proc. SIAM Int. Conf. Data Mining*. Philadelphia, PA, USA: SIAM, 2017, pp. 90–98.

[22] S. Zhai, Y. Cheng, W. Lu, and Z. Zhang, "Deep structured energy based models for anomaly detection," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1100–1109.

[23] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *Proc. Int. Conf. Inf. Process. Med. Imag.*, 2017, pp. 146–157.

[24] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "GANomaly: Semi-supervised anomaly detection via adversarial training," in *Proc. Asian Conf. Comput. Vis.* Ney York, NY, USA: Springer, 2018, pp. 622–637.

[25] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection–A new baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6536–6545.

[26] M. Ye, X. Peng, W. Gan, W. Wu, and Y. Qiao, "AnoPCN: Video anomaly detection via deep predictive coding network," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1805–1813.

[27] L. Ruff *et al.*, "Deep one-class classification," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 4393–4402.

[28] L. Bergman and Y. Hoshen, "Classification-based anomaly detection for general data," in *Proc. 8th Int. Conf. Learn. Represent. (ICLR)*, Addis Ababa, Ethiopia, Apr. 2020.

[29] S. Goyal, A. Raghunathan, M. Jain, H. V. Simhadri, and P. Jain, "DROCC: Deep robust one-class classification," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 3711–3721.

[30] G. Pang, L. Cao, and C. Aggarwal, "Deep learning for anomaly detection: Challenges, methods, and opportunities," in *Proc. 14th ACM Int. Conf. Web Search Data Mining*, Mar. 2021, pp. 1127–1130.

[31] 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*, Dec. 2011, pp. 1050–1055.

[32] A. Marcos Alvarez, M. Yamada, A. Kimura, and T. Iwata, "Clustering-based anomaly detection in multi-view data," in *Proc. 22nd ACM Int. Conf. Conf. Inf. Knowl. Manage. (CIKM)*, 2013, pp. 1545–1548.

[33] T. Iwata and M. Yamada, "Multi-view anomaly detection via robust probabilistic latent variable models," in *Proc. NIPS*, 2016, pp. 1136–1144.

[34] 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, vol. 33, no. 1, pp. 4894–4901.

[35] S. Li, M. Shao, and Y. Fu, "Multi-view low-rank analysis with applications to outlier detection," *ACM Trans. Knowl. Discovery from Data*, vol. 12, no. 3, pp. 1–22, Jun. 2018.

[36] Y.-X. Ji *et al.*, "Multi-view outlier detection in deep intact space," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2019, pp. 1132–1137.

[37] S. Wang, Y. Liu, L. Chen, and C. Zhang, "Cross-aligned and gumbel-refactored autoencoders for multi-view anomaly detection," in *Proc. IEEE 33rd Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2021, pp. 1368–1375.

[38] Z. Wang and C. Lan, "Towards a hierarchical Bayesian model of multi-view anomaly detection," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 1–7.

[39] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*. Bellevue, WA, USA: Omnipress, Jun./Jul. 2011, pp. 689–696.

[40] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 2949–2980, Oct. 2014.

[41] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1933–1941.

[42] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-RNN)," 2014, *arXiv:1412.6632*.

[43] S. Sun, W. Dong, and Q. Liu, "Multi-view representation learning with deep Gaussian processes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 12, pp. 4453–4468, Dec. 2021.

[44] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Empirical Methods Natural Lang. Process. (EMNLP)*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 1103–1114.

[45] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Bagher Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2247–2256.

[46] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in Statistics*. Springer, New York, NY, USA, 1992, pp. 162–190.

[47] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

[48] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1083–1092.

[49] A. Benton, H. Khayrallah, B. Gujral, D. A. Reisinger, S. Zhang, and R. Arora, "Deep generalized canonical correlation analysis," in *Proc. 4th Workshop Represent. Learn. NLP (RepL4NLP)*, 2019, pp. 1–6.

[50] A. Frome *et al.*, "Devise: A deep Visual-Semantic embedding model," in *Proc. 27th Annu. Conf. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2013, pp. 2121–2129.

[51] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 7–16.

[52] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 154–162.

[53] J. Gu, J. Cai, S. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7181–7189.

[54] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037–4058, Nov. 2021.

[55] D. Liu, K.-T. Lai, G. Ye, M.-S. Chen, and S.-F. Chang, "Sample-specific late fusion for visual category recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 803–810.

[56] X. Liu *et al.*, "Late fusion incomplete multi-view clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2410–2423, Oct. 2019.

[57] J. Liu, X. Liu, Y. Yang, X. Guo, M. Kloft, and L. He, "Multiview subspace clustering via co-training robust data representation," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Apr. 9, 2021, doi: 10.1109/TNNLS.2021.3069424.

[58] J. Liu, X. Liu, Y. Yang, S. Wang, and S. Zhou, "Hierarchical multiple kernel clustering," in *Proc. 35th AAAI Conf. Artif. Intell. (AAAI)*, Feb. 2021.

[59] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[60] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015.

[61] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[63] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[64] B. Ramachandra, M. Jones, and R. R. Vatsavai, "A survey of single-scene video anomaly detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2293–2312, May 2022.

[65] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2462–2470.

[66] G. Yu *et al.*, "Cloze test helps: Effective video anomaly detection via learning to complete video events," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 583–591.

[67] O. Kopuklu, J. Zheng, H. Xu, and G. Rigoll, "Driver anomaly detection: A dataset and contrastive learning approach," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 91–100.

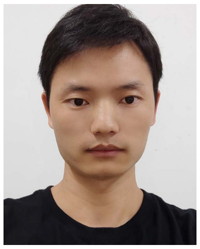
[68] X. Liu *et al.*, "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, 2021, doi: 0.1109/TKDE.2021.3090866.

[69] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, "A review on generative adversarial networks: Algorithms, theory, and applications," 2020, *arXiv:2001.06937*.

[70] K. Han *et al.*, "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022, doi: 10.1109/TPAMI.3152247.

**Siqi Wang** is currently an Assistant Research Professor with the College of Computer, National University of Defense Technology, Changsha, China. He has authored or coauthored leading conferences and journals, such as Annual Conference on Neural Information Processing Systems (NeurIPS), AAAI Conference On Artificial Intelligence (AAAI), International Joint Conference on Artificial Intelligence (IJCAI), ACM Multimedia (ACM MM), IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), and IEEE Transactions on Image Processing (TIP). His research interests include outlier/anomaly detection and unsupervised learning.

He serves as a Program Committee (PC) Member and Reviewer for top-tier conference, such as NeurIPS and AAAI and several prestigious journals.



**Jiyuan Liu** is currently pursuing the Ph.D. degree with the National University of Defense Technology (NUDT), Changsha, China.

He has authored or coauthored papers in journals and conferences, such as the IEEE Transactions on Knowledge and Data Engineering (TKDE), IEEE Transactions on Neural Networks and Learning Systems (TNNLS), International Conference on Machine Learning (ICML), IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR), IEEE International Conference on Computer Vision (ICCV), ACM Multimedia (ACM MM), AAAI Conference on Artificial Intelligence (AAAI), and International Joint Conference on Artificial Intelligence (IJCAI). His current research interests include multiview clustering, deep clustering, and anomaly detection.

**Guang Yu** received the bachelor's degree in computer science and technology from Sichuan University, Chengdu, China, in 2018. He is currently pursuing the Ph.D. degree with the College of Computer, National University of Defense Technology, Changsha, China.

His research interests include anomaly/outlier detection and self-supervised/unsupervised learning.

**Xinwang Liu** (Senior Member, IEEE) received the Ph.D. degree from the National University of Defense Technology (NUDT), Changsha, China, in 2013.

He is currently a Professor with the School of Computer, NUDT. He has authored or coauthored more than 60 peer-reviewed papers in journals and conferences, such as the IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), IEEE Transactions on Knowledge and Data Engineering (TKDE), IEEE Transactions on Image Processing (TIP), IEEE Transactions on Neural Networks and Learning Systems (TNNLS), IEEE Transactions on Multimedia (TMM), IEEE Transactions on Information Forensics and Security (TIFS), Annual Conference on Neural Information Processing Systems (NeurIPS), IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR), IEEE International Conference on Computer Vision (ICCV), AAAI Conference on Artificial Intelligence (AAAI), and International Joint Conference on Artificial Intelligence (IJCAI). His current research interests include kernel learning and unsupervised feature learning.

**Sihang Zhou** received the Ph.D. degree from the National University of Defense Technology (NUDT), Changsha, China, in 2019.

He is currently a Lecturer with the College of Intelligence Science and Technology, NUDT. He has authored or coauthored more than 20 peer-reviewed papers, including the IEEE Transactions on Image Processing (TIP), IEEE Transactions on Neural Networks and Learning Systems (TNNLS), IEEE Transactions on Medical Imaging (TMI), *Information Fusion*, *Medical Image Analysis*, AAAI Conference on Artificial Intelligence (AAAI), and International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). His current research interests include machine learning and medical image analysis.

**En Zhu** received the Ph.D. degree from the National University of Defense Technology (NUDT), Changsha, China, in 2005.

He is currently a Professor with the School of Computer Science, NUDT. He has authored or coauthored more than 60 peer-reviewed papers, including the IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), IEEE Transactions on Neural Networks and Learning Systems (TNNLS), *Pattern Recognition* (PR), AAAI Conference on Artificial Intelligence (AAAI), and International Joint Conference on Artificial Intelligence (IJCAI). His research interests include pattern recognition, image processing, machine vision, and machine learning.

Dr. Zhu was a recipient of the China National Excellence Doctoral Dissertation.

**Yuexiang Yang** received the B.S. degree in mathematics from Xiangtan University, Xiangtan, China, in 1986, and the M.S. degree in computer application and the Ph.D. degree in computer science and technology from the National University of Defense Technology, Changsha, China, in 1989 and 2008, respectively.

His research interests include information retrieval, network security, and data analysis.

**Jianping Yin** received the Ph.D. degree from the National University of Defense Technology (NUDT), Changsha, China, in 1990.

He is currently a Distinguished Professor with the Dongguan University of Technology, Dongguan, China. He has authored or coauthored more than 150 peer-reviewed papers, including the IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), IEEE Transactions on Neural Networks and Learning Systems (TNNLS), *Pattern Recognition* (PR), AAAI Conference on Artificial Intelligence (AAAI), and International Joint Conference on Artificial Intelligence (IJCAI). His research interests include pattern recognition and machine learning.

**Wenjing Yang** received the Ph.D. degree in multiscale modeling from Manchester University, Manchester, U.K., in 2014.

She is currently an Associate Research Fellow with the State Key Laboratory of High Performance Computing, Institute for Quantum Information, National University of Defense Technology, Changsha, China. Her research interests include deep learning, multiagent reinforcement learning, and high-performance computing.