Multi-view Deep Anomaly Detection: A Systematic Exploration

Siqi Wang, Jiyuan Liu, Guang Yu, Xinwang Liu, Sihang Zhou, En Zhu, Yuexiang Yang, Jianping Yin, 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 multi-view deep learning. However, it has not been formally discussed by the literature and remains under-explored. Motivated by this blank, this paper makes four-fold contributions: First, to our best knowledge, this is the first work that formally identifies and formulates the multi-view deep AD problem. Second, we take recent advances in relevant areas into account and systematically devise various baseline solutions, which lays the foundation for multi-view deep AD research. Third, to remedy the problem that limited benchmark datasets are available for multi-view deep AD, we extensively collect existing public data and process them into more than 30 multi-view benchmark datasets via multiple means, so as to provide a better evaluation platform for multi-view deep AD. Finally, by comprehensively evaluating the devised solutions on different types of multi-view 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 to the new multi-view deep AD topic.**

***Index Terms*—Deep anomaly detection, multi-view deep learning, multi-view deep anomaly detection**

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 datum 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, whilst abnormal events like 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 like information retrieval [2], fault detection [3], authorship verification [4], enhanced multi-class classification [5] and etc. In literature, AD is sometimes referred as outlier detection [6], one-class learning [7], novelty detection [8] or out-of-distribution detection [9] and etc. 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 *outlier detection* (OD) [10]. We distinguish OD from AD because their terms are often interchangeably used in literature and cause confusion.

Compared with fully supervised binary/multi-class 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 like 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 Sec. 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 paper, 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 birth to a large amount of multi-view 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 multi-view data. To handle the above two challenges posed by modern data, an emerging realm named *multi-view deep learning* has attracted surging attention from researchers. Specifically, multi-view 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, multi-view deep learning methods usually leverage multi-view fusion or

S. Wang, J. Liu, G. Yu, X. Liu, S. Zhou, E. Zhu, Y. Yang and W. Yang are with College of Computer, National University of Defense Technology (NUDT), Changsha, 410073, China. E-mail: {wangsiqi10c, liujiyuan13, xinwangliu, enzhu, yyx, wenjing.yang}@nudt.edu.cn, {yuguangnudt, sihangjoe}@gmail.com. J. Yin is with Dongguan University of Technology, Dongguan, 523808, China. E-mail: jpyin@dgut.edu.cn. Corresponding author: Xinwang Liu. S. Wang and J. Liu contributed equally to this work.

Manuscript received June 4, 2021.

multi-view alignment techniques [13] to exploit inter-view information embedded in multi-view data. Multi-view deep learning has already been successfully applied to many tasks [13], [14], [15]. Therefore, it is quite natural for us to consider the intersection of AD and multi-view deep learning, i.e. *multi-view deep AD*. Multi-view 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 multi-view deep learning methods have been thoroughly studied in literature, the problem of multi-view deep AD has not been formally defined and systematically explored to our best knowledge. Such a blank constitutes to the biggest motivation of this paper. There are three major obstacles when looking into multi-view deep AD: 1) *Above all, the lack of formal formulation of the problem.* Despite its huge application potential in various real-world scenarios, multi-view deep AD has not been formally identified and formulated, which prevents researchers from giving sufficient attention to this novel but challenging problem. 2) *Second, the lack of baseline methods*. Although many attempts have been made to approach multi-view deep learning, they are typically designed for other tasks and therefore not explored for AD. In the meantime, existing AD approaches are merely applicable to the single-view case. 3) *Third, the lack of proper benchmark datasets for evaluation*. Previous researches usually evaluate AD models by the "one *vs.* all" protocol [16]. For any binary/multi-class 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, existing multi-view 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 multi-view deep AD is still insufficient.

To bridge the above gaps, this paper for the first time formally identifies and formulates the problem of multi-view deep AD, and carries out a systematic study on this new area. Our contributions can be summarized as follows:

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

II. Related Work

In this section, we will focus on reviewing deep AD, multi-view OD and multi-view deep learning, which are the most relevant areas to multi-view deep AD. In Sec. 1 and Sec. 2 of supplementary material, we also briefly review classic methods for AD and multi-view learning due to the page limit.

A. Deep Anomaly Detection

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 auto-encoder (AE) into a deep one. For example, stacked denoising auto-encoder (SDAE) [19] and deep convolutional auto-encoder (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 energy based model [22]. In addition to AE based methods, other popular generative neural networks like 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 SVDD into deep SVDD (DSVDD), which learns to map latent representations of positive data into a hypersphere with minimal radius. Golan et al. [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 et al. [28] further extends 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 are typically limited to the single-view setting.

B. Multi-view Oultier (Anomaly) Detection

Multi-view OD is a relevant but essentially different area from multi-view deep AD in this paper. As a comparison, multi-view 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 multi-view OD is often termed as "multi-view anomaly detection" in some prior works like [32], [33], [34], but their setup is evidently different from AD or multi-view deep AD in this context (see Sec. I). The pioneer work of multi-view OD is proposed by Gao et al. [31], while a series of improved solutions are developed [32], [10], [33], [35], [34]. Most multi-view OD methods spot outliers by the cluster structure of given unlabeled data, which are obtained by classic techniques like spectral clustering [32] or outlierness estimation [34], while only very recent works [36], [37] begin to explore DNNs to perform multi-view deep OD. We notice that the latest work [37] for the first time leverages autoencoder 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: Autoencoders for multi-view OD [37] are fed with contaminated unlabelded data, while pure data from a single class are used to train autoencoders in this work. Second, our work places more emphasize on designing a generic framework rather than a specific solution like [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] leverage a hierarchical Bayesian model to address "semi-supervised multi-view anomaly detection", which is a multi-view 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 multi-view deep AD in this paper.

C. Multi-view Deep Learning

As classic multi-view learning does not involve representation learning and lacks the ability to handle with complex data, multi-view deep learning has rapidly become an emerging topic. Current multi-view deep learning methods are usually categorized into two groups, i.e. *multi-view fusion* and *multi-view alignment* based methods. Multi-view fusion based methods fuse the learned representations from different views into a joint representation, which can be realized by either simple operations like max/sum/concatenation [13], or sophisticated means like a neural network. Specifically, the pioneer work of Ngiam et al. [39] proposes a multi-modal deep auto-encoder for multi-view deep fusion, while Srivastava et al. [40] perform the fusion by multi-modal deep boltzman machine (DBM). Such neural network based multi-view fusion can also be conducted on modern neural network architecture like convolutional neural networks (CNNs) [41] and recurrent neural networks (RNNs) [42]. Latest work from Sun et al. [43] employs a multi-view 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 multi-view case by low-rank decomposition. Unlike multi-view fusion, multi-view 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 multi-view 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 auto-encoders (DCCAE), 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 multi-view 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 multi-view representation learning by learning modality-invariant representations [52] or cross-view transformation [53]. Consequently, many solutions have been proposed for multi-view deep learning, and they are widely adopted to serve many tasks like action recognition, sentiment analysis and image captioning. However, none of those works has considered the marriage of AD and multi-view deep learning, which motivates this paper.

III. Problem Formulation

To tackle the first obstacle mentioned in Sec. I, we will provide a formal problem formulation of multi-view deep AD in the first place. Given the normal class $\mathcal{C}_n$, a multi-view datum $\{\mathbf{x}_{train}^{(v)}\}_{v=1}^{V}$ is sampled from $\mathcal{C}_n$ for training, where $V \geq 2$ is the number of views and $\mathbf{x}_{train}^{(v)} \in \mathbb{R}^{d_1^{(v)} \times d_2^{(v)} \times \cdots d_{M_v}^{(v)}}$ is a M_v-dimensional tensor with the shape $d_1^{(v)} \times d_2^{(v)} \times \cdots d_{M_v}^{(v)}$. To be more specific, $\mathbf{x}_{train}^{(v)}$ denotes the observation from the v_{th} view, while $M_v = 1$ and $M_v > 1$ corresponds 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 datum $\{\mathbf{x}_{train}^{(v)}\}_{v=1}^{V}$, the goal of multi-view deep AD is to obtain a DNN model:

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

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

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

In practice, $\mathcal{M}_{\mathbf{\Theta}}$ is usually supposed to obtain a score $\mathcal{S}(\{\mathbf{x}_{test}^{(v)}\}_{v=1}^{V})$, which indicates the likelihood that $\{\mathbf{x}_{test}^{(v)}\}_{v=1}^{V}$ 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

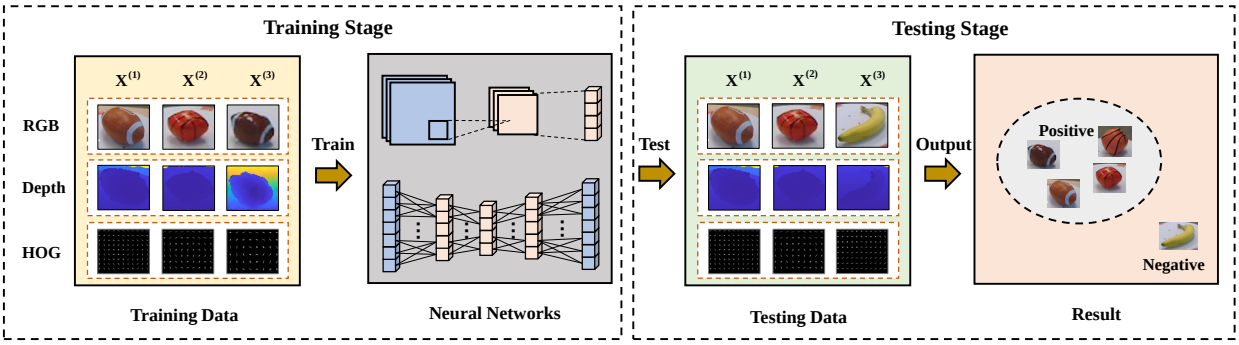


Fig. 1: An overview of multi-view deep AD. The example shows soccer balls described by data from three views (RGB, depth and HOG), and the goal is to determine whether incoming multi-view data are from the soccer ball class or not.

learning and classification), we will focus on discussing the the model that consists of pure DNNs. In other words, we discuss the case where $\mathcal{M}_{\Theta}$ is able to perform end-to-end AD. An overview of multi-view deep AD task is presented in Fig. 1.

IV. The Proposed Baselines

Having provided a formal problem formulation, we will address the second issue in Sec. I by designing baseline solutions to multi-view 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 anomaly detection methods and self-supervision based solutions.

A. Fusion based Solutions

A core issue for multi-view 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 multi-view 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 below.

1) Framework: Given a multi-view datum $\{\mathbf{x}_{train}^{(v)}\}_{v=1}^{V}$ 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 $Enc^{(v)} : \mathbb{R}^{d_1^{(v)} \times d_2^{(v)} \times \cdots d_{M_v}^{(v)}} \mapsto \mathbb{R}^{d_l^{(v)}}$ encodes $\mathbf{x}_{train}^{(v)}$ into a latent embedding $\mathbf{h}^{(v)}$ with a dimension $d_l^{(v)}$:

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

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 multi-view datum:

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

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 multi-view 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, an decoder $Dec^{(v)} : \mathbb{R}^D \mapsto \mathbb{R}^{d_1^{(v)} \times d_2^{(v)} \times \cdots d_{M_v}^{(v)}}$ intends to map $\mathbf{h}$ back to v_{th} view's original input $\mathbf{x}_{train}^{(v)}$:

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

where $\hat{\mathbf{x}}_{train}^{(v)}$ 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:

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

In addition to the mean square errors (MSE) above, other types of reconstruction loss $\ell(\cdot)$ are also applicable, such as L_1-norm reconstruction loss. During testing, an incoming multi-view datum $\{\mathbf{x}_{test}^{(v)}\}_{v=1}^{V}$ is fed into the network to obtain the reconstructed datum $\{\hat{\mathbf{x}}_{test}^{(v)}\}_{v=1}^{V}$ by Eq. 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 datum belongs to $\mathcal{C}_n$. In other words, a baseline score for the v_{th} view can be directly obtained by $\mathcal{S}^{(v)}(\mathbf{x}_{test}^{(v)}) = -\ell(\mathbf{x}_{train}^{(v)}, \hat{\mathbf{x}}_{train}^{(v)})$. Finally, we can obtain a score function by the reconstruction errors of all views:

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

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 multi-view 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* (abbreviated as SUM, and the abbreviation of other methods are similarly given). Summation based fusion combines latent embeddings from different views

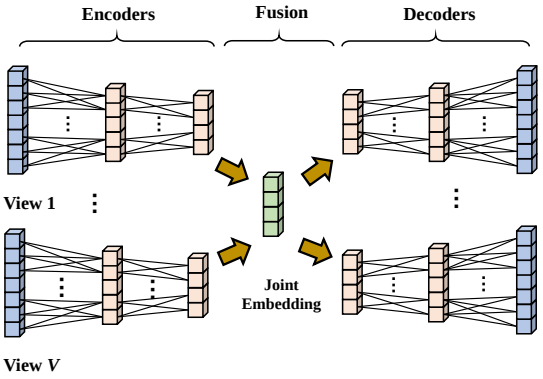


Fig. 2: Fusion based solutions for multi-view deep AD.

by 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:

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

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 datum 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 paper.

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 $\mathbf{h}$:

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

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:

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

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 multi-layer fully-connected network to perform DNN based fusion.

4) *Tensor based fusion* (TF). Tensor based fusion [44] is an emerging method in multi-view 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. Afterwards, $\mathcal{Z}$ 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 $\mathbf{h}$:

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

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 $k = 1, \cdots, D$. 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 multi-view fusion, multi-view alignment is another popular category of methods in multi-view 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 multi-view deep AD solutions below.

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

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

where $\mathcal{F}_a$ is the set of available alignment functions. As shown in Eq. 14, a key difference between alignment based solutions and fusion based solutions is that fusion usually occurs within one multi-view datum, while the alignment of two views can involve multiple multi-view 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 $\{Dec^{(v)}\}_{v=1}^V$ 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 Eq. 6 and α is the weight of alignment loss. Given a testing datum

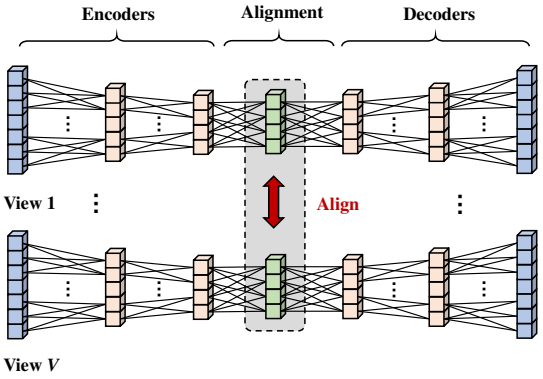


Fig. 3: Alignment based solutions for multi-view deep AD.

$\{\mathbf{x}_{test}^{(v)}\}_{v=1}^{V}$, we also leverage the reconstruction errors as baseline scores, which is the same as Eq. 7.

2) Implementations: The core issue of alignment based solutions is the design of alignment function F_a. Inspired by the literature of multi-view 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:

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

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

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 like Eq. 16. Nevertheless, such an alignment function only considers the view similarity within one multi-view datum. To consider the view similarity across different datum, we are inspired by [50] and propose to adopt a more sophisticated similarity measure between different views: For i_{th} and j_{th} view, a similarity loss $Sim(i,j)$ is computed based on $s(\cdot)$ and a hinge loss:

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

where m is a margin. The above similarity loss encourages the embeddings from the same multi-view datum to be similar, while embeddings from two different multi-view 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:

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

3) *Correlation based alignment* (DCCA). Canonical correlation analysis (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, deep CCA [47]. To conduct correlation based alignment, we intend to maximize the correlation between two views. Specifically, we stack v_{th} view's embeddings of N multi-view data in a training batch into a $d_l^{(v)} \times N$ embedding matrix: $\mathbf{H}^{(v)} = [\mathbf{h}_1^{(v)}, \cdots, \mathbf{h}_N^{(v)}]$, while $\mathbf{H}^{(v)}$ can be centered by $\bar{\mathbf{H}}^{(v)} = \mathbf{H}^{(v)} - \frac{1}{N}\mathbf{H}^{(v)} \cdot \mathbf{1}$, where $\mathbf{1}$ is a $N \times N$ 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 $\sum_{ii} = \frac{1}{N-1}\bar{\mathbf{H}}^{(i)} \cdot \bar{\mathbf{H}}^{(i)\top} + r\mathbf{I}$, $\sum_{ij} = \frac{1}{N-1}\bar{\mathbf{H}}^{(i)} \cdot \bar{\mathbf{H}}^{(j)\top}$ and $\sum_{jj} = \frac{1}{N-1}\bar{\mathbf{H}}^{(j)} \cdot \bar{\mathbf{H}}^{(j)\top} + r\mathbf{I}$, 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]:

$$Corr(i,j) = ||T_{ij}||_{tr} = tr(T_{ij}^{\top} \cdot T_{ij})^{\frac{1}{2}} \quad (19)$$

The final alignment function can be calculated by:

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

C. Deep Anomaly Detection Tailored Solutions

Apart from baselines based on multi-view deep learning, we design the third type of baseline solutions by tailoring 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 below.

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

$$\mathcal{S}^{(v)} = \mathcal{M}^{(v)}(\mathbf{x}_{test}^{(v)}), \quad v = 1, \cdots, V \quad (21)$$

The final score for the multi-view datum $\{\mathbf{x}_{test}^{(v)}\}_{v=1}^{V}$ is computed by a late fusion function $F_l(\cdot)$:

$$\mathcal{S}(\{\mathbf{x}_{test}^{(v)}\}_{v=1}^{V}) = F_l(\mathcal{S}^{(1)}(\mathbf{x}_{test}^{(1)}), \mathcal{S}^{(2)}(\mathbf{x}_{test}^{(2)}) \cdots, \mathcal{S}^{(V)}(\mathbf{x}_{test}^{(V)})) \quad (22)$$

2) Implementations: The choice of deep AD model plays a center role in designing tailored deep AD solutions. This paper introduces two representative deep AD methods in the literature to construct baseline models for multi-view deep AD: Standard deep autoencoders (DAE) and the recent deep support vector data description (DSVDD) [27]:

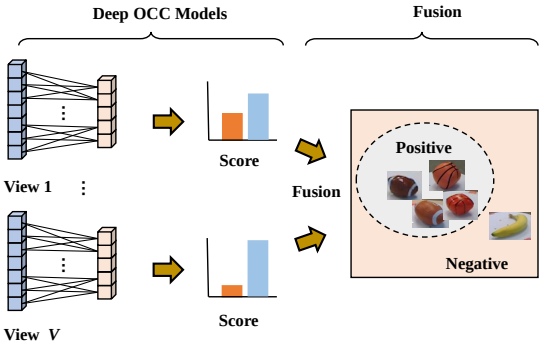


Fig. 4: Framework of deep AD tailored solutions.

1) *DAE based solution* (DAE). DAE leverages DNNs as the encoder $Enc^{(v)}$ and the decoder $Dec^{(v)}$ to reconstruct input data from a low-dimensional embedding. Formally, given N training data $\{\mathbf{x}_n^{(v)}\}_{n=1}^N$ from the v_{th} view, DAE requires to minimize the reconstruction loss:

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

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

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

DAE based baseline is also viewed as the most fundamental baseline for multi-view 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 hyper-sphere 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 multi-view 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:

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

where $\boldsymbol{\theta}^{(v)}$ are the learnable parameters of DSVDD, and λ is the weight of the L_2-norm regularization term. The encoder $Enc^{(v)}$ 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 above optimization problem can be efficiently solved by gradient descent. During inference, one can score a test datum $\mathbf{x}_{test}^{(v)}$ by calculating the distance between its embedding $\mathbf{h}_{test}^{(v)}$ and the center:

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

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 multi-view deep AD, creating self-supervision can be an appealing solution. Multi-view data intrinsically contains 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 multi-view 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 a multi-view datum $\{\mathbf{x}^{(v)}\}_{v=1}^V$, we partition the view indices into two subsets $\mathcal{P}$ and $\mathcal{Q}$, which satisfy:

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

Note that the intersection of $\mathcal{P}$ and $\mathcal{Q}$ may not be empty. By $\mathcal{P}$ and $\mathcal{Q}$, we can partition the multi-view 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 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 Sec. 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:

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

where $\boldsymbol{\theta}_D^{(j)}$ is the set of learnable weights for decoder $Dec^{(j)}$. 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 $\ell(Dec^{(j)}(\mathbf{h}^{\mathcal{P}}),\mathbf{x}^{(j)})$ can be used for scoring during inference.

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

1) *Plain prediction* (PPRD), where $\mathcal{Q}=\{v\}$ and $\mathcal{P}=\{1,2,\cdots,V\}-\mathcal{Q}$. 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), where $\mathcal{P}=\{v\}$ and $\mathcal{Q}=\{1,2,\cdots,V\}$. 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 above, 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 like 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 multi-view learning [57], [58], we also adopt the simple averaging strategy for late fusion in our baseline solutions above, namely:

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

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 multi-view deep AD. Eleven baseline solutions presented above 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 multi-view/multi-modal deep anomaly detection [37], which is essentially multi-view/multi-modal deep outlier detection (MDOD) by our definition in Sec. II-B. We customize them to learn from pure normal training data and design two additional solutions: Multi-view outlier detection based on deep intact space (MODDIS) [36] and cross-aligned autoencoder (CAAE) [37]. Their performances are also reported in experiments as a reference.

V. Benchmark Datasets

A. Limitations of Existing Datasets

Public benchmark datasets play an pivotal role in prompting the development of machine learning algorithms. However, as we briefed in Sec. I (the third issue), existing datasets basically suffer from some important limitations when they are used for evaluating multi-view deep AD: 1) *First, existing multi-view datasets are not adequate for multi-view deep AD.* To be more specific, frequently-used multi-view benchmark datasets (e.g. *Flower17/Flower102*[1]) are originally designed to evaluate traditional multi-view 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 multi-view datasets can be directly adopted for multi-view deep AD. 2) *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, multi-view deep learning is still a relatively new area with much less applicable benchmark datasets. 3) *Most importantly, very few benchmark dataset is speficially designed for the background of multi-view deep AD*. As we introduced in Sec. I, multi-view deep AD actually enjoys broad applications in many realms, such as vision based anomaly detection and fault detection, but benchmark datasets in such background are quite rare. In literature, many works adopt the "one *v.s.* all" protocol to convert a binary/multi-class dataset into an AD dataset, which is non-comprehensive for the evaluation of multi-view deep AD.

To this end, we need to build more benchmark datasets for multi-view deep AD. However, collecting multi-view 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 multi-view datasets, so as to construct abundant benchmark datasets in a highly efficient manner. Collected data and processing techniques will be elaborated below.

B. More Multi-view Benchmark Datasets

We intend to build our new multi-view 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 multi-view datasets* and *video based multi-view datasets*.

1) Image based Multi-view Datasets: Image data are the most fundamental data type in deep learning. We process image data into multi-view 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 multi-view data by describing it with different image descriptors. In this paper, 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 multi-view data,

[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/

we extract six types of features, i.e. color histogram, GIST, HOG2x2, HOG3x3, LBP and SIFT, which are implemented by a feature extraction Toolbox[7]. Second, *multiple pre-trained DNN models*: Classic image descriptors often find it hard to describe high-resolution images effectively. To convert high-resolution images to multi-view 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 like 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 fifteen 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 multi-view Deep AD.

2) Video based Multi-view Datasets: Compared with image data, video data contain both spatial and temporal information, so it is even more natural to transform them into multi-view representation. Since anomaly detection is a representative application of AD, we simply collect video data from benchmark datasets that are designed for the video anomaly detection (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. Afterwards, 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 labelled as 0, other patches are labelled 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 labelled as 1. In this way, we can yield video based multi-view datasets, which are readily applicable to evaluate multi-view deep AD with real-world application background.

C. Existing Multi-view/Multi-modal Datasets

Besides, we collect some existing multi-view/multi-modal datasets for more comprehensive evaluation. We collect 13 multi-view/multi-modal 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 multi-view/multi-modal datasets that are not specifically designed for AD, we adopt the "one *v.s.* all" protocol to evaluate multi-view 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 multi-view 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 multi-view/multi-modal benchmark datasets used in this paper is given in supplementary material.

VI. Empirical Evaluations

Having established formulation, baselines and benchmark datasets for multi-view 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 in-depth analysis on the characteristics of each model.

A. Experimental Setup

For multi-view datasets that are specifically designed for AD (*MvTecAD*, video based multi-view 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 multi-class multi-view datasets, we apply the "one *v.s.* all" protocol (detailed in Sec. V-C) for training and evaluating the AD performance. For multi-class datasets that possess more than 10 classes, we select the first 10 qualified classes (≥ 300 training data) for experiments. For those multi-view 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

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

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 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 multi-view deep AD solutions on both our new multi-view datasets and selected existing multi-view 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 multi-view datasets, video based multi-view datasets and selected existing multi-view datasets are given in Table 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 is reported in supplementary material). From those results, we can draw the following observations:

1) *In some cases, most of 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 multi-view datasets that are widely-used in the literature, e.g. *BBC*, *Caltech20*, *Cora* and *Reuters*. On many image based multi-view 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 multi-view datasets, like 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) attains 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

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

TABLE IV: AUROC (%) of different baselines on existing multi-view/multi-modal 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

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 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 deep autoencoder, 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 multi-view 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 multi-view/multi-modal datasets in Table III). Meanwhile, CAAE is generally better than MODDIS, which validates the use of AE in multi-view deep AD.

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

C. Further Analysis

1) Comparison with Single-view Performance: To enable a better insight to devised multi-view 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 deep autoencoder with data from one single views and selecting the best performer among the obtained deep autoencoders. 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 multi-view deep AD. Therfore, it is merely used as a reference to reflect how existing baselines exploit multi-view information. The results are shown in Fig. 5, and we can come to an interesting conclusion: *Despite of*

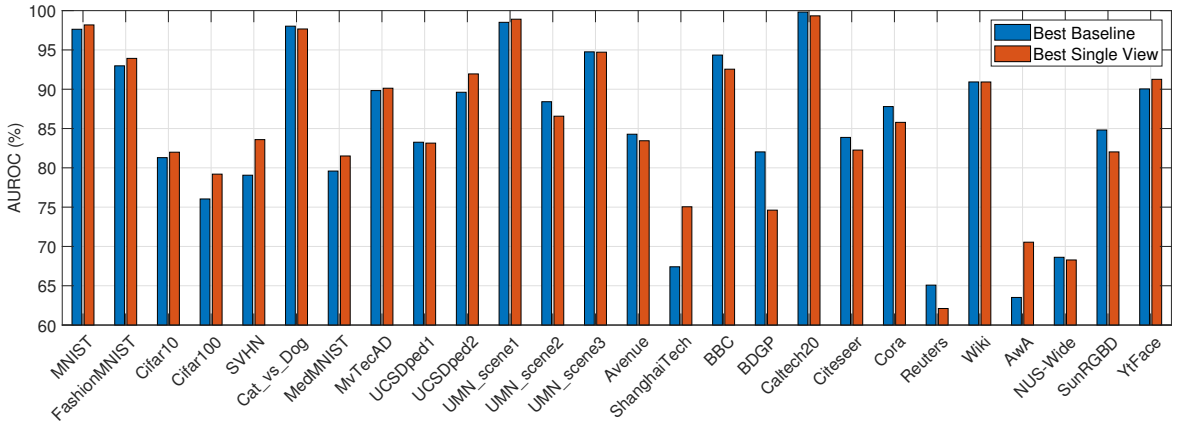


Fig. 5: AUROC (%) between the best baseline and best single view.

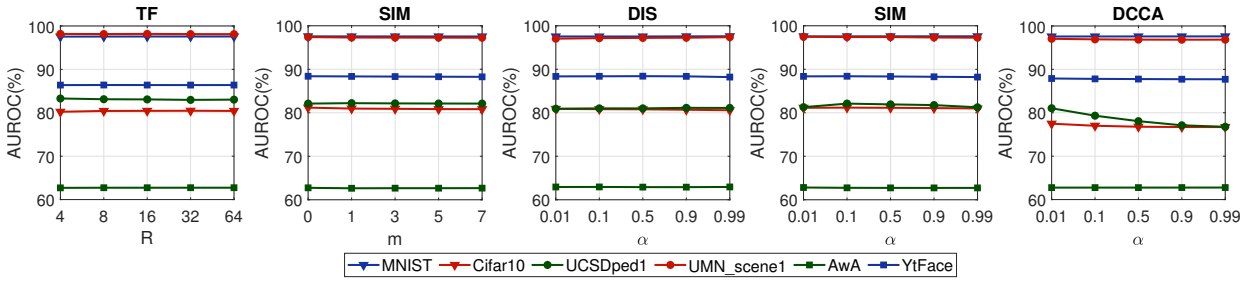


Fig. 6: Sensitivity analysis of typical hyperparameters in multi-view deep AD.

TABLE V: AUROC (%) of different late fusion strategies on selected existing multi-view 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 multi-view information for multi-view 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 multi-view 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 existing baselines are still unable to find a perfect way to exploit the contributing information embedded in each view. Second, redundant information in multi-view data could be detrimental to the multi-view deep AD performance. As a consequence, there is a large room for developing improved multi-view 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 α for alignment based baselines (DIS, SIM and DCCA). We choose the R, m and α value 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 multi-view deep AD. Since we assume that no datum 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 existing multi-view datasets and show the results in Table V. As shown by Table V, the averaging strategy almost constantly outperforms max-value and min-value strategy (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 an abnormal datum. However, it is noted that the max-value strategy can produce very poor fusion results, e.g. on *Citeseer* and *Cora* dataset. Therefore, the averaging strategy could still be an informative baseline late fusion strategy for

multi-view deep AD, which is somewhat similar to the case of multi-view learning.

VII. Discussion

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

1) A non-trivial "killer" approach to multi-view deep AD still requires exploration. As we have shown in Sec. 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 multi-view 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 multi-view 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 Sec. 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 multi-view 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 paper 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 like unsupervised representation learning. Naturally, a brand-new learning paradigm may be a good remedy to multi-view deep AD.

4) Newly-emerging DNN models can be explored for enhancing multi-view deep AD. In this paper, most baselines are developed based on the classic encoder-decoder like DNN models. This is due to the fact that deep autoencoder and its variants are the most commonly-used tool for deep AD, and they can be good reference to understand multi-view 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 multi-view deep AD. For example, it will be interesting to leverage the self-attention mechanism of transformers to capture the inter-view correspondence within multi-view data.

5) Multi-view deep AD is a relevant but different topic from other realms like multi-view deep OD (MDOD). The effectiveness of two customized MDOD solutions on several datasets (e.g. *ShanghaiTech*, *DriverAD*) suggests that multi-view deep AD also benefits from the progress in other related realms like MDOD. However, the severe performance degradation of those MDOD solutions in some other cases (e.g. BDGP, SunRGBD) also shows that they are not universally applicable solutions to multi-view deep AD. Thus, mutl-view deep AD cannot be simply equalized to multi-view deep OD.

VIII. Conclusion

This paper investigates a pervasive but unexplored problem: Multi-view deep AD. Within the scope of our best knowledge, we are the first to formally identify and formulate multi-view deep AD. In order 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 multi-view datasets by processing public data via various means. Together with some existing multi-view 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.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62006236, 61922088 and 61906020), NUDT Research Project (ZK20-10), HPCL Autonomous Project (202101-15), National Key R&D Program of China (2020AAA0107100) and Education Ministry-China Mobile Research Funding (MCM20170404).

References

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[2] L. M. Manevitz and M. Yousef, "One-class svms for document classification," *Journal of machine Learning research*, vol. 2, no. Dec, pp. 139–154, 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," *Computers & Industrial Engineering*, vol. 48, no. 2, pp. 395–408, 2005.

[4] M. Koppel and J. Schler, "Authorship verification as a one-class classification problem," in *Proceedings of the twenty-first international conference on Machine learning*, 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 Recognition*, vol. 48, no. 12, pp. 3969–3982, 2015.

[6] V. Chandola and V. Kumar, "Outlier detection : A survey," *Acm Computing Surveys*, vol. 41, no. 3, 2007.

[7] D. M. J. Tax, "One-class classification," *Ph.D. Thesis*, June 2001.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[9] A. Sedlmeier, R. Müller, S. Illium, and C. Linnhoff-Popien, "Policy entropy for out-of-distribution classification," in *International Conference on Artificial Neural Networks*. Springer, 2020, pp. 420–431.

[10] H. Zhao and Y. Fu, "Dual-regularized multi-view outlier detection," in *International Joint Conference on Artificial Intelligence*, 2015.

[11] S. S. Khan and M. G. Madden, "One-class classification: taxonomy of study and review of techniques," *The Knowledge Engineering Review*, vol. 29, no. 3, pp. 345–374, 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 transactions on knowledge and data engineering*, vol. 31, no. 10, pp. 1863–1883, 2018.

[14] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[15] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63 373–63 394, 2019.

[16] I. Golan and R. El-Yaniv, "Deep anomaly detection using geometric transformations," in *Advances in Neural Information Processing Systems*, 2018, pp. 9758–9769.

[17] S. Rayana, "ODDS library," 2016. [Online]. Available: http://odds.cs.stonybrook.edu

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

[19] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion," *Computer Vision and Image Understanding*, vol. 156, pp. 117–127, 2017.

[20] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 733–742.

[21] J. Chen, S. Sathe, C. Aggarwal, and D. Turaga, "Outlier detection with autoencoder ensembles," in *Proceedings of the 2017 SIAM international conference on data mining*. SIAM, 2017, pp. 90–98.

[22] S. Zhai, Y. Cheng, W. Lu, and Z. Zhang, "Deep structured energy based models for anomaly detection," *international conference on machine learning*, pp. 1100–1109, 2016.

[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 *International Conference on Information Processing in Medical Imaging*, 2017, pp. 146–157.

[24] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," in *Asian conference on computer vision*. 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 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 *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1805–1813.

[27] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in *International conference on machine learning*, 2018, pp. 4393–4402.

[28] L. Bergman and Y. Hoshen, "Classification-based anomaly detection for general data," in *International Conference on Learning Representations*, 2019.

[29] S. Goyal, A. Raghunathan, M. Jain, H. V. Simhadri, and P. Jain, "Drocc: Deep robust one-class classification," in *International Conference on Machine Learning*. PMLR, 2020, pp. 3711–3721.

[30] G. Pang, L. Cao, and C. Aggarwal, "Deep learning for anomaly detection: Challenges, methods, and opportunities," in *The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event*. ACM, 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 *2011 IEEE 11th International Conference on Data Mining*. IEEE, 2011, pp. 1050–1055.

[32] A. Marcos Alvarez, M. Yamada, A. Kimura, and T. Iwata, "Clustering-based anomaly detection in multi-view data," in *ACM conference on Information & Knowledge Management*, 2013, pp. 1545–1548.

[33] T. Iwata and M. Yamada, "Multi-view anomaly detection via robust probabilistic latent variable models." in *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 *AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 4894–4901.

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

[36] Y.-X. Ji, L. Huang, H.-P. He, C.-D. Wang, G. Xie, W. Shi, and K.-Y. Lin, "Multi-view outlier detection in deep intact space," in *IEEE International Conference on Data Mining*, 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 *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2021, pp. 1368–1375.

[38] Z. Wang and C. Lan, "Towards a hierarchical bayesian model of multi-view anomaly detection," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 2020.

[39] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *ICML*, 2011.

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

[41] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *IEEE conference on computer vision and pattern recognition*, 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)," *arXiv preprint arXiv:1412.6632*, 2014.

[43] S. Sun, W. Dong, and Q. Liu, "Multi-view representation learning with deep gaussian processes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[44] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *EMNLP*, 2017.

[45] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *56th Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 2247–2256.

[46] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in statistics*. Springer, 1992, pp. 162–190.

[47] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International conference on machine learning*. PMLR, 2013, pp. 1247–1255.

[48] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *International conference on machine learning*. PMLR, 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 *4th Workshop on Representation Learning for NLP*, 2019, pp. 1–6.

[50] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," 2013.

[51] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 7–16.

[52] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 154–162.

[53] J. Gu, J. Cai, S. R. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7181–7189.

[54] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[55] D. Liu, K.-T. Lai, G. Ye, M.-S. Chen, and S.-F. Chang, "Sample-specific late fusion for visual category recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 803–810.

[56] X. Liu, X. Zhu, M. Li, L. Wang, C. Tang, J. Yin, D. Shen, H. Wang, and W. Gao, "Late fusion incomplete multi-view clustering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 10, pp. 2410–2423, 2018.

[57] J. Liu, X. Liu, Y. Yang, X. Guo, M. Kloft, and L. He, "Multiview subspace clustering via co-training robust data representation," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2021.

[58] J. Liu, X. Liu, Y. Yang, S. Wang, and S. Zhou, "Hierarchical multiple kernel clustering," in *Proceedings of the Thirty-fifth AAAI Conference on Artificial Intelligence, (AAAI-21), Virtually, February 2-9, 2021*, 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 *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[60] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[61] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *IEEE*, pp. 2818–2826, 2016.

[62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

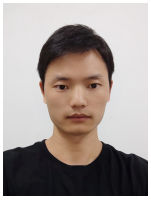
[63] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[64] B. Ramachandra, M. Jones, and R. R. Vatsavai, "A survey of single-scene video anomaly detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2462–2470.

[66] G. Yu, S. Wang, Z. Cai, E. Zhu, C. Xu, J. Yin, and M. Kloft, "Cloze test helps: Effective video anomaly detection via learning to complete video events," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 583–591.
[67] O. Kopuklu, J. Zheng, H. Xu, and G. Rigoll, "Driver anomaly detection: A dataset and contrastive learning approach," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 91–100.
[68] Xiao, Liu, Fanjin, Zhang, and Tang, "Self-supervised learning: Generative or contrastive," 2020.
[69] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, "A review on generative adversarial networks: Algorithms, theory, and applications," *arXiv preprint arXiv:2001.06937*, 2020.
[70] K. Han, Y. Wang, H. Chen, X. Chen, and D. Tao, "A survey on visual transformer," 2020.

Siqi Wang is currently an assistant research professor in College of Computer, NUDT. His main research include outlier/anomaly detection and unsupervised learning. His works have been published on leading conferences and journals, such as NeurIPS, AAAI, IJCAI, ACM MM, TPAMI and TIP. He serves as a PC member and reviewer for top-tier conference like NeurIPS and AAAI and several prestigious journals.



Jiyuan Liu is a PhD student in National University of Defense Technology (NUDT), China. His current research interests include multi-view clustering, deep clustering and anomaly detection. He has published papers in journals and conferences such as IEEE T-KDE, IEEE T-NNLS, ICML, CVPR, ICCV, ACMMM, AAAI, IJCAI, etc.

Guang Yu received the bachelor's degree in computer science and technology from Sichuan University, Chengdu, China, in 2018. He is currently working toward the Ph.D. degree at the College of Computer, National University of Defense Technology, Changsha, China. His main research interests include anomaly/outlier detection and self-supervised/unsupervised learning.

Xinwang Liu received his PhD degree from National University of Defense Technology (NUDT), China. He is now Professor at School of Computer, NUDT. His current research interests include kernel learning and unsupervised feature learning. Dr. Liu has published 60+ peer-reviewed papers in journals and conferences such as IEEE T-PAMI, IEEE T-KDE, IEEE T-IP, IEEE T-NNLS, IEEE T-MM, IEEE T-IFS, NeurIPS, CVPR, ICCV, AAAI, IJCAI, etc.

Sihang Zhou received his PhD degree from National University of Defense Technology (NUDT), China. He is now lecturer at College of Intelligence Science and Technology, NUDT. His current research interests include machine learning and medical image analysis. Dr. Zhou has published 20+ peer-reviewed papers, including IEEE T-IPEFERENCES, IEEE T-NNLS, IEEE T-MI, Information Fusion, Medical Image Analysis, AAAI, MICCAI, etc.weight

En Zhu received his PhD degree from National University of Defense Technology (NUDT), China. He is now Professor at School of Computer Science, NUDT, China. His research interests are pattern recognition, image processing, machine vision and machine learning. Dr. Zhu has published 60+ peer-reviewed papers, including IEEE T-CSVT, IEEE T-NNLS, PR, AAAI, IJCAI, etc. He was awarded China National Excellence Doctoral Dissertation.

Yuexiang Yang received the B.S. degree in Mathematics from Xiangtan University, Xiangtan, China, in 1986, the M.S. degree in Computer Application and the PHD degree in Computer Science and Technology from 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 his PhD degree from National University of Defense Technology (NUDT), China. He is now the distinguished Professor at Dongguan University of Technology. His research interests include pattern recognition and machine learning. Dr. Yin has published 150+ peer-reviewed papers, including IEEE T-CSVT, IEEE T-NNLS, PR, AAAI, IJCAI, etc.

Wenjing Yang Wenjing Yang received the Ph.D. degree in multi-scale modeling from Manchester University, Manchester, U.K. She is currently an Associate Research Fellow with Institute for Quantum Information & State Key Laboratory of High Performance Computing, National University of Defense Technology, Changsha, China. Her research interests include deep learning, multiagent reinforcement learning and high-performance computing.