Elastic One-Class Domain-Adversarial Neural Network for Trustworthy Anomaly Detection

1st Yang Liu
School of Computer and Information Engineering
Henan Normal University
Xinxiang, China
Tdhdhfhg@qq.com

2nd Jianliang He
School of Computer and Information Engineering
Henan Normal University
Xinxiang, China
hejianliang@htu.edu.cn

3rd Long Chao
School of Computer and Information Engineering
Henan Normal University
Xinxiang, China
chaolong@stu.htu.edu.cn

4th Wentao Mao*
School of Computer and Information Engineering
Henan Normal University
Engineering Lab of Intelligence Business &
Internet of Things, Henan Province, China
Xinxiang, China
maowt@htu.edu.cn

***Abstract*—Deep transfer learning has shown substantial potential in developing the decision capability of one-class anomaly detection. However, existing SVDD-like methods adopting hypersphere-formed decision mechanism with hard boundary generally neglect the transitional characteristics from weak anomalies (e.g., incipient faults of mechanical supporting element) to genuine anomalies. Also, these methods lack uncertainty quantification for weak anomalies near decision boundary, thus leading to inadequate decision confidence. To address these concerns, this paper proposes an elastic one-class domain-adversarial neural network for trustworthy anomaly detection. By adopting the classical Deep SVDD as the baseline architecture, a deep elastic SVDD model with hypersphere annulus is built to capture weak anomalies' transitional characteristics, while the anomaly probability of samples within the annulus is calculated using the Gumbel Copula function. On this basis, an elastic one-class domain-adversarial neural network with hypersphere annulus is developed to seek domain-invariant representation of one-class discriminative information. By forcing geometric overlap of the hyperspheres in source and target domains, not only the detection accuracy for genuine anomalies but also the identification capability for weak ones are both enhanced. Experimental evaluation is conducted on two typical anomaly detection problems, i.e., image recognition detection on the MNIST∼USPS dataset and early fault detection on the IEEE PHM Challenge 2012 bearing dataset. The results show that the proposed model is capable of reaching higher detection accuracy on data of various forms with lower false alarm rate and better detection confidence for weak anomalies.**

***Index Terms*—Anomaly detection, Transfer learning, One-class classification, Decision confidence, Anomaly probability**

I. Introduction

With the rapid development of machine learning techniques, early fault detection of key components in high-end equipment based on condition monitoring data has become critical for operational safety. However, under actual industrial environment, degradation process of the components is often accompanied by signals of weak fault that are easily submerged by background noise. As a result, weak faults are difficult to be separated in a stable and reliable manner, leading to the continuous accumulation of safety risk. Trustworthy anomaly detection just aims to identify genuine anomalies and quantify weak anomalies in uncertainty environment. Achieving trustworthy anomaly detection from dynamically collected condition monitoring data with early warning signs precisely captured remains a major technical bottleneck in both academic and industrial fields.

In recent years, deep anomaly detection has achieved remarkable progress in equipment health management [1], and predictive maintenance [2]. Compared with traditional statistical approaches and shallow models, deep learning methods are capable of end-to-end extraction of discriminative information, which improves anomaly detection performance [3]–[5]. Nevertheless, due to the scarcity and high acquisition cost of anomaly samples, supervised learning methods relying on both normal and anomaly data are impractical in real industrial scenarios. To overcome this limitation, early studies mainly focused on unsupervised or one-class anomaly detection methods that use only normal samples [6]. But under industrial environments, varying operating conditions lead to distribution shift, making conventional one-class classification methods prone to cut the generalization capability down. To address this issue, transfer learning has been further introduced into anomaly detection, referred to as *one-class transfer learning* or *cross-domain anomaly detection*. Representative strategies include sample selection [7], feature mapping [8], and domain-adversarial training [9], et al. Nevertheless, most methods primarily emphasize statistical alignment of feature distributions. When faced with complex samples with heavy noise interference, they fail to effectively capture the transitional characteristics of weak anomalies and still struggle to deliver trustworthy detection results.

From the mentioned-above analysis, it can be inferred that the critical challenge of cross-domain anomaly detection lies in extracting and transferring discriminative information from source domain, especially the information related to weak anomalies. Existing one-class anomaly detection methods rely on the compactness assumption of normal samples in feature space [3], [10]. In industrial applications, anomalies do not always occur abruptly. Instead, it is typically a gradual and accumulative process [9], [11]. Consequently, conventional hard boundary-formed detection rules are not always effective and often fail to identify weak anomalies in an arbitrary manner, leading to alarm delay. Here we take Deep SVDD [3] as an example for demonstration. As illustrated in Fig. 1(a), clear detection boundary can be learned when features are well separated (e.g., standard handwritten digit 0). For ambiguous samples such as poorly written digits whose shapes lie between 0 and 6, an annular zone emerges near the decision boundary. In such case, the hard hypersphere boundary learned by Deep SVDD cannot accommodate transitional variations, making ambiguous samples easily to be misclassified. An ideal approach, as illustrated in Fig. 1(b), is to explicitly define an annular zone between the safe zone and the anomaly zone, i.e., realizing trustworthy anomaly detection. It is referred to as a hypersphere annulus, which identifies samples in the annular zone as uncertain and assigns them anomaly probability for better decision credibility.

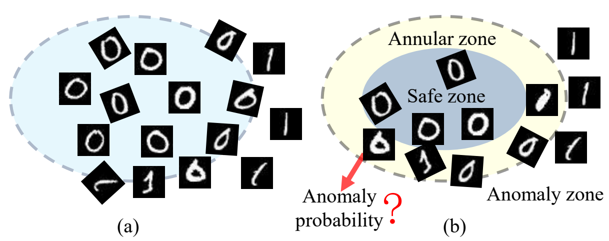

Fig. 1. Schematic illustration of trustworthy anomaly detection with a hypersphere annulus, where (a) denotes the hard boundary method represented by Deep SVDD, and (b) denotes the proposed method with a hypersphere annulus.

This study just focuses on trustworthy cross-domain anomaly detection that aims to improve the detection of weak anomalies by adapting a detection rule with hypersphere annulus learned from source domain to target domain. A key challenge lies in establishing an effective transfer pathway for one-class discriminative information defined by the hypersphere annulus. To address this challenge, this paper first builds an elastic Deep SVDD model with a hypersphere annulus as the baseline. Then an elastic one-class domain-adversarial neural network (EO-DANN) is further developed for trustworthy anomaly detection. The proposed method learns shared one-class discriminative representation between source and target domains using domain-adversarial training, while geometrically forcing hyperspheres of two domains to overlap as much as possible. In this mechanism, the adaptation of hypersphere facilitates the transfer of one-class discriminative information, while the adaptation of hypersphere annulus enables the transmission of transitional characteristics of weak anomalies. Experimental results on image recognition and early fault detection show the superior performance of the proposed method. Particularly, the proposed method necessarily provides anomaly probability for weak anomalies with higher detection accuracy, which exhibits feasibility for industrial-scale applications.

The main contributions of this paper are summarized as follows: 1) An elastic Deep SVDD model with hypersphere annulus is proposed. The proposed model overcomes the limitation on the hard boundary by introducing elastic decision boundary with annular zone. Furthermore, the Gumbel Copula function is employed to quantify the anomalous probability for samples in annular zone; 2) A one-class domain-adversarial training method with center alignment is developed. By aligning the hypersphere centers, this method promotes the geometric overlap between two domains' hyperspheres. This enables reliable transfer of discriminative information related to weak anomalies and enhances trusting results of anomaly detection under an uncertainty environment.

II. Elastic Deep SVDD

To capture the transitional characteristics of weak anomalies, this section proposes an elastic Deep SVDD model with hypersphere annulus structure. The core motivation is to learn a compact representation of normal samples in feature space and establish a hypersphere annulus based on distance distribution. This partitions the feature space into three zones: safe zone, annular zone and anomaly zone. Furthermore, the anomaly probability is assigned to samples located in the annular zone to explicitly quantify their uncertainty.

This section adopts Deep SVDD as the basic model architecture. Specifically, the variational autoencoder (VAE) encoder is employed as the feature generator E. The aim of E is to learn a neural network $\phi(\cdot; W_f) : X \to \mathcal{F} \in \mathbb{R}^m$ (where $W_f = (w^1, \ldots, w^L)$ denotes the weights of the l-th hidden layer) that projects the input data into a m-dimensional feature space. Similar to Deep SVDD, the elastic Deep SVDD seeks a minimum hypersphere enclosing normal samples in the feature space $\mathcal{F}$. Different from conventional formulations, an elastic hypersphere annulus is further constructed by introducing an inner radius and an outer radius around the hypersphere boundary. The optimization problem is formulated as

$$\min_{R_{\text{out}}, W_f} \mathcal{L}_{HA} + \beta \mathcal{L}_{KL} \tag{1}$$

In (1):

- $$\begin{aligned} \mathcal{L}_{HA} = & R_{\text{out}}^2 + \frac{1}{\nu n} \sum_{i=1}^{n} \max\left\{0, \|\phi(x_i; W_f) - c\|^2 - R_{\text{out}}^2\right\} \\ & + \frac{\lambda}{2} \sum_{l=1}^{L} \|W_f^l\|_F^2 \end{aligned} \tag{2}$$

  $\mathcal{L}_{HA}$ is the hypersphere annulus loss, where $c = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)$ denotes the hypersphere center, ν is the regularization parameter that controls the tightness of the hypersphere boundary, and n is the number of training samples.

- $$\mathcal{L}_{KL} = \frac{1}{n}\sum_{i=1}^{n}(\phi(x_i)\|\mathcal{N}(0,1)) \tag{3}$$
  $\mathcal{L}_{KL}$ is the Kullback-Leibler (KL) divergence loss used to force the features extracted by E to approximate a standard normal distribution. Minimizing (3) will prevent hypersphere collapse in the feature space.

After solving (1), the inner radius can be determined jointly by the outer radius and the maximum distance between training samples and the hypersphere center, as:

$$R_{in} = 2R_{out} - \max_{i=1,\ldots,n} \|\phi(x_i) - c\| \tag{4}$$

Using R_{in} and R_{out}, the feature space is divided into three non-overlapping zones: safe zone, annular zone, and anomaly zone.

Considering weak anomalies often exhibit synchronized deviations across multiple feature dimensions, the proposed elastic Deep SVDD model further assigns an anomaly probability to each test sample by adopting the Gumbel Copula function [12] due to its capability of capturing upper-tail dependence. The implementation is as follows. For each test sample, the marginal probability distribution of each feature dimension is estimated using the empirical cumulative distribution function (ECDF), as:

$$u^j = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(x_i^j \le x_{\text{test}}), \quad j = 1,\ldots,m \tag{5}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function and j represents the feature dimension. Subsequently, the dependence among feature dimensions is characterized by the Gumbel Copula function, yielding the joint anomaly probability as:

$$\text{Copula}(u^1, u^2, \ldots, u^m) = \exp\left\{-\left[\sum_{j=1}^{m}(-\ln u^j)^\theta\right]^{1/\theta}\right\} \tag{6}$$

where θ denotes the correlation parameter estimated via Kendall's Tau method [13].

Based on (6), the final decision probability of the proposed elastic Deep SVDD model for recognizing the test sample x_{test} as anomalous is output as:

$$P(x_{\text{test}}) = \begin{cases} 0, & D(x_{\text{test}}) < R_{\text{in}}^2 \\ \text{Copula}(\mathbf{u}), & R_{\text{in}}^2 \le D(x_{\text{test}}) \le R_{\text{out}}^2 \\ 1, & D(x_{\text{test}}) > R_{\text{out}}^2 \end{cases} \tag{7}$$

where $D(x_{\text{test}}) = \|\phi(x_{\text{test}}) - c\|^2$, $\mathbf{u} = (u^1, u^2, \ldots, u^m)$.

From (7), the proposed elastic Deep SVDD model works on the hypersphere annulus structure. If the test sample is located within the safe zone, its anomaly probability is 0, indicating a high confidence of normal state. If the sample falls into the anomaly zone, the anomaly probability is 1 and the sample will be classified as high-confidence anomaly. Once the sample lies within the annular zone, it will be considered as an uncertain state with the anomaly probability calculated by (7). By explicitly modeling upper-tail dependence via the Gumbel Copula function, the proposed elastic Deep SVDD model is capable of recognizing the occurrence of abnormal state more sensitively, thus providing trustworthy probability for weak anomalies.

III. Elastic One-class Domain-adversarial Neural Network

Based on the elastic Deep SVDD, this section elaborates the algorithmic details of EO-DANN for trustworthy cross-domain anomaly detection. As illustrated in Fig. 2, EO-DANN mainly consists of three parts: 1) feature extractor, 2) anomaly detector with a hypersphere annulus, and 3) hypersphere annulus adaptation based on center alignment. Among them, hypersphere annulus adaptation plays a vital role in transferring one-class discriminative information from source domain to target domain.

The overall architecture of EO-DANN follows the classic domain-adversarial training framework, with modifications for the one-class problem. Specifically, source domain data and target domain data are fed into the feature extractor. A decoder is also utilized to preserve reconstruction capability. The extracted features are then forwarded to one-class anomaly detector and domain discriminator. Through iterative back-propagation, the model finally converges to domain-invariant representation that retains one-class discriminative information.

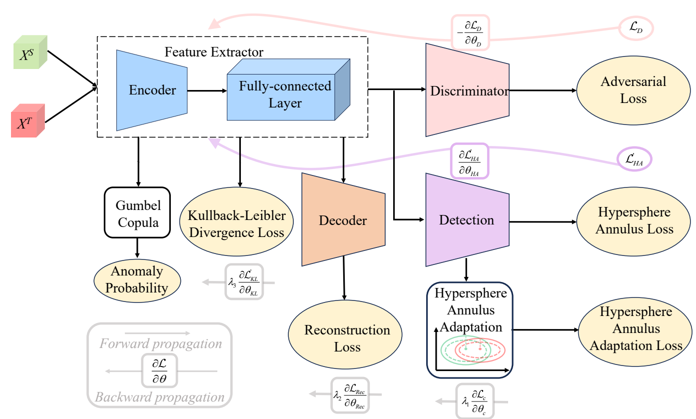

Fig. 2. Structure of the proposed EO-DANN model

A. Feature Extractor

Consistent with the elastic Deep SVDD, a VAE encoder is adopted as the domain-shared feature extractor. It projects source and target domains' data into a common subspace to extract domain-invariant feature representation. The inputs are normal samples from both domains: $X^S = \{x_i^S\}_{i=1}^{n^S} \subset \mathbb{R}^d$, $X^T = \{x_j^T\}_{j=1}^{n^T} \subset \mathbb{R}^d$, where S and T denote source domain and target domain respectively. Due to the distribution discrepancy between the two domains, i.e., $P^S(\phi(x)) \neq P^T(\phi(x))$, the hypersphere annulus learned from the source domain is supposed not to be applicable to the target domain. That is just the initial motivation of this study.

To prevent losing critical information of target domain, reconstruction loss is first built as:

$$\mathcal{L}_{\text{Rec}} = \mathbb{E}_{x\sim X^S} \left\|x - \hat{x}\right\|^2 + \mathbb{E}_{x\sim X^T} \left\|x - \hat{x}\right\|^2 \tag{8}$$

where $\hat{x}$ denotes the decoder output reconstructed on the latent features, i.e., $\hat{x} = Decoder(\phi(x))$.

B. Anomaly Detector with Hypersphere Annulus

Based on $\phi(x_i^S; W_f)$ and $\phi(x_j^T; W_f)$, the hyperspheres on the normal samples from the source and target domains are trained using the elastic Deep SVDD shown in Section II. EO-DANN minimizes the volumes of the two hyperspheres in feature space $\mathcal{F}$ with $R_{\text{out}} = \{R_{\text{out}}^S, R_{\text{out}}^T\}$ and $c = \{c^S, c^T\}$, yielding the hypersphere annulus loss $\mathcal{L}'_{HA}$ and the KL divergence loss $\mathcal{L}'$, as:

$$\begin{aligned}
\mathcal{L}'_{HA} &= (R_{\text{out}}^S)^2 + (R_{\text{out}}^T)^2 \\
&\quad + \frac{1}{\nu n^S}\sum_{i=1}^{n^S} \max\left\{0, \|\phi(x_i^S; W_f) - c^S\|^2 - (R_{\text{out}}^S)^2\right\} \\
&\quad + \frac{1}{\nu n^T}\sum_{j=1}^{n^T} \max\left\{0, \|\phi(x_j^T; W_f) - c^T\|^2 - (R_{\text{out}}^T)^2\right\} \\
&\quad + \frac{\lambda}{2}\sum_{l=1}^{L} \|W_f^l\|_F^2
\end{aligned} \tag{9}$$

$$\begin{aligned}
\mathcal{L}'_{KL} &= \frac{1}{n^S}\sum_{i=1}^{n^S} D\left(\phi(x_i^S)\|\mathcal{N}(0,1)\right) \\
&\quad + \frac{1}{n^T}\sum_{j=1}^{n^T} D\left(\phi(x_j^T)\|\mathcal{N}(0,1)\right)
\end{aligned} \tag{10}$$

C. Hypersphere Annulus Adaptation

To transfer the one-class discriminative information, an efficient solution is to enforce geometric consistency between hypersphere annuli across the two domains. It is equivalent to aligning the centers of two hyperspheres, i.e., center alignment. We believe that the adaptation of hypersphere annulus can well support the information transfer about the transitional characteristic of weak anomalies beyond the classical strategy of feature adaptation. The hypersphere annulus adaptation loss $\mathcal{L}_c$ is designed as:

$$\mathcal{L}_c(W_f) = \|c^S - c^T\|^2 \tag{11}$$

where c^S, c^T are the center of two hyperspheres respectively. Minimizing (11) forces the overlap of two hypersphere annuli, thereby promoting consistent distribution characteristic between the two domains in the feature space.

In addition, the domain discriminator D is used to distinguish which domain the incoming data belong to. Its loss function is expressed as [14]

$$\mathcal{L}_D(W_d) = \mathbb{E}_{x\sim X^S}\left[\log D(\phi(x))\right] + \mathbb{E}_{x\sim X^T}\left[\log(1 - D(\phi(x)))\right] \tag{12}$$

where W_d is the network weight of D. Maximizing (12) aims to confuse features from the target domain with those from the source domain. If the feature mapping $\phi(\cdot)$ extracted by E cannot distinguish between two domains' samples, $\phi(\cdot)$ is then considered as a domain-invariant feature representation.

In summary, the overall optimization function of EO-DANN is

$$\mathcal{L}_{total} = -\mathcal{L}_D + \mathcal{L}'_{HA} + \lambda_1\mathcal{L}_c + \lambda_2\mathcal{L}_{Rec} + \lambda_3\mathcal{L}'_{KL} \tag{13}$$

where λ_1, λ_2, λ_3 are the regularization parameters. Solving (13) requires minimizing $\mathcal{L}'_{HA}$, $\mathcal{L}_c$, $\mathcal{L}_{\text{Rec}}$, $\mathcal{L}'_{KL}$ while maximizing $\mathcal{L}_D$ iteratively. The ADAM algorithm is employed for optimization, details of which are omitted here due to space limitation. This process can be summarized as: seeking the optimal $\phi^*(\cdot)$ that enables anomaly detectors in both domains to achieve the best one-class discriminative performance, while with $\phi^*(\cdot)$, conducting feature adaptation and hypersphere annulus adaptation in the domain-adversarial training. Through this joint optimization, EO-DANN learns domain-invariant feature representations while forcing geometric overlap of the two hypersphere annuli. This process not only facilitates cross-domain transfer of one-class discriminative information, but also enables effective learning of transitional characteristics of weak anomalies. The anomaly probability output of EO-DANN is identical to that in (7).

IV. Experiments

In the section, EO-DANN is evaluated on two anomaly detection tasks: 1) image recognition detection [?] on the MNIST∼USPS datasets and 2) early fault detection of rolling bearings [15] on the IEEE PHM Challenge 2012 dataset. For both tasks, only normal samples are used for training.

A. Image Recognition Detection

1) Experimental Setting: Some typical handwritten digit images in the MNIST and USPS datasets are shown in Fig. 3 for straightforward understanding. MNIST serves as the source domain with 100 clear images per digit, while USPS is the target domain with 25 blurry images per digit. In ten experiments, each digit (0-9) is treated as the normal class in turn. The EO-DANN model is trained on normal samples from both domains and tested on unused USPS normal images along with next-digit anomalies.

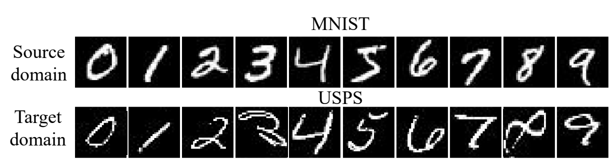

Fig. 3. Examples of the MNIST∼USPS dataset

2) Experimental Results: Fig. 4 shows the feature distribution before and after the transfer achieved by EO-DANN. Before the transfer, the feature distribution of the target domain is relatively scattered and much separated from the source domain. In contrast, after the transfer by EO-DANN, the features of the two domains achieve good alignment well, indicating that the model successfully learns domain-shared features.

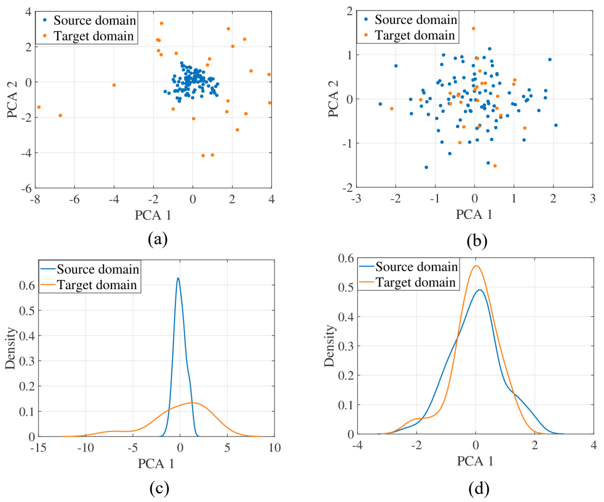

Fig. 4. Feature distribution and the corresponding probability density curve of the two domains' data, where (a) and (c) are before the transfer, (b) and (d) are after the transfer by EO-DANN.

Fig. 5 visualizes some samples with different risk levels represented by the obtained anomaly probability. The first row shows clear normal samples assigned low probability, while the bottom row shows clear anomalies with high probability. It is clear that the samples in the annular zone get intermediate probabilities. For instance, a sloppily digit 6 receives p=0.4, and a visually distorted 0 is assigned p=0.96. These results confirm that EO-DANN not only detects anomalies but also effectively quantifies the uncertainty for ambiguous digit images.

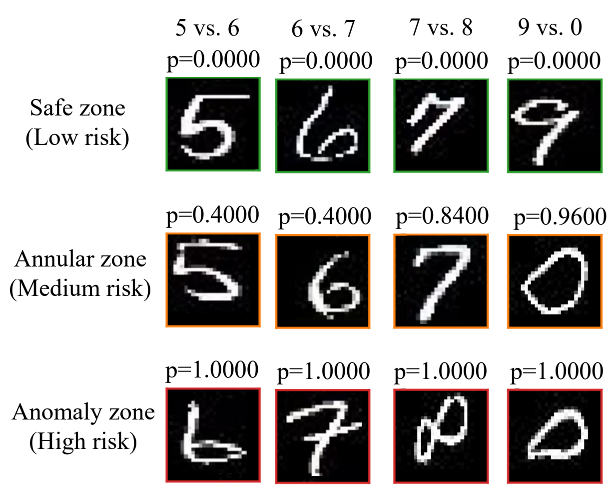

Fig. 5. Visualization of anomaly detection results with different risk level. Each column represents one experiment (e.g., digit 6 as normal vs. 7 as anomaly), while p indicates anomaly probability)

3) Competing Experiments: Eight anomaly detection methods are compared to verify the effectiveness of EO-DANN, as listed in Table I. Among them, OC-SVM, SVDD and iForest are shallow models without transfer learning; the others are deep learning models. DCAE, SSL and DAAD are SOTA methods for anomaly detection transfer learning. Deep SVDD (1) is trained only on target data, while Deep SVDD (2) is trained on both source and target data.

Please note that the classification decision in Table I is based on anomaly probability, with the decision boundary set at 0.5. The samples with anomaly probability below 0.5 are classified as normal, while those above are identified as anomaly. From Table I, EO-DANN achieves the highest accuracy in almost all groups.

B. Early Fault Detection of Rolling Bearings

Rolling bearings are critical components whose health status directly affects operational safety. Under high-speed and heavy-load conditions, faults are inevitable and typically begin with weak degradation. Early fault detection is crucial for preventing severe faults and scheduling maintenance. As shown in Fig. 6, the whole-life degradation process includes three states: normal, early fault, and fast degradation, with early fault detection aiming to identify incipient faults from normal data.

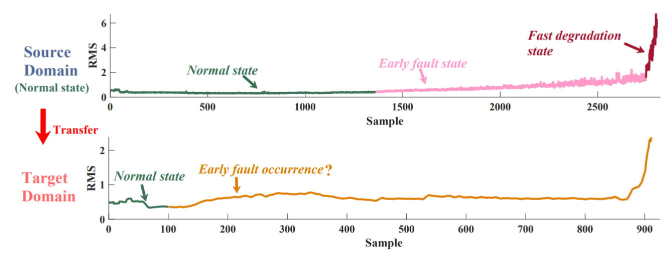

Fig. 6. Schematic diagram of early fault detection of rolling bearings with one-class transfer learning.

1) Experimental Settings: This section evaluates the proposed method on the IEEE PHM Challenge 2012 dataset (PHM for short), which contains whole-life data of 17 bearings from normal to fault under three speeds (1800 rpm, 1650 rpm, 1500 rpm) with corresponding loads (4000N, 4200N, 5000N); see [21] for details. Due to distribution differences, Condition 1 is the source domain and Condition 2 the target domain. Source training data consist of the first 500 normal samples from multiple source bearings (verified in [22]), and target training data consist of the first 100 samples of bearing B2_1. Table II and Fig. 7 summarize this setup.

2) Experimental Results: In Fig. 8(a), the hypersphere annulus structure constructs a clear decision boundary: dark blue samples inside the green ellipse represent the normal state, with anomaly probability 0; dark red sample points outside the red ellipse represent severe anomalies, with probability 1. The zone between the two ellipses (as shown in the local zoom box) corresponds to early fault occurrence (approximately samples 100-150) marked by the red rectangle in Fig. 8(b). Within this interval, the distance scores rise above the safety line but remain under the fault line. The corresponding samples in the feature space show a transitional spread. At this stage,

TABLE I
CLASSIFICATION ACCURACY ON MNIST∼USPS

Normal	Anomalous	OC-SVM [16]	SVDD [17]	iForest [18]	Deep SVDD(1)	Deep SVDD(2)	DCAE (pre-train) [19]	SSL (pre-train) [4]	DAAD [20]	Our
0	1	0.57	0.59	0.48	0.58	0.90	0.85	0.67	0.80	0.87
1	2	0.86	0.38	0.61	0.80	0.78	0.71	0.76	0.78	0.91
2	3	0.79	0.36	0.69	0.69	0.83	0.72	0.73	0.71	0.75
3	4	0.70	0.49	0.55	0.65	0.70	0.76	0.72	0.64	0.93
4	5	0.83	0.54	0.64	0.69	0.83	0.67	0.78	0.69	0.83
5	6	0.76	0.52	0.64	0.73	0.73	0.70	0.68	0.66	0.91
6	7	0.71	0.56	0.53	0.64	0.82	0.75	0.69	0.58	0.88
7	8	0.78	0.40	0.66	0.69	0.77	0.66	0.70	0.73	0.81
8	9	0.62	0.48	0.59	0.61	0.61	0.60	0.59	0.65	0.70
9	0	0.71	0.48	0.58	0.73	0.75	0.69	0.76	0.67	0.68

TABLE II
DATASET CONFIGURATION ACROSS DOMAINS

Domain	Bearing	Condition	Name	Number
Source domain	B1_1	Condition 1	Source 1	500
	B1_2	Condition 1	Source 2	500
	B1_3	Condition 1	Source 3	500
	B1_4	Condition 1	Source 4	500
Target domain	B2_1	Condition 2	Target 1	100

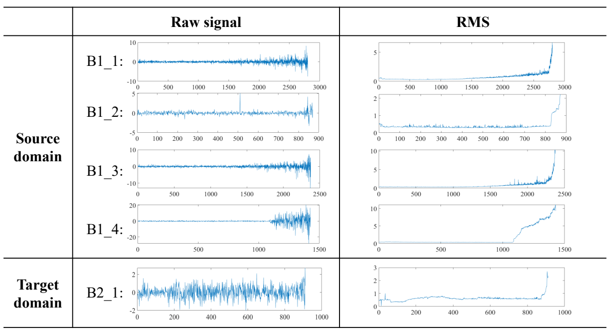

Fig. 7. Data from the PHM dataset. The left column shows raw vibration signals, and the right column shows the corresponding RMS curves.

EO-DANN maps the distance scores to anomaly probabilities (corresponding to the color mapping from blue to red in the figure), achieving uncertainty quantification of the early fault samples. Also, the results verify that EO-DANN overcomes the limitations of hard boundary.

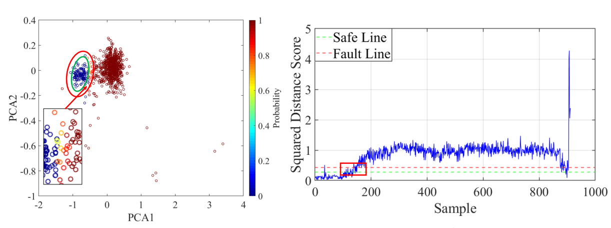

Fig. 8. (a) Feature distribution and (b) distance score curve for target bearing B2_1. In (a), the legend color represents the anomaly probability computed by the Gumbel Copula (see (6)).

To validate the effectiveness of EO-DANN, two ablation experiments are set up: 1) removing the hypersphere annulus structure, i.e., the decision boundary retains only the outer radius, and the probability model operates over the entire space; 2) removing the hypersphere annulus adaptation, i.e., training only with source domain data. The ablation results are shown in Fig. 9. EO-DANN maintains high stability during the normal state and exhibits noticeable probability fluctuations at the initial part of early fault. Despite the alarm position is earlier than EO-DANN, too many false alarms are raised in Fig. 9(b). It is clear that introducing the hypersphere annulus effectively suppresses noise interference, while the hypersphere annulus adaptation guarantees robust transfer of one-class discriminative knowledge across varying operating conditions.

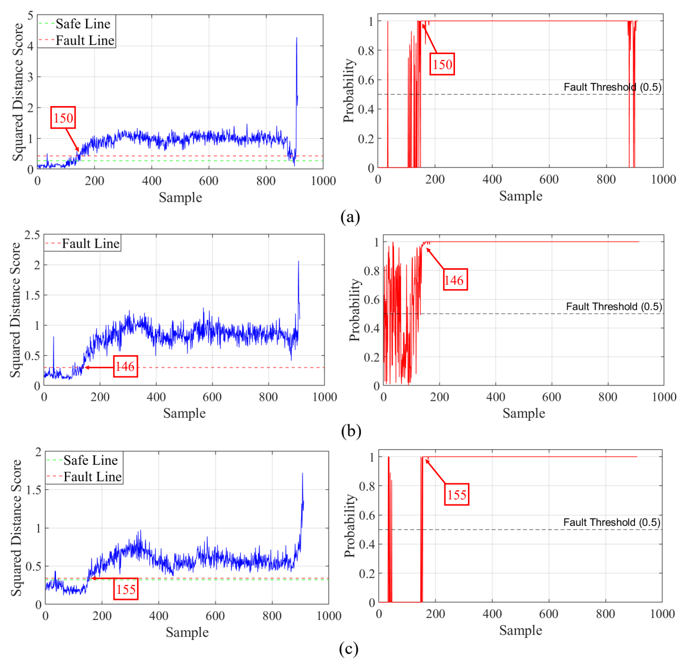

Fig. 9. Ablation experimental results, where (a) is the result of the proposed EO-DANN model, (b) is the result without hypersphere annulus structure, and (c) is the result without hypersphere annulus adaptation.

3) Comparative Experiments: To further validate the proposed method, 17 representative anomaly detection approaches are compared in Table III, following [26] with detection location and false alarm count as evaluation metrics.

From Table III, the proposed method identifies early faults at a relatively earlier location with a much lower number of false alarms. The reason lies in two parts. At the feature level, the domain-adversarial training algorithm achieves hypersphere-center alignment of source and target domain features in the latent space. Moreover, the hypersphere annulus effectively

TABLE III

EXPERIMENTAL RESULTS OF TOTAL 18 METHODS ON THE PHM DATASET

Category	Method	Loc.	Alarms
Signal analysis	1. BEMD + AMMA [23]	185	–
Anomaly detection (no transfer)	2. Kurtosis + LOF	606	35
	3. SDAE + LOF	280	17
	4. Kurtosis + SVDD	819	53
	5. SDAE + SVDD	880	4
	6. Kurtosis + iFOREST	311	36
	7. SDAE + iFOREST	876	20
	8. Deep SVDD (1)	253	2
	9. Deep SVDD (2)	201	0
	10. SSL	885	7
Anomaly detection (with transfer)	11. DCAE pre-train [19]	271	8
	12. KD pre-train [24]	223	31
	13. SSL pre-train [4]	150	5
Early fault detection (no transfer)	14. SDFM [25]	175	7
	15. FDDA [26]	879	4
Early fault detection (with transfer)	16. DAFD [27]	225	13
	17. DAAD [20]	179	6
	18. EO-DANN	150	1

shields interference from steady-state noise, while the anomaly probability sensitively perceives subtle deviations within the initial part of early fault. This enables the model not only to maintain safety margin but also to capture the characteristics of early fault for the target bearing.

V. CONCLUSION

To realize trustworthy anomaly detection, this paper proposes the EO-DANN model by leveraging hypersphere annulus structure and probabilistic quantization. Through capturing and transferring the transitional characteristics of weak anomalies, EO-DANN achieves high-confidence one-class detection on domain adversarial training. This model is simple but effective, and can provide a practically-feasible risk assessment solution for industrial applications.

Future work will focus on the mathematical modeling of the transitional characteristics of weak anomalies, which will establish a theoretical guaranty beyond empirical validation. Additionally, as the proposed model yields probabilistic output, it is significant practical value to integrate with large language models (LLMs) for zero-shot inference.

VI. ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62472146, and in part by the Key R&D Special Project of Henan Province, China under Grant No. 261111213000.

REFERENCES

[1] R. Zhao, R. Yan, Z. Chen, and G. Liu, "Deep learning and its applications to machine health monitoring," Mech. Syst. Signal Process., vol. 115, pp. 213–237, 2019.

[2] Z. Li, Q. He, and J. Li, "A survey of deep learning-driven architecture for predictive maintenance," Eng. Appl. Artif. Intell., vol. 133, p. 108285, 2024.

[3] L. Ruff, R. Vandermeulen, N. Goernitz, and M. Kloft, "Deep one-class classification," in Proc. Int. Conf. Mach. Learn., 2018, pp. 4393–4402.

[4] K. Sohn, C. L. Li, J. Yoon, and D. Min, "Learning and evaluating representations for deep one-class classification," unpublished, arXiv:2011.02578, 2020.

[5] P. Zheng, S. Yuan, X. Wu, and J. Li, "One-class adversarial nets for fraud detection," in Proc. AAAI Conf. Artif. Intell., 2019, pp. 1286–1293.

[6] G. Michau and O. Fink, "Unsupervised transfer learning for anomaly detection: Application to complementary operating condition transfer," Knowl.-Based Syst., vol. 216, p. 106816, 2021.

[7] R. Xia, C. Zong, X. Hu, and E. Cambria, "Feature ensemble plus sample selection: Domain adaptation for sentiment classification," IEEE Intell. Syst., vol. 28, no. 3, pp. 10–18, May/Jun. 2013.

[8] Q. Wan, L. Gao, X. Li, and L. Wang, "Unsupervised image anomaly detection and segmentation based on pretrained feature mapping," IEEE Trans. Ind. Informat., vol. 19, no. 3, pp. 2330–2339, Mar. 2022.

[9] X. Han and S. Yuan, "Unsupervised cross-system log anomaly detection via domain adaptation," in Proc. 30th ACM Int. Conf. Inf. Knowl. Manage., 2021, pp. 3068–3072.

[10] P. Perera and V. M. Patel, "Learning deep features for one-class classification," IEEE Trans. Image Process., vol. 28, no. 11, pp. 5450–5463, Nov. 2019.

[11] Y. Lei, N. Li, L. Guo, and N. Li, "Machinery health prognostics: A systematic review from data acquisition to RUL prediction," Mech. Syst. Signal Process., vol. 104, pp. 799–834, 2018.

[12] G. Horváth, E. Kovacs, R. Molontay, and M. Csernel, "Copula-based anomaly scoring and localization for large-scale, high-dimensional continuous data," ACM Trans. Intell. Syst. Technol., vol. 11, no. 3, pp. 1–26, 2020.

[13] P. K. Sen, "Estimates of the regression coefficient based on Kendall's tau," J. Amer. Stat. Assoc., vol. 63, no. 324, pp. 1379–1389, 1968.

[14] L. Chen, H. Chen, Z. Wei, and Z. Jin, "Reusing the task-specific classifier as a discriminator: Discriminator-free adversarial domain adaptation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022, pp. 7181–7190.

[15] W. Mao, S. Tian, J. Fan, and X. Liang, "Online detection of bearing incipient fault with semi-supervised architecture and deep feature representation," J. Manuf. Syst., vol. 55, pp. 179–198, 2020.

[16] J. Saari, D. Strömbergsson, J. Lundberg, and A. Gustafsson, "Detection and identification of windmill bearing faults using a one-class support vector machine," Measurement, vol. 137, pp. 287–301, 2019.

[17] Y. Zhao, S. Wang, and F. Xiao, "Pattern recognition-based chillers fault detection method using support vector data description," Appl. Energy, vol. 112, pp. 1041–1048, 2013.

[18] Y. Chen, Z. Zhao, H. Wu, and Z. Li, "Fault anomaly detection of synchronous machine winding based on isolation forest and impulse frequency response analysis," Measurement, vol. 188, p. 110531, 2022.

[19] Y. Zhou, X. Liang, W. Zhang, L. Zhang, and X. Song, "VAE-based deep SVDD for anomaly detection," Neurocomputing, vol. 453, pp. 131–140, 2021.

[20] W. Mao, G. Wang, L. Kou, and X. Liang, "Deep domain-adversarial anomaly detection with one-class transfer learning," IEEE/CAA J. Autom. Sinica, vol. 10, no. 2, pp. 524–546, Feb. 2023.

[21] P. Nectoux, R. Gouriveau, K. Medjaher, and N. Zerhouni, "PRONOSTIA: An experimental platform for bearings accelerated degradation tests," in Proc. IEEE Int. Conf. Prognostics Health Manage., 2012, pp. 1–8.

[22] W. Mao, L. Ding, S. Tian, and X. Liang, "Online detection for bearing incipient fault based on deep transfer learning," Measurement, vol. 152, p. 107278, 2020.

[23] Y. Li, M. Xu, X. Liang, and W. Huang, "Application of bandwidth EMD and adaptive multiscale morphology analysis for incipient fault diagnosis of rolling bearings," IEEE Trans. Ind. Electron., vol. 64, no. 8, pp. 6506–6517, Aug. 2017.

[24] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 4183–4192.

[25] W. Mao, J. Chen, X. Liang, and S. Li, "A new online detection approach for rolling bearing incipient fault via self-adaptive deep feature matching," IEEE Trans. Instrum. Meas., vol. 69, no. 2, pp. 443–456, Feb. 2019.

[26] W. Lu, Y. Li, Y. Cheng, and D. Meng, "Early fault detection approach with deep architectures," IEEE Trans. Instrum. Meas., vol. 67, no. 7, pp. 1679–1689, Jul. 2018.

[27] W. Lu, B. Liang, Y. Cheng, and D. Meng, "Deep model based domain adaptation for fault diagnosis," IEEE Trans. Ind. Electron., vol. 64, no. 3, pp. 2296–2305, Mar. 2016.