IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 5 December 2023, accepted 23 December 2023, date of publication 2 January 2024,
date of current version 10 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3349132

SURVEY

A Comprehensive Survey of Deep Transfer Learning for Anomaly Detection in Industrial Time Series: Methods, Applications, and Directions

PENG YAN[1,2], AHMED ABDULKADIR[1], PAUL-PHILIPP LULEY[1], MATTHIAS ROSENTHAL[4], GERRIT A. SCHATTE[5], BENJAMIN F. GREWE[2,3], AND THILO STADELMANN[1,6], (Senior Member, IEEE)

[1]Centre for Artificial Intelligence, ZHAW School of Engineering, 8400 Winterthur, Switzerland
[2]Faculty of Science, University of Zurich, 8057 Zürich, Switzerland
[3]Institute of Neuroinformatics, ETH Zürich, 8057 Zürich, Switzerland
[4]Institute of Embedded Systems, ZHAW School of Engineering, 8401 Winterthur, Switzerland
[5]Innovation Lab, Kistler Instrumente AG, 8408 Winterthur, Switzerland
[6]European Centre for Living Technology (ECLT), 30123 Venice, Italy

Corresponding author: Peng Yan (yanp@zhaw.ch)

This work was supported by Innosuisse grant 101.787 IP-ENG "DISTRAL".

ABSTRACT Automating the monitoring of industrial processes has the potential to enhance efficiency and optimize quality by promptly detecting abnormal events and thus facilitating timely interventions. Deep learning, with its capacity to discern non-trivial patterns within large datasets, plays a pivotal role in this process. Standard deep learning methods are suitable to solve a specific task given a specific type of data. During training, deep learning demands large volumes of labeled data. However, due to the dynamic nature of the industrial processes and environment, it is impractical to acquire large-scale labeled data for standard deep learning training for every slightly different case anew. Deep transfer learning offers a solution to this problem. By leveraging knowledge from related tasks and accounting for variations in data distributions, the transfer learning framework solves new tasks with little or even no additional labeled data. The approach bypasses the need to retrain a model from scratch for every new setup and dramatically reduces the labeled data requirement. This survey first provides an in-depth review of deep transfer learning, examining the problem settings of transfer learning and classifying the prevailing deep transfer learning methods. Moreover, we delve into applications of deep transfer learning in the context of a broad spectrum of time series anomaly detection tasks prevalent in primary industrial domains, e.g., manufacturing process monitoring, predictive maintenance, energy management, and infrastructure facility monitoring. We discuss the challenges and limitations of deep transfer learning in industrial contexts and conclude the survey with practical directions and actionable suggestions to address the need to leverage diverse time series data for anomaly detection in an increasingly dynamic production environment.

INDEX TERMS Deep transfer learning, time series analysis, anomaly detection, manufacturing process monitoring, predictive maintenance.

LIST OF ACRONYMS

AI	Artificial Intelligence.
CNN	Convolutional Neural Network.
DAN	Deep Adaptation Networks.
DNN	Deep Neural Network.
FCL	Fully Connected Layer.
GAN	Generative Adversarial Network.
LSTM	Long Short Term Memory.
ML	Machine Learning.
MMD	Maximum Mean Discrepancy.

The associate editor coordinating the review of this manuscript and approving it for publication was Yu Liu.

RNN Recurrent Neural Network.
SAE Sparse Auto-Encoder.

I. INTRODUCTION

A. MOTIVATION AND CONTRIBUTION

The fourth industrial revolution – Industry 4.0 [1], that is characterized by increasing efficiency through the digitization of production, automation, and horizontal integration across companies [2], and the advent of connected cyber-physical systems – referred to as internet of things [3], [4], [5], increases the need for autonomous and intelligent process monitoring. This can be exemplified by the use case of a smart factory in which industrial processes are transformed to be more flexible, intelligent, and dynamic [6], or the use case of decentralized energy production with wind and solar [7]. In these examples, AI-powered anomaly detection integrates the analysis of time series data to detect unusual patterns in the recorded data. To achieve this, a deep learning architecture is modeled to capture indicators of normal and abnormal operation. The learning process involves the analysis of historic time series sensor data of normal and possibly abnormal operations. This data is for example used for *representation-* or *reconstruction-based* learning. After training, the deep learning model *represents* or *reconstructs* normal data in a certain way. The model is designed in a way that abnormal data–because it is different–is either *represented* differently from the normal data or *reconstructed* poorly and thus recognized as an anomaly. By identifying operational parameters that fall outside a window of normal interval, operators can trigger interventions and adjustments to ensure high product quality and safe operations. To achieve this, physical properties such as pressure or temperature are monitored and analyzed in real-time applications. Changes in these variables capture drifting and abrupt faults caused by process failures or malfunctions [8]. The production process must adapt quickly to changes in production and the environment to meet the requirements for flexibility and dynamics. Further use cases exist in a wide range of diverse fields, such as manufacturing monitoring including automatic quality control [9], [10], predictive maintenance of goods and services [11], [12], [13], [14], [15], [16], infrastructure monitoring of building systems [17], [18] and power plant [19], digital agriculture [20], petrochemical process optimization [21], computer network intrusion detection [22], or aircraft flight monitoring [23], to name a few.

Artificial Intelligence, particularly deep learning, provides competent frameworks with underlying deep neural networks to automate intelligent monitoring and provide valuable assistance to operators and high-level control systems. Leveraging the power of deep learning, informative features of the data – technically referred to as *representations* [24] – can be captured in a machine-learned model and thereby enable a detailed understanding of variations in standard operations.

However, the task or underlying data may change under non-trivial and non-stationary conditions. For instance, the monitoring system of a milling machine may be assigned the task of identifying a blunt tool based on vibration in one scenario, and in a different scenario, it may utilize the same vibration measurements to detect insufficient cooling lubricant. Knowledge acquired to solve one task in one setting with a given tool, machined part, and type of machine may be transferred to solve the same or similar task in another setting with a different tool, machined part, or type of machine. Slowly changing conditions (drifts), abrupt mode changes (for instance, due to tool change), and new tasks (such as the detection of another failure mode) may require adjustments to the deep learning model. In these cases, it is desirable to adjust the analysis model without retraining from scratch, as it is costly or impractical to acquire sufficient training data to learn the full manifold [25].

Transfer learning is a machine learning framework to achieve this [26], [27], [28], [29], [30]. As depicted in Fig. 1, data and algorithms from one task may be leveraged in a new related one. By accounting for changes in data distributions and tasks or leveraging existing models, knowledge learned from related tasks can be used to improve performance on new tasks instead of retraining a model for each individual application from scratch. This transfer-learning-boosted modeling forms the basis for identifying anomalies that deviate from established patterns in a non-trivial manner without full re-training.

Deep transfer learning [29], [31] extends the transfer learning paradigm by leveraging deep learning. In industrial contexts, it ensures optimal production even as production conditions shift. This dynamic adaptability is key in maintaining the effectiveness of anomaly detection systems in the dynamic environment that characterizes industrial applications including the broad categories of manufacturing process monitoring, predictive maintenance, energy management, and infrastructure facility monitoring as detailed in Section IV.

This survey is a non-systematic yet application-oriented review with a narrow focus on deep transfer learning for anomaly detection in time series in the industry. Our main contributions are as follows:

- We categorize transfer learning problem settings and then systematically summarize deep transfer learning approaches into four categories. With the foundations of deep transfer learning, we equip the reader with a working knowledge of the main principles and intuitions.
- We analyze the recent literature and provide a comprehensive overview of the current state of the art of deep transfer learning approaches for time series anomaly detection for main industrial applications.
- We discuss potential challenges and limitations and then give directions for future work with actionable recommendations for AI practitioners and decision-makers.

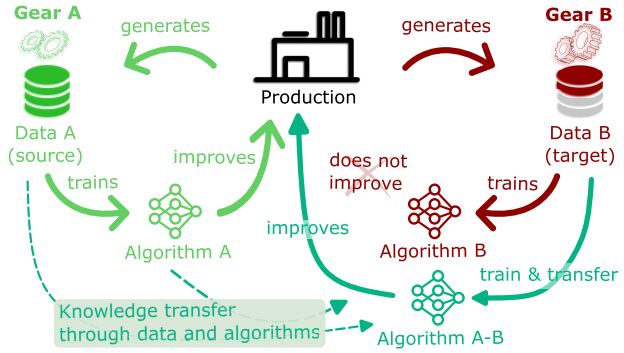

FIGURE 1. Transfer learning is useful when changes in production take place and sufficient data for full retraining is not available as shown here for a hypothetical production of two types of gears. In the production of gear A, a lot of data is available to train a deep learning model that helps improve production. In the production of gear B, data is more limited, and the traditionally trained deep learning model fails to improve production. With suitable transfer learning methods, however, data and algorithms acquired during the production of gear A can be leveraged to support improving the production of gear B because the data and tasks in the production of both gears are related.

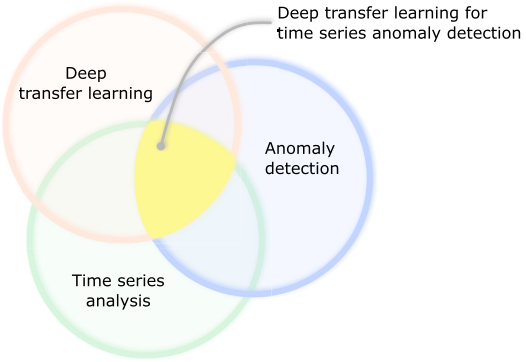

FIGURE 2. Venn diagram of this survey's focus on the intersection of transfer learning, anomaly detection, and time series analysis.

To our knowledge, this is the first survey of deep transfer learning in the narrow context of industrial time series anomaly detection. The review describes the underlying methodological principles and methods within a generic taxonomy and discusses practical implications for AI practitioners to make informed decisions. We cover multiple areas of application, including manufacturing monitoring, maintenance prediction, and infrastructure monitoring.

The rest of the paper is organized as follows. First, we provide an overview of transfer learning by introducing a taxonomy of transfer learning problem settings and further categorizing deep transfer learning approaches (Section II). Then, we describe the task of anomaly detection in time series (Section III) in selected industrial applications (Section IV). To conclude, we discuss current challenges, limitations, and future research directions (Sections V–VI) in the field.

B. SURVEY METHODOLOGY

We seek to identify application-oriented peer-reviewed literature in the intersection of transfer learning as the learning framework, time series as the data domain, and anomaly detection as the task (Fig. 2). To execute the selection process of literature, we search related terms on Google Scholar, Scopus, Elsevier, and IEEE databases. Based on the title, we pick those papers that may fit the narrow topic into a preselection list. Eventually, we included publications matching the topic according to the abstract and screening of the content.

Along the reviewed topical papers, we include contextually relevant papers such as deep learning approaches that are agnostic to data types and tasks. For deep transfer learning in general, we searched the keywords "transfer learning" and "deep transfer learning". Specifically, we focus more on deep transfer learning approaches. Then, we switch to the application-oriented cases where deep transfer learning is applied to tackle time series anomaly detection in the main industrial applications. To achieve this, we search queries like "deep transfer learning for time series anomaly detection" and "deep transfer learning for predictive maintenance". After searching in the database, we carefully check and screen out the most relevant literature based on the following inclusion/exclusion criteria: (1) We only include the applications that utilize deep transfer learning approaches, instead of traditional transfer learning; (2) We only include publications after 2013; (3) We cover all three main topics in Fig. 2 (highlighted with cycles), but we specifically focus on the intersection of the three aforementioned domains. After carefully screening out, we select 45 papers for deep transfer learning in general and 37 papers for deep transfer learning for anomaly detection in industrial time series.

Fig. 3 illustrates the taxonomy in this survey to categorize reviewed studies based on different aspects, including deep transfer learning, time series anomaly detection, industrial applications, current challenges, and future directions.

II. DEEP TRANSFER LEARNING

A. OVERVIEW OF THE FIELD

Transfer learning in the setting of industrial time series analysis for anomaly detection is a tool to increase the flexibility of autonomous process monitoring. It addresses the challenge of adapting the algorithm, and thus the decision process, to a related but previously unseen setting where limited training data is available. The transfer eliminates the need to train a deep learning model from scratch, which in turn reduces the amount of necessary data and compute required to solve a new task or adjust to a new data domain. In either case, knowledge is transferred from a source to a target domain, as described below. The transfer learning problem settings can be categorized as inductive or transductive transfer depending on the data and task conditions. We categorize deep learning-based transfer learning approaches into instance transfer, parameter transfer, mapping transfer and domain-adversarial transfer. We illustrate them by using two intuitive examples in

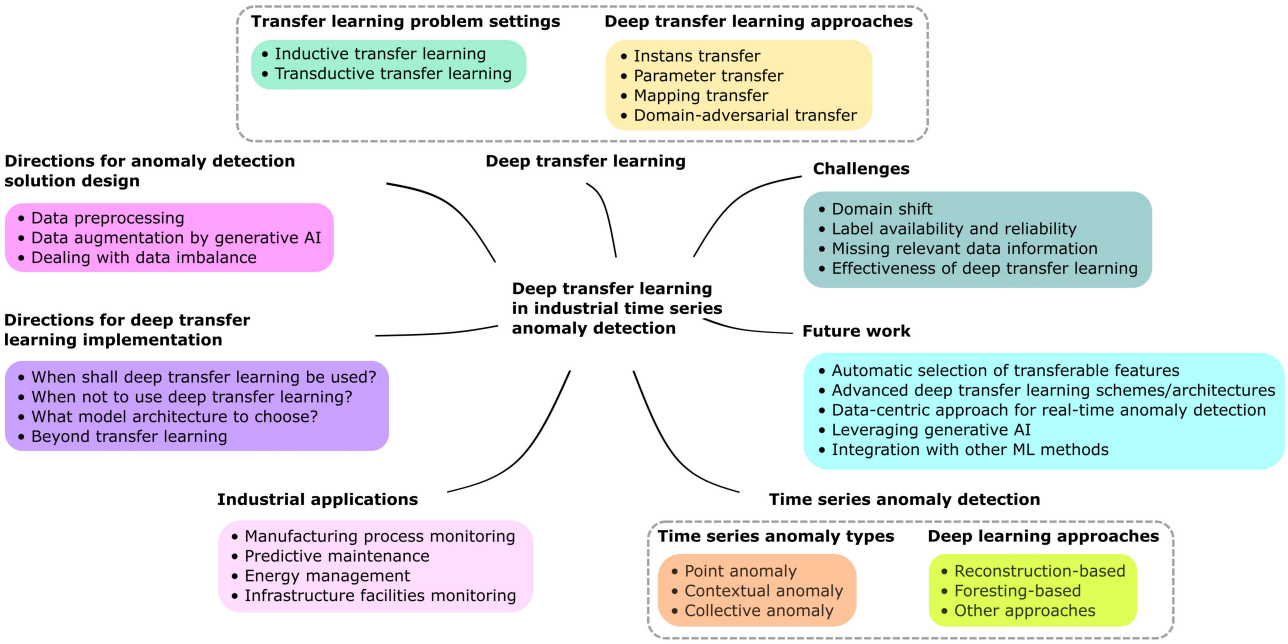

FIGURE 3. A generic taxonomy in this paper to analyze deep transfer learning for industrial time series anomaly detection.

Fig. 4, with more details being elaborated in the following sections.

B. FORMAL DESCRIPTION OF DEEP TRANSFER LEARNING

Domain $\mathcal{D}$ includes the domain feature space $\mathcal{X}$ and marginal data distribution $P(X)$ as $\mathcal{D} = \{\mathcal{X}, P(X)\}$, where X is the domain data, $X = \{x_1, \ldots, x_n\} \in \mathcal{X}$. Similarly, a learning task is defined as $\mathcal{T} = \{\mathcal{Y}, f_{\mathcal{T}}(\cdot)\}$, where $\mathcal{Y}$ denotes the task space and usually represents class label. For anomaly detection tasks, $\mathcal{Y}$ is the set of the two classes "normal" and "abnormal". The function $f_{\mathcal{T}}(\cdot)$ can be used to predict the corresponding label of a new instance x_i. The objective predictive function $f_{\mathcal{T}}(\cdot)$ learned from domain data can be interpreted as a form of conditional probability. Thus, the learning task can be rewritten as $\mathcal{T} = \{\mathcal{Y}, P(Y|X)\}$, where $P(Y|X)$ is used as a likelihood measure to determine how well a given data set X fits with a corresponding class label set Y.

In the surveyed literature on transfer learning for anomaly detection in industrial applications, transfer learning methods from other fields, such as computer vision and natural language processing, were adopted. We therefore use a generic classification scheme for transfer learning methods. We largely follow the definition of transfer learning in literature [27], [28]. Given a source domain $\mathcal{D}_S$ and learning task $\mathcal{T}_S$, as well as a target domain $\mathcal{D}_T$ and learning task $\mathcal{T}_T$, transfer learning aims to improve the performance of the predictive function $f_{\mathcal{T}}(\cdot)$ in $\mathcal{D}_T$ by transferring knowledge from $\mathcal{D}_S$ and $\mathcal{T}_S$, where $\mathcal{D}_S \neq \mathcal{D}_T$ and/or $\mathcal{T}_S \neq \mathcal{T}_T$. Usually, the size of source dataset is much smaller than target dataset.

This definition of transfer learning can be broadened, i.e., the target task can benefit from multiple source domains. Transfer learning is thus the idea of making the best use of related source domains to solve new tasks. In contrast, traditional machine learning (ML) methods learn each task separately from scratch, and each respective model can only be applied to the corresponding task.

We define a taxonomy of transfer learning problem settings as shown in Fig. 4 mainly depending on the label availability in the two domains to be easily applicable to the requirements of a case at hand (compare different definitions for other purposes in the literatures [27], [28], [29]).

We differentiate it into inductive and transductive transfer learning [28]. Inductive transfer learning is applied when the target task is different from the source task, i.e., $\mathcal{T}_S \neq \mathcal{T}_T$ (meaning that $\{\mathcal{Y}_S \neq \mathcal{Y}_T\}$ or $\{P(Y_S|X_S) \neq P(Y_T|X_T)\}$). The conditional probability distribution is induced with labeled training data in the target domain [34]. A corresponding example is illustrated as Scenario A in Fig. 4, where the learning tasks are different and the goal of transfer learning is to recognize point anomaly from the collective anomaly task. Related areas of inductive transfer learning are multi-task learning [35], [36] and sequential learning, depending on whether tasks are learned simultaneously or sequentially.

Transductive transfer learning is applied when the source and target tasks are the same, while the source and target domain are different, i.e., $\mathcal{T}_S = \mathcal{T}_T$ and $\mathcal{D}_S \neq \mathcal{D}_T$ (meaning that $\{\mathcal{X}_S \neq \mathcal{X}_T\}$ or $\{P(X_S) \neq P(X_T)\}$. A subcategory is domain adaptation [37] when the feature space of source and target data are the same but the corresponding marginal distributions are different (i.e., $\{\mathcal{X}_S = \mathcal{X}_T\}$ and $\{P(X_S) \neq P(X_T)\}$). Scenario B in Fig. 4 is an example of transductive transfer learning where the learning tasks are identical,

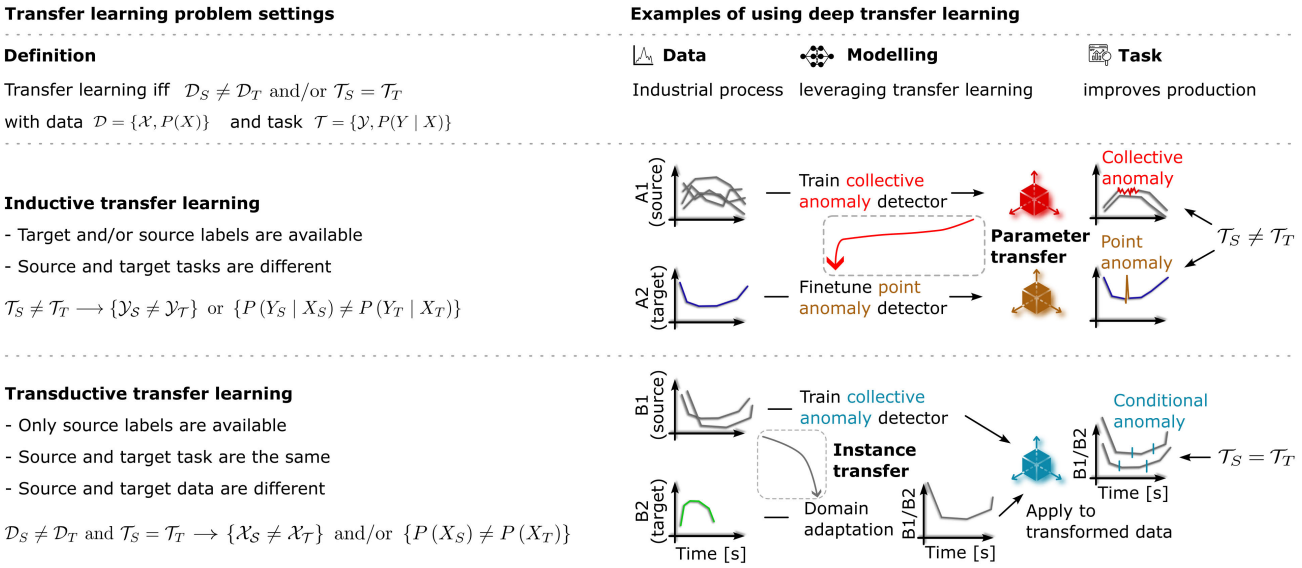

FIGURE 4. Taxonomy of transfer learning problem settings (left; see Section II-B for the definition of terms) and corresponding examples using deep transfer learning approaches (right). On the left, we classify transfer learning problems as inductive or transductive transfer settings Correspondingly, we provide two examples using deep transfer learning methods: In the inductive transfer setting, we collect time series data from screw production and wrench production. Labeled screw data (A1) is used to detect collective anomalies (a set of data points behaving differently compared to the entire time series [32], [33], further explained in Section III). Then, parameter transfer (Section II-C2) is applied to transfer knowledge by fine-tuning the pre-trained model from labeled screw data to detect point anomalies (further explained in Section III) on labeled wrench data (A2). For the transductive transfer setting in the lower panel, we present a different situation for contextual anomaly detection (further explained in Section III). In this case, we have two datasets, B1 and B2, analyzed using the same model. However, the data in B2 significantly differs in appearance from the data in B1. To address this problem, instance transfer (further explained in Section II-C1) is used. Through this learning process, the data in B2 is transformed in a way that makes it compatible with the model that has been trained exclusively on data from B1. Transfer learning, in this case, is thus achieved by adapting the data to fit the model through domain adaptation rather than adjusting the model to fit new data.

TABLE 1. Overview of deep transfer learning approaches with references.

Deep transfer learning approach	Short description	References
Instance transfer	Augmenting target data by transforming data instance from the source domain to the target domain	[42]–[44]
Parameter transfer	Transfering learned parameters of a pre-trained model from source domain and adapting the model for target domain	[17], [45]–[53]
Mapping transfer	Reducing feature discrepancies between source and target domains by minimizing the distance between mapped features in the latent space	[27], [54]–[59]
Domain-adversarial transfer	Extracting an indiscriminative feature representation between source and target domain through adversarial training	[60]–[65]

and the goal of transfer learning is to recognize contextual anomalies in an unlabelled data set.

C. DEEP TRANSFER LEARNING APPROACHES

Since deep neural networks (DNNs) can learn useful feature representations from large amounts of data through back-propagation [24], they have been widely adopted for tackling complex problems in practice [38], [39], [40], [41], which involve large-scale and high-dimensional data. Deep transfer learning methods implement transfer learning principles within DNN and, among other things, enable deep learning based analysis pipelines to be applied to new datasets.

Based on the transferring techniques in the surveyed literature, we access how knowledge is shared across domains and help increase the performance in the target task or domain. we divide deep transfer learning approaches further into 4 categories: instance transfer, parameter transfer, mapping transfer, and domain-adversarial transfer, as illustrated in Table 1. Furthermore, instance transfer, mapping transfer, and domain-adversarial transfer can be described as data-driven approaches. They focus on transferring knowledge by leveraging a large amount of data. It usually involves transforming and adjusting the data instances or manipulating data from different domains by feature alignment, feature mapping, etc. On the other hand, parameter transfer is

a model-driven approach, which places more emphasis on understanding the underlying structure and dynamics of the data. It usually involves transferring the parameters of pre-trained model from source domain to target domain.

1) INSTANCE TRANSFER

The intuition of instance transfer is that although source and target domains differ, it is still possible to transform and reuse source data together with a few labeled target samples. A typical approach is to re-create some labeled data from the source domain. For example, He et al. propose an instance-based deep transfer learning model with an attention mechanism to predict stock movement [42]. They first create new samples from the source dataset that are similar to the target samples by using an attention network and then train another network on the created samples and target training samples for prediction tasks. Since two networks are trained separately for different tasks, it needs further investigation to what extent the generated samples can contribute to the prediction task. Amirain et al. introduce an innovative instance transfer method for domain adaptation [43]. They propose an effective auto-encoder model with a pseudo-label classifier to reconstruct new data instances that obtain general features across different datasets for medical image analysis. Taking another avenue, Wang et al. exclude the source data that negatively impacts training target data. Specifically, they choose a pre-trained model from a source domain, estimate the impact of all training samples in the target domain, and remove samples that lower the model's performance. Then, the optimized training data is used for fine-tuning. The experiments are conducted on large image datasets [44]. Instead of transferring the data, the approach excludes certain samples based on the pre-trained model's predictions. Additional validation is required in industrial environments, especially when only a few data are available in some industrial settings.

2) PARAMETER TRANSFER

Parameter transfer adapts the learned parameters of a pre-trained model to a new model. This assumes that DNNs can get similar feature representations from similar domains. Thus, through transferring parts of the DNN layers together with pre-trained parameters and/or hyperparameters, the pre-trained model is used as a base model to further train on target domain data and solve different learning tasks. Particularly, parameter transfer has gained popularity in computer vision and natural language processing, where large models are pre-trained on large datasets [45]. In natural language processing, for example, BERT [46] and GPT-3 [53] are based on the Transformer architecture [48] which can be fine-tuned for a variety of downstream tasks, including content generation [49], language translation [66], question answering [67], and summarization [50]. In computer vision, Yosinski et al. investigate the general transferability of CNNs in image recognition [30]. They analyze the transferring effect by fine-tuning or freezing a certain amount of layers in the networks. Experimental results show that transferring features from source to target domain improves network generalization compared to those trained solely on the target dataset. Additionally, they quantify the model performance by assessing how features at what layers transfer from one task to another. It is surprising to find that transferring a pre-trained network from any number of layers can produce a boost for fine-tuning on a new dataset. However, the experiments are only conducted on certain image datasets, and Tuggener et al. show [68] the limits of parameter transfer when the chosen architecture is overfitted on the particularities of certain large-scale datasets.

Unlike the typical way of fine-tuning a pre-trained model, Guo et al. propose an adaptive fine-tuning approach SpotTune to find the optimal fine-tuning strategy for the target task [51]. Specifically, a policy network is used to make routing decisions on whether to pass the target instance through the pre-trained model. The results show SpotTune is effective in most cases by using a hybrid of parameter and instance transfer. Sager et al. propose an unsupervised domain adaptation for vertebrae detection in 3D CT volumes by transferring knowledge across domains during the training process [52].

3) MAPPING TRANSFER

Mapping transfer refers to learning a related feature representation for the target domain by feature transformation, which includes feature alignment, feature mapping, and feature encoding [27]. The goal is to reduce feature discrepancies between source and target domains by minimizing the distance between the distribution of latent feature representation. There are various criteria to measure the distribution difference, including Wasserstein distance [69], Kullback-Leibler Divergence [70], etc. Among them, Maximum Mean Discrepancy (MMD) [55] is most frequently adopted in mapping transfer from the surveyed papers. The MMD is calculated as the difference between the mean embeddings of the samples in a reproducing kernel Hilbert space associated with a chosen kernel function. Added to the target loss function, it serves as a powerful tool for comparing the similarity of complex, high-dimensional datasets using a wide variety of kernel functions.

Previous work has focused on transferred feature extraction/dimensionality reduction using MMD. Wang et al. focus more on the subdomain of the same subcategory instead of the alignment of the global distribution between source and target domain [54]. Specifically, they first use the attention mechanism to extract discriminative features that are most related to the fault signal. Then, local MMD is applied to transfer knowledge to adjust the distribution of related subdomains under the same category. Long et al. propose their Joint Adaptation Network [56] based on MMD, in which the joint distributions of multiple domain-specific layers across domains are aligned. In addition, an adversarial training

version was adopted to make distributions of the source and target domains more distinguishable. Similarly, Long et al. adopt multi-layer adaptation and proposed Deep Adaptation Networks (DAN) [57]. The first three convolutional layers are used in DAN models to extract general features. For the last three layers, multi-kernel MMD bridges the cross-domain discrepancy and learns transferable features. Zhang et al. propose a Deep Transfer Network in which two types of layers are used to obtain domain invariant features across domains by adding MMD loss. The shared feature extraction layers learn a shared feature subspace between the source and the target samples, and the discrimination layer is then used to match conditional distributions by classifier transduction [58]. Venkateswara et al. propose Deep Adaptation Hash network [59], which is fine-tuned from the VGG-F [71] network. Multi-kernel MMD loss is employed to train the Deep Adaptation Hash network to learn feature representations that align the source and target domains.

4) DOMAIN-ADVERSARIAL TRANSFER

Inspired by Generative Adversarial Networks (GANs) [72], [73], the goal of domain-adversarial transfer is to extract a transferable feature representation that is indiscriminative between source and target domain through adversarial training. Adversarial transfer is primarily concerned with addressing domain adaptation problems.

Soleimani and Nazerfard utilize the GANs framework to perform cross-subject transfer learning [60]. The generator is used to generate samples that are similar to the target data. Meanwhile, the discriminator distinguishes the fake samples from the target samples. The classifier is trained to discriminate the labeled source data and fake samples to learn generalized features invariant to source and target domains. It is important to note that in real-world applications, training GANs can be unstable due to mode collapse, especially in the case when the source data and target data are unbalanced, and the generator may fail to generate fake samples that can confuse the discriminator. Tzeng et al. adopt a domain confusion loss across the source and target domains to learn a domain invariant representation [61]. Ganin et al. propose a new domain adaptation architecture by adding a domain classifier after feature extraction layers [63]. A gradient reversal layer is used to ensure the similarity of the feature distributions over source and target domains. Similarly, Ozyurt et al. develop a novel framework for unsupervised domain adaptation of time series data by using contrastive learning and domain-adversarial transfer learning [62]. A domain classification loss is applied to extract domain invariant features. The drawback is that the experiments are designed in a way that the source and target data sizes are similar, whereas in practice, the target data is usually much fewer than the source data. Ajakan et al. propose a domain adversarial DNN in which a domain regressor is applied to learn a domain invariant feature representation [64]. Tzeng et al. use an unsupervised domain adaptation method that combines adversarial learning with discriminative feature learning [65].

D. RELATED LEARNING PARADIGMS

Besides the dedicated transfer learning approaches discussed above, there are methods that represent alternative ways to solve tasks across domains or are complementary to the native transfer learning methods.

- *Multi-task learning* is a machine learning technique where a single model is trained on multiple tasks simultaneously. The idea is to improve the performance of the model by learning a shared representation that captures the features between all tasks. Because the network learns to solve multiple tasks, it may generalize better to new data and tasks.
- *Continuous learning* [74] is a learning process where the model continuously learns new tasks from previous tasks over time without forgetting how to solve previous tasks. To some extent, continuous learning can be seen as a sequential transfer learning process, with the constraint to preserve the performance of the previous tasks, which leads to an accumulation of knowledge over time.
- *Few-shot learning* [75] is a type of machine learning where a model can learn and perform well on a new task with only a limited number of labeled samples. In extreme cases, the model can learn with one label [76] and without any label [77]. Whereas, transfer learning usually involves reusing the model from relevant tasks and continuing training on the target dataset.
- *Domain generalization* [78], [79] focuses on developing a generalized model from one or multiple distinct domains to detect unseen target domain data. The main goal is to overcome the domain shift problem. Domain generalization and transfer learning are both applied to transfer knowledge from source domain to target domain. The major difference between transfer learning and domain generalization lies in the utilization of target domain data. Transfer learning leverages knowledge from source domain and target domain. In contrast, domain generalization solely learns from source domain, without access to the target data.
- *Meta-learning* [80], [81] is known as "learning to learn". For meta-learning, models are trained on a different set of tasks instead of a set of data in the traditional machine learning setting. In this sense, meta-learning can be seen as a form of transfer learning because it involves transferring knowledge from task to task.
- *Knowledge distillation* [82] effectively learns a small model trained to mimic the behavior of a larger, more complex model. The knowledge learned by the larger model can be transferred to the smaller model, which can then be used for the target task.

- *Self-supervised learning* [83], [84] involves training a model to predict some aspect of the input data without any external supervision. The learned representations can be used for various downstream tasks, including those that involve transferring knowledge from one domain to another.

III. TIME SERIES ANOMALY DETECTION IN INDUSTRY

Time series anomaly detection encompasses statistical techniques to analyze and interpret sequential temporal data. In the context of industrial processes, time series anomaly detection plays a crucial role in automating monitoring, effectively scheduling maintenance, and controlling the efficiency, quality, and performance of these processes. For example, after the detection of an anomaly, another model that captures the relationship between time course and different failure modes or drifts may be exploited for predictive maintenance. For example, in injection molding process monitoring, anomaly detection models are used to analyze recorded sensor data from injection molding machines to detect bad parts and identify the root cause of anomalies [85]. There are two basic ways to detect anomalies: for supervised anomaly detection, labels (normal/abnormal) are needed per time series to build a binary classifier [86]. For unsupervised anomaly detection, an anomaly score or confidence value that is conditioned purely on normal data can be used to differentiate abnormal from normal instances [87], [88].

A. ANOMALY TYPES

According to the literature [89], an outlier is an observation that deviates significantly from other observations in a way that it is likely that it was generated by a different mechanism. In this survey, we focus on time series data collected from machine sensor readings in the context of industrial applications, either univariate (only one variable is recorded over time) or multivariate (several simultaneously recorded measurements). Time series anomalies might occur for various reasons, including internal factors (e.g., temporary sensor error, machinery malfunction) and external factors (e.g. human error, ambient temperature). They can be divided into three categories [32], [33]: point anomalies, contextual anomalies, and collective anomalies. Point anomalies are isolated samples that deviate significantly from the normal behavior of that time series, which can be seen on the left of Fig. 5, e.g., a sudden spike in a pressure reading from a manufacturing machine sensor. These point anomalies can be caused by temporal sensor error, human error, or abnormal machinery operations. Contextual anomalies represent data points that deviate from normal ones only in their current context, and an example can be seen in the middle of Fig. 5. Collective anomalies are a set of data points that in their entirety (but not individually) are abnormal with respect to the entire time series, as shown on the right of Fig. 5.

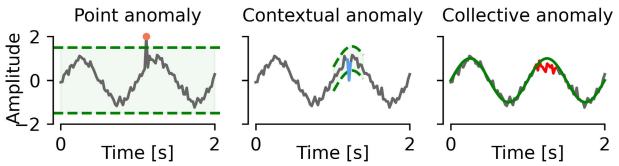

FIGURE 5. Three time series anomaly types. Gray lines represent recorded time series signals, and dashed green lines are *a priori* set thresholds of normal operations. The red dots and the red line indicate anomalies. Point anomalies are single values that fall outside of a pre-set range (left panel). Contextual anomalies are samples that deviate from the current context (middle panel). Collective anomalies are defined as a series of data points that all fall within the range of operation but jointly are not expected (right panel).

B. CHALLENGES

Challenges regarding detecting time series anomalies persist due to two specific properties: (1) The *complexity of time series data*. As the automation level of industrial processes and the complexity of industrial systems increases, univariate time series data become insufficient and inefficient in representing any industrial process in its entirety. Hence, more sensors are installed to monitor the whole process, making it necessary to detect anomalies from multivariate time series, which poses particular challenges since it requires consideration of temporal dependencies and relationships between variables and modalities. Many researchers work on discovering generalized patterns from spatial and temporal correlated multivariate time series data. Zhang et al. propose a Deep Convolutional Autoencoding Memory network [87], where they build an autoencoder to capture spatial dependency of multi-variant data using MMD to distinguish noisy, normal and abnormal data. Zhu et al. propose an interpretable model agnostic multivariate time-series anomaly detection method for applications of cyber physical systems [90]. The new method considers both the temporal and feature dimensions through an adaptive mask based series saliency module to produce accurate anomaly detection results and reasonable interpretations in the form of a mask matrix. (2) The *dynamic variability in industrial processes*. Industrial processes often have high dynamic variability and can be affected by a wide range of conditions, such as changes in temperature, pressure, and humidity. These conditions can cause fluctuations in the process outputs, which leads to data shift and domain shift. This can make it challenging to detect anomalies and maintain control over the industrial process.

C. ANOMALY DETECTION METHODS

Time series anomaly detection has been investigated for decades, and various types of methods have been proposed [91]. This paper exclusively discusses the time series anomaly detection techniques using deep learning, leveraging its robust representation learning capabilities. Current deep learning methods can be mainly divided into reconstruction-based, forecasting-based, and other methods. Fig. 6 illustrates the two main methods. In deep reconstruction-based anomaly

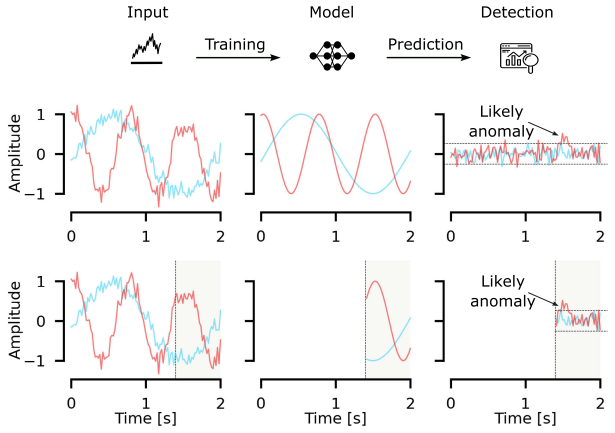

FIGURE 6. Illustration of deep learning-based anomaly detection (top row) with reconstruction-based (center row) and forecasting-based (bottom row) anomaly detection in time series. The first column represents two time series. The second column shows the reconstructed (top) and forecasted (bottom) time series. The third column shows the difference between reconstructed/forecasted time series. Deviations from the reconstructed or forecasted time series are indicative of an anomaly. In deep reconstruction-based anomaly detection, the entire sequence is reconstructed in a decoder-encoder architecture, and the reconstructed sequence is used to compare with the actual sequence. In deep forecasting-based anomaly detection, the end of a sequence is predicted using the start of the sequence and only the forecasted sequence is used to assess the similarity to the ground truth. In this example, the red time series has a likely anomaly at about 1.5 seconds. (Best viewed in color.)

detection, the reconstructed sequence is used to compare with the actual sequence. Differently, in deep forecasting-based anomaly detection, only the forecasted sequence is used to assess the similarity to the ground truth.

1) RECONSTRUCTION-BASED METHODS

Reconstruction-based methods aim to learn the data distribution of the normal time series and differentiate the abnormalities from the normal ones by computing the reconstruction errors. Audibert et al. propose a fast and stable method – Unsupervised anomaly detection for multivariate time series [92], based on adversely trained autoencoders. The encoder-decoder architecture within an adversarial training framework combines the advantages of autoencoders and adversarial training while compensating for the limitations of each technique. After training two autoencoders, the anomaly score is defined by balancing the reconstructed errors from the two autoencoders with two hyperparameters. However, the challenge arises in selecting these two hyperparameters of the anomaly score when the testing dataset is unavailable.

Malhotra et al. also formulate an anomaly score based on reconstruction error [93]. They first train the LSTM encoder-decoder model to reconstruct the normal time series. Subsequently, they leverage the reconstruction errors to calculate the probability by using Maximum Likelihood Estimation to detect a specific point within a time series as an anomaly. They set a window to detect anomalies. The window will be labeled as anomalous if the probability exceeds a threshold. Similarly, Wei et al. also propose an LSTM-based encoder-decoder model to detect multivariant time series sequences based on the reconstruction error [94]. The major difference is the anomaly detection criteria. They assume the reconstruction error of train/test data follows the normal distribution and detect anomalies by using the 2-sigma rule of the normal distribution as a threshold. Zeng et al. propose an adversarial transformer structure to detect multivariate time series anomalies effectively [95]. Here, two-stage adversarial training is applied for the transformer. In the first stage, two transformers are trained by minimizing the reconstruction error to capture the temporal trends in the time series. In the second stage, the reconstruction error serves as prior knowledge in the adversarial training process, enabling the model to distinguish anomalies from normal time series. Then, an anomaly score is defined by combining the anomaly probability and reconstruction error. Again, a threshold has been chosen to differentiate anomalies from normal ones.

GANs, as effective unsupervised learning methods, have been used in time series anomaly detection. Anomaly detection methods based on GANs focus on extracting features by adversarial training on normal samples. Consequently, features from the abnormal samples diverge from those of the normal ones, reflecting in reconstruction error and discrimination value. Li et al. use LSTM-RNN as a base model for building generator and discriminator in GAN [96]. The proposed framework considers multiple variables to capture the temporal correlation of multi-time series distributions. Additionally, they proposed a novel anomaly score, which can detect anomalies through discrimination and reconstruction. More specifically, the score is a combination of the reconstruction difference between generated data and original data and the discrimination results from the discriminator. Similarly, Niu et al. and Bashar et al. both propose an LSTM-based VAE-GAN for time series anomaly detection, where LSTM networks are used as the generator, and discriminator [97], [98]. When it comes to anomaly scores, setting an optimal threshold is usually a critical step. However, using a small portion of the test set to decide the optimal threshold may not be practical in real-world scenarios [97]. Additionally, it is important to note that the method has been only tested for point anomaly detection, further investigation is required when they are applied to detect other anomaly types.

2) FORECASTING-BASED METHODS

Forecasting-based methods predict the value of the following timestamps and predict temporal anomalies according to the prediction error. Kim et al. propose a forecasting-based unsupervised time-series anomaly detection method using transformer architecture [99]. The idea is to train a transformer-like model by forecasting a fixed-length time series based on the previous timestamps. The trained model is used to predict time series with an anomaly score such that an instance where the anomaly score is larger than a static

threshold is defined as an anomaly. A dynamic thresholding technique is also mentioned but not explicitly discussed in the paper.

Deng and Hooi propose a novel attention-based graph neural network approach [100] that learns a graph of dependence relationships between multi-variant time series signals by forecasting the behavior based on past time series. Then, a graph deviation scoring is defined for each sensor to detect and explain anomalies. Tang et al. propose an interpretable multivariate time series anomaly detection method based on graph neural networks and gated recurrent units [101]. The feature representation is learned through forecasting the future time series segment. An abnormal score is set for each time series to detect anomalies. The feature embedding is then used for 2D visualization through t-SNE plots to interpret the clusters within time series from different sensors.

3) OTHER METHODS

Ding et al. propose a joint network to integrate the advantages of reconstruction and forecasting/prediction [102]. First, they propose a multimodal graph attention network to tackle the spatial-temporal dependencies for multimodal time series. Further, they optimize the reconstruction and prediction modules simultaneously to predict anomalies. Himeur et al. take advantage of annotated data and directly use a DNN as a classifier to classify normal and abnormal energy consumption types [103]. The enormous imbalance of real anomaly patterns is one concern in the approach. Thus, a normalized technique of power consumption data is applied to deal with this problem. The normalized data represent the difference in power consumption rates of each current time sample and the previous one. It can provide information on how fast the consumption reacts to the time evolution. However, any further evaluation of this technique is not discussed, and it is still an open problem regarding anomaly detection for other datasets. Yang et al. propose a contrastive learning structure with dual attention to learn a permutation invariant representation of the data with superior discrimination characteristics between normal points and anomalies [104]. Unlike most reconstruction-based models, their model is a self-supervised framework based on representation learning. The new method achieves state-of-the-art comparable performance on six multivariate and one univariate time series anomaly detection benchmark datasets. However, the extensive framework with two multi-head-attention blocks may be prone to overfitting. This concern is amplified by the absence of training details, leaving only evaluation details disclosed.

In principle, these anomaly detection approaches are applicable to all types of anomalies. Reconstruction-based methods are typically applied to the entire or a portion of the time series. Long-time series are commonly segmented into subsequences using a predefined sliding window. In the case of detecting context/collective anomalies, the reconstructed loss of the time series sequence is evaluated within the predefined sliding window, if the reconstruction loss is larger than an acceptable threshold, then that time series sequence is classified as an anomaly. In the case of point anomalies, reconstruction is performed at each single time stamp, akin to a regression problem, and then the reconstruction loss of each single timestamp is evaluated to determine whether the single timestamp is anomalous or not. It is also applied to forecasting-based anomaly detection methods, instead of computing reconstruction error, forecasting-based methods predict the value in the next time stamp for point anomalies or the next time series sequence for context/collective anomalies. The anomalies will be detected based on the deviation between the predicted value and the normal value. Other anomaly detection approaches usually combine the reconstruction-based and foresting-based methods.

To sum up, these methods are applicable to each type of anomaly. However, the effectiveness of these anomaly detection approaches may vary depending on the anomaly detection tasks at hand, which are characterized by the granularity at which the time series data is observed and analyzed.

IV. INDUSTRIAL APPLICATIONS

A. OVERVIEW

Deep transfer learning techniques have gained prominence in computer vision and natural language processing, primarily due to the abundance of available datasets. However, their adoption in the context of industrial time series data has been comparatively limited. This hesitancy can be attributed to the limited public availability of such datasets and the unique domain-specific characteristics they possess, which complicate generalized advancements. Encouragingly, there has been a recent uptick in the application of deep transfer learning for anomaly detection within the industry such as fault diagnosis [105], quality management [106], manufacturing process monitoring [85], network/software security [107], and infrastructure monitoring [108]. These can be mapped onto the core industrial domains of manufacturing process and infrastructure monitoring, predictive maintenance, and energy management. Table 2 presents a compact comparison of the related works using deep transfer learning approaches to solve these tasks.

Fig. 7 illustrates the Sankey diagram of the connections between industrial applications and the deep transfer learning approaches based on our literature survey. The diagram shows every path that connects the four dimensions of the methodology-problem-landscape within the surveyed literature. The broader the path is, the more papers are related to the linked topics. The goal is to give an overview of how deep transfer learning is applied to industrial problems in the recent literature and specifically show with these four dimensions: (1) which deep transfer learning approaches are actually used in practice; (2) what the main industrial domains for time series anomaly detection are; (3) what deep transfer learning category these domains belong to; (4) what labels are available in source and target domain.

TABLE 2. A compact overview of industrial applications that used deep transfer learning for time series anomaly detection.

Reference	Industrial task	Industrial domain	Deep transfer learning approach	Transfer learning problem setting	Deep learning framework	Source type	Source label	Target label
[11]	Industrial metal forming anomaly detection	Predictive maintenance	Parameter transfer	Inductive	CNN	Multiple	✓	✓
[12]	Monitoring systems Anomaly detection	Predictive maintenance	Parameter transfer	Inductive	U-Net	Multiple	✓	✓
[13]	Car body-side production line fault diagnosis	Predictive maintenance	Parameter transfer	Inductive	SAE	Multiple	✓	✓
[14]	Rotation bearings fault detection	Predictive maintenance	Mapping transfer	Transductive	Auto-encoder	Multiple	✓	✗
[15]	Industrial control systems anomaly detection	Predictive maintenance	Parameter transfer	Inductive	ResNet8	Single	✓	✓
[16]	Service elevator fault detection	Predictive maintenance	Parameter transfer	Inductive	CNN, RNN	Multiple	✓	✓
[19]	Nuclear power plants fault detection	Predictive maintenance	Parameter transfer	Inductive	CNN	Multiple	✓	✓
[109]	Building energy systems fault diagnosis	Predictive maintenance	Parameter transfer	Inductive	CNN	Multiple	✓	✓
[110]	Press machine production prediction	Predictive maintenance	Parameter transfer	Inductive	CNN	Single	✓	✓
[111]	Wind turbine fault detection	Predictive maintenance	Parameter transfer	Inductive	CNN	Single	✓	✓
[112]	Industrial machine operating fault detection	Predictive maintenance	Parameter transfer	Inductive	CNN, LSTM	Single	✓	✓
[9]	Machine turning operations classification	Manufacturing process monitoring	Parameter transfer	Inductive	VGG, ResNet	Single	✓	✓
[10]	Injection molding process quality control	Manufacturing process monitoring	Parameter transfer	Inductive	FCN	Multiple	✓	✓
[85]	Injection molding process quality control	Manufacturing process monitoring	Parameter transfer	Inductive	FCN	Single	✓	✓
[113]	Injection molding process anomaly detection	Manufacturing process monitoring	Parameter transfer	Inductive	FCN	Multiple	✓	✓
[114]	Injection molding process anomaly detection	Manufacturing process monitoring	Parameter transfer	Inductive	FCN	Multiple	✓	✓
[115]	Aluminum gravity die casting quality prediction	Manufacturing process monitoring	Parameter transfer	Inductive	FCN	Single	✓	✓
[20]	Agriculture/manufacturing systems anomaly detection	Manufacturing process monitoring	Parameter transfer	Inductive	LSTM	Single	✓	✓
[116]	Manufacturing testbeds anomaly detection	Manufacturing process monitoring	Parameter transfer	Inductive	LSTM, RNN	Single	✓	✓
[117]	Industrial metal (pump) forming anomaly detection	Manufacturing process monitoring	Parameter transfer	Inductive	LSTM	Single	✓	✓
[118]	Industrial control systems anomaly detection	Manufacturing process monitoring	Parameter transfer	Inductive	Auto-encoder	Single	✓	✓
[21]	Petrochemical production process anomaly detection	Energy saving	Parameter transfer	Inductive	LSTM, CNN	Single	✓	✓
[119]	Electricity consumption anomaly detection	Energy saving	Parameter transfer	Inductive	FCN	Single	✗	✓
[120]	Building's energy consumption anomaly detection	Energy saving	Parameter transfer	Inductive	AlexNet-40	Single	✓	✓
[121]	Power consumption anomaly detection	Energy saving	Mapping transfer	Transductive	DAN	Single	✓	✗
[122]	Building's power consumption anomaly detection	Energy saving	Parameter transfer	Inductive	LSTM	Single	✓	✓
[23]	Aircraft flight anomaly detection	Infrastructure facilities monitoring	Parameter transfer	Inductive	LSTM	Single	✓	✓
[108]	Anomaly identification for bridge groups	Infrastructure facilities monitoring	Parameter transfer	Inductive	CNN	Single	✓	✓
[22]	Network intrusion detection	Infrastructure facilities monitoring	Parameter transfer	Inductive	CNN, LSTM	Single	✓	✓
[17]	Building occupation detection	Infrastructure facilities monitoring	Parameter transfer	Inductive	CNN	Single	✓	✓
[18]	Building occupation detection	Infrastructure facilities monitoring	Parameter transfer	Inductive	CNN	Single	✓	✓

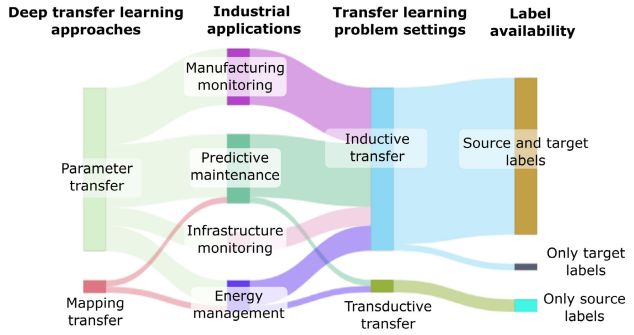

FIGURE 7. Overview of Sankey diagram of transfer learning problem setting, deep transfer learning approach categories, and label availability in the surveyed industrial domains.

Key observations from Fig. 7 are: (1) parameter transfer is much more frequently used than any other deep transfer learning approach across all surveyed industrial applications since fine-tuning a pre-trained model on target data is more straightforward to implement by taking advantage of the pre-trained model on the source dataset and usually without fundamental modification on the model architecture. It is noteworthy that instance transfer and adversarial transfer do not appear in the diagram. Apparently, these two deep transfer learning approaches are not considered effective in time series anomaly detection tasks in industry. The difficulty lies in implementing and training these scarcely researched approaches in the industrial field, as indicated by the findings. (2) Hybrid approaches of parameter and mapping transfer can be seen in predictive maintenance. (3) Most industrial applications use inductive transfer learning, indicating they focus on leveraging labeled source and target data to solve the target task, i.e., use supervised learning.

B. MANUFACTURING PROCESS MONITORING

Manufacturing process monitoring is crucial to ensure high-quality products and low rejection rates. For example, in injection molding machines, sensors are installed to detect molding conditions in the cavities, such as cavity pressure and temperature. These signals are used to analyze particularly the mold filling and solidification process for each produced part. Such cyclic processing data can also be seen in metal machining (cutting force signal) or joining of parts (joining force signal). Currently, parameter transfer is predominantly used for manufacturing processes [10], [11], [20], [85], [113], [114], [116], [117], [123].

Park et al. propose a transfer learning technique to detect time series anomalies for different industrial control systems [118]. First, they apply principal components analysis to reduce the dimension of source and target data. A DNN model is then trained on the compressed source data, and after a reasonable mapping algorithm is adopted to map the features of source to target domain, the pre-trained model is further trained on the target data. The model achieves good performance even when a model is retrained with only a proportion of target data. For the experiments, they only test on two comparatively larger datasets and fail to show that the transferred model performs better than the one without retraining for one dataset. Further investigation needs to explain the negative transfer. Additionally, even when they only take a small proportion of target data for transfer learning purposes, the sample size still exceeds 5000, which exceeds most industrial applications. Abdallah et al. apply parameter transfer to monitor the operation status of manufacturing testbeds with vibration sensor data [20], [116]. Hsieh et al. transfer knowledge across three chambers in a production line to detect anomalous time series data [123]. Results show reduced training time and improved detection accuracy through transfer learning. In injection molding, parameter transfer is applied to transfer the knowledge from one or more source domains to solve tasks in a target domain [10], [113], [114]. Specifically, they employ simple fully connected neural networks and transfer knowledge from one product to another by freezing the first few layers and fine-tuning only the last few layers. Instead of evaluating the time series data directly from sensors, they represent the industrial process by the parameters of the machine settings. However, they can still provide useful insight for the case of the time-series data. Tercan et al. build a bridge between simulated data and real data using parameter transfer in injection molding [85]. Here, a fully connected neural network is trained on simulated data and then partially or fully reused to further train on real data. Results show that the transferred model performs better than a network trained from scratch on real experimental data. In manufacturing processes, a simulation model/process hence can play a critical role, but deeper analysis is needed to further understand and reduce the gap between simulation and real data. Additionally, Lockner et al. explore the impact of transfer learning with varying amounts of source data and assess how performance is influenced by different configurations of frozen layers [114]. Maschler et al. compare different DNNs for anomaly detection tasks on metal forming datasets [11]. Further, they propose a deep transfer learning framework aiming to transfer knowledge between tasks. However, the proposed architecture is not validated by experiments. Later, Maschler et al. apply continuous learning on the same dataset by transferring knowledge from several source tasks to a target task to train a deep learning algorithm capable of solving both source and target tasks [117]. Specifically, they use regularization approaches using altered loss functions to solve related tasks that appear best suited.

C. PREDICTIVE MAINTENANCE

Predictive maintenance aims to predict the necessity of maintenance before production is negatively impacted by a failure. Tasks involve monitoring equipment to anticipate maintenance requirements (i.e., predict probable future failure) to optimize maintenance schedules [124]. Time series anomaly detection is often used in respective systems to identify abnormal behaviors in operation that may indicate the need for maintenance, such as increasing noise, vibrations, etc.

Mao et al. use mapping transfer with a Sparse Auto-Encoder (SAE) for motor vibration anomaly detection [14]. A transformation from the source and target data to a common latent feature space is learned with MMD loss to make the feature distribution of two domains as identical as possible. Similarly, Wen et al. also use mapping transfer with an SAE architecture for fault detection of rotation bearings, using an MMD regularizer to extract a common feature representation [125]. Subsequently, they propose a new MU-Net architecture to detect multivariate time series anomalies [12]. First, they pre-train a U-Net [126] on a large time series dataset for an anomaly detection task. Then, they propose a new model MU-Net, which is built upon U-Net. In MU-Net, each channel can leverage a pre-trained U-Net through fine-tuning to transfer knowledge for multivariate time series anomaly detection.

In a different application, parameter transfer is used to predict the remaining useful life for tools in manufacturing [127]. An SAE network is first trained to predict the remaining useful life of a cutting tool on retrospectively acquired data in an offline process. The trained network is then transferred to production with a new tool for online remaining useful life prediction. The result shows that transfer-learning based hybrid deep learning significantly reduces the training time and is highly suitable for real-time industrial fault diagnosis/prediction in various environments. Similarly, parameter transfer is implemented to reduce the gap between different industrial environments [13], [112]. Xu et al. use a stacked SAE to extract general features from source data and a digital-twin-assisted fault diagnosis approach is presented to transfer knowledge from virtual space to physical space for real-time use [13]. Here, a DNN model is first fully trained in virtual space and then migrated to the physical space using parameter transfer for real-time use.

The above-mentioned literature proves that deep transfer learning is a research field that could simplify the life cycle of predictive maintenance systems and facilitate DNN model reusability by reducing the required data and training time, helping adapt them to solve similar tasks.

D. ENERGY MANAGEMENT

Energy management deals with systems that detect abnormal excessive consumption caused by end-users' unusual behavior or malfunction of faulty devices or systems [120]. The goal is to develop automatic, quick-responding, accurate, and reliable fault detection to save energy and build environmentally friendly systems. Energy anomaly detection systems monitor data during energy generation, transmission, and utilization, to ensure normal energy consumption.

Xu et al. design a cluster-based deep adaptation layer to improve a deep adaptation network, effectively reducing the mismatch in transfer learning of spinning power consumption anomaly detection [121]. The basic architecture consists of five convolutional layers and three fully connected layers. The weight parameters of the convolutional layers are shared between source and target domains. The cluster-based deep adaptation layer is designed across the feature layers of two networks to cluster feature representations of the source and target domains respectively. The proposed method shows superiority over fine-tuning and DAN because the adaptation layer can minimize the distance between the nearest neighbor clusters across the source and target domains to match the most similar distribution of feature representations. It is important to note that the anomalies are defined and tagged by human experts as different types, thus the problem becomes a classification task. However, in real-world industrial applications, it's almost impossible to enumerate unknown anomaly types because of the highly dynamic environment. Liang et al. successfully build an electricity consumption time series anomaly detection method in aluminum extrusion [119]. Parameter transfer is applied to transfer domain knowledge from another data-sufficient domain. First, they train on sufficient extruding machine data in an unsupervised way and then use only a few data samples from different extruding machines to adapt the model by transfer learning. It is important to note that when the target data is already sufficient, transferring knowledge can be detrimental as it can decrease prediction accuracy on the final task. Copiaco et al. aims to detect anomalies for building energy consumption via transfer learning from pre-trained CNN models [128]. First, they convert 1D time series signals to 2D image representations. These serve as inputs for pre-trained vision models to capture inherent spatially invariant features. In the end, a SVM is applied to classify anomaly types. The SVM classifier obtained optimal results when operating upon a pre-trained ALexNet model with normalized grayscale graphical representations. However, a deeper discussion regarding the effect of the different pre-trained models is not presented. Additionally, converting 1D time series to 2D images by creating a matrix representation of the sensor readings may lead to information loss during the transformation process, which should be further investigated.

E. INFRASTRUCTURE FACILITIES MONITORING

Infrastructure facilities monitoring refers to monitoring and maintaining the conditions of infrastructure facilities, such as bridges, buildings [129], and networks. This can include detecting potential problems or failures. The goal is to minimize the impact of failures on the public or the environment. This application commonly uses parameter transfer to transfer knowledge from facility to facility to take advantage of similar data and tasks.

Dhillon et al. present a parameter transfer approach towards building a network intrusion detection system based on CNN and LSTM [22]. Specifically, They extract and learn patterns by mapping the input data into a lower dimensional representation by convolutional layers. Then, they employ the LSTM layer to enhance learning and recognizing patterns across time. In the end, a fully connected layer is used as a classifier to predict normal and malicious data. To do

the parameter transfer, they reuse the model architecture and freeze most weight parameters for the target domain so that they do not need a large training dataset to retrain the model. However, they do not mention implementation details, like which layers are frozen in the transfer learning stage. Observing how transfer learning performs with different frozen layers would be interesting. Pan et al. apply parameter transfer to fully use the similarity of the anomalous patterns across different bridges [108]. They train a CNN model on one bridge data, then transfer the knowledge obtained by the CNN model to a small part of the target data. They update the last three fully connected layers while keeping the convolutional layers intact. The experimental results show transfer learning achieves higher accuracy anomaly detection across bridges. Weber et al. takes advantage of simulation data by training on synthetic environmental data, then fine-tunes the pre-trained model and transfers the knowledge from simulation data for real-time online building occupancy detection [17]. Although the results show the effectiveness of transfer learning, the availability, and reliability of the simulation data for other industrial applications is still an open issue. Sayed et al. adapt parameter transfer using pre-trained CNN models, such as AlexNet and GoogLeNet, pre-trained on ImageNet [18]. The pre-trained model is then further used for downstream tasks. The results show the pre-trained models outperform their customed CNN model, which is not pre-trained. However, it is important to note that the transfer effect may not be entirely convincing due to the fact that the customer CNN model is not pre-trained on the same dataset.

F. APPLICATION-INDEPENDENT CONSIDERATIONS

Data scarcity, as well as domain shift, stand out as the two main common problems independent of the industrial field of application. The same problems have originally prompted the use of transfer learning in general, and respectively, general techniques are applied widely across domains. Regarding data scarcity, this un-surprisingly involves leveraging pre-trained models as a starting point for further training. Regarding domain shift, mapping transfer and parameter transfer are the most often-used approaches. Unlike parameter transfer, mapping transfer incorporates the source and target data in the training process. Instance transfer and domain-adversarial transfer learning were not employed in the surveyed literature – researchers seem to not see huge value in these methods for the surveyed fields.

Another common aspect across time series anomaly detection applications is the choices of model architecture to facilitate the training process by capturing temporal dependencies and recognizing patterns over varying time scales: Favourite architectures include CNNs, LSTMs, and auto-encoders that have sets of assumptions (inductive biases) about the data they analyze that make them excel in understanding the sequential nature of time series data. CNNs assume local (in time) connectivity, stationary statistics, and hierarchical structure, and induce certain translation-invariance. LSTMs are still given preference in many applications over the more modern deep learning architecture of choice for sequence learning, the transformer. The reason is their stronger inductive bias, leading to less data (and compute) hunger. Both CNN and LSTM networks can be built as e.g. classifiers, but also auto-encoders. These latter architectures have the advantage of learning low-dimensional representations of the high-dimensional time series with minimal loss of signal in an unsupervised way. The analysis of the data in the low-dimensional latent space facilitates anomaly detection. The concrete choice of architecture does not depend on the field of application but on the data and task, and thus the most suitable inductive bias.

As an interim conclusion, the most striking application-independent finding is that across the surveyed literature, predominantly simple, tried-and-tested design patterns for transfer learning are used in industry. The field of deep transfer learning would offer a much wider variety of approaches.

V. DISCUSSION

A. POTENTIAL

The automation of industrial process monitoring stands as a transformative step toward increasing efficiency and optimizing quality. While standard deep learning training is sufficient in discerning intricate patterns from vast datasets, its application in the dynamic industrial landscape is not without challenges. Chief among them is the impracticality of continuously obtaining large-scale labeled data to train models afresh for every nuanced variation in processes. Deep transfer learning has shown promise with its adaptive capabilities. By mitigating the need for extensive labeled data and eliminating the necessity to train models from scratch for every distinct setup. However, adopting deep transfer learning beyond simple parameter transfer is still a challenge.

B. CHALLENGES

1) DOMAIN SHIFT

Different from the i.i.d assumption in most machine learning problem settings, many industrial processes suffer from substantial domain shift due to dynamic changes in industrial settings, e.g., change of products or measuring sensors. Domain shift lies at the heart of the deep transfer learning problem. Particularly, the dynamic changes in many industrial processes, up to an apparent dissimilarity of source and target data, make the transfer learning task particularly challenging. Here we list the key challenges associated with domain shift:

- *Covariate shift* occurs when the marginal distribution of the features changes from source domain to target domain. The distribution mismatch poses challenges in transferring the knowledge from source to target domain.
- *Concept shift* refers to the changes in the relationships between features and labels. The relationship can change

from source domain to target domain, leading to bias and error in the model.
- *Label shift* refers to the label distribution in the target domain that can be different from the source domain, whether the marginal distribution changes or not.

2) LABEL AVAILABILITY AND RELIABILITY

Deep transfer learning is built upon deep learning, which usually requires a large amount of labeled data, the more data a model has available for training, the better it can generalize to new examples. In real-world industrial time series anomaly detection tasks, collecting data is probably easy, but collecting labels is much more expensive and time-consuming, sometimes prohibitively so, leading to the unavailability of sufficient labeled data. Self-supervised learning can be used to re-label a large amount of unlabeled data and thus anomaly detection models usually need to learn in an unsupervised or semi-supervised mode [130]. In industrial cases, another significant concern is to ensure the data quality. Due to the high cost associated with obtaining reliable and precise labels, usually self-supervised learning is applied to create pseudo labels or relabel the unlabeled data, thus facilitating the transfer learning process. Additionally, a data-centric process with humans in the loop can be involved in improving label reliability. However, unreliable labels can still affect the transfer learning training process.

3) MISSING RELEVANT DATA INFORMATION

Missing relevant data poses significant challenges for transfer learning since it can affect the model's ability to generalize and transfer knowledge from source to target domain.

- *Imbalanced data*: Even if the labels can be collected, anomalies can be extremely rare by design, which poses the risk of training with extremely imbalanced data. A practical problem for anomaly detection in industry is the extremely imbalanced data distribution, in which normal samples dominate in data and abnormal samples only share a small percentage in the whole dataset. Prior research has proven that the effect of class imbalance on classification performance by using deep learning is detrimental [131]. However, most research studies still ignore such problems, which can result in poor performance regarding the minority class, i.e., abnormal data are misclassified as normal.
- *Information loss*: missing data can lead to lost important features. For example, some information that has a significant effect on the process from case to case is not even recorded or is too complex to record (i.e. part geometry, machine geometry, or environmental conditions in injection molding processes).

Various approaches have been developed to address these challenges to reduce the domain gap between the source and target domains, aiming at mitigating domain shift. These techniques involve domain generalization, contrastive learning, and adversarial examples. The domain shift problem is far from being solved. To tackle this problem, transfer learning requires a deep understanding of the target data's characteristics and appropriate transferable strategies to effectively bridge the gap between the source and target domains.

4) EFFECTIVENESS OF DEEP TRANSFER LEARNING

The general effectiveness of deep transfer learning is limited by the difficulty of determining which knowledge or to what extent the knowledge should be transferred from source to target task. Unlike natural language processing, pre-training a language model on a large corpus of text data can help the model learn the statistical patterns and semantic and syntactic representations of words and sentences, which can be used for new natural language processing tasks with a few or even without data. Due to data privacy, large available public datasets usually do not exist for industrial time series, or they cannot be used because of a large domain gap between different datasets and tasks. In this case, transferring all of the knowledge may not be beneficial, as it may be irrelevant. In the worst case, this can lead to negative transfer [28], [132], in which the extracted knowledge harms the new task-learning. This requires assessing how source and target tasks are related, carefully selecting the knowledge to be transferred, and selecting the proper means to implement this transfer. Glorot et al. attempt to analyze and quantify the gained knowledge from source to target domain [133]. For example, they define transfer error, transfer loss, transfer ratio, and in-domain ratio, which provide metrics to interpret the transferring performance.

C. DIRECTIONS FOR ANOMALY DETECTION SOLUTION DESIGN

1) DATA PREPROCESSING

How data preprocessing should be conducted is an open question. For industrial applications, some researchers contend that using raw time series data directly as input for training may not be the most efficient. Hence, they propose deriving or selecting features from time series data by statistical methods or human experience. This can significantly decrease the complexity of the dataset. On the other hand, this crops a lot of potentially useful information, e.g., the time series trend. To reduce the dimensionality, some researchers use machine parameters as features in the manufacturing process instead of the processing data collected by sensors [10], [85], [113], [114]. Others try different transformations of raw time series data, a common way being to transform 1D time series data to 2D image data [9], [15], [112] or transforming time domain signals otherwise into the frequency domain [39]. However, as large-scale computation power and storage become cheaper and more accessible, it is becoming increasingly common to use deep learning techniques to process time series data directly [11], [13].

2) DATA AUGMENTATION BY GENERATIVE AI

Data augmentation is useful for deep learning models because it can help to prevent overfitting. For deep transfer learning, when a model becomes too closely adapted to the specifics of the source domain, it may not be able to generalize well to some examples in the task domain. One important technique is to acquire effective synthetic data, e.g., using a simulation process or model to explore potential anomalous conditions by simulating industrial processes under parameters that cannot yet be experienced in the real world. High fidelity and reliable simulation data can provide training data at low cost and mitigate the problem of insufficient samples for deep transfer learning [13]. Another way to generate effective synthetic data is to use generative models, such as GANs. GANs are only trained on normal data to generate indistinguishable normal samples so that abnormal samples can be distinguished during the testing stage of the overarching anomaly detection system, as they deviate from the normal data distribution [134]. To increase the number of anomalous samples and thus the robustness of the anomaly detection model, the technique of adversarial perturbation known from computer vision [135] can be used.

3) DEALING WITH DATA IMBALANCE

DNNs perform well when they are trained on balanced datasets. However, in practice, it is difficult to get sufficient anomalous data for anomaly detection tasks. For example, the manufacturing process is usually in a healthy state due to the pre-designed and optimized operation. Several ways exist to address the imbalanced dataset for time series anomaly detection. One way is to oversample the minority class, e.g., by randomly replicating samples from the minority class to equalize the number of samples from each class in each batch. The Synthetic Minority Over-sampling Technique is an advanced method that creates synthetic samples to force the decision region of the minority class to become more general [136]. This technique is widely used in anomaly detection tasks in industry [137], [138]. Apart from oversampling, resampling strategies are frequently used to assign a higher probability to abnormal samples and evenly select the same amount of samples from both classes in each batch. Moreover, a weighted loss can be implemented to balance the loss between the abnormal and normal class in supervised anomaly detection [131].

D. DIRECTIONS FOR DEEP TRANSFER LEARNING IMPLEMENTATION

1) WHEN SHALL DEEP TRANSFER LEARNING BE USED?

(1) Limited data availability: It poses a significant challenge in machine learning, particularly when aiming to train models for specific tasks. Pre-training a model on a larger or more diverse dataset, even if unrelated to the specific task at hand, enables the acquisition of generalizable features and representations. These generalized features, learned from a broader context, can then be effectively transferred to analyze the target domain with limited data. This can effectively provide a practical solution to the challenges posed by data scarcity. (2) Similar domains: Deep transfer learning is well suited when tackling source and target domains with a high degree of similarity. In such instances, knowledge can be derived either from a model pre-trained on a similar dataset or one trained on both source and target data. In both cases, the model can efficiently transfer relevant features and representations within the domain, facilitating a more robust adaptation to the target dataset, and ultimately optimizing the model's ability to discern and detect patterns within the target domain. (3) Limited resources (encompassing both time and computational power): When faced with resource constraints, it is recommended to employ parameter transfer, especially if a pre-trained model is readily available. As described in [34], the transfer might improve learning in three distinct ways: (a) a higher performance at the very beginning of learning, (b) a steeper slope in the learning curve, or (c) a higher asymptotic performance. Parameter transfer leverages the learned parameters and weights of a pre-trained model, often trained on a larger dataset. By doing so, the resource-intensive process of training a model from the ground up is circumvented, and the computational burden is significantly alleviated.

2) WHEN NOT TO USE DEEP TRANSFER LEARNING?

(1) Irrelevant data: If the target data is vastly different from the source data, deep transfer learning may not be appropriate, sometimes even leading to negative transfer. For example, if one wants to train a model for natural language processing on a new dataset, using a pre-trained model that has been trained on image data may not yield meaningful results. This is due to the vast dissimilarity in data modalities and features between images and text. (2) Task-specific models: In scenarios where the target task is well-defined and specific, and pre-trained models do not align closely with the task requirements, it is usually more effective to build a task-specific model from scratch. (3) High domain shift: If there is a large difference between the source and the target domain, deep transfer learning may not be effective. This can happen when the data distributions, features, or labels are vastly different. (4) Abundance of labeled data available for target task: If there are enough data for the new task, it may be more effective to train a model from scratch [119].

3) WHAT MODEL ARCHITECTURE TO CHOOSE?

We recommend choosing the model architecture mainly based on data size and label availability, starting from a relatively small network and moving gradually to more complex DNNs. CNNs also effectively extract time series features [19], [111]. For semi-supervised settings, CNN-based auto-encoders are trained to reconstruct the original data [110]. It is important to effectively capture the temporal dependencies and extract features of time series data. LSTMs are extensively employed for this purpose, as they excel in

detecting temporal dependencies in time series data [21], [23], [112].

Exploring hybrid architectures that combine the advantages of CNNs, RNNs, and LSTMs for tasks involving both spatial and temporal dependencies can be beneficial. Cao et al. propose a multi-head CNN–RNN architecture for multi-time series anomaly detection [16]. A CNN is used to extract meaningful features from raw data and then an RNN is applied to learn temporal patterns simultaneously. Similarly, Dhillon et al. utilize LSTM layers to model the time series signals after obtaining the features from a CNN. An alternative way to benefit from different models is the use of ensemble approaches, combining the strengths of different model architectures to enhance performance, especially in situations where the target task requires capturing diverse features. In the future, we expect more applications to use transformer-based approaches as pre-trained models become available and public datasets get open-sourced.

4) BEYOND TRANSFER LEARNING

Foundation models like SAM [139] or others, using for example transformer architectures [48] or diffusion models [140], demonstrate emerging properties such as in-context learning [53] and complex cross-modality conditioning. This is achieved by training complex and often auto-regressive models with massive amounts of data, although the precise mechanisms that lead to this are not well understood. Some of those models generalize to new settings and tasks, without an explicit element of transfer learning. Thus, the application of foundation models in industrial time series analysis has the potential to reduce and eventually eliminate the need to explicitly account for changes in the domain within the modeling, by instead having the foundation model provide the transfer capability (see examples in Sec. VI-b). To not only detect anomalies but also identify failure modes, analyze root causes, and elicit an appropriate intervention, AI systems must implicitly or explicitly model causal relations. Counterfactual inference incorporates causal relations between observations and interventions, which allows predictions of outcomes never seen during training [141].

Another aspect of deep learning implementations is the limited computing power of hardware platforms, such as embedded systems in industry. Sensor data are typically acquired using resource-constrained edge processing devices that struggle with computationally intensive tasks, especially when training a DNN model. Federated learning stands out as a leading solution, with its ability to utilize data while preserving privacy [142], [143]. The technology enables a more collaborative approach to ML while preserving user privacy by storing data decentralized on distributed devices rather than on a central server. Combining deep transfer learning with federated learning is a promising and powerful combination in the abovementioned industrial applications.

VI. CONCLUSION

In this survey, we presented a comprehensive overview of deep transfer learning by defining transfer learning problem settings and categorizing the state-of-the-art deep transfer learning approaches based on the surveyed papers. Then, we review and emphasize on investigating deep transfer learning approaches for time series anomaly detection in different industrial settings. Equipped with this foundation, we selected representative examples of the landscape of fielded applications to provide practitioners with a guide to the field and possibilities of industrial time series anomaly detection.

The main finding of this survey is that only a limited variety of deep transfer learning methods are employed in anomaly detection in industrial time series analysis – mainly simple ones. Almost all applications employ parameter transfer, arguably the most straightforward transfer approach. In its simplest implementation, it only involves fine-tuning a pre-trained model. Accordingly, the employed network architectures are simple, none of the reviewed research papers used advanced DNN building blocks like Transformer, which are common in computer vision and language modeling. We expect this type of architecture with suitable modifications and/or pre-trained parameters to spread to more niche fields. Despite this, the survey suggests that deep transfer learning approaches have huge potential and promise for solving more complex and dynamic anomaly detection tasks in industry. As the field is still in an early stage, more R&D is expected to fully realize the potential of deep transfer learning in increasingly complex settings.

In the end, we highlight the importance of considering feasibility, reliability, explainability, and real-time data stream when designing a transfer learning system for time series anomaly detection. After carefully discussing open challenges, we gave practical directions for time series anomaly detection solution design and deep transfer learning implementation. In our view, the following directions hold the greatest potential for future work:

1) AUTOMATIC SELECTION OF TRANSFERABLE FEATURES [57]

It refers to methods for selecting and transferring only the relevant knowledge for the new tasks from the base model. This could involve the use of techniques such as selective fine-tuning and distillation to identify the most important features learned from source domains [30], [144].

2) INVESTING INTO ADVANCED DEEP TRANSFER LEARNING SCHEMES AND DNN MODELS

The conceptionally simplest parameter transfer approach has the advantage of being readily applicable by interdisciplinary teams without ML research experience. However, it seems promising to invest in testing more sophisticated deep transfer learning approaches according to different use cases, such as mapping transfer, adversarial transfer, etc. The same applies to testing diverse DNN models besides

straightforward ones. Recently, large models have been used in time series anomaly detection. For example, Xu et al. propose the Anomaly Transformer with a new anomaly-attention mechanism to compute the association discrepancy [145]. A minimax strategy is devised to amplify the normal-abnormal distinguishability of the association discrepancy. On the other hand, Pintilie et al. leverage diffusion models for multivariate time series anomaly detection [146]. They train two diffusion-based models that outperform strong transformer-based methods on synthetic datasets and are competitive on real-world data. Additionally, their DiffusionAE model is more robust to different levels and the number of anomaly types. These large models have proven to be effective and advantageous given certain data and tasks. It's important to note that their effectiveness also depends on the characteristics of the time series data and the requirements of the anomaly detection task. Additionally, model computational efficiency and interpretability should be considered, especially in real-time or resource-constrained industrial applications.

3) DATA-CENTRIC APPROACH TO REAL-TIME ANOMALY DETECTION

The data-centric approach focuses on improving ML models by ensuring high-quality labeled data [147] using techniques such as re-labeling, re-weighting, or data augmentation [148]. Currently, a human-in-the-loop solution is still needed. Frameworks have been proposed to assist annotators with graph-based algorithms such as nearest neighbor graphs [84], decision trees [149], or factor graphs [150]. Although these methods have proven to be effective, a more automated process is a goal for future research.

4) LEVERAGING GENERATIVE AI

Generative models like GANs and diffusion models can generate synthetic time series data, making them valuable for data augmentation. Augmenting the original data with synthetic samples can enhance the deep learning models' robustness, especially in real-world applications where target data are limited. These models can also be leveraged to examine anomalies and generate anomalies to help alleviate the imbalance within the data [151].

5) INTEGRATION WITH OTHER ML METHODS

To develop robust AI solutions for time series anomaly detection in the industry, relying solely on transfer learning is insufficient. Future strategies should integrate other ML approaches, including continuous learning, meta-learning, and federated learning.

ACKNOWLEDGMENT

This work has been supported financially by Innosuisse grant 101.787 IP-ENG "DISTRAL". The authors would like to acknowledge Claudio Riginio for his constructive comments on an earlier draft of this paper and for his assistance with the illustrations.

REFERENCES

[1] H. Kagermann, W.-D. Lukas, and W. Wahlster, "Industrie 4.0: Mit dem internet der dinge auf dem weg zur 4. Industriellen revolution," *VDI Nachrichten*, vol. 13, no. 1, pp. 2–3, 2011.

[2] V. Roblek, M. Meško, and A. Krapež, "A complex view of industry 4.0," *SAGE Open*, vol. 6, no. 2, Apr. 2016, Art. no. 215824401665398.

[3] L. Wang, M. Törngren, and M. Onori, "Current status and advancement of cyber-physical systems in manufacturing," *J. Manuf. Syst.*, vol. 37, pp. 517–527, Oct. 2015.

[4] S. Jeschke, C. Brecher, T. Meisen, D. Özdemir, and T. Eschert, "Industrial Internet of Things and cyber manufacturing systems," in *Industrial Internet of Things*. Cham, Switzerland: Springer, 2017, pp. 3–19.

[5] L. S. Dalenogare, G. B. Benitez, N. F. Ayala, and A. G. Frank, "The expected contribution of industry 4.0 technologies for industrial performance," *Int. J. Prod. Econ.*, vol. 204, pp. 383–394, Oct. 2018.

[6] H. Kagermann, "Chancen von industrie 4.0 nutzen," in *Handbuch Industrie 4.0 Bd.4*. Berlin, Germany: Springer, 2017, pp. 237–248.

[7] S. M. A. A. Abir, A. Anwar, J. Choi, and A. S. M. Kayes, "IoT-enabled smart energy grid: Applications and challenges," *IEEE Access*, vol. 9, pp. 50961–50981, 2021.

[8] Y.-J. Park, S.-K.-S. Fan, and C.-Y. Hsu, "A review on fault detection and process diagnostics (don't short) in industrial processes," *Processes*, vol. 8, no. 9, p. 1123, Sep. 2020.

[9] Y. Liao, I. Ragai, Z. Huang, and S. Kerner, "Manufacturing process monitoring using time-frequency representation and transfer learning of deep neural networks," *J. Manuf. Processes*, vol. 68, pp. 231–248, Aug. 2021.

[10] Y. Lockner and C. Hopmann, "Induced network-based transfer learning in injection molding for process modelling and optimization with artificial neural networks," *Int. J. Adv. Manuf. Technol.*, vol. 112, nos. 11–12, pp. 3501–3513, Feb. 2021.

[11] B. Maschler, T. Knodel, and M. Weyrich, "Towards deep industrial transfer learning for anomaly detection on time series data," in *Proc. 26th IEEE Int. Conf. Emerg. Technol. Factory Autom. (ETFA)*, Sep. 2021, pp. 1–8.

[12] T. Wen and R. Keyes, "Time series anomaly detection using convolutional neural networks and transfer learning," 2019, *arXiv:1905.13628*.

[13] Y. Xu, Y. Sun, X. Liu, and Y. Zheng, "A digital-twin-assisted fault diagnosis using deep transfer learning," *IEEE Access*, vol. 7, pp. 19990–19999, 2019.

[14] W. Mao, D. Zhang, S. Tian, and J. Tang, "Robust detection of bearing early fault based on deep transfer learning," *Electronics*, vol. 9, no. 2, p. 323, Feb. 2020.

[15] W. Wang, Z. Wang, Z. Zhou, H. Deng, W. Zhao, C. Wang, and Y. Guo, "Anomaly detection of industrial control systems based on transfer learning," *Tsinghua Sci. Technol.*, vol. 26, no. 6, pp. 821–832, Dec. 2021.

[16] M. Canizo, I. Triguero, A. Conde, and E. Onieva, "Multi-head CNN–RNN for multi-time series anomaly detection: An industrial case study," *Neurocomputing*, vol. 363, pp. 246–260, Oct. 2019.

[17] M. Weber, C. Doblander, and P. Mandl, "Towards the detection of building occupancy with synthetic environmental data," 2020, *arXiv:2010.04209*.

[18] A. N. Sayed, Y. Himeur, and F. Bensaali, "From time-series to 2D images for building occupancy prediction using deep transfer learning," *Eng. Appl. Artif. Intell.*, vol. 119, Mar. 2023, Art. no. 105786.

[19] Y. Yao, D. Ge, J. Yu, and M. Xie, "Model-based deep transfer learning method to fault detection and diagnosis in nuclear power plants," *Frontiers Energy Res.*, vol. 10, Mar. 2022, Art. no. 823395.

[20] M. Abdallah, W. J. Lee, N. Raghunathan, C. Mousoulis, J. W. Sutherland, and S. Bagchi, "Anomaly detection through transfer learning in agriculture and manufacturing IoT systems," 2021, *arXiv:2102.05814*.

[21] C. Panjapornpon, S. Bardeeniz, M. A. Hussain, and P. Chomchai, "Explainable deep transfer learning for energy efficiency prediction based on uncertainty detection and identification," *Energy AI*, vol. 12, Apr. 2023, Art. no. 100224.

[22] H. Dhillon and A. Haque, "Towards network traffic monitoring using deep transfer learning," in *Proc. IEEE 19th Int. Conf. Trust, Secur. Privacy Comput. Commun. (TrustCom)*, Dec. 2020, pp. 1089–1096.

[23] P. Xiong, Y. Zhu, Z. Sun, Z. Cao, M. Wang, Y. Zheng, J. Hou, T. Huang, and Z. Que, "Application of transfer learning in continuous time series for anomaly detection in commercial aircraft flight data," in *Proc. IEEE Int. Conf. Smart Cloud (SmartCloud)*, Sep. 2018, pp. 13–18.

[24] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[25] B. Maschler, H. Vietz, H. Tercan, C. Bitter, T. Meisen, and M. Weyrich, "Insights and example use cases on industrial transfer learning," *Proc. CIRP*, vol. 107, pp. 511–516, Jan. 2022.

[26] B. Maschler and M. Weyrich, "Deep transfer learning for industrial automation: A review and discussion of new techniques for data-driven machine learning," *IEEE Ind. Electron. Mag.*, vol. 15, no. 2, pp. 65–75, Jun. 2021.

[27] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021.

[28] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[29] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *Proc. ICANN*, 2018, pp. 270–279.

[30] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. NeurIPS*, vol. 2, 2014, pp. 3320–3328.

[31] F. Yu, X. Xiu, and Y. Li, "A survey on deep transfer learning and beyond," *Mathematics*, vol. 10, no. 19, p. 3619, Oct. 2022.

[32] K. Choi, J. Yi, C. Park, and S. Yoon, "Deep learning for anomaly detection in time-series data: Review, analysis, and guidelines," *IEEE Access*, vol. 9, pp. 120043–120065, 2021.

[33] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," 2019, *arXiv:1901.03407*.

[34] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of Research on Machine Learning*. Hershey, PA, USA: IGI Global, Jan. 2009.

[35] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, Jul. 1997.

[36] S. Ruder, "An overview of multi-task learning in deep neural networks," 2017, *arXiv:1706.05098*.

[37] W. M. Kouw and M. Loog, "An introduction to domain adaptation and transfer learning," 2018, *arXiv:1812.11806*.

[38] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.

[39] Y. Lukic, C. Vogt, O. Dürr, and T. Stadelmann, "Speaker identification and clustering using convolutional neural networks," in *Proc. IEEE 26th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2016, pp. 1–6.

[40] T. Stadelmann, M. Amirian, I. Arabaci, M. Arnold, G. F. Duivesteijn, I. Elezi, and M. Geiger, "Deep learning in the wild," in *Proc. 8th ANNPR*, 2018, pp. 17–38.

[41] J. Schmidhuber, "Annotated history of modern AI and deep learning," 2022, *arXiv:2212.11279*.

[42] Q.-Q. He, S. W. I. Siu, and Y.-W. Si, "Instance-based deep transfer learning with attention for stock movement prediction," *Int. J. Speech Technol.*, vol. 53, no. 6, pp. 6887–6908, Jul. 2022.

[43] M. Amirian, J. A. Montoya-Zegarra, J. Gruss, Y. D. Stebler, A. S. Bozkir, M. Calandri, F. Schwenker, and T. Stadelmann, "PrepNet: A convolutional auto-encoder to homogenize CT scans for cross-dataset medical image analysis," in *Proc. 14th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Oct. 2021, pp. 1–7.

[44] T. Wang, J. Huan, and M. Zhu, "Instance-based deep transfer learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 367–375.

[45] R. Bommasani et al., "On the opportunities and risks of foundation models," 2021, *arXiv:2108.07258*.

[46] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. ACL Conf. NAACL HLT*, Jun. 2019, pp. 4171–4186.

[47] K. Zhang, S. Wang, S. Wang, and Q. Xu, "Anomaly detection of control moment gyroscope based on working condition classification and transfer learning," *Appl. Sci.*, vol. 13, no. 7, p. 4259, Mar. 2023.

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, vol. 30, 2017, pp. 6000–6010.

[49] Y. Dou, M. Forbes, R. Koncel-Kedziorski, N. Smith, and Y. Choi, "Is GPT-3 text indistinguishable from human text? Scarecrow: A framework for scrutinizing machine text," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 7250–7274.

[50] N. Alexandr, O. Irina, K. Tatyana, K. Inessa, and P. Arina, "Fine-tuning GPT-3 for Russian text summarization," in *Data Science and Intelligent Systems*. Cham, Switzerland: Springer, 2021, pp. 748–757.

[51] Y. Guo, H. Shi, A. Kumar, K. Grauman, T. Rosing, and R. Feris, "SpotTune: Transfer learning through adaptive fine-tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4800–4809.

[52] P. Sager, S. Salzmann, F. Burn, and T. Stadelmann, "Unsupervised domain adaptation for vertebrae detection and identification in 3D CT volumes using a domain sanity loss," *J. Imag.*, vol. 8, no. 8, p. 222, Aug. 2022.

[53] T. B. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Sys.*, vol. 33, 2020, pp. 1877–1901.

[54] Y. Wang, J. Yan, X. Ye, Q. Jing, J. Wang, and Y. Geng, "Few-shot transfer learning with attention mechanism for high-voltage circuit breaker fault diagnosis," *IEEE Trans. Ind. Appl.*, vol. 58, no. 3, pp. 3353–3360, May 2022.

[55] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," 2014, *arXiv:1412.3474*.

[56] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, Aug. 2017, pp. 2208–2217.

[57] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. ICML*, vol. 37, Jul. 2015, pp. 97–105.

[58] X. Zhang, F. Xinnan Yu, S.-F. Chang, and S. Wang, "Deep transfer network: Unsupervised domain adaptation," 2015, *arXiv:1503.00591*.

[59] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5385–5394.

[60] E. Soleimani and E. Nazerfard, "Cross-subject transfer learning in human activity recognition systems using generative adversarial networks," *Neurocomputing*, vol. 426, pp. 26–34, Feb. 2021.

[61] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4068–4076.

[62] Y. Ozyurt, S. Feuerriegel, and C. Zhang, "Contrastive learning for unsupervised domain adaptation of time series," 2022, *arXiv:2206.06243*.

[63] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," in *Domain Adaptation in Computer Vision Applications*. Cham, Switzerland: Springer, 2017, pp. 189–209.

[64] H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, and M. Marchand, "Domain-adversarial neural networks," 2014, *arXiv:1412.4446*.

[65] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2962–2971.

[66] Z. Sun, M. Wang, and L. Li, "Multilingual translation via grafting pre-trained language models," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 2735–2747. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.233/

[67] M. Glass, A. Gliozzo, R. Chakravarti, A. Ferritto, L. Pan, G. P. S. Bhargav, D. Garg, and A. Sil, "Span selection pre-training for question answering," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2773–2782.

[68] L. Tuggener, J. Schmidhuber, and T. Stadelmann, "Is it enough to optimize CNN architectures on ImageNet?" *Frontiers Comput. Sci.*, vol. 4, Nov. 2022, Art. no. 1041703.

[69] J. Shen, Y. Qu, W. Zhang, and Y. Yu, "Wasserstein distance guided representation learning for domain adaptation," in *Proc. AAAI*, Feb. 2018, pp. 4058–4065.

[70] W. Dai, G.-R. Xue, Q. Yang, and Y. Yu, "Co-clustering based classification for out-of-domain documents," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2007, pp. 210–219.

[71] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 6.1–6.12.

[72] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NeurIPS*, vol. 27, 2014, pp. 2672–2680.

[73] J. Schmidhuber, "Generative adversarial networks are special cases of artificial curiosity (1990) and also closely related to predictability minimization (1991)," *Neural Netw.*, vol. 127, pp. 58–66, Jul. 2020.

[74] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Netw.*, vol. 113, pp. 54–71, May 2019.

[75] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surv.*, vol. 53, no. 3, p. 63, Jun. 2020.

[76] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594–611, Apr. 2006.

[77] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 951–958.

[78] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4396–4415, Apr. 2023.

[79] G. Blanchard, G. Lee, and C. Scott, "Generalizing from several related classification tasks to a new unlabeled sample," in *Proc. NeurIPS*, vol. 24, 2011, pp. 2178–2186.

[80] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Lean. (ICML)*, Jun. 2017, pp. 1126–1135.

[81] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5149–5169, Sep. 2022.

[82] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, Jun. 2021.

[83] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 857–876, Jan. 2023.

[84] H. Bai, M. Cao, P. Huang, and J. Shan, "Self-supervised semi-supervised learning for data labeling and quality evaluation," in *Proc. NeuIPS*, Nov. 2021, pp. 1–6. [Online]. Available: https://machinelearning.apple.com/research/self-supervised-semi-supervised-learning

[85] H. Tercan, A. Guajardo, J. Heinisch, T. Thiele, C. Hopmann, and T. Meisen, "Transfer-learning: Bridging the gap between real and simulation data for machine learning in injection molding," *Proc. CIRP*, vol. 72, pp. 185–190, Jan. 2018.

[86] N. Goernitz, M. Kloft, K. Rieck, and U. Brefeld, "Toward supervised anomaly detection," *J. Artif. Intell. Res.*, vol. 46, pp. 235–262, Feb. 2013.

[87] Y. Zhang, Y. Chen, J. Wang, and Z. Pan, "Unsupervised deep anomaly detection for multi-sensor time-series signals," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 2, pp. 2118–2132, Feb. 2023.

[88] T. Stadelmann, V. Tolkachev, B. Sick, J. Stampfli, and O. Dürr, "Beyond ImageNet: Deep learning in industrial practice," in *Applied Data Science*. Cham, Switzerland: Springer, 2019, pp. 205–232.

[89] D. M. Hawkins, *Identification of Outliers*, vol. 11. Dordrecht, The Netherlands: Springer, 1980.

[90] H. Zhu, C. Yi, S. Rho, S. Liu, and F. Jiang, "An interpretable multivariate time-series anomaly detection method in cyber-physical systems based on adaptive mask," *IEEE Internet Things J.*, vol. 14, no. 8, pp. 1–13, Aug. 2021.

[91] S. Schmidl, P. Wenig, and T. Papenbrock, "Anomaly detection in time series: A comprehensive evaluation," *Proc. VLDB Endowment*, vol. 15, no. 9, pp. 1779–1797, May 2022.

[92] J. Audibert, P. Michiardi, F. Guyard, S. Marti, and M. A. Zuluaga, "USAD: UnSupervised anomaly detection on multivariate time series," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 3395–3404.

[93] P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, and G. Shroff, "LSTM-based encoder–decoder for multi-sensor anomaly detection," 2016, *arXiv:1607.00148*.

[94] Y. Wei, J. Jang-Jaccard, W. Xu, F. Sabrina, S. Camtepe, and M. Boulic, "LSTM-autoencoder-based anomaly detection for indoor air quality time-series data," *IEEE Sensors J.*, vol. 23, no. 4, pp. 3787–3800, Feb. 2023.

[95] F. Zeng, M. Chen, C. Qian, Y. Wang, Y. Zhou, and W. Tang, "Multivariate time series anomaly detection with adversarial transformer architecture in the Internet of Things," *Future Gener. Comput. Syst.*, vol. 144, pp. 244–255, Jul. 2023.

[96] D. Li, D. Chen, B. Jin, L. Shi, J. Goh, and S.-K. Ng, "MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks," in *Proc. ICANN*, 2019, pp. 703–716.

[97] Z. Niu, K. Yu, and X. Wu, "LSTM-based VAE-GAN for time-series anomaly detection," *Sensors*, vol. 20, no. 13, p. 3738, Jul. 2020.

[98] M. A. Bashar and R. Nayak, "TAnoGAN: Time series anomaly detection with generative adversarial networks," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2020, pp. 1778–1785.

[99] J. Kim, H. Kang, and P. Kang, "Time-series anomaly detection with stacked transformer representations and 1D convolutional network," *Eng. Appl. Artif. Intell.*, vol. 120, Apr. 2023, Art. no. 105964.

[100] A. Deng and B. Hooi, "Graph neural network-based anomaly detection in multivariate time series," in *Proc. AAAI*, May 2021, vol. 35, no. 5, pp. 4027–4035.

[101] C. Tang, L. Xu, B. Yang, Y. Tang, and D. Zhao, "GRU-based interpretable multivariate time series anomaly detection in industrial control system," *Comput. Secur.*, vol. 127, Apr. 2023, Art. no. 103094.

[102] C. Ding, S. Sun, and J. Zhao, "MST-GAT: A multimodal spatial–temporal graph attention network for time series anomaly detection," *Inf. Fusion*, vol. 89, pp. 527–536, Jan. 2023.

[103] Y. Himeur, A. Alsalemi, F. Bensaali, and A. Amira, "A novel approach for detecting anomalous energy consumption based on micro-moments and deep neural networks," *Cognit. Comput.*, vol. 12, no. 6, pp. 1381–1401, Nov. 2020.

[104] Y. Yang, C. Zhang, T. Zhou, Q. Wen, and L. Sun, "DCdetector: Dual attention contrastive representation learning for time series anomaly detection," in *Proc. 29th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2023, pp. 3033–3045.

[105] W. Li, R. Huang, J. Li, Y. Liao, Z. Chen, G. He, R. Yan, and K. Gryllias, "A perspective survey on deep transfer learning for fault diagnosis in industrial scenarios: Theories, applications and challenges," *Mech. Syst. Signal Process.*, vol. 167, Mar. 2022, Art. no. 108487.

[106] J. Ma, J. C. P. Cheng, C. Lin, Y. Tan, and J. Zhang, "Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques," *Atmos. Environ.*, vol. 214, Oct. 2019, Art. no. 116885.

[107] I. Rosenberg, G. Sicard, and E. David, "End-to-end deep neural networks and transfer learning for automatic analysis of nation-state malware," *Entropy*, vol. 20, no. 5, p. 390, May 2018.

[108] Q. Pan, Y. Bao, and H. Li, "Transfer learning-based data anomaly detection for structural health monitoring," *Struct. Health Monitor.*, vol. 22, no. 5, pp. 3077–3091, Jan. 2023.

[109] G. Li, L. Chen, J. Liu, and X. Fang, "Comparative study on deep transfer learning strategies for cross-system and cross-operation-condition building energy systems fault diagnosis," *Energy*, vol. 263, Jan. 2023, Art. no. 125943.

[110] O. Serradilla, E. Zugasti, J. R. de Okariz, J. Rodriguez, and U. Zurutuza, "Adaptable and explainable predictive maintenance: Semi-supervised deep learning for anomaly detection and diagnosis in press machine data," *Appl. Sci.*, vol. 11, no. 16, p. 7376, Aug. 2021.

[111] J. Zgraggen, M. Ulmer, E. Jarlskog, G. Pizza, and L. G. Huber, "Transfer learning approaches for wind turbine fault detection using deep learning," in *Proc. PHM Soc. Eur. Conf.*, vol. 6, no. 1, p. 12, Jun. 2021.

[112] M. Zabin, H.-J. Choi, and J. Uddin, "Hybrid deep transfer learning architecture for industrial fault diagnosis using Hilbert transform and DCNN–LSTM," *J. Supercomput.*, vol. 79, no. 5, pp. 5181–5200, Mar. 2023.

[113] H. Tercan, A. Guajardo, and T. Meisen, "Industrial transfer learning: Boosting machine learning in production," in *Proc. IEEE 17th Int. Conf. Ind. Informat. (INDIN)*, vol. 1, Jul. 2019, pp. 274–279.

[114] Y. Lockner, C. Hopmann, and W. Zhao, "Transfer learning with artificial neural networks between injection molding processes and different polymer materials," *J. Manuf. Processes*, vol. 73, pp. 395–408, Jan. 2022.

[115] S. Gellrich, M.-A. Filz, A.-S. Wilde, T. Beganovic, A. Mattheus, T. Abraham, and C. Herrmann, "Deep transfer learning for improved product quality prediction: A case study of aluminum gravity die casting," *Proc. CIRP*, vol. 104, pp. 912–917, Jan. 2021.

[116] M. Abdallah, B.-G. Joung, W. J. Lee, C. Mousoulis, N. Raghunathan, A. Shakouri, J. W. Sutherland, and S. Bagchi, "Anomaly detection and inter-sensor transfer learning on smart manufacturing datasets," *Sensors*, vol. 23, no. 1, p. 486, Jan. 2023.

[117] B. Maschler, T. T. Huong Pham, and M. Weyrich, "Regularization-based continual learning for anomaly detection in discrete manufacturing," *Proc. CIRP*, vol. 104, pp. 452–457, Jan. 2021.
[118] J. Park, B. Kim, and H. Kim, "MENDEL: Time series anomaly detection using transfer learning for industrial control systems," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Feb. 2023, pp. 261–268.
[119] P. Liang, H.-D. Yang, W.-S. Chen, S.-Y. Xiao, and Z.-Z. Lan, "Transfer learning for aluminium extrusion electricity consumption anomaly detection via deep neural networks," *Int. J. Comput. Integr. Manuf.*, vol. 31, nos. 4–5, pp. 396–405, Apr. 2018.
[120] A. Copiaco, Y. Himeur, A. Amira, W. Mansoor, F. Fadli, S. Atalla, and S. S. Sohail, "An innovative deep anomaly detection of building energy consumption using energy time-series images," *Eng. Appl. Artif. Intell.*, vol. 119, Mar. 2023, Art. no. 105775.
[121] C. Xu, J. Wang, J. Zhang, and X. Li, "Anomaly detection of power consumption in yarn spinning using transfer learning," *Comput. Ind. Eng.*, vol. 152, Feb. 2021, Art. no. 107015.
[122] F. D. Simone and F. Amigoni, "Analysis of machine learning methods for anomaly detection of power consumption in buildings," M.S. thesis, Dept. Electron., Inf. Bioeng., Politecnico di Milano, Milan, Italy, 2021. [Online]. Available: https://www.politesi.polimi.it/handle/10589/183344
[123] R.-J. Hsieh, J. Chou, and C.-H. Ho, "Unsupervised online anomaly detection on multivariate sensing time series data for smart manufacturing," in *Proc. IEEE 12th Conf. Service-Oriented Comput. Appl. (SOCA)*, Nov. 2019, pp. 90–97.
[124] O. Serradilla, E. Zugasti, J. Rodriguez, and U. Zurutuza, "Deep learning models for predictive maintenance: A survey, comparison, challenges and prospects," *Appl. Intell.*, vol. 52, no. 10, pp. 10934–10964, Aug. 2022.
[125] L. Wen, L. Gao, and X. Li, "A new deep transfer learning based on sparse auto-encoder for fault diagnosis," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 1, pp. 136–144, Jan. 2019.
[126] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.
[127] C. Sun, M. Ma, Z. Zhao, S. Tian, R. Yan, and X. Chen, "Deep transfer learning based on sparse autoencoder for remaining useful life prediction of tool in manufacturing," *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2416–2425, Apr. 2019.
[128] A. Copiaco, Y. Himeur, A. Amira, W. Mansoor, F. Fadli, and S. Atalla, "Exploring deep time-series imaging for anomaly detection of building energy consumption," in *Proc. IEEE Asia–Pacific Conf. Comput. Sci. Data Eng. (CSDE)*, Dec. 2022, pp. 1–5.
[129] A. N. Sayed, Y. Himeur, and F. Bensaali, "Deep and transfer learning for building occupancy detection: A review and comparative analysis," *Eng. Appl. Artif. Intell.*, vol. 115, Oct. 2022, Art. no. 105254.
[130] M. Goldstein and S. Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data," *PLoS ONE*, vol. 11, no. 4, Apr. 2016, Art. no. e0152173.
[131] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Netw.*, vol. 106, pp. 249–259, Oct. 2018.
[132] K. A. Smith-Jentsch, E. Salas, and M. T. Brannick, "To transfer or not to transfer? Investigating the combined effects of trainee characteristics, team leader support, and team climate," *J. Appl. Psychol*, vol. 86, pp. 279–292, May 2001.
[133] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, Nov. 2011, pp. 513–520.
[134] W. Jiang, Y. Hong, B. Zhou, X. He, and C. Cheng, "A GAN-based anomaly detection approach for imbalanced industrial time series," *IEEE Access*, vol. 7, pp. 143608–143619, 2019.
[135] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *Proc. ICLR*, 2015, pp. 1–11.
[136] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
[137] M. Ijaz, G. Alfian, M. Syafrudin, and J. Rhee, "Hybrid prediction model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, synthetic minority over sampling technique (SMOTE), and random forest," *Appl. Sci.*, vol. 8, no. 8, p. 1325, Aug. 2018.
[138] S. Mokhtari, A. Abbaspour, K. K. Yen, and A. Sargolzaei, "A machine learning approach for anomaly detection in industrial control systems based on measurement data," *Electronics*, vol. 10, no. 4, p. 407, Feb. 2021.
[139] A. Kirillov, A. Kirillov, E. Mintun, N. Ravi, H. Mao, and C. Rolland, "Segment anything," *Proc. ICCV*, 2023, pp. 4015–4026.
[140] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10674–10685.
[141] A. Vlontzos, B. Kainz, and C. M. Gilligan-Lee, "Estimating categorical counterfactuals via deep twin networks," *Nature Mach. Intell.*, vol. 5, no. 2, pp. 159–168, Feb. 2023.
[142] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "When edge meets learning: Adaptive control for resource-constrained distributed machine learning," in *Proc. IEEE Conf. Comput. Commun.*, Apr. 2018, pp. 63–71.
[143] M. M. Amiri and D. Gündüz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 1432–1436.
[144] W. Ge and Y. Yu, "Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 10–19.
[145] J. Xu, H. Wu, J. Wang, and M. Long, "Anomaly Transformer: Time series anomaly detection with association discrepancy," in *Proc. ICLR*, 2022, pp. 1–20.
[146] I. Pintilie, A. Manolache, and F. Brad, "Time series anomaly detection using diffusion-based models," 2023, *arXiv:2311.01452*.
[147] T. Stadelmann, T. Klamt, and P. H. Merkt, "Data centrism and the core of data science as a scientific discipline," *Arch. Data Sci., A*, vol. 8, no. 2, pp. 1–16, Mar. 2022.
[148] P.-P. Luley, J. M. Deriu, P. Yan, G. A. Schatte, and T. Stadelmann, "From concept to implementation: The data-centric development process for AI in industry," in *Proc. 10th IEEE Swiss Conf. Data Sci. (SDS)*, Jun. 2023, pp. 73–76.
[149] Z. Y.-C. Liu, S. Roychowdhury, S. Tarlow, A. Nair, S. Badhe, and T. Shah, "AutoDC: Automated data-centric processing," in *Proc. NeuIPS*, Nov. 2021, pp. 1–6.
[150] D. Kang, N. Arechiga, S. Pillai, P. D. Bailis, and M. Zaharia, "Finding label and model errors in perception data with learned observation assertions," in *Proc. Int. Conf. Manage. Data*, Jun. 2022, pp. 496–505.
[151] M. Salem, S. Taheri, and J. S. Yuan, "Anomaly generation using generative adversarial networks in host-based intrusion detection," in *Proc. 9th IEEE Annu. Ubiquitous Comput., Electron. Mobile Commun. Conf. (UEMCON)*, Nov. 2018, pp. 683–687.

PENG YAN received the M.Sc. degree in informatics from the University of Zurich, Zürich, Switzerland, in 2022, where he is currently pursuing the Ph.D. degree in data science. He is a Research Assistant with the ZHAW School of Engineering, Winterthur, Switzerland. He is a member of the Centre for Artificial Intelligence and the Machine Vision, Perception and Cognition Group. His research interests include deep learning, time series analysis, computer vision, and intelligent algorithms. He is working on applying deep learning to industrial applications.

AHMED ABDULKADIR received the Doctor of Engineering degree from Albert-Ludwigs-Universität, Freiburg-im-Breisgau, Germany. He studied life sciences with École Polytechnique Féderale de Lausanne, Switzerland. He is currently a Senior Research Scientist with the Center of Artificial Intelligence, ZHAW School of Engineering, Winterthur, Switzerland. Before his appointment, he researched machine learning for medical and biological applications. He is also researching application-oriented methods for industrial and medical domains.

PAUL-PHILIPP LULEY received the B.Sc. degree in biomedical engineering from UAS Technikum Wien. He is currently pursuing the M.Sc. degree in computer science with the University of Zurich. He was Research Assistant at the Centre of Artificial Intelligence, ZHAW School of Engineering, Winterthur, Switzerland. Currently, he is a Research Scientist with Julius Bär Gruppe AG, specializing in robust deep learning techniques with a focus on MLOps. His research interests include the intersection of robust deep learning and MLOps, where he aims to develop cutting-edge solutions for operationalizing and maintaining machine learning models and ensuring their reliability and effectiveness in practical applications.

MATTHIAS ROSENTHAL received the Ph.D. degree from ETH Zürich, Switzerland, in 1997. Since 2014, he has been a Senior Lecturer at the ZHAW Zurich University of Applied Sciences. Currently, he is a Professor of Multiprocessor-Systems and the Head of the Research Area of Realtime-Platforms at the Institute of Embedded Systems, ZHAW Zurich University of Applied Sciences. He specialized in digital signal processing and led various research and development teams in the industry for several years. His research interests include multiprocessors, hybrid multicore systems, distributed digital signal processing, real-time embedded computing, and edge-AI.

GERRIT A. SCHATTE received the Ph.D. degree from TUM, experimentally investigating heat transfer to supercritical fluids. Following the Ph.D. degree, his interest in measuring systems led him to join Kistler Instrumente AG, the world market leader for dynamic mechanical measuring systems. Throughout his roles in research and industry, he has worked with quantitative methods in both research and development and production settings with a large variety of differently conditioned data depending on the application. As a Team Leader of business development/project management with the Innovation Laboratory, Kistler Instrumente AG, he oversees multiple projects to develop Kistler's data-based digital services and the corresponding business. He is currently a Mechanical Engineer with a passion for process measuring and control technology.

BENJAMIN F. GREWE received the Ph.D. degree in neuroinformatics from ETH Zürich, Switzerland, in 2010. He was a Postdoctoral Researcher in systems neuroscience with Stanford University, USA, from 2011 to 2016. He is currently a Professor with the Institute of Neuroinformatics, University of Zurich, and ETH Zürich. His research interests include the intersection of biological and artificial intelligence. His long-term vision is to extract fundamental principles of network learning from real biological networks and then reverse-engineer their functionality as logical, reproducible algorithms.

THILO STADELMANN (Senior Member, IEEE) received the Doctor of Science degree from Marburg University, Germany, in 2010, with a focus on multimedia analysis and voice recognition. He studied computer science in Giessen and Marburg. He held engineering and leadership roles in the automotive industry for several years before his appointment with the ZHAW School of Engineering, Winterthur, Switzerland. He is currently a Professor of AI/ML with the ZHAW School of Engineering, the Director of the ZHAW Centre for Artificial Intelligence, and the Head of the Machine Perception and Cognition Group, ZHAW School of Engineering. His current research interest includes robust deep learning to solve diverse pattern recognition tasks, such as document analysis or industrial or medical computer vision.