Engineering Research Express



PAPER

RECEIVED
19 September 2025

REVISED
21 November 2025

ACCEPTED FOR PUBLICATION
28 November 2025

PUBLISHED
9 December 2025

# SSKD: self-supervised knowledge distillation for reliable remote surface defect detection under wireless transmission attenuation

**Dong Wu**[1], **Peng Chen**[2,3,*], **Qingsheng Wei**[2], **Juxiao Ma**[2], **Junyu Qi**[4], **Chunhua Guo**[5], **Renpu Li**[6] and **Hai Jiang**[7]

[1] Department of Information Engineering, Changzhou Vocational Institute of Industry Technology, Changzhou, 213164, Jiangsu, People's Republic of China
[2] College of Engineering, Shantou University, Shantou, 515063, Guangdong, People's Republic of China
[3] Key Laboratory of Intelligent Manufacturing Technology, Ministry of Education, Shantou, 515063, Guangdong, People's Republic of China
[4] Electronics & Drives, Reutlingen University, Reutlingen, 72762, Germany
[5] Basic teaching department, Changzhou Vocational Institute of Industry Technology, Changzhou, 213164, Jiangsu, People's Republic of China
[6] School of Economics and Management, Changzhou Vocational Institute of Engineering, Changzhou, 213164, Jiangsu, People's Republic of China
[7] Jiangsu Lizhuo Information Technology Co. Ltd., 213000, Changzhou, Jiangsu, People's Republic of China
[*] Author to whom any correspondence should be addressed.

**E-mail:** pengchen@alu.uestc.edu.cn **and** dr.pengchen@foxmail.com

**Keywords:** ball screw drives, fault diagnosis, wireless transmission, image restoration, deep learning

## Abstract

Wireless sensor networks (WSNs) enable scalable, remote surface-defect inspection in smart manufacturing, but attenuation during wireless image transmission—manifesting as noise, compression artifacts, and packet loss—undermines diagnostic reliability. Existing solutions often assume high-fidelity inputs, rely on large paired datasets of pristine and degraded images, or require costly hardware upgrades, and they degrade sharply under adverse channel conditions. We propose a Self-Supervised Knowledge Distillation (SSKD) framework that reconstructs attenuated observations on the fly without historical pairings or rigid data correlations, avoiding hardware changes. SSKD couples a self-supervised learning architecture with cross-domain knowledge transfer to autonomously restore defect-salient cues and stabilize downstream detection. Across diverse attenuation scenarios and domains (including Ball Screw Drive monitoring), experiments show that SSKD consistently preserves diagnostic fidelity and robustness relative to strong baselines under severe degradation, enabling reliable remote inspection with commodity WSN deployments.

## 1. Introduction

The Industrial Internet of Things (IIOT) has enabled sensorized monitoring and predictive maintenance in machine tools [1]. In CNC systems, ball screw drives (BSDs) are essential for precise axis motion but operate in harsh conditions—contamination, overheating, and vibration accelerate wear and can degrade diagnostic reliability. Prior work has explored both traditional and deep learning–based fault detection for BSDs [2–5]. This paper focuses on robust BSD health assessment under realistic industrial variability to reduce downtime and improve maintainability.

BSDs are frequently regarded as typical components in machine tools, with numerous diagnostic and health monitoring techniques applied to them [6]. Most of these approaches can be broadly divided into physical model-based methods and data-driven methods. Many investigations have been performed on the dynamic modeling and vibration analysis of ball screw pairs [7–9]. These investigations aimed to identify potential sources of fault and deterioration as well as to better understand the underlying mechanisms of ball screw pair dynamics under varied operating scenarios. When examining the contact deformation process of a ball, nut, and screw shaft using the Hertz contact theory, Xu *et al* [7] discovered an association between the axial restoring

force and the axial displacement of a single ball. Duan *et al* [10] applied the Multi-Point Constraints (MPCs) approach to explore the mechanical joints of a twin ball screw feed system and evaluate its dynamic behavior. Chen *et al* [11] studied the characteristics of the dynamic contact and dynamic stiffness of ball screw pairs to figure out the relationship between the screw speed and the typical nut contact force. Through separating the predicted response and the actual response, physical modeling allows for the simulation and evaluation of the physical deterioration process of structural components [12]. Most parameters have clear physical meanings, and their selection directly influences the simulation's precision. Yet, due to the intricacy of the operation conditions and the physical structure, several parameters are extremely challenging to establish under the real-world engineering application.

In contrast to physical models, data-based technique, encompassing traditional machine learning and deep learning, are frequently adopted for the surveillance of industrial equipment health [13–15]. These traditional machine learning based techniques have proven to be effective in this domain and have continually evolved for enhanced performance. Huang *et al* [16], using an SVM, extracted features from inputs like vibration, current, speed, and encoders. Riaz *et al* [17] derived failure attributes of ball screw pairs leveraging a synergy of multi-class support vector machine (Mc-SVM) and knowledge base-Remnant-PCA (Kb-Rem-PCA). In order to monitor the preload loss of single nut ball screws, Denkena *et al* [18] explored into the potential of sensor fusion based on principal component analysis. Cross-sensor domain adaptation was investigated by Pandhare *et al* [19] for the purpose of identifying discrete differences in preload level and backlash. Deep learning-based techniques have been demonstrated to be efficient in assisting with the identification of possible concerns with industrial assets and enabling prompt maintenance and repair. A Convolutional Neural Network (CNN) is employed by Benker and Zaeh [20] to distinguish between various failure states and preload levels. Azamfar *et al* [21] conduct cross-domain classification for discrete ball screw health states through using highest mean discrepancy metric integrated in a CNN. Similarly, Li *et al* [22] examined towards how deep neural networks performed across domains, and they propose using class-level domain adaptation in addition to domain-level adaptation for improved results. Pandhare *et al* [19] developed a convolutional neural network-based domain adaptation approach for conducting defect detection on a ball screw pair. As previously stated, these methods involve data processing and analysis that isn't conducted on the machine physically, but rather at a different location where the machine's measurement data may be evaluated. These approaches rely heavily on the availability and quality of the collected machine data. If the data is not being properly recorded, it may be difficult to accurately detect defects or assess the condition of the machine. Moreover, these off-site approaches rely on gathering and analyzing data that has already been recorded, which prevents them from offering real-time feedback and analysis.

However, diagnostic methods for BSDs could benefit from leveraging data gathered across multiple ball screw pairs to enable rapid and efficient diagnosis, whether performed remotely or on-site. Additionally, the remote and onsite technique allows engineers to access the system and measure data without requiring physical access to the machine. The condition monitoring data can be transmitted through wired or wireless networks in a remote health evaluation system. The condition monitoring data can be transmitted through a wired or wireless communication network in a remote health evaluation system. Wireless communication systems, such those based on Wireless Sensor Networks (WSNs), provide the benefits of rapid deployment, quicker installation, and lower costs as compared with traditional communication systems [23]. Yet, WSNs could experience data loss issues because of radio interference, shoddy construction, improper antenna alignment, inclement weather, or long transmission distances [24].

The issue of data missing has been addressed using a variety of methods, including developing an improved communication protocols [13, 25] and increasing the hardware reliability of the WSNs [26]. However, these methods can only partially resolve the problem of data loss and may also increase hardware cost, system delay, and power consumption. Therefore, a missing data recovery technique should be taken into consideration to prevent data loss during wireless transmission. The existing missing data recovery technique has not been developed in the field of fault diagnosis and condition monitoring of BSDs. However, it has been investigated utilizing statistical [27], artificial intelligence-based [28], and filter-based methodologies [29] in other industrial applications like wind turbine. Yet, these methods have limitations, such as the requirement for a substantial portion of historical data that isn't truly available in practical applications, the utilization of strong data correlations, and the accumulation of prediction error.

Thus, the challenge of the aforementioned has been summarized as below.

(i) Data Loss During Wireless Transmission: While existing solutions like improved protocols and hardware can help, they increase costs and only partially solve the problem. Our Self-Supervised Knowledge Distillation(SSKD) approach enables effective image reconstruction even with substantial transmission data loss.

(ii) Historical Data Dependency and Strong Data Correlation Requirements: Current missing data recovery techniques require large amounts of historical data and strong data correlations, which are often unavailable in practical settings. Our SSKD methodology overcomes these limitations by operating without prior data or additional correlational data.

(iii) Lack of Missing Data Recovery Solutions for BSDs: While missing data recovery has been studied in other industrial applications, there was no solution specifically developed for fault diagnosis and condition monitoring of Ball Screw Drives. Our work is the first to address this gap with a specialized SSKD-based solution.

The Masked Auto Encoder (MAE) [30] is a self-supervised learning model that operates by masking random patches of an input image and then reconstructing the entire image from the unmasked portion. During training, it learns to predict the masked regions based on the visible parts, effectively understanding the global structure and context of images. Knowledge Distillation [31] is a model compression technique where a smaller model, known as the student, learns from a larger and more complex model, referred to as the teacher. The student model aims to mimic the behavior of the teacher by learning from its outputs, such as softened probability distributions over classes, thereby inheriting its knowledge while remaining computationally efficient for deployment. To address the challenges previously outlined, the research proposes a SSKD approach for addressing the problem of data loss in remote diagnostics and condition monitoring systems. This methodology is comprised of key elements, including an imaging component equipped with a camera system installed within the CNC machine, an SSKD mechanism utilizing a model that has been pretrained in a source domain and fine-tuned for the target setting to recover images, and a subsystem for detecting and diagnosing faults. The imaging is facilitated by a Raspberry Pi V2 micro-controller camera to gather the initial pictures. During the transmission process from the device through intermediary stages to cloud-based storage, there exists a risk of losing image pixels. Despite this, the methodology reconstructs images at the receiving end by managing pixel loss and recreating the original image using available data, even with significant transmission loss. In contrast with other data recovery methods, the proposed technique doesn't necessitate prior data. To the best of the authors' knowledge, this is the pioneering research to address data loss in remote fault diagnostics and condition monitoring for BSDs, introducing an innovative technique for fault diagnosis that demonstrates resilience to pixel loss without reliance on additional correlational data.

Contemporary advancements in Cross-Domain Knowledge Distillation (CDKD) have increasingly prioritized the mitigation of disparate feature representations between source and target domains, primarily by implementing sophisticated alignment strategies that rigorously address the underlying domain shift. Illustrating this trend, the KDFuse framework [32] addresses global compatibility by constructing an intermediate collaborative domain that facilitates coherent feature interaction, whereas complementary approaches like CroAtiSA [33] focus on granularity by designing fine-grained subdomain alignment losses to ensure precise sample-level adaptation. Furthermore, the integration of visual prompting with dynamic distillation mechanisms, as demonstrated by VPSP [34], offers a novel method to adaptively guide student networks; taken together, these distinct yet convergent methodologies substantiate the argument that efficacious CDKD requires a paradigm shift beyond superficial feature imitation toward more semantically robust and adaptive knowledge transfer protocols capable of operating across heterogeneous tasks.

The primary contributions of this paper can be summarized as follows:

(i) Paradigm Innovation in Self-Supervised Learning Architecture: The manuscript introduces a pioneering SSKD framework that fundamentally revolutionizes the conventional approach to missing data recovery in industrial diagnostics. By integrating knowledge distillation with self-supervised learning mechanisms, it establishes a novel theoretical foundation that transcends the traditional constraints of historical data dependency and correlation requirements.

(ii) Cross-Domain Knowledge Transfer Enhancement: The research advances the theoretical understanding of domain adaptation in industrial diagnostics by proposing a sophisticated pre-training and fine-tuning mechanism. This contribution bridges the gap between source and target domains in Ball Screw Drive monitoring, establishing a robust methodology for knowledge transfer that maintains diagnostic fidelity despite data degradation.

(iii) Autonomous Data Recovery Paradigm: The work establishes a groundbreaking theoretical framework for autonomous image reconstruction in industrial monitoring systems. By leveraging self-supervised learning principles, it introduces a novel approach to pixel-level data recovery that operates independently of external data correlations, fundamentally advancing the field of remote diagnostic systems in smart manufacturing environments.

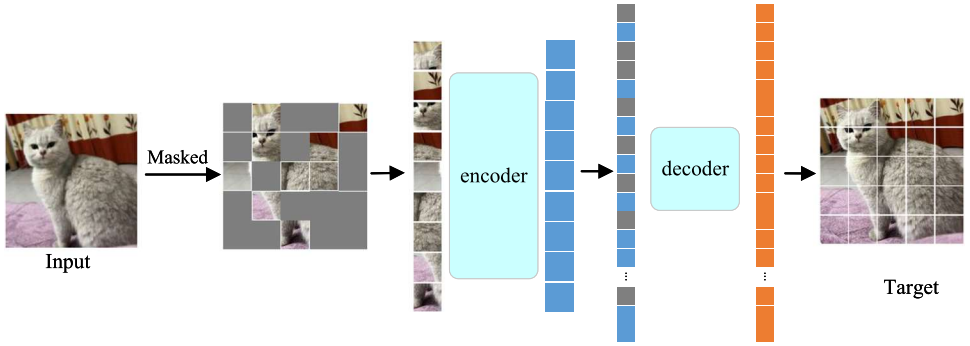


**Figure 1.** The architecture of masked autoencoder. © (2024) IEEE. Reprinted, with permission, from [9].

The structure of this research is outlined as follows: First, section 2 delves into the foundational theory of MAE. This is followed by section 3 which introduces the proposed self-supervised knowledge distillation approach. The application of this methodology for remote fault diagnosis in machine tool components is detailed in section 4. Within this, section 4.1 provides insights on the experimental test rig, and the data acquisition setup, while section 4.2 presents the experimental results related to the remote diagnosis of defective ball screw drives. The research culminates with conclusions in section 6.

## 2. Related theory

The field of remote diagnostics for machine tool elements, particularly BSDs, has seen significant advancements with the integration of the IIoT and deep learning techniques. The core challenge lies in ensuring reliable data transmission and accurate fault diagnosis, especially when using WSNs that are susceptible to data loss. To address these challenges, researchers have developed various methodologies that can be broadly categorized into physical model-based approaches [35] and data-driven methods [36]. Physical model-based approaches focus on understanding the mechanical behavior and dynamics of BSDs through theoretical models and simulations. These methods aim to predict potential faults based on the physical parameters and operating conditions of the system. However, they often require precise knowledge of the system's structure and can be limited by complex real-world conditions. Deep learning-based methods have been demonstrated to be effective in fault detection and condition monitoring. However, the data-driven approach is extremely dependent on both the quantity and quality of data.When the data is incomplete or missing, the accuracy of the model judgment will be greatly affected. The field of image processing and computer vision continues to make great strides, with the MAE [30] representing a breakthrough innovation in image inpainting. Based on the MAE model, the researchers have made some applications in fault diagnosis[37]. Image inpainting involves filling in missing or corrupted parts of an image, a challenging task that MAE tackles through its unique architecture. While traditional autoencoders commonly encode the full input image into a latent representation for tasks like compression and denoising, MAE's innovation lies in its selective encoding of only the visible, undamaged image patches. By exclusively encoding the available uncorrupted regions, MAE is forced to learn semantic representations capturing information about the complete image. This compact latent code is then fed to a lightweight decoder which, together with the visible patches, reconstructs the full undamaged image by reasonably filling in the missing areas. This clever approach allows MAE to leverage the visible patches to generate the complete image representation. To enable this mechanism, MAE utilizes a Vision Transformer (ViT) [38] as the backbone for encoding visible patches.

As depicted in figure 1, the holistic MAE architecture comprises three pivotal components: an encoder module primed for processing visible patches, a masking module for strategically masking out patches, and a decoder module responsible for generating a full-fledged image from the latent code and the remaining visible patches. The MAE's operation commences with the application of a mask onto the input image. By randomly setting a portion of the image's pixels to zero, a defined fraction of the image becomes obscured. The unobscured, non-zero pixels delineate the visible patches of the image. This collection of visible patches is then subject to encoding through the encoder module, resulting in a lower-dimensional latent representation. This latent representation is subsequently harnessed by the decoder module, alongside the visible patches, to intricately reconstruct the original image, encompassing even the sections that were originally masked out.

The Masked auto-encoders methodology represents an innovative stride in image inpainting, leveraging its ingenious design to elegantly address the challenge of missing or corrupted image segments. Through its

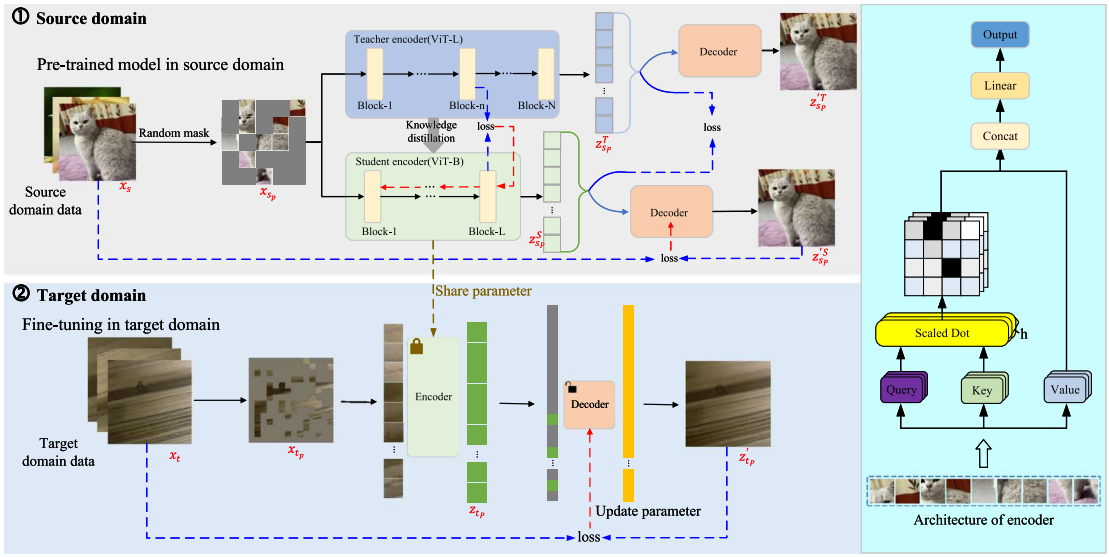


**Figure 2.** The proposed architecture of self-supervised knowledge distillation model. © (2024) IEEE. Reprinted, with permission, from [9].

strategic encoding of solely available, uncorrupted patches and a sophisticated decoder mechanism, MAE stands as a robust solution for image reconstruction in scenarios characterized by partial image damage.

## 3. The proposed self-supervised knowledge distillation methodology

The SSKD model, as depicted in figure 2, operates through a two-stage process: initially pre-training on the source domain data, specifically the ImageNet [39] dataset which contains over 14 million hand-labeled images in approximately 22,000 categories, and subsequently adapting to the target domain. This novel approach enhances the reliability of remote health condition monitoring for machine tool components, improving both efficiency and convenience. During the initial phase, the encoder harnesses knowledge distillation—a method wherein a smaller model is trained to replicate the feature-extraction capabilities of a more comprehensive counterpart. Recent advances in cross-modal knowledge distillation have demonstrated the effective transfer of knowledge from Vision Transformers to Convolutional Neural Networks for tackling complex vision tasks [40, 41]. Inspired by the efficacy of such cross-architecture learning, our SSKD framework adapts and extends this paradigm to address the unique challenge of wireless transmission attenuation in industrial diagnostics. As a result of extensive pre-training on the source database, the SSKD model excels in identifying underlying feature sets. Upon achieving satisfactory pre-training, the model's shared encoder and decoder adapt their weights for the target domain, rendering the system adept at generalizing across different sets of data pertinent to remote diagnostics. We adopt the notation summarized in table 1, which provides a comprehensive list of the mathematical symbols and variables used to describe the source and target domains, the shared encoder–decoder structure, the feature representations, and the associated loss functions within the SSKD framework. This table serves as a reference for all symbols appearing in the subsequent formulation and analysis of our method.

Specifically, the methodology collects images from the screw ball drive using the built-in camera monitoring system within the CNC machine. The image data is then wirelessly transmitted to a cloud server and subsequently to the remote user and operations control center, as depicted in figure 3. Domain adaptation is applied to the data to rectify any discrepancies, enabling accurate retrieval of images and measurements. Essentially, the proposed technique first learns robust representations via pre-training on source data, then adapts them to the target domain. This provides precise and reliable remote condition monitoring of machine tool components. The ability to remotely assess component states with precision presents immense potential for advancing machine tool maintenance and diagnostics.

### 3.1. Pre-trained process in source domain

The pre-training stage of the architecture is intricately designed, incorporating multiple essential components. These include the initial input image representation, embeddings for masked and unmasked patches, an encoder implementing knowledge distillation techniques, and a decoder for image reconstruction. During the knowledge distillation stage, a larger ViT is employed as the teacher model, while a smaller ViT serves as the

**Table 1.** Summary of mathematical symbols and notations used in the paper.

| Symbol | Definition |
|---|---|
| $\mathbf{x}_s$ | Input image from the source domain, $xs \in \mathbb{R}^{H\times W\times C}$ |
| $\mathbf{x}_t$ | Input image from the target domain |
| $\mathbf{x}_{s_p} = \{\mathbf{x}_{s_p}^i\}_{i=1}^N$ | Set of non-overlapping patches extracted from input image |
| $N$ | The number of tokens |
| $H, W$ | Height and width of the input image |
| $C$ | Number of color channels in the input image |
| $P$ | Patch size |
| $P'$ | Decoder patch size, where $P' > P$ |
| $\mathbf{M} = \{\mathbf{m}^1, \mathbf{m}^2, \ldots, \mathbf{m}^N\}$ | Random binary mask, $m_i \in \{0, 1\}$ |
| $\mathbf{m}^i \in \{0, 1\}$ | Indicates that patch $i$ is masked or unmasked |
| $\mathbf{x}_{sm}$ | Set of masked patches |
| $\mathbf{x}_{su}$ | Set of unmasked patches |
| $\mathbf{z}_{sm}$ | Embedding of masked patches |
| $\mathbf{z}_{sp}$ | Embedding of unmasked patches |
| $\mathbf{e}$ | Positional embedding vector, $e \in \mathbb{R}^D$ |
| $D$ | Dimensionality of the latent embedding space |
| $\mathbf{z}'_{sP}$ | Latent representation output by the encoder for unmasked patches |
| $W_E, b_E$ | Learnable weight matrix and bias vector of the encoder |
| $g(\cdot)$ | Activation function used in the encoder |
| $f_t$ | Teacher model (pre-trained Vision Transformer, e.g., ViT-L) |
| $f_s$ | Student model (lighter Vision Transformer, e.g., ViT-B) |
| $\mathbf{Q}^h, \mathbf{K}^h, \mathbf{V}^h$ | Query, Key, and Value matrices for head $h$ in multi-head self-attention |
| $\mathbf{W}^{Q^h}, \mathbf{W}^{K^h}, \mathbf{W}^{V^h}$ | Learnable projection matrices for queries, keys, and values |
| $d_k$ | Dimension of the key vectors, $d_k = D/H$ |
| $H$ | Number of attention heads |
| $A$ | Attention weight matrix, computed via softmax |
| $O$ | Output of a single attention head |
| $O_h$ | Final output of the multi-head self-attention layer |
| $W_o$ | Output projection matrix for concatenating attention heads |
| $\mathbf{O}'$ | First linear transformation output in MLP sub-layer |
| $\mathbf{O}''$ | Final output of MLP sub-layer after GELU activation |
| $\mathbf{W}_1, \mathbf{W}_2$ | Learnable weight matrices in the MLP sub-layer |
| $\mathbf{b}_1, \mathbf{b}_2$ | Bias terms in the MLP sub-layer |
| $GELU(\cdot)$ | Gaussian Error Linear Unit activation function |
| $\mathbf{L_{KL}}(\mathbf{p}, \mathbf{t})$ | Kullback-Leibler divergence loss between teacher and student outputs |
| $\mathbf{t}$ | Softened probability distribution (target) from the teacher model |
| $\mathbf{p}$ | Prediction from the student model |
| $\mathbf{c_t}, \mathbf{c_s}$ | Class token features from teacher and student models, respectively |
| $\mathbf{L}_{ck}$ | Class token distillation loss |
| $\mathbf{Ft}$ | Feature representation from the teacher network |
| $\mathbf{Fs}$ | Feature representation from the student network |
| $\mathbf{L}_{fd}$ | Feature distillation loss |
| $Normal(\cdot)$ | Whitening operation using layer normalization |
| $\mathbf{L}_1(\cdot, \cdot)$ | Smooth L1 loss function |
| $\beta$ | Threshold parameter in smooth L1 loss, typically set to 2.0 |
| $MSE$ | Mean Squared Error loss used in fine-tuning stage |

student model. The objective is to facilitate the training of a randomly initialized student model by replicating the teacher model's output through knowledge distillation. After the stage of encoder training, a decoder is utilized to decode the vector derived from the encoder, aiding in image recovery. This sub-module initializes the learning process by leveraging a pre-trained model on a large, diverse dataset like ImageNet. The extensive training on such datasets enables the model to learn rich feature representations that are generalizable across various domains. It forms the foundational knowledge base upon which subsequent fine-tuning is built.

(a) **Input image representation**

For a given input image $\mathbf{x}_s \in \mathbb{R}^{H\times W\times C}$ from the source domain, it is tokenized into non-overlapping patches as $\mathbf{x}_{s_p} = \{\mathbf{x}_{s_p}^i\}_{i=1}^N$, where $\mathbf{x}_{s_p}^i \in \mathbb{R}^{N\times(P^2C)}$. Here, $H$, $W$ are image height and width, $N$ is the number of tokens, $C$ is the number of channels, and $P$ is the patch size. A random binary mask $\mathbf{M}$ is defined as $\{\mathbf{m}^1, \mathbf{m}^2, \ldots, \mathbf{m}^N\}$, where $\mathbf{m}^i \in \{0, 1\}$.

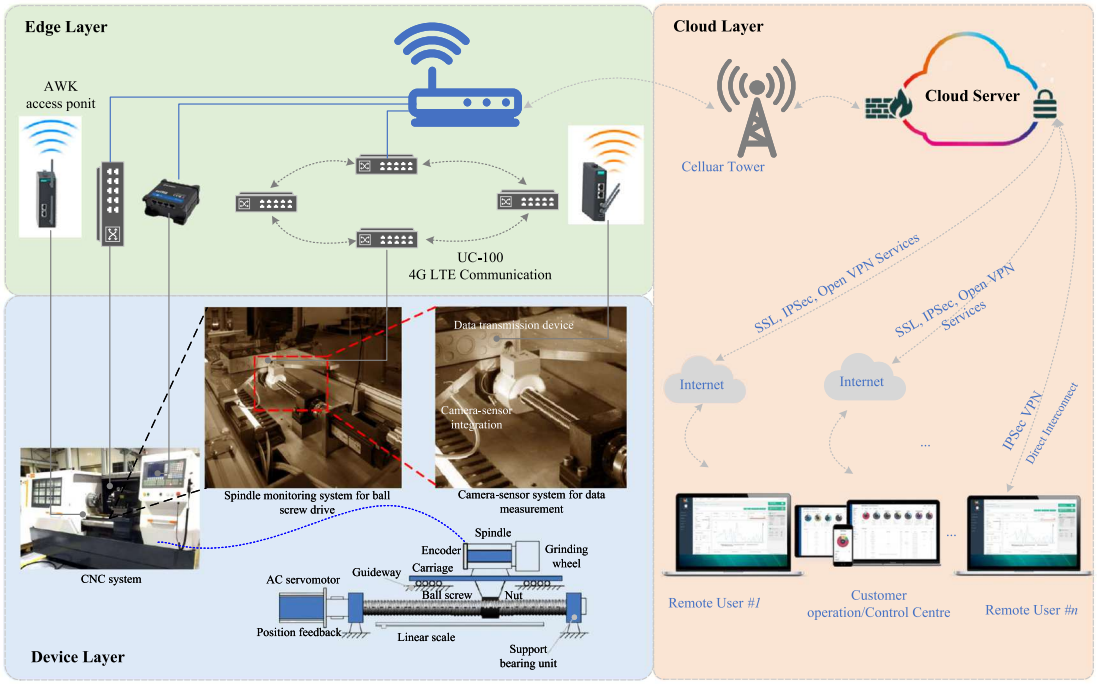


**Figure 3.** Test rig schematic and configuration overview. © (2024) IEEE. Reprinted, with permission, from [9] and © (2025) IEEE. Reprinted, with permission, from [15].

(i) **Masked patch embedding**

The masked patches $\mathbf{x}_{s_m}$ are flattened into a vector $\mathbf{x}_{s_m}^F \in \mathbb{R}^{P^2C}$, and are replaced with the mask token embedding $\mathbf{m} \in \mathbb{R}^D$:

$$\mathbf{z}_{s_m} = \{\mathbf{x}_{s_p}^i | \mathbf{m}^i = 1\}_{i=1}^N \tag{1}$$

where $\mathbf{x}_{s_p}$ represents all the patches of the input image, and the patch $\mathbf{x}_{s_p}^i$ is masked if the corresponding $\mathbf{m}^i = 1$. Following this, positional embeddings $\mathbf{e} \in \mathbb{R}^D$ are added to $\tilde{\mathbf{z}}_{s_p} = \mathbf{z}_{s_p} + \mathbf{e}$.

(ii) **Unmasked patch embedding**

The unmasked patches $\mathbf{x}_{s_u}$ are flattened into a vector $\mathbf{x}_{s_u}^F \in \mathbb{R}^{P^2C}$ and projected into an embedding $\mathbf{z}_{s_p} \in \mathbb{R}^D$:

$$\mathbf{z}_{s_P} = \{\mathbf{x}_{s_p}^i | \mathbf{m}^i = 0\}_{i=1}^N \tag{2}$$

where $\mathbf{x}_{s_p}$ represents all the patches of the input image, and the patch $\mathbf{x}_{s_p}^i$ is unmasked if the corresponding $\mathbf{m}^i = 0$.

(b) **Encoder**

The encoder processes unmasked patches of the image to generate a latent representation that captures the underlying semantic information. It is crucial for encoding the visible parts of the image into a compact yet informative feature space. The unmasked patches $\mathbf{z}_{s_p}$ are fed into the encoder to obtain a latent representation $\mathbf{z}'_{s_p} \in \mathbb{R}^D$:

$$\mathbf{z}'_{s_p} = g(\mathbf{W}_E \mathbf{z}_{s_P} + \mathbf{b}_E) \tag{3}$$

where $\mathbf{W}_E \in \mathbb{R}^{D \times HWC}$ and $\mathbf{b}_E \in \mathbb{R}^D$ are learned weight and bias parameters, $g(\cdot)$ is the activation function, and $D$ is the dimension of $\mathbf{z}'_{s_p}$. The ViT framework is adopted by both the teacher model and the student model. In general, the student network has a lower number of heads compared to the teacher network. For instance, the teacher model utilizes the Vision Transformer Large (ViT-L), incorporating 16 heads, while the student Model utilizes the Vision Transformer Base (ViT-B), which includes 12 heads. The encoder adopts a multi-head self-attention structure.

(i) *Multi-head self-attention (MHSA) sub-layer.* For each head $h$:

$$\begin{aligned}\mathbf{Q}^h &= \mathbf{Z}\mathbf{W}^{Q^h}\\ \mathbf{K}^h &= \mathbf{Z}\mathbf{W}^{K^h}\\ \mathbf{V}^h &= \mathbf{Z}\mathbf{W}^{V^h}\end{aligned} \tag{4}$$

where $\mathbf{Z}$ is the input feature, $\mathbf{W}^{Q^h}$, $\mathbf{W}^{K^h}$, $\mathbf{W}^{V^h} \in \mathbb{R}^{D\times D_h}$ are learned projections, and $D_h = D/H$ for $H$ heads. Compute the attention scores between all pairs of positions in the sequence by taking the dot product of the queries and keys, and applying a softmax function to obtain attention weights:

$$A = \text{softmax}\left(\frac{Q^h K^{hT}}{\sqrt{d_k}}\right) \tag{5}$$

where $d_k$ is the dimensionality of the keys.Compute the output feature $O$ as a weighted sum of the values, where the weights are given by the attention scores:

$$O = AV^h \tag{6}$$

In the case of multi-head self-attention, this process is performed multiple times, each with its own set of weight matrices $W^Q$, $W^K$, and $W^V$. The outputs of the different attention heads are concatenated and linearly transformed to produce the final output sequence $O_h$.

$$O_h = \text{Concat}(\text{A}_1, \text{A}_2, \ldots, \text{A}_h)\, W_o \tag{7}$$

where $W_o$ is output weight matrix. It is a weight matrix of shape $d_{v'} \times d_v$ used to project the concatenated outputs of all attention heads back to the original dimensionality.

(ii) *MLP sub-layer propagates information:*

$$\begin{aligned}\mathbf{O}' &= \mathbf{O}\mathbf{W}_1 + \mathbf{b}_1\\ \mathbf{O}'' &= \text{GELU}(\mathbf{O}')\mathbf{W}_2 + \mathbf{b}_2\end{aligned} \tag{8}$$

where $\mathbf{W}_1 \in \mathbb{R}^{D\times D_F}$, $\mathbf{W}_2 \in \mathbb{R}^{D_F\times D}$, $\mathbf{b}_1 \in \mathbb{R}^{D_F}$, $\mathbf{b}_2 \in \mathbb{R}^D$ are learned weights and biases.

(c) **Knowledge Distillation**

In the knowledge distillation stage, various distillation techniques are integrated to correspond to the outputs from different stages within the model framework. This exploration highlights several knowledge distillation loss functions, each crafted to fulfill specific aims of the distillation process. This mechanism facilitates the transfer of knowledge from a larger teacher model to a smaller student model. The teacher model, having been trained extensively, provides guidance through its predictions or intermediate representations, allowing the student model to mimic its behavior effectively. Consider an input image symbolized by $x$, where $f_t$ represents the teacher model, and $f_s$ denotes the student model. The core of knowledge distillation involves transferring knowledge from $f_t$ to $f_s$ by optimizing $f_s$ while maintaining the static state of $f_t$. Primarily, this training process is governed by the Kullback-Leibler $KL$ divergence, expressed as:

$$\mathcal{L}_{KL}(\boldsymbol{t}, \boldsymbol{p}) = \boldsymbol{t}\log\frac{\boldsymbol{t}}{\boldsymbol{p}} \tag{9}$$

where $\boldsymbol{t}$ denotes the target generated by $f_t(x)$, and $\boldsymbol{p}$ is the prediction produced by $f_s(x)$. When applying Class Token Distillation to the encoder output, the class token features of $f_s$ and $f_t$ are denoted as $\boldsymbol{c_s}$ and $\boldsymbol{c_t}$, respectively. The loss function for class token distillation $\mathcal{L}_{ck}$ is as follows:

$$\mathcal{L}_{ck} = \mathcal{L}_{KL}(\boldsymbol{c_t}, \boldsymbol{c_s}) \tag{10}$$

In order to tackle the prevalent challenge of mismatched feature dimensions between the teacher network and the student network, an extra linear layer is incorporated into the student network's output. This additional layer aims to synchronize the feature dimension of the student's prediction (referred to as $\boldsymbol{F_s}$) with the desired feature (referred to as $\boldsymbol{F_t}$) of the teacher network. Consequently, the formulation of the loss function for feature distillation $\mathcal{L}_{fd}$ can be expressed in the following manner:

$$\mathcal{L}_{fd} = \mathcal{L}_1(\boldsymbol{F_t}, \boldsymbol{F_s}) \tag{11}$$

where $Normal(\cdot)$ is the whitening operation implemented by layer norm without affiliation, and $\mathcal{L}_1$ is the smooth $L_1$ loss defined as:

$$\mathcal{L}_1(y, \hat{y}) = \begin{cases} \frac{1}{2}(\hat{y} - y)^2/\beta, & |\hat{y} - y| \leqslant \beta \\ \left(|\hat{y} - y| - \frac{1}{2}\beta\right), & \text{otherwise} \end{cases} \tag{12}$$

where $\beta$ is set to 2.0.

(d) **Decoder**

The decoder takes the latent representation produced by the encoder and reconstructs the full image, including the masked-out regions. It integrates the visible patches with the generated latent code to produce a complete and coherent image. The decoder follows a similar architecture but with larger patch size $P' > P$. It takes as input the concatenated latent representations of the unmasked patches from the encoder along with the shared mask token embedding:

$$\mathbf{Z}'_s = \mathbf{z}'_{s_p} \oplus \mathbf{z}_{s_m} \tag{13}$$

where $\oplus$ represents the concatenation operation. $\mathbf{z}'_{s_p}$ represents the unmasked patch embeddings from the encoder output and $\mathbf{z}_{s_m}$ denotes the masked patch embeddings. The decoder Transformer blocks aim to reconstruct the original image dimensions by first flattening the reconstructed patch embeddings $\hat{\mathbf{z}}_{sp} \in \mathbb{R}^D$ into vectors $\hat{\mathbf{z}}^F_{sp} \in \mathbb{R}^{P'^2C}$ and then reshaping into patches.

A final linear projection layer maps the reconstructed patches to generate the output image:

$$\hat{\mathbf{x}}_s = \mathbf{W}_D[\mathbf{Z}'_s] + \mathbf{b}_D \tag{14}$$

where $\mathbf{W}_D \in \mathbb{R}^{HWC \times NP'^2C}$ and $\mathbf{b}_D \in \mathbb{R}^{HWC}$. The model is trained end-to-end to minimize the reconstruction loss between $\mathbf{x}_s$ and $\hat{\mathbf{x}}_s$.

$$\mathcal{L}_{\text{rec\_s}} = ||\mathbf{x}_s - \hat{\mathbf{x}}_s||_2^2 \tag{15}$$

### 3.2. Domain adaptation in target scenario

This module adjusts the model's parameters to adapt to the specific characteristics of the target domain, compensating for differences in data distribution between the source and target environments. By employing the aforementioned pre-training approach of knowledge distillation, a high-capacity ViT with powerful representation capabilities can be achieved, as well as a compact ViT model that has been trained based on the former. The subsequent task involves transferring the resulting compact ViT to the target domain. The key idea is to use the learned representations from the source domain to initialize the model, and then further fine-tune the model on target domain data. This allows the model to leverage the features learned from ImageNet in the source domain, while also adapting to the unique characteristics of the target domain. Fine-tuning the pre-trained weights enables transferring knowledge from the large labeled ImageNet dataset to the target domain, providing an initialization for the model before optimizing it for the target data.

(a) **Domain adaptation**

The target domain images are processed in a manner analogous to the source domain. The adaptation process can be summarized in the following steps:

(I) **Input image representation in target domain**

Given an input image $\mathbf{x}_t \in \mathbb{R}^{H \times W \times C}$ in target domain, it is divided into $N$ non-overlapping patches $\mathbf{x}_{t_p} \in \mathbb{R}^{P \times P \times C}$. Subsequently, a subset of these patches, $\mathbf{x}_{t_m}$, is masked by setting their values to zero, mimicking the pre-processing done in the source domain.

(II) **Masked and unmasked patch embedding in target domain**

The masked and unmasked patches are then transformed into token embeddings in the same way as was done for the source domain data.

(i) **Masked patch embedding**

The masked patches $\mathbf{x}_{t_m}$ are flattened and converted into embeddings $\mathbf{z}_p$. This embedding is initialized with a mask token embedding $\mathbf{s}_m$, which indicates the absence of original information. The pre-trained

encoder is used to encode the embeddings from the target domain data. The weights of the encoder have been pre-trained on the source domain data and are now being fine-tuned on the target domain data.

(ii) **Unmasked patch embedding**

Unmasked patches are similarly flattened and projected into embedding $\mathbf{z}_p$ using a linear transformation. These embeddings contain information about the image's actual content. Both embeddings have added positional embeddings $\mathbf{e}$ to account for the spatial layout of the patches within the image.

(III) **Encoder and decoder adaptation**

The pre-trained encoder, initially trained on $\mathbf{x}_s$, is now fine-tuned on the target domain data. The token embeddings are fed into the encoder, which consists of Transformer blocks with both MHSA and MLP sub-layers. Within the MHSA, the token embeddings are transformed into query, key, and value projections. The scaled dot-product attention mechanism computes relationships across multiple attention heads, capturing long-range dependencies and intricate patterns. The MLP sub-layer further refines these embeddings, using a feed-forward network enriched with GELU activation, which introduces non-linearity. The encoder's role is to capture both local and global information from the image, producing a set of latent representations. The pre-trained decoder, which reconstructs the image from the latent representations, is also fine-tuned. The decoder receives the latent representations of unmasked patches along with a shared mask token embedding. Using its Transformer blocks, the decoder attempts to predict the content of the masked patches, leveraging the global context information it has. The reconstructed patches are then converted back into image format. This involves flattening the embeddings into vectors and reshaping them into patches. A final linear projection then generates the output image. The decoder reconstructs the image from the encoder's latent representations. The goal is to predict the content of the masked patches using the global context. The weights of the decoder have been pre-trained on the source domain data and are now being fine-tuned on the target domain data.

(b) **Fine-tuning the model in the target domain**

Since the data in the source domain and target domain may have different distributions, the model may not perform as expected on the target domain due to these differences. The goal of fine-tuning in the target domain is to adapt the model to these differences, thereby improving its performance on the target domain data. The idea is to leverage the shared features and representations learned from the source domain while also learning the unique characteristics of the target domain. A domain alignment step should be implemented, which involves aligning the source domain and target domain in a common latent space. This can be achieved using techniques like Maximum Mean Discrepancy (MMD). Specifically, if $\mathbf{Z}_s$ and $\mathbf{Z}_t$ are the encoded representations of the source and target domains respectively, the aim is to minimize the difference between their distributions:

$$\mathcal{L}_{\text{domain}} = \text{MMD}(\mathbf{Z}_s, \mathbf{Z}_t) \tag{16}$$

Once the domain alignment is achieved, the next step involves fine-tuning the model on a subset of labeled target domain data. This helps in making slight adjustments to the model such that it starts to understand the nuances of the target domain. The objective here would be to minimize the reconstruction loss for the target domain data:

$$\mathcal{L}_{\text{rec\_t}} = ||\mathbf{x}_t - \hat{\mathbf{x}}_t||_2^2 \tag{17}$$

Consistency regularization can be introduced to ensure that the model predictions remain consistent across small perturbations in the input. For instance, if $\mathbf{x}_t'$ is a slightly perturbed version of $\mathbf{x}_t$, the reconstructed outputs should be similar:

$$\mathcal{L}_{\text{consistency}} = ||\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_t'||_2^2 \; \mathcal{L}_{\text{consistency}} \tag{18}$$

The overall objective for adaptation can be a combination of the aforementioned losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec\_t}} + \lambda_1 \mathcal{L}_{\text{domain}} + \lambda_2 \mathcal{L}_{\text{consistency}} \tag{19}$$

where $\lambda_1$ and $\lambda_2$ are hyper-parameters that control the importance of the domain alignment and consistency losses respectively.

### 3.3. Domain alignment in self-supervised knowledge distillation

The successful adaptation of machine learning models from a source domain to a target domain is a critical challenge, especially when the data distributions differ significantly. In this chapter, we delve into the domain alignment strategies employed in the SSKD method. This process ensures that the model trained on a large, general dataset like ImageNet can effectively generalize to specialized tasks such as recovering images with missing pixels from BSD in CNC machines. We will explore the theoretical foundations, practical implementation, and evaluation metrics used to achieve robust domain alignment.

*3.3.1. Theoretical foundations*

**Domain discrepancy:** Differences in data distributions between source and target domains can hinder model performance. Addressing these discrepancies ensures better adaptation. **Feature-Level Alignment:** Techniques like Maximum Mean Discrepancy (MMD), adversarial training, and instance normalization help align feature representations across domains, making them more invariant to distributional shifts. **Self-Supervised Learning:** Leveraging unlabeled data through tasks like masked autoencoding enables the model to learn meaningful representations without explicit labels, which is vital for handling incomplete BSD images.

*3.3.2. Practical implementation*

The SSKD method employs a structured approach to domain alignment:

(i) **Pre-training phase**: Train the teacher and student models on the source domain using knowledge distillation and masked autoencoding tasks. Optimize hyper-parameters such as learning rates, batch sizes, and masking ratios during this phase.

(ii) **Initialization for Target Domain**: Initialize the encoder weights of the student model with pre-trained weights from the source domain. Ensure effective mask token embedding for BSD images.

(iii) **Adaptive fine-tuning**: c

- Using a learning rate schedule(linear learning rate scaling) that starts high for rapid adaptation and gradually lowers for fine-tuning.
- Gradually unfreezing layers, starting with the decoder and top encoder layers.
- Incorporating domain adversarial loss and reconstruction loss to encourage feature-level alignment and accurate image recovery.

(iv) **Iterative refinement**: Continuously evaluate and refine the model using a validation set from the target domain. Quantify distributional closeness using metrics like MMD or Wasserstein distance and adjust normalization layers and activation functions as needed.

The methodology outlined in algorithm 1 progresses through a series of steps, divided into two distinct stages. The initial stage, comprising steps (1) to (11), is dedicated to the pre-training of the image retrieval model using data from the source domain. The subsequent stage, encompassing steps (12) to (22), deals with the fine-tuning of the model within the target domain and employs transfer learning techniques specifically adapted to the ball screw dataset.

**Algorithm 1.** Self-supervised knowledge distillation framework

**Require:** Source domain data $\mathcal{D}_s = \{(\mathbf{x}_s^i)\}_{i=1}^{n_s}$, where $\mathbf{x}_s^i$ is the input image
**Require:** Target domain data $\mathcal{D}_t = \{(\mathbf{x}_t^j)\}_{j=1}^{n_t}$, where $\mathbf{x}_t^j$ is the input image
**Require:** Pre-trained model $f_\theta(\cdot)$ with teacher encoder (ViT-L), student encoder (ViT-B), and shared decoder
**1: Pre-training on source domain:**
**2:** Initialize the model $f_\theta(\cdot)$ with random weights
**3: for** $(\mathbf{x}_s^i) \in \mathcal{D}_s$ **do**
**4:** Extract the features of the teacher encoder $F_t$ and student encoder $F_s$
**5:** Encode $\mathbf{x}_s^i$ using the teacher encoder (ViT-L) to obtain $\mathbf{z}_{sp}^{\prime T}$
**6:** Encode $\mathbf{x}_i^s$ using the student encoder (ViT-B) to obtain $\mathbf{z}_{sp}^{\prime S}$
**7:** Pass $\mathbf{z}_{sp}^{\prime T}$ and $\mathbf{z}_{sp}^{\prime S}$ to the shared decoder
**8:** Compute the knowledge distillation loss $\mathcal{L}_{\mathrm{KD}} = \mathcal{L}_{\mathrm{KL}}(\mathbf{z}_{sp}^{\prime T}, \mathbf{z}_{sp}^{\prime S}) + \mathcal{L}_1(Normal(\boldsymbol{F_t}), \boldsymbol{F_s})$
**9:** Compute the reconstruction loss $\mathcal{L}_{\mathrm{rec\_s}}(\mathrm{Decoder}(\mathbf{z}_{sp}^{\prime S}), y_i^s) = ||\mathrm{Decoder}(\mathbf{z}_{sp}^{\prime S}) - \mathbf{x}_s^i||_2^2$
**10:** Update the model parameters $\theta$ using the combined loss $\mathcal{L} = \mathcal{L}_{\mathrm{KD}} + \lambda \mathcal{L}_{\mathrm{rec\_s}}$
**11: end for**
**12: Fine-tuning on target domain:**
**13:** Freeze the student encoder (ViT-B) for fine-turning on target domain
**14: for** $(\mathbf{x}_t^j) \in \mathcal{D}_t$ **do**
**15:** Encode $\mathbf{x}_t^j$ using the frozen part of the student encoder (ViT-B) to obtain $\mathbf{z}_t^j$
**16:** Pass $\mathbf{z}_t^j$ to the decoder
**17:** Compute the gap between the source and target domains $\mathcal{L}_{\mathrm{domain}}$
**18:** Compute the reconstruction loss $\mathcal{L}_{\mathrm{rec\_t}}(\mathrm{Decoder}(\mathbf{z}_s^j), \mathbf{x}_t^j) = ||\mathrm{Decoder}(\mathbf{z}_s^j) - \mathbf{x}_t^j||_2^2$
**19:** Compute the consistency of source and destination decoders $\mathcal{L}_{\mathrm{consistency}}$
**20:** Compute total loss $\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{rec\_t}} + \lambda_1 \mathcal{L}_{\mathrm{domain}} + \lambda_2 \mathcal{L}_{\mathrm{consistency}}$
**21:** Update the parameters of the decoder using $\mathcal{L}_{\mathrm{total}}$
**22: end for**
**23: return** Fine-tuned model $f_\theta'(\cdot)$

## 4. Case study on remote diagnosis of machine tool components

This section presents case studies on the fault diagnosis of machine tool components, illustrating the effectiveness of the proposed SSKD methodology. Initially, section 4.1 elaborates on the experimental test rig and the data acquisition setup, which are instrumental in gathering data from the monitored machine tool components. Subsequent to the detailed exposition of the experimental framework, section 4.2 analyzes the results derived from the deployment of the SSKD on the acquired dataset, emphasizing its reliability in fault diagnosis. The case studies specifically focus on remote condition monitoring of machine tools, demonstrating how the proposed method adeptly utilizes pixel-level loss adaptations learned during the process. This capability enables the precise identification of component faults through data captured remotely. The comprehensive analysis, spanning from setup to results, provides solid validation of the SSKD's efficiency. It demonstrates that the methodology not only enhances the robustness but also ensures reliable fault diagnosis in real-world remote monitoring environments.

### 4.1. Test rig and configuration for experimental data acquisition setup

The experimental data acquisition setup consists of three main layers, as illustrated in figure 3: the device layer, edge layer, and cloud layer. The device layer collects images from the screw ball drive using a built-in camera monitoring system within the CNC machine. The image data is then wirelessly transmitted to a cloud server and subsequently relayed to the remote user and operations control center. The edge layer aggregates and pre-processes the image data streamed from the machine. Finally, the cloud layer provides computing resources for further data analysis and storage. This section details each of these layers and their interconnections that enable collection of the experimental images, as outlined in the architecture diagram in figure 3. Specifically, the composition of the device layer is described, along with the functionality of the edge device for data handling and the role of the cloud back-end for storage and analysis.

The device layer of the CNC machine comprises various components, including the AC servomotor, guide-way, ball screw, spindle, tool, and carriage. These components form the typical structure of a spindle and car-riage-ball screw system. The architecture of the device layer, depicting these specific machining components, is illustrated in figure 3. The servomotor drives the ball screw to move the carriage and tool. The spindle rotates the workpiece while the tool machines it. The guideway provides support for linear motion of the carriage. Together, these elements enable the machining operations performed by the CNC machine.

The edge layer plays a crucial role in the remote transmission of sensor images obtained from the device layer. This layer is composed of various components, such as an AWK access point for connectivity within the local network, a UC-100 4G LTE communication module for extended range wireless transmission, and others, as depicted in the 'edge layer' section of figure 3. Sensor images from the device layer are collected at the edge device, which might be an industrial PC or gateway. This data is then wirelessly relayed using the AWK access point within the local network domain. For transmitting the data over larger distances to the cloud servers, the 4G LTE module facilitates cellular communication. By handling the collection, processing, and transmission of images, the edge layer establishes an essential link between physical devices and cloud-based services. This architecture obviates the need for direct access to the sensors, thereby facilitating cloud-level analysis. The components and data flow within the edge layer are depicted graphically in the 'edge layer' section of figure 3.

In the cloud layer, a connection is established between the cellular tower and cloud server, enabling secure transmission of image data to remote users through services like SSL, IPsec, and OpenVPN. As a result, remote users can securely access and receive the transmitted images. Additionally, the Customer Operation Control Center receives a continuous information flow through a direct interconnect using an IPsec VPN. This setup ensures secure, real-time communication, granting the control center instant access to the data. This seamless access plays a vital role in maintaining efficient operations and enabling well-informed decision-making within the organization. The cloud layer connectivity establishes the critical link between the edge layer and remote stakeholders, facilitating secure data transmission and instantaneous monitoring of machine conditions.

As illustrated in figure 3, the camera system integrated within the device layer has generated the image measurement results depicted in figure 4. This figure highlights a spindle displaying clear signs of surface spalling, otherwise known as pitting. This particular spindle has been subjected to a series of durability tests, managing to meet its projected service life entirely, which consequently led to the identification of preload loss. Areas experiencing spalling are underscored by the red dashed markers on the ball screw. The progression of the spindle's damage is recorded via the image data. Understanding the position of the image on the ball screw, the resolution of the image, and the geometric characteristics of the ball screw, facilitates a meticulous evaluation of the ball screw spindle's current state. Merging this image data with in-depth knowledge concerning the wear patterns that take place on ball screws and the signs of their deficiencies, enables the detection of the fault's exact location, type, and intensity. This augmented information strengthens the potential of remote fault diagnosis, providing an inclusive perspective of the system's degradation and damage.

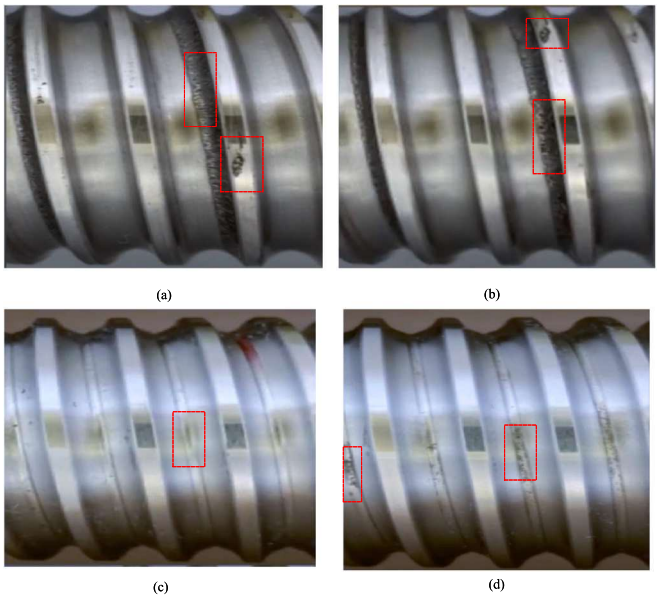


**Figure 4.** The sections of screw ball drives exhibiting pitting and flaking damage.

### 4.2. Experimental results and analysis

In order to assess the efficacy of the proposed SSKD methodology, this research utilizes two ball screw drive systems labeled as #1 and #2. These systems are employed to demonstrate the efficiency of the proposed approach in detecting and diagnosing faults across various degrees of pixel loss scenarios. The visual representations provided in the sections referred to as Case I and Case II (sections 4.2.1 and 4.2.2 respectively) serve as illustrative examples, demonstrating the results discussed in the text.

To illustrate the effectiveness of the proposed methodology for image data retrieval and damage detection, an initial quantitative comparative analysis is conducted. This analysis employs indicators such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM), along with the model's capability to detect surface damage, which is assessed using the mean Intersection over Union (mIoU) as the evaluative metric. The focus is specifically on addressing the challenges associated with remote monitoring of machine tool health, particularly the issues of pixel data loss during image transmission. The study examines the performance across various Pixel Loss Rates (PLR), spanning from 5% to 75%. This approach provides a structured evaluation of the effectiveness of the proposed method under different conditions of image degradation.

The evaluation of the effectiveness of various image retrieval methods, such as MAE (ViT-B) [30], MAE (ViT-L) [30], CAE (ViT-B) [42], PeCo (ViT-B) [43],LocalMIM [44], CL-MAE [45] and PCP-MAE [46] is crucial for applications including image retrieval and surface damage detection. For a comprehensive analysis, each method is trained using an Nvidia RTX3090Ti GPU with CUDA 11.4 and implemented in PyTorch 1.10. Specifically, During the pre-training phase, we utilized the ImageNet dataset as the training set and adopted the ViT-L model with layer depths of ($L = 2, 4, 22, 24$) as the teacher model and the ViT-B model with layer depths of ($L = 2, 4, 10, 12$) as the student model. An initial learning rate of $1e-4$ was set, and a linear learning rate decay strategy was employed to dynamically adjust the learning rate throughout training. Batch Size is set to 2048. The mask strategy uses a random mask, and the mask ratio is from 5% to 75%. While at the stage of fine-tuning, we use the BSD dataset as the training set, The rest of the hyperparameters are consistent with the pre-training phase. Each model's decoder consists of a single transformer block with an embedding dimension of 256 and eight attention heads.

To assess image retrieval quality under different conditions of pixel data loss, ranging from 5% to 75%, performance metrics such as PSNR and SSIM are used, as depicted in table 2. In the pixel loss rate of 5%, the proposed method SSKD surpasses MAE (ViT-B) in terms of PSNR, scoring 39.11 against 38.67. Additionally, the SSIM values indicate a advantage for SSKD (ViT-L), with a score of 0.96 compared to PeCo's 0.95.As the pixel loss rate increases to 35%, the models encounter greater challenges, with MAE (ViT-B) recording its lowest PSNR of 35.44. However, the SSKD methods demonstrate significant resilience, maintaining robust performance across severe conditions. Specifically, the SSKD achieves a high PSNR of 30.93 and an SSIM of 0.88 at a 75% pixel loss rate, outpacing PeCo's scores of 30.24 PSNR and 0.84 SSIM.

Furthermore, the mean Intersection over Union (mIoU) metric, essential for evaluating semantic segmentation, shows that SSKD models consistently achieve the highest accuracy levels across all stages of pixel data

**Table 2.** Results of comparative analysis for image retrieval.

| Model | PSNR/SSIM | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 5% | 15% | 25% | 35% | 45% | 55% | 65% | 75% |
| MAE(ViT-B) [30] | 38.67/0.95 | 37.68/0.95 | 36.70/0.94 | 35.80/0.93 | 34.87/0.92 | 33.88/0.91 | 31.50/0.88 | 30.36/0.86 |
| MAE(ViT-L) [30] | 39.02/0.96 | **38.13/0.95** | 37.82/0.96 | 36.11/0.94 | 36.38/0.94 | **34.69/0.92** | 33.09/0.91 | 30.85/0.88 |
| CAE(ViT-B) [42] | 38.45/0.94 | 37.56/0.94 | 36.73/0.94 | 35.72/0.93 | 34.79/0.92 | 33.81/0.91 | 31.69/0.89 | 30.33/0.85 |
| PeCo(ViT-B) [43] | 38.65/0.95 | 37.63/0.94 | 36.89/0.94 | 35.44/0.92 | 34.83/0.93 | 33.75/0.90 | 31.91/0.89 | 30.24/0.84 |
| LocalMIM [44] | 38.64/0.95 | 37.73/0.95 | 36.79/0.94 | 36.11/0.93 | 34.71/0.93 | 33.86/0.92 | 31.79/0.89 | 30.53/0.87 |
| CL-MAE [45] | 39.01/0.95 | 37.95/0.95 | 37.54/0.95 | 37.83/0.94 | 35.23/0.94 | 34.24/0.92 | 32.65/0.90 | 30.64/0.87 |
| PCP-MAE [46] | 38.95/0.96 | 37.63/0.94 | 37.43/0.95 | 37.65/0.93 | 35.29/0.93 | 33.99/0.93 | 31.95/0.89 | 30.43/0.86 |
| SSKD | **39.11/0.96** | 38.11/0.96 | **37.84/0.97** | **37.84/0.97** | **36.43/0.94** | 34.63/0.92 | **33.29/0.92** | **30.93/0.88** |
| Pixel lost | 28.21/0.88 | 23.40/0.81 | 21.05/0.71 | 20.09/0.68 | 18.96/0.65 | 18.47/0.61 | 17.30/0.61 | 16.80/0.56 |
| Ground truth | | | | | inf/1 | | | |

**Table 3.** Results of comparative analysis for damage detection.

| Model | mIoU | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 5% | 15% | 25% | 35% | 45% | 55% | 65% | 75% |
| MAE(ViT-B) [30] | 0.72 | 0.72 | 0.68 | 0.68 | 0.67 | 0.65 | 0.65 | 0.46 |
| MAE(ViT-L) [30] | 0.78 | **0.77** | 0.73 | 0.71 | 0.72 | 0.69 | 0.71 | 0.51 |
| CAE(ViT-B) [42] | 0.71 | 0.70 | 0.68 | 0.66 | 0.67 | 0.65 | 0.64 | 0.43 |
| PeCo(ViT-B) [43] | 0.72 | 0.71 | 0.69 | 0.68 | 0.70 | 0.66 | 0.66 | 0.45 |
| LocalMIM [44] | 0.74 | 0.72 | 0.70 | 0.69 | 0.69 | 0.68 | 0.68 | 0.48 |
| CL-MAE [45] | 0.77 | 0.74 | 0.75 | 0.73 | 0.69 | **0.71** | 0.70 | 0.51 |
| PCP-MAE [46] | 0.76 | 0.74 | **0.76** | 0.71 | 0.70 | 0.69 | 0.68 | 0.50 |
| SSKD | **0.79** | 0.76 | 0.74 | **0.75** | **0.73** | 0.70 | **0.72** | **0.54** |
| Pixel lost | 0.58 | 0.37 | 0.20 | 0.15 | 0.16 | 0.15 | 0.09 | 0.06 |
| Ground truth | 0.83 | | | | | | | |

loss. This underscores their superior ability in accurately identifying surface damage even when image quality is compromised. Detailed performance results are documented in the respective tables, with PSNR and SSIM data presented in table 2 and mIoU comparisons outlined in table 3.

In our comparison experiment, we evaluated multiple indicators for instance segmentation in figure 5: figure 5(a) shows Precision, figure 5(b) shows Accuracy, figure 5(c) shows Recall, and figure 5(d) shows F1 Score. Our method, SSKD, demonstrated excellent performance across all tested pixel loss rates (5% to 75%). Specifically, at a 75% pixel loss rate, SSKD achieved a Precision of 0.6135, Accuracy of 0.9876, and F1 Score of 0.5236. These results significantly outperformed methods like MAE (ViT-B/L), CAE (ViT-B), PeCo (ViT-B), LocalMIM, CL-MAE and PCP-MAE. For example, at 75% loss, MAE (ViT-B) had a Precision of 0.558, Accuracy of 0.896, and F1 Score of 0.477. SSKD showed greater robustness and higher overall performance across the entire test range, proving its superiority in handling image missing information. These findings indicate that SSKD can provide near-perfect accuracy and recall at low loss rates and maintain strong performance at high loss rates, highlighting its practical potential and reliability.

#### *4.2.1. Case I: fault detection and diagnosis in CNC system driven by #1 ball screw*

The study analyzes the effectiveness of fault detection and diagnosis as demonstrated in figures 6 and 7, focusing specifically on the identification of pitting damage under various conditions of pixel loss. Utilizing technique such as Mask R-CNN [47], the procedure achieves noteworthy fault detection precision. In figure 6, the subfigures (a1) ~ (a8) present the original images which have undergone varying degrees of pixel loss, specifically at rates of 5%, 15%, 25%, 35%, 45%, 55%, 65%, and 75%. Correspondingly, the retrieved images affected by varying degrees of pixel loss are shown in figure 7(A1) ~ (A8). A hallmark of this approach is its ability to maintain a consistent accuracy rate of 99% in detecting instances of pitting, irrespective of pixel loss within the image. At a Pixel Loss Rate (PLR) of 5%, these detection capabilities remain robust, however there is a noted decline in accuracy as the rate of pixel depletion intensifies. When the pixel loss reaches 25%, the original image becomes inadequate for identifying pitting damage. Conversely, applying the proposed image reconstruction

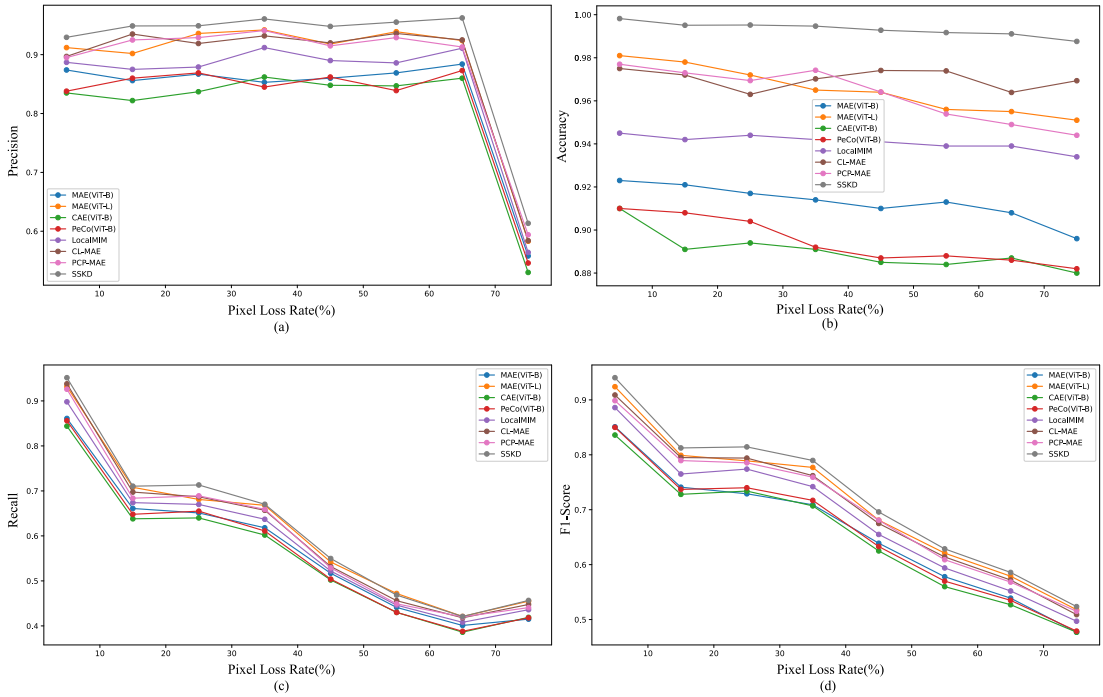


**Figure 5.** Comparison results of surface damage identification. (a) Precision. (b) Accuracy. (c) Recall. (d) F1-score.

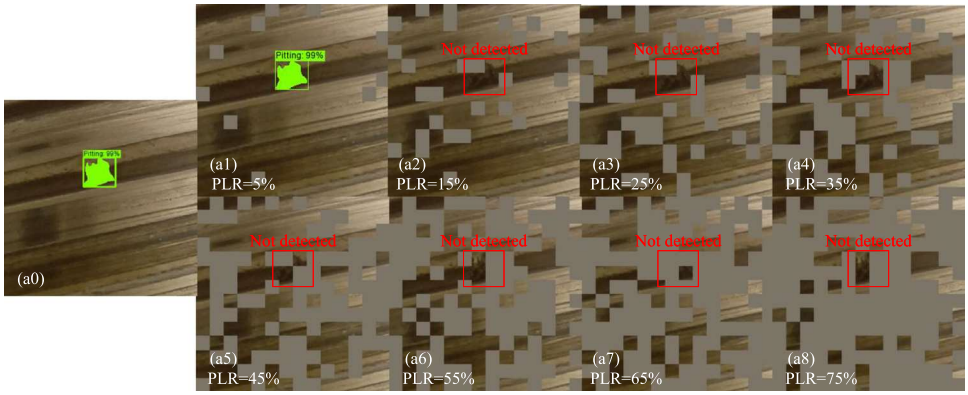


**Figure 6.** Results of damage detection and diagnosis for case I without image retrieval across varying degrees of Pixel Loss Rate (PLR). (a0) Reference image. (a1) ∼ (a8) Damage detection results with different levels of pixel loss ranging from 5% to 75%.

method significantly enhances fault detection efficacy. Notably, even with a 35% loss of pixels, the recognition accuracy for pitting in the reconstructed image remains almost at 99%.

The study further reveals that in more extreme scenarios, where pixel loss rates escalate to 45% and 75%, identifying pitting-related defects using the original images becomes unfeasible. Despite these challenges, the reconstructed images prove advantageous, upholding a 98% detection accuracy for the damaged area. This demonstrates the superior performance of using reconstructed images for fault diagnosis in severe conditions of pixel loss.

In summary, this study provides a compelling insight into the efficacy of fault detection and diagnosis techniques in identifying pitting damage, even under conditions of substantial pixel loss. It employs a methodology where reconstructed images are utilized to enhance the accuracy of fault detection compared to the reliance on original images that have suffered pixel loss. The results reveal that even when the pixel loss reaches as high as 75%, the reconstructed images maintain the capability for precise fault diagnosis. Such results underscore the robustness of the proposed method in accurately detecting faults across a range of compromised image conditions.

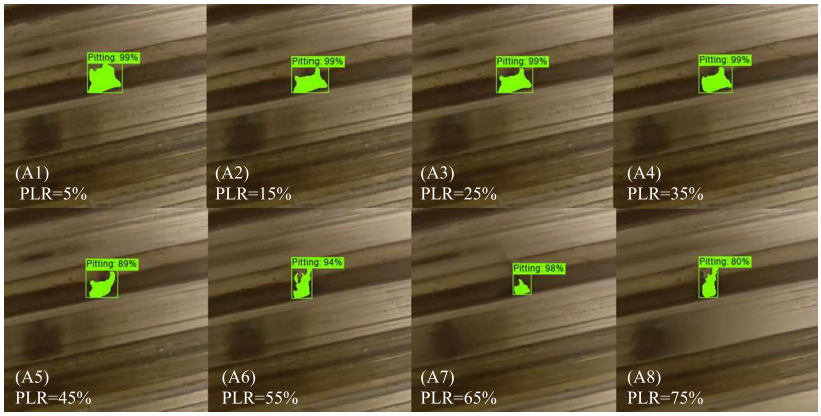


**Figure 7.** Results of damage detection and diagnosis for case I following image retrieval under varying degrees of PLR. (A1) ~ (A8) Damage detection results subsequent to image retrieval under different levels of pixel loss, spanning from 5% to 75%.

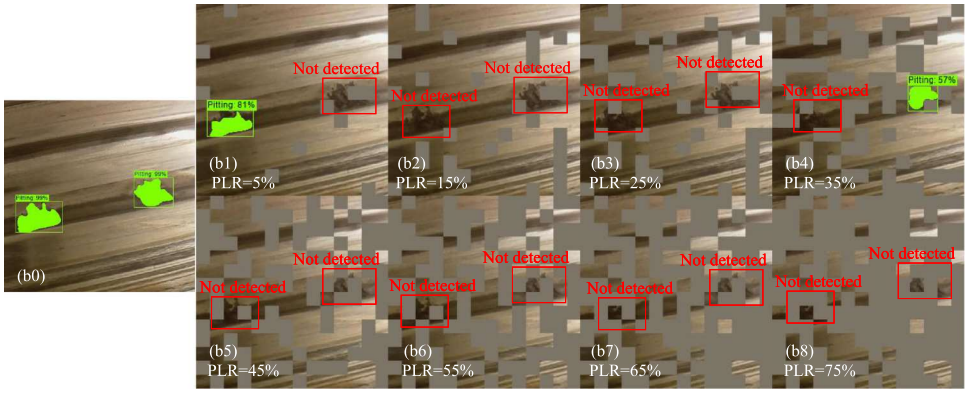


**Figure 8.** Results of damage detection and diagnosis for case II without image retrieval across varying degrees of Pixel Loss Rate (PLR). (b0) Reference image. (b1) ~ (b8) Damage detection results with different levels of pixel loss ranging from 5% to 75%.

*4.2.2. Case II: fault detection and diagnosis in CNC system driven by #2 ball screw*

To further explore the effectiveness of the proposed method for identifying pitting damage in a ball screw component, another experimental case study has been conducted. This study specifically focuses on identifying two types of damage areas within the CNC system. As discussed in section 4.2.1 for Case I, the results of image retrieval for Case II under varying degrees of pixel loss are presented in figure 9. The original images with 5%, 15%, 25%, 35%, 45%, 55%, 65%, and 75% pixel loss are shown in sub-figures (b1)-(b8) of figure 8, while the corresponding retrieved images are displayed in sub-figures (B1)-(B8) of figure 9. Across all levels of pixel loss, the reconstructed images closely match the ground truth, demonstrating the effectiveness of the proposed approach for degraded images in Case II.

At a pixel loss of 5%, only one of the defective regions could be identified with an accuracy of 81%, while the other pitting damage could not be identified. With pixel loss increasing to 35%, the accuracy of pitting recognition for one of the defective regions significantly dropped to 57% within the original image, rendering the second faulty area completely indiscernible. In contrast, using a reconstructed image increased the pitting recognition accuracy for these two defective areas to 68% and 98%, even under 35% pixel loss conditions. As the pixel loss rates escalated to 65% and 75%, the original image with pixel loss failed to identify the two pitting-related defective areas. However, the reconstructed image maintained an almost 99% accuracy in recognizing one of these flawed regions, highlighting its superior fault recognition performance. Notably, even at 75% pixel loss, identifying pitting in the original image became impossible, but the reconstituted image still achieved an accuracy of 98%. This once again emphasizes the advantage of image reconstruction in ensuring reliable fault diagnosis despite image degradation, aligning with the conclusions drawn from Case I.

*4.2.3. Cross-domain generalization analysis and extended experimental validation*

In order to comprehensively assess the generalization capabilities and domain adaptability of our proposed methodology, we conducted an extensive series of experimental evaluations across multiple cross-domain datasets. These datasets were specifically selected because they exhibit substantially different characteristics from our training

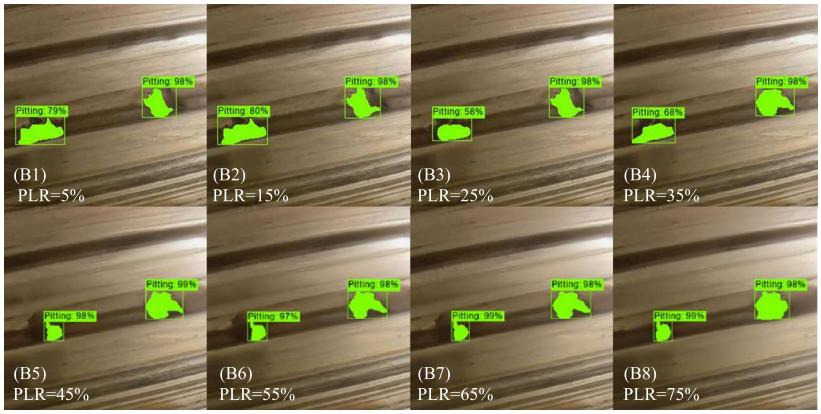


**Figure 9.** Results of damage detection and diagnosis for case II following image retrieval under varying degrees of PLR. (B1) ~ (B8) Damage detection results subsequent to image retrieval under different levels of pixel loss, spanning from 5% to 75%.

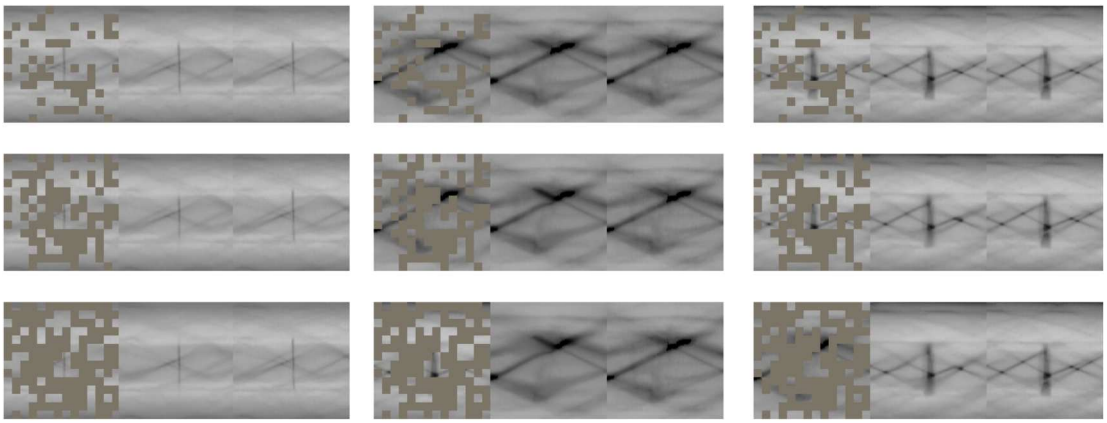

**Figure 10.** Example results on the ACCC dataset images. For each triplet, the pixel-lost image is shown on the left, the reconstruction in the middle, and the ground-truth on the right. From top to bottom, the pixel loss rates are 25%, 45% and 65% respectively.

**Table 4.** Results of comparative analysis for different dataset.

| Dataset | PSNR | | | SSIM | | |
|---|---|---|---|---|---|---|
| | 25% | 45% | 65% | 25% | 45% | 65% |
| ACCC [48] | 37.72 | 36.46 | 33.31 | 0.97 | 0.94 | 0.92 |
| CrackForest [49] | 37.68 | 36.41 | 33.28 | 0.97 | 0.95 | 0.92 |
| NEU surface defect [50] | 37.70 | 36.44 | 33.25 | 0.96 | 0.93 | 0.90 |

domain, particularly in terms of their texture patterns, structural complexity, and visual manifestations. Our experimental validation specifically encompassed three distinct datasets: the ACCC dataset [48], which focuses on concrete crack detection; the CrackForest dataset [49], which contains diverse pavement crack imagery; and the NEU surface defect dataset [50], which comprises various industrial surface anomalies. These datasets were chosen deliberately as they present significantly different challenges in terms of texture complexity, structural layout, illumination conditions, and overall visual characteristics when compared to our original training data distribution.

The qualitative results of our cross-domain experiments are extensively documented in figures 10 through 12. Through careful examination of these results, we observed that our method demonstrates remarkable capability in accurately reconstructing both macro-level structural elements and micro-level textural details, even when confronted with previously unseen domains and challenging visual scenarios. Furthermore, to provide a quantitative foundation for our analysis, we conducted rigorous performance evaluations using two widely accepted image quality metrics: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM). The comprehensive results of these evaluations are presented in table 4, which reveals consistent and robust performance metrics across all tested domains. This consistency in performance strongly suggests that our model has successfully avoided

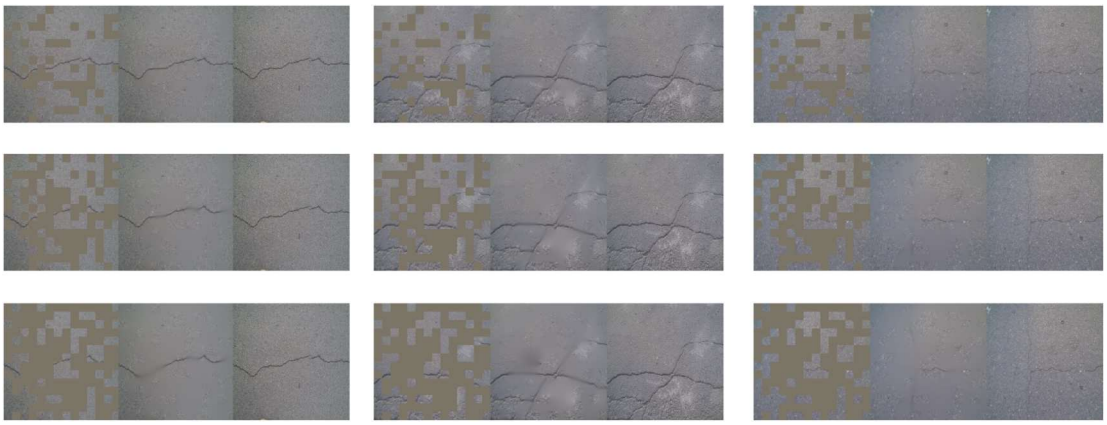

**Figure 11.** Example results on the CrackForest dataset images. For each triplet, the pixel-lost image is shown on the left, the reconstruction in the middle, and the ground-truth on the right. From top to bottom, the pixel loss rates are 25%, 45% and 65% respectively.

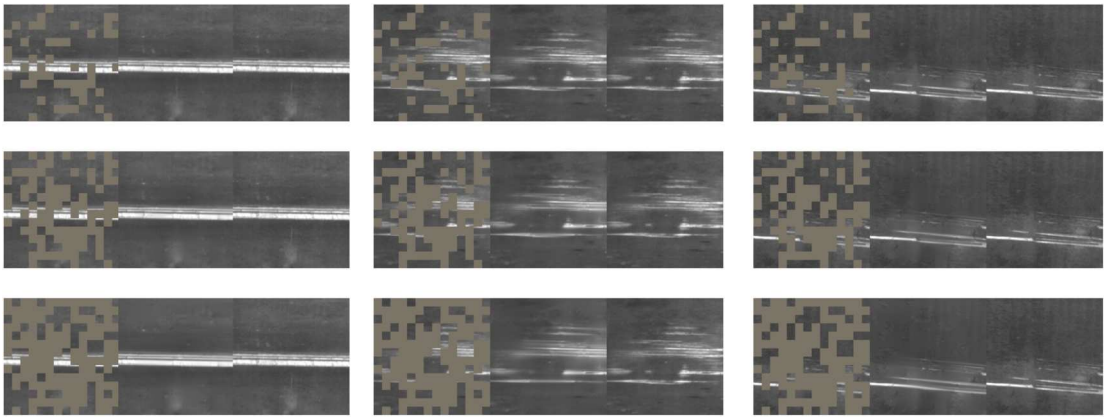

**Figure 12.** Example results on the NEU surface defect dataset images. For each triplet, the pixel-lost image is shown on the left, the reconstruction in the middle, and the ground-truth on the right. From top to bottom, the pixel loss rates are 25%, 45% and 65% respectively.

overfitting to the training data distribution and has developed a genuinely generalizable capability for reconstructing missing pixel-level information across diverse real-world scenarios.

The empirical evidence gathered through these extensive cross-domain experiments provides substantial support for the robustness and adaptability of our proposed approach. These findings are particularly significant in the context of practical real-world applications, where input data frequently deviates from the initial training distribution in terms of visual characteristics, noise patterns, and structural complexity. Moreover, the consistent performance across varied domains indicates that our method has successfully learned to capture and reconstruct fundamental image features that are invariant across different types of surface textures and structural patterns, thereby demonstrating its potential for broad practical applicability in diverse real-world scenarios.

The results of these extended experiments not only validate the effectiveness of our approach but also provide valuable insights into its behavior when confronted with out-of-distribution data, which is a critical consideration for the deployment of such systems in real-world applications. This comprehensive evaluation framework thus serves to establish the practical utility and reliability of our method across a broad spectrum of application scenarios.

## 5. Ablation study

The ablation study, presented in table 5, systematically evaluates the contribution of our proposed self-supervised knowledge distillation (KD) framework and the subsequent transfer learning (TL) for domain adaptation. The results clearly demonstrate the effectiveness of our KD framework. Comparing Configuration

**Table 5.** Ablation study.

| Configuration | PSNR / SSIM | mIoU | Params (M) |
|---|---|---|---|
| ViT-L (w/o KD) | 36.38/0.94 | 0.72 | 86.56 |
| ViT-B (w/o KD) | 34.87/0.92 | 0.67 | 53.24 |
| ViT-L (w/ KD) + TL | 36.68/0.96 | 0.77 | 86.56 |
| ViT-B (w/ KD) + TL | 36.43/0.94 | 0.73 | 53.24 |

**Configurations:**
'ViT-L/B' indicates the encoder architecture.
'w/ KD' means the model is trained with our self-supervised knowledge distillation framework.
'+ TL' denotes the application of domain adaptation via transfer learning and fine-tuning on the target dataset.

1 (ViT-L w/o KD) and Configuration 3 (ViT-L w/ KD + TL), we observe a significant improvement in both reconstruction quality (PSNR/SSIM from 36.38/0.94 to 36.68/0.96) and segmentation accuracy (mIoU from 0.72 to 0.77), despite using the same large ViT-L encoder and no target domain fine-tuning in the first case. This improvement is directly attributable to the knowledge distillation process, which enables the student encoder to learn richer, more robust representations by mimicking the teacher's latent features. Furthermore, the importance of transfer learning is highlighted by comparing Configurations 1 and 3. Configuration 1, which uses the powerful ViT-L encoder without KD or TL, performs worse than Configuration 3. This shows that simply using a larger encoder is insufficient; the knowledge distilled from the teacher and the fine-tuning on the target domain are crucial for optimal performance. Finally, the study validates the efficiency of our compact student model. Configuration 4 (ViT-B w/ KD + TL) achieves performance (36.43/0.94, mIoU 0.73) very close to the larger model in Configuration 3, while reducing the parameter count from 86.56M to 53.24M. This confirms that our knowledge distillation framework successfully transfers the knowledge from the large teacher to the compact student, resulting in a high-performance, resource-efficient model suitable for practical deployment.

## 6. Conclusion

In the rapidly advancing field of IIoT, this research introduces a pioneering SSKD framework that transforms data integrity and diagnostic reliability in industrial systems. The framework uniquely combines knowledge distillation with self-supervised learning, establishing a robust methodology that overcomes traditional historical data dependencies, particularly for BSD monitoring in CNC machines. The research's key innovation lies in its novel approach to missing data recovery, where the SSKD framework achieves autonomous image reconstruction without external data correlations. The advanced pre-training and fine-tuning mechanisms enable effective cross-domain knowledge transfer, maintaining diagnostic accuracy even under significant data degradation. Through pixel-level data recovery operating independently of external correlations, this framework addresses critical challenges in remote diagnostic systems for smart manufacturing.

Future research opportunities include: (1) Expanding SSKD applications across diverse industrial sectors. (2) Evaluating framework scalability in larger industrial ecosystems. (3) Developing comprehensive industrial-scale validation methods.

## Acknowledgments

This work received support from the National Natural Science Foundation of China (Grant 52105111), the Guangdong Basic and Applied Basic Research Foundation (Grant 2025A1515012256), the Scientific Research Foundation of Changzhou Vocational Institute of Engineering (No. 11130900120001), the Scientific Research Foundation of Changzhou Vocational Institute of Industry Technology (No. BS202213101002), the Professional Virtual Teaching and Research Room Construction Practices and Exploration Studies (No. XHYBLX2023004), and the Changzhou Intelligent Connected Vehicle Driverless Driving and Network Security Technology Key Laboratory (No. CM2024007).

## Data availability statement

The data cannot be made publicly available upon publication because they are owned by a third party and the terms of use prevent public distribution. The data that support the findings of this study are available upon reasonable request from the authors.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## References


[1] Wang H, Xie Y, Wang Y, Zhang C, Ni P, Lu Y and Wang Y 2025 Surface defect segmentation of steel pipes based on superpixel prior knowledge *Eng. Res. Express* **7** 035570
[2] Pandhare V, Miller M, Vogl G W and Lee J 2022 Ball screw health monitoring with inertial sensors *IEEE Trans. Ind. Inf.* **19** 7323–34
[3] Liu X, Li Y, Cheng Y and Cai Y 2023 Sparse identification for ball-screw drives considering position-dependent dynamics and nonlinear friction *Robotics and Computer-IntegratedManufacturing* **81** 102486
[4] Chen P, Ma J, He C, Jin Y and Fan S 2025 Progressive contrastive representation learning for defect diagnosis in aluminum disk substrates with a bio-inspired vision sensor *Expert Syst. Appl.* **289** 128305
[5] Wang Y, Chen P, Wei Q, Qi J, He C and Zhou C 2025 Multi-channel fusion scale transformed signals with magnetic leakage for damage detection in steel wire ropes *Nondestruct. Test. Eval.* 1–26
[6] Yin Z, Yang Y, Shen G, Li Y, Huang L and Hu N 2023 Dynamic modeling, analysis, and experimental study of ball screw pairs with nut spalling faults in electromechanical actuators *Mech. Syst. Signal Process.* **184** 109751
[7] Xu M, Cai B, Li C, Zhang H, Liu Z, He D and Zhang Y 2020 Dynamic characteristics and reliability analysis of ball screw feed system on a lathe *Mech. Mach. Theory* **150** 103890
[8] Xu C, Chen P, Gao J, Jin Y and Rao M 2025 Semi-supervised transfer learning preserving spatial homogeneity for gearbox diagnostics in extraneous transient noise *Nondestruct. Test. Eval.* 1–29
[9] Chen P, Ma Z, Xu C, Jin Y and Zhou C 2024 Self-supervised transfer learning for remote wear evaluation in machine tool elements with imaging transmission attenuation *IEEE Internet of Things Journal* **11** 23045–54
[10] Duan M, Lu H, Zhang X, Li Z, Zhang Y, Yang M and Liu Q 2021 Dynamic modeling and experimental research on position-dependent behavior of twin ball screw feed system *The International Journal of Advanced Manufacturing Technology* **117** 3693–703
[11] Chen Y, Zhao C, Li Z and Lu Z 2020 Analysis on dynamic contact characteristics and dynamic stiffness estimating method of single nut ball screw pair based on the whole rolling elements model *Applied Sciences* **10** 5795
[12] Bertolino A C, Sorli M, Jacazio G and Mauro S 2019 Lumped parameters modelling of the emas' ball screw drive with special consideration to ball/grooves interactions to support model-based health monitoring *Mech. Mach. Theory* **137** 188–210
[13] Chen P, Xu C, Ma Z and Jin Y A mixed samples-driven methodology based on denoising diffusion probabilistic model for identifying damage in carbon fiber composite structures *IEEE Trans. Instrum. Meas.* **72** 3513411
[14] Chen P, Ma J, He C, Jin Y and Fan S 2025 Semi-supervised consistency models for automated defect detection in carbon fiber composite structures with limited data *Meas. Sci. Technol.* **36** 046109
[15] Chen P, Ma Z, Xu C, Zhang M, Li H, Zheng K and Jin Y 2025 Scale-aware domain adaptation for surface defects detection on machine tool components in contaminant measurements *IEEE Trans. Instrum. Meas.* **74** 1–9
[16] Huang Y-C, Kao C-H and Chen S-J 2018 Diagnosis of the hollow ball screw preload classification using machine learning *Applied Sciences* **8** 1072
[17] Riaz N, Shah S I A, Rehman F and Khan M J 2021 An intelligent hybrid scheme for identification of faults in industrial ball screw linear motion systems *IEEE Access* **9** 35136–50
[18] Denkena B, Bergmann B and Schmidt A 2021 Preload monitoring of single nut ball screws based on sensor fusion *CIRP J. Manuf. Sci. Technol.* **33** 63–70
[19] Pandhare V, Li X, Miller M, Jia X and Lee J 2020 Intelligent diagnostics for ball screw fault through indirect sensing using deep domain adaptation *IEEE Trans. Instrum. Meas.* **70** 1–11
[20] Benker M and Zaeh M F 2022 Condition monitoring of ball screw feed drives using convolutional neural networks *CIRP Ann* **71** 313–6
[21] Azamfar M, Li X and Lee J 2020 Intelligent ball screw fault diagnosis using a deep domain adaptation methodology *Mech. Mach. Theory* **151** 103932
[22] Li X, Zhang W, Ma H, Luo Z and Li X 2020 Deep learning-based adversarial multi-classifier optimization for cross-domain machinery fault diagnostics *J. Manuf. Syst.* **55** 334–47
[23] Gungor V C, Lu B and Hancke G P 2010 Opportunities and challenges of wireless sensor networks in smart grid *IEEE Trans. Ind. Electron.* **57** 3557–64
[24] Bao Y, Li H, Sun X, Yu Y and Ou J 2013 Compressive sampling–based data loss recovery for wireless sensor networks used in civil structural health monitoring *Structural Health Monitoring* **12** 78–95
[25] Saravanan S and Karthikeyan E 2011 A protocol to improve the data communication over wireless network *International Journal of Wireless & Mobile Networks* **3** 95
[26] Mietzner J, Schober R, Lampe L, Gerstacker W H and Hoeher P A 2009 Multiple-antenna techniques for wireless communications-a comprehensive literature survey *IEEE Communications Surveys & Tutorials* **11** 87–105
[27] Chen P, Wu Y, Xu C, Huang C-G, Zhang M and Yuan J 2025 Interference suppression of nonstationary signals for bearing diagnosis under transient noise measurements *IEEE Trans. Reliab.* **11** 1–15
[28] Qiao W, Harley R G and Venayagamoorthy G K 2008 Fault-tolerant indirect adaptive neurocontrol for a static synchronous series compensator in a power network with missing sensor measurements *IEEE Trans. Neural Netw.* **19** 1179–95
[29] Demirhan H and Renwick Z 2018 Missing value imputation for short to mid-term horizontal solar irradiance data *Appl. Energy* **225** 998–1012

[30] He K, Chen X, Xie S, Li Y, Dollár P and Girshick R 2022 Masked autoencoders are scalable vision learners *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 16000–9
[31] Hinton G 2015 Distilling the knowledge in a neural network arXiv:1503.02531
[32] Yang C, Luo X, Zhang Z, Chen Z and Wu X-J 2025 Kdfuse: a high-level vision task-driven infrared and visible image fusion method based on cross-domain knowledge distillation *Information Fusion* **118** 102944
[33] Huang Y, Zhang K, Xia P, Wang Z, Li Y and Liu C 2024 Cross-attentional subdomain adaptation with selective knowledge distillation for motor fault diagnosis under variable working conditions *Adv. Eng. Inf.* **62** 102948
[34] Li X, Huang G, Cheng L, Zhong G, Liu W, Chen X and Cai M 2024 Cross-domain visual prompting with spatial proximity knowledge distillation for histological image classification *Journal of Biomedical Informatics* **158** 104728
[35] Yu X, Sun Y, Li H and Wu S 2022 An improved meshing stiffness calculation algorithm for gear pair involving fractal contact stiffness based on dynamic contact force *European Journal of Mechanics-A/Solids* **94** 104595
[36] Huang Y-C and Hsieh Y-K 2022 Applying a support vector machine for hollow ball screw condition-based classification using feature extraction *Proc. Inst. Mech. Eng. Part B J. Eng. Manuf.* **236** 1839–52
[37] Zhang H, Liu W, Shi J, Chang S, Wang H, He J and Huang Q 2022 Maefe: masked autoencoders family of electrocardiogram for self-supervised pretraining and transfer learning *IEEE Trans. Instrum. Meas.* **72** 1–15
[38] Dosovitskiy A *et al* 2020 An image is worth 16x16 words: transformers for image recognition at scale arXiv:2010.11929
[39] Deng J, Dong W, Socher R, Li L-J, Li K and Fei-Fei L 2009 Imagenet: a large-scale hierarchical image database *IEEE conference on computer vision and pattern recognition* (IEEE) pp 248–55
[40] Song H, Xie J, Duan Y, Xie X, Zhou Y and Wang W 2025 Cmkd-net: a cross-modal knowledge distillation method for remote sensing image classification *Adv. Space Res.* **75** 8515–34
[41] Song H, Xie J, Liang L, Su Y, Xiao Y, Zhang X, Ouyang Y, Li X, Chen S and Li Y 2025 Symmetrical learning and transferring: Efficient knowledge distillation for remote sensing image classification *Symmetry* **17** 1–34
[42] Chen X, Ding M, Wang X, Xin Y, Mo S, Wang Y, Han S, Luo P, Zeng G and Wang J 2023 Context autoencoder for self-supervised representation learning *Int. J. Comput. Vis.* **132** 208–23
[43] Dong X, Bao J, Zhang T, Chen D, Zhang W, Yuan L, Chen D, Wen F, Yu N and Guo B 2023 Peco: Perceptual codebook for bert pre-training of vision transformers *Proceedings of the AAAI Conference on Artificial Intelligence* 37 552–60
[44] Wang H, Tang Y, Wang Y, Guo J, Deng Z-H and Han K 2023 Masked image modeling with local multi-scale reconstruction *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp 2122–31
[45] Madan N, Ristea N-C, Nasrollahi K, Moeslund T B and Ionescu R T 2024 Cl-mae: Curriculum-learned masked autoencoders 2492–502
[46] Zhang X, Zhang S and Yan J 2024 Pcp-mae: learning to predict centers for point masked autoencoders *Advances in Neural Information Processing Systems* **37** 80303–27
[47] He K, Gkioxari G, Doll'ar P and Girshick R 2017 Mask r-cnn *Proceedings of the IEEE International Conference On Computer Vision* pp 2961–9
[48] Wei R, Wei H, Chen D, Xie L, Wang Z and Hu Y 2020 Defect detection for aluminium conductor composite core x-ray image with deep convolution network *J. Phys. Conf. Ser.* **1633** 012166
[49] Shi Y, Cui L, Qi Z, Meng F and Chen Z 2016 Automatic road crack detection using random structured forests *IEEE Trans. Intell. Transp. Syst.* **17** 3434–45
[50] Bao Y, Song K, Liu J, Wang Y, Yan Y, Yu H and Li X 2021 Triplet-graph reasoning network for few-shot metal generic surface defect segmentation *IEEE Trans. Instrum. Meas.* **70** 1–11