Hindawi
Journal of Sensors
Volume 2022, Article ID 1807103, 14 pages
https://doi.org/10.1155/2022/1807103

Research Article

A Novel Data Analytics Oriented Approach for Image Representation Learning in Manufacturing Systems

Yue Liu,[1,2] **Junqi Ma**,[1] **Xingzhen Tao**,[1] **Jingyun Liao**,[3] **Tao Wang**,[4] **and Jingjing Chen**[4,5]

[1]*Jiangxi College of Applied Technology, China*
[2]*Guangzhou University, China*
[3]*Guangzhou Panyu Polytechnic, China*
[4]*Zhijiang College, Zhejiang University of Technology, China*
[5]*School of Economics, Fudan University, China*

Correspondence should be addressed to Tao Wang; wt@zzjc.edu.cn

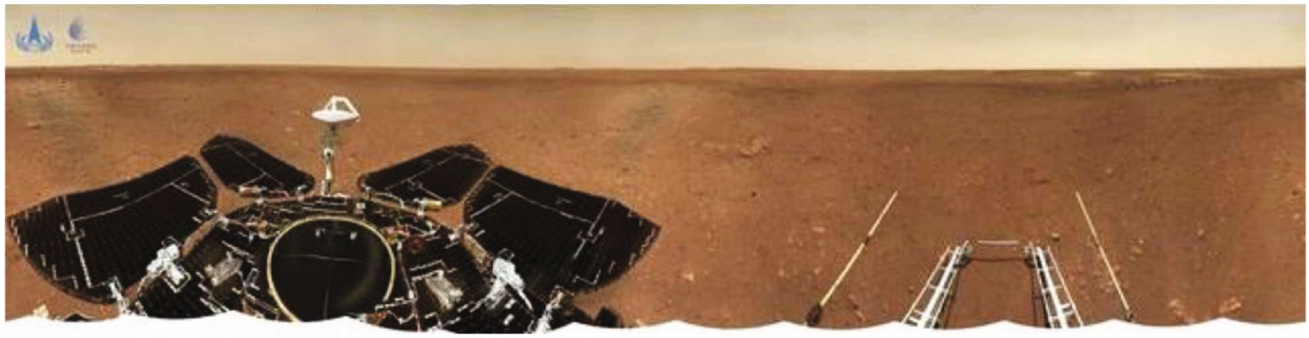
Received 2 November 2021; Revised 3 December 2021; Accepted 17 December 2021; Published 13 January 2022

Academic Editor: Haidong Shao

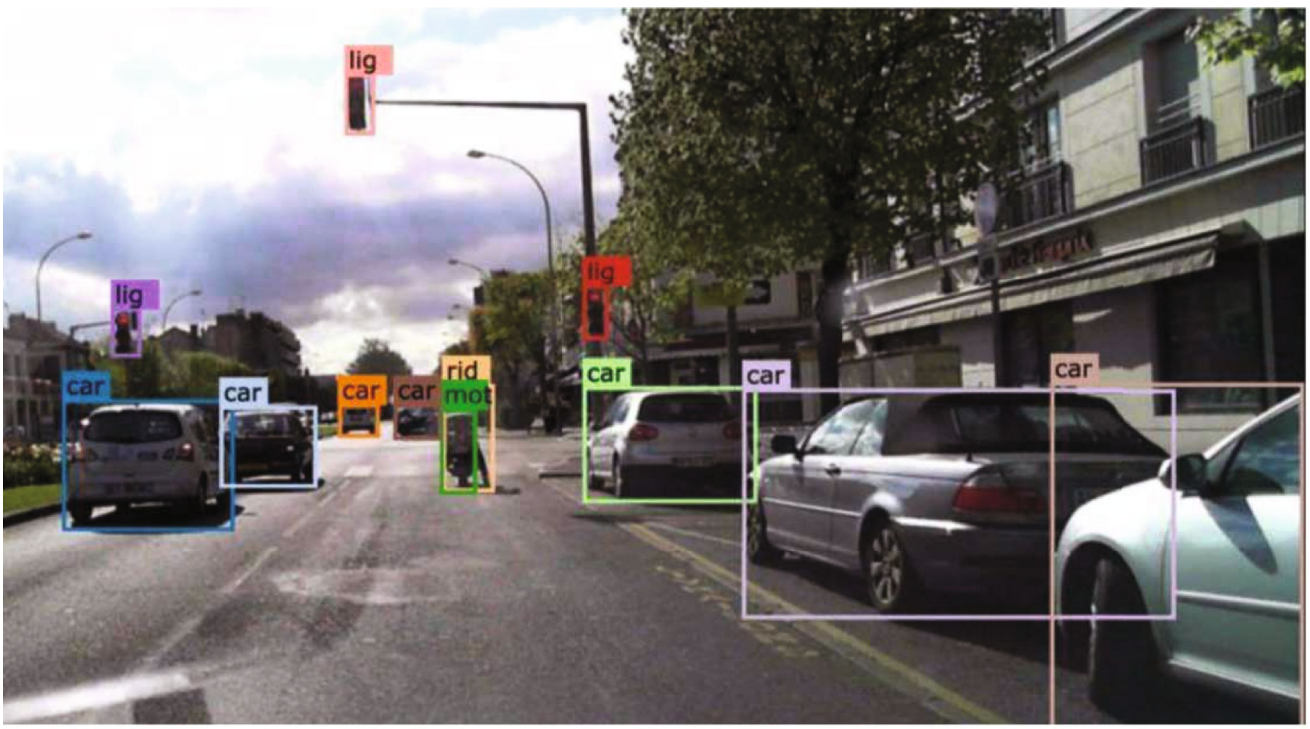
In the era of digital manufacturing, huge amount of image data generated by manufacturing systems cannot be instantly handled to obtain valuable information due to the limitations (e.g., time) of traditional techniques of image processing. In this paper, we propose a novel self-supervised self-attention learning framework—TriLFrame for image representation learning. The TriLFrame is based on the hybrid architecture of Convolutional Network and Transformer. Experiments show that TriLFrame outperforms state-of-the-art self-supervised methods on the ImageNet dataset and achieves competitive performances when transferring learned features on ImageNet to other classification tasks. Moreover, TriLFrame verifies the proposed hybrid architecture, which combines the powerful local convolutional operation and the long-range nonlocal self-attention operation and works effectively in image representation learning tasks.

1. Introduction

The researchers in the field of computer vision have already achieved great progress in the techniques for image recognition; most of these achievements are based on supervised learning methods. For example, ImageNet [1] acts as a large-scale labelled image dataset applicable for all kinds of image learning tasks, among which supervised methods, e.g., ResNet [2] and AlexNet [3], are dominating and providing the state-of-the-art performances. Although with the thriving of semisupervised learning, unsupervised learning, and self-supervised learning, some competitive methods are emerging, e.g., fast-SWA [4], VAT [5], CPC [6], DIM [7], AMDIM [8], and IIC [9]. These methods show that the performance gap between reduced supervised and supervised methods is shrinking; the amount of labels required for training a competitive unsupervised or self-supervised method is dramatically decreasing. It is noted that certain amount of labels as guiding reference for learning methods is too valuable to ignore that usually results in the gradually decreased adoptions of fully unsupervised methods [10]. All these studies above imply that self-supervised methods are becoming more and more promising in the area of image representation learning, yet we have not seen any self-supervised learning method that surpasses the performance of supervised methods in a general perspective. On the other hand, considering the huge amount of image data being generated every day by manufacturing systems, it is reasonable to rethink the methodology of image learning. Specifically, in the environment of manufacturing systems, applications such as robotics, autopilot systems, medical diagnosis, smart home, and smart city systems are generating significant amount of data every day. It is notable that a large portion of data produced in manufacturing systems is image, as shown in Figure 1.

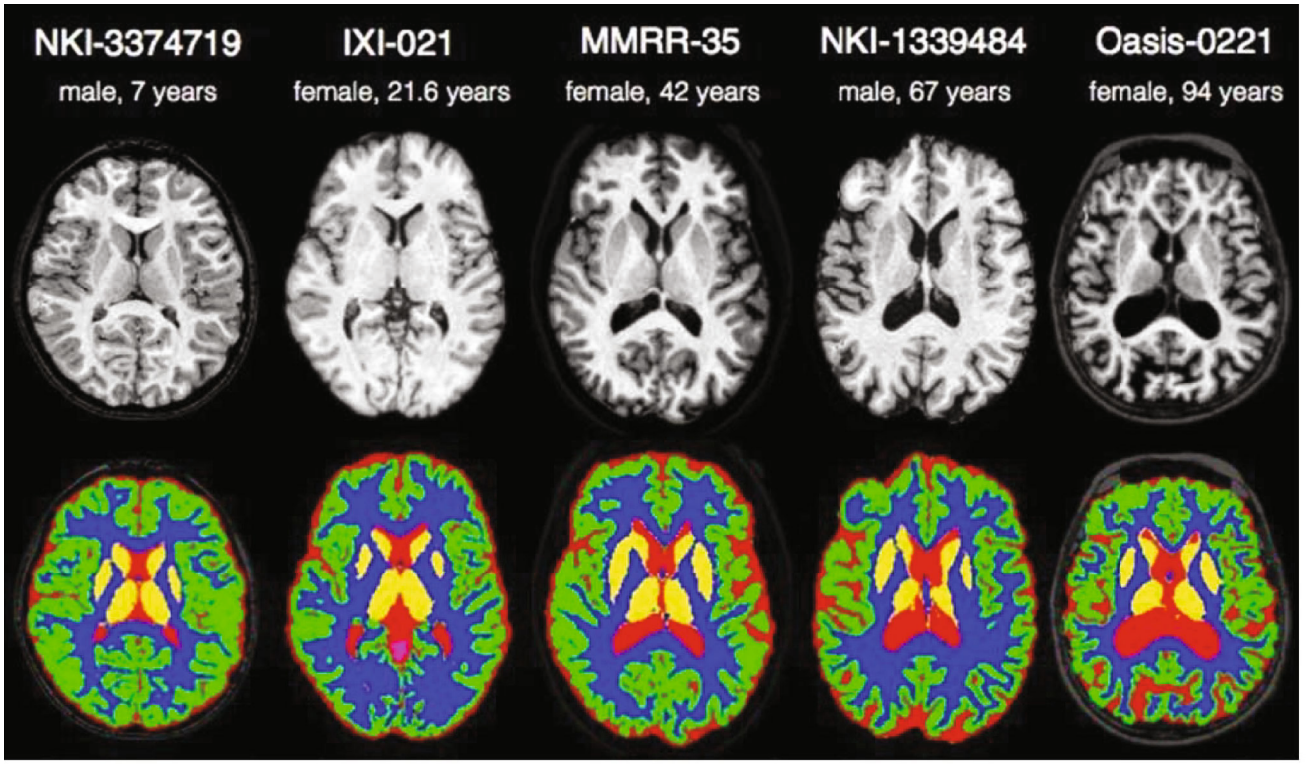


(a)

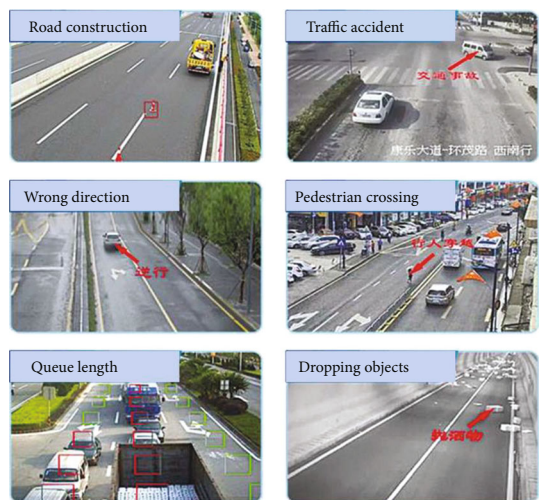


(b)

Figure 1: Continued.



(c)



(d)

Figure 1: (a) A picture by a navigation terrain camera; (b) the object recognition result of pictures captured by camera mounted on an autopilot vehicle; (c) the diagram of individual differences in brain area and shape, captured for further medical diagnosis; (d) images from surveillance cameras of smart city system being processed for incident detection.

Images are generated by various kinds of manufacturing systems in a sharp speed and are ready to be analyzed though technologies like image classification, object detection, image segmentation, image filtering, denoising, etc. However, it is noted that image data cannot be instantly processed due to performance limitations by the manufacturing systems. Although with the research focus being transferred to subdivision fields of image learning, e.g., medical image processing [11], face recognition [12], image analysis in autonomous vehicle [13], and fault diagnosis in manufacturing system [14–17], the capability is gradually catching up with the explosive growth of image data being generated in manufacturing environments. It should be also well noted that, in order to equip models with specific processing capability in subdivision fields, many task-specific image datasets with accurate labelling are created for supervised training. The fact that, it demands tremendous effort to label every image according to the training target, directly strangles the development of supervised learning methods. Thus, many researchers attempt to

exploring alternatives of supervised learning, for example, self-supervised methods using the structural information of the image data to supervise the learning process. In this paper, we propose a general solution for image representation learning based on the self-supervised method, with the capability of transferring to subdivision fields with trivial effort.

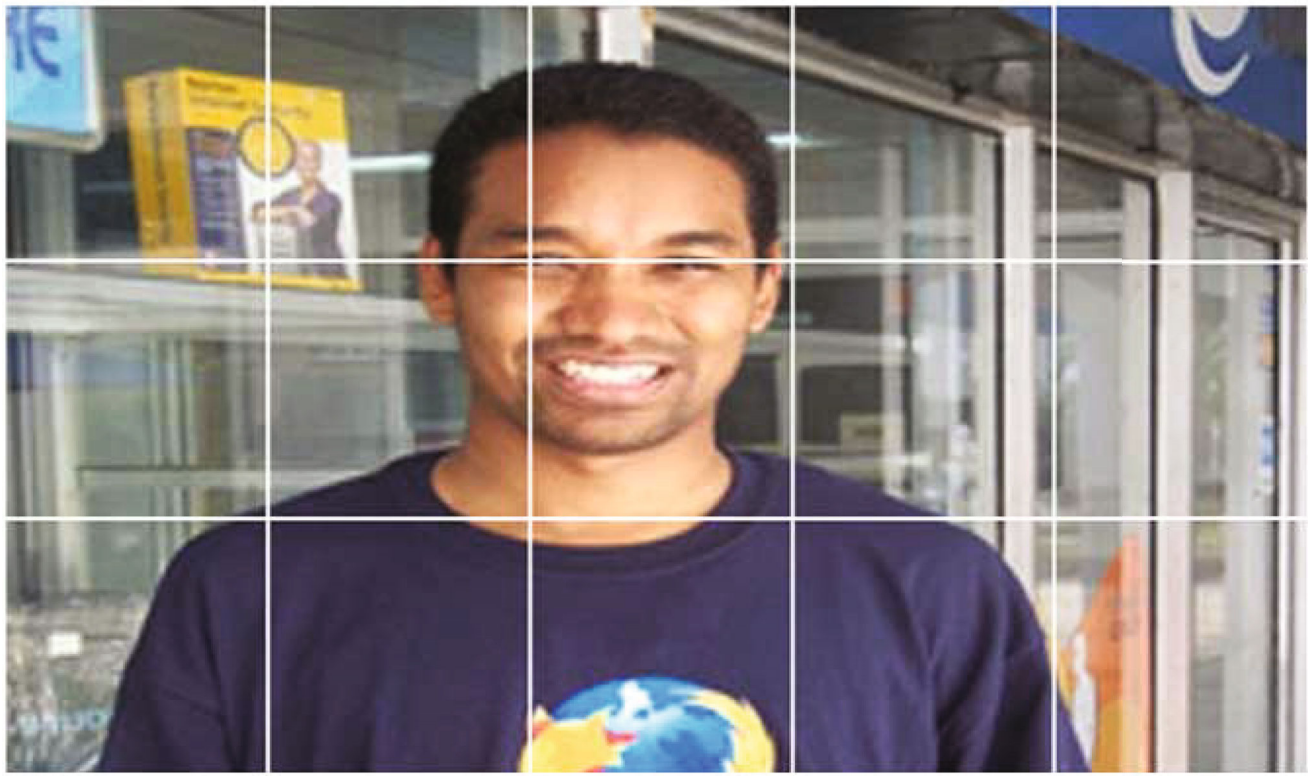
Current methods of self-supervised learning for image processing can roughly be divided into the following two categories: generative methods and contrastive methods. Generative methods based on Auto-Encoder [18–20] and generative adversarial networks (GAN) [21–23] rely on reconstruction error in pixel-level to learn image representation. Relying on pixel-based objectives significantly reduces the capability to model correlations or complex structures and makes model heavily focus on low-level features instead of abstract representative features. Contrastive methods [6, 7, 24–27] learn image representations by contrasting positive and negative samples in the latent space, which forces model to discard pixel-wise information and focus on the structure and correlation of the image as a whole. While executing an image learning task, the aim is to get a semantic structural embedding of image which is generalized and can be transferred to subdivision tasks which do not dependent on pixel-level details; thus, contrastive methods better fit our purpose. We are inspired by Contrastive Predictive Coding (CPC) introduced in [6] which utilizes a probabilistic contrastive loss (called "the InfoNCE") to force the model to learn the underlying semantic information that is shared among the input sequence. However, when applying CPC to image representation learning, a major issue needs to be addressed properly; as shown in Figure 2, it is difficult to predict image patches which contain objects that never appear in previous content as CPC lacks of knowledge of long-range dependencies across the entire image. In this paper, we equip the CPC framework with the power of self-attention [28] which is skilled in capturing nonlocal long-range dependencies, in order to learn a better semantic structural representation of image. The intuition is that if sending the latent embeddings of image patches through the self-attention framework (i.e., a transformer encoder architecture), then each patch embedding will have an impression of the content of other patches, and features of more correlation with others will be emphasized as a result. This process facilitates the learning of nonlocal high-level representation of image.

In this paper, we propose a novel data analytics oriented approach for image representation learning with self-supervised learning and self-attention for manufacturing systems—TriLFrame. The framework is aimed at learning nonlocal semantic features of image with the ability to predict the missing patches of image in latent space. Although following the idea of contrastive self-supervision, TriLFrame is different from [6, 25] in major aspects: first, TriLFrame applies self-attention mechanism on the top of backbone convolutional operations to capture long-range dependencies; second, TriLFrame makes use of nonlocal image patches with no overlap to construct positive and negative samples for contrastive learning; and third, we introduce a progressive prediction strategy instead of the simple linear transformation used in [6]. It should also be noted that this paper is an extension of our previous work CPCTR [29], which is a self-supervised self-attention framework for video representation learning. CPCTR utilizes self-attention operations to encode long-range spatio-temporal correlations of video data in order to capture "slow features" in video. Different from CPCTR, this paper makes the following novel contributions: first, specifically for image data processing, we design a new self-supervision pretext, which is the first few to introduce self-attention to contrastive image representation learning; second, a novel positive and negative sample construction is designed for contrastive learning which only requires spatial information of data; and third, under the context of manufacturing systems, we conduct experiments on different image datasets to show the effectiveness of the proposed method.

The contribution in this paper is summarized as follows: (1) we propose the self-supervised self-attention coding framework for image learning in manufacturing environment; (2) we apply the transformer encoder to TriLFrame to learn nonlocal spatial dependencies to better learn the semantic representation of image, and we experiment on the self-attention module in TriLFrame to reveal its effectiveness; and (3) we evaluate TriLFrame on the ILSVRC ImageNet competition dataset [30] as many authors [31–33]. With unlabeled image data, we show that a pretrained TriLFrame can be easily transferred to image classification tasks with competitive performances.

2. Related Literatures

2.1. Contrastive Learning. Based on the theory of Noise Contrastive Estimation (NCE) [34], contrastive learning uses classification tasks to discriminate positive sample from negative sample. The learning process is greatly dependent on operation on the latent space of input data (i.e., input data is preprocessed to reduce dimension), which forces contrastive learning to pay more attention to semantic structural representations while less attention to low-level pixel-wise features. In order to improve the efficiency of contrastive learning, users are required to carefully select positive and negative samples. Generally, negative samples that are hard to discriminate can improve the learning quality greatly. Contrastive learning has been proven competitive in the contexts of natural language processing [35, 36], audio processing [6], image processing [24, 27], video understanding [37, 38], etc., and a number of researches have been investigating the prospect of contrastive learning using no negative pairs [39] and no momentum encoder [40]. Its performance shows a promising prospect of contrastive learning. Recently, a new contrastive learning approach, Contrastive Predictive Coding (CPC) in [6], proposes an effective framework that can be applied to sequenced data modality, e.g., natural language, audio, video, or image (an image can be cut into a spatial sequence of image patches). CPC encodes underlying shared features that is slowly varying across data sequences and discarding local information. These shared features are called "slow features," which refers to these



(a)

Figure 2: Continued.

(b)

Figure 2: (a) The upper part of a full image. (b) The full image labelled as "Mobile Phone" in the ImageNet dataset.

features that are slowly varying across time, e.g., the identity of a speaker in an audio signal, an activity carried out in a video, and an object in an image.

2.2. Self-Supervised Learning for Image. With the development of self-supervised learning, especially the wide adoption of contrastive learning, self-supervised methods have shown a promising prospect for the images learning [6, 24, 25, 27, 31, 41]. CPC [6] using self-supervised training on unlabeled ImageNet dataset and fine-tuned with linear classification already outperform the supervised AlexNet [3]; Data-Efficient CPC [25] as an extended work has scaled up CPC and achieved Top 1 accuracy of 71.5% on the image classification task on ImageNet; it also exhibits high data efficiency when fine-tuning with labelled data compared with fully supervised methods. Deep InfoMax (DIM) [7] learns image representations through the internal structure information. A follow-up work of DIM and Augment Multiscale DIM (AMDIM) [8] utilizes invariant features across data augmentations, e.g., color jittering and random cropping; it gets Top 1 accuracy of 68.4% on ImageNet with unsupervised pretraining and evaluated by linear classification task. Contrastive Multiview Coding (CMC) [26] learns representations using different versions of the same image,

e.g., image of different angles, as data transformations which the representation should cope with. In the conventional formulation of contrastive learning, the size of minibatch restricts the total number of negative samples; Momentum Contrast (MoCo) [24] effectively lifts this restriction by maintaining a long queue of all negative samples, when training the negative encoder does not update with the positive encoder. The experiment results show that MoCo outperforms supervised models in several downstream tasks on different image datasets. Typically, these image downstream tasks need supervised training with labelled image to achieve good results; however, MoCo shows that the performance gap between supervised and unsupervised methods has largely been closed.

2.3. Self-Attention. TriLFrame also adopts the idea of self-attention mechanism [28], in which we know the emerging transformer architecture. A self-attention operation calculates the response of a position in an input sequence by paying attention to every position in the sequence and uses the average in the representation space, resulting each response being embedded with correlations with every other position regardless of their distance. Self-attention also carries a major merit, i.e., self-attention module can be calculated simultaneously, which dramatically accelerate the training process. Self-attention, or transformer, already becomes the de facto standard for natural language processing (NLP) tasks [42, 43], and recently, many researches explore their application in computer vision, e.g., object detection [44], image classification [45], video classification [46], and video segmentation [47]. Vision Transformer (ViT) [45] is constructed with pure transformer encoders and is applied directly to a sequence of image patches; ViT achieves competitive results on image classification tasks. After training ViT on large-scale image datasets and transferred to image recognition benchmarks, ViT gets remarkable results compared to state-of-the-art CNNs. However, there is a major drawback of ViT, i.e., it requires substantially more image data and computational resources to train compared to CNNs. Thus, the self-attention architecture (just the transformer encoder) is applied to a sequence of patch embeddings of image (preprocessed by CNNs, e.g., ResNet [2]), aiming to learn nonlocal correlations to implement the learning of semantic structural representation image.

2.4. Image Classification. Traditional image classification architectures [2, 3, 48] take advantage of convolutional networks for processing images and get remarkable performance. Convolutional network is still the de facto standard for image processing tasks and has been implemented in many practical applications. In recent research, new architectures, e.g., networks using transformer [45, 49, 50] or multilayer perceptron (MLP) [51–53], are challenging the leading position of CNNs. We also note that hybrid CNN-Transformer architecture [54–57] argues that the combination of local convolutional operation and nonlocal self-attention operation is the optimal solution for computer vision tasks. All recent works are trying to break the limitation of CNNs. Here, in our work, RGB image data is utilized to train a hybrid CNN-Transformer architecture in a self-supervised manner, and then, the model is fine-tuned for image classification tasks; it is also an attempt to explore new framework of image processing.

3. Self-Supervised Self-Attention Learning Method

In this section, the core components and implementations of TriLFrame which include the learning framework, the sample construction, and the self-attention module are presented.

3.1. Framework. The aim of TriLFrame is to learn a nonlocal semantic representation of image. The image is first processed by a convolutional operation, and then, patches of latent representations are unfolded for self-attention operation. As illustrated in Figure 3, TriLFrame takes an RGB image as input and unfolds the latent embedding into a number of patches (16 patches as in the experiment), given the former part of patches the TriLFrame predicts the latter part of patches. We use the latter part patches and the predicted patches to construct positive and negative samples for contrastive learning. First step, the RGB image is preprocessed, and a convolutional operation $f(.)$ computes the image embedding X:

$$X = f(\text{image}), \tag{1}$$

where X has dimension $\mathbb{R}^{C\times H\times W}$. As same as ViT, we break apart the latent embedding X along the spatial dimension $\mathbb{R}^{H\times W}$ to get patch 1 to patch 16, where each patch is named as $x_i \in \mathbb{R}^C$, and we have $X = \{x_i \in \mathbb{R}^C, i \in \{1, 2, \cdots, H \times W\}\}$. These patches are sent to the self-attention function transEnc$(.)$ to compute nonlocal correlations:

$$Z = \text{transEnc}(X), \tag{2}$$

where $Z = \{z_i \in \mathbb{R}^C, i \in \{1, 2, \cdots, H \times W\}\}$.

Afterwards, patches are accumulated along the patch sequence by an aggregation function $g(.)$ to get a context c_i:

$$c_i = g(z_1, z_2, \cdots, z_i). \tag{3}$$

The accumulated context c_i is of the same dimension as x_i and z_i. In our initial settings, $x_i, z_i, c_i \in \mathbb{R}^{256}$.

If feature vector $z_1, z_2, \cdots, z_i$ have embedded with semantic structural features, e.g., the key features of image patch and have been aware of the correlation with other patches, then the accumulated context c_i can predict embeddings of the rest patches by using a simple inference function $\varphi(.)$:

$$\hat{z}_{i+1} = \varphi(c_i) = \varphi(g(z_1, z_2, \cdots, z_i)), \tag{4}$$

$$\hat{z}_{i+2} = \varphi(c_{i+1}) = \varphi(g(z_1, z_2, \cdots, z_i, \hat{z}_{i+1})), \tag{5}$$

where $\hat{z}_{i+1}$ is the predicted embedding of patch $i+1$. As

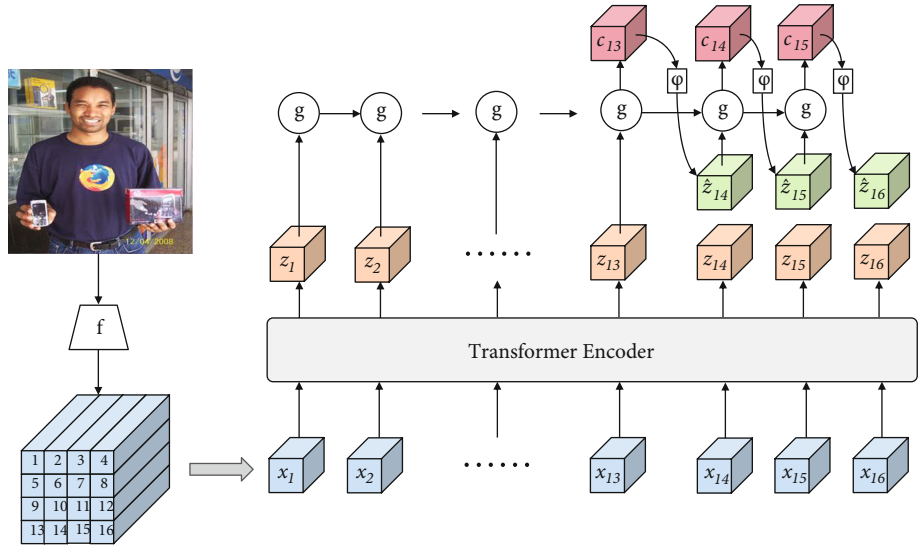


Figure 3: A diagram of self-supervised self-attention learning framework.

instructed in Seq2Seq [58], we infer future embeddings in a sequential mode. For the prediction of patch $i+2$, the context c_{i+1} which accumulates every past embeddings including the latest predicted $\hat{z}_{i+1}$, as illustrated in Equation (5). We totally infer q patches for one image.

3.2. Contrastive Learning. In TriLFrame, contrastive learning is implemented by discriminating positive "Predicted Patch - Ground-truth Patch" sample pair (named "Pred-GT" sample pair) and negative Pred-GT sample pair. Follow the idea of NCE [34] and CPC [6], a NCE variant is adopted as our contrastive loss for contrastive image learning. The proposed contrastive loss draws the predicted patches closer to the ground-truth patches while the Pred-GT sample pair does not need to be completely the same, i.e., model just needs to learn nonlocal semantic structural representation without paying attention to pixel-level details or noises.

As illustrated in Figure 4, the red arrow line connects the only positive sample pair, and the dashed black arrow line shows two negative sample pairs constructed by (1) the predicted patch embedding and a random ground-truth patch embedding of the same image and (2) the predicted patch embedding and the patch embedding of an image from the same minibatch. In TriLFrame, we break down an image representation X of dimension $\mathbb{R}^{C\times H\times W}$ into $H \times W$ patches, for i-th image patch; the ground-truth latent embedding z_i is couple with its predicted latent embedding $\hat{z}_i$; both embeddings are of the same dimension $\mathbb{R}^C$. As illustrated in Figure 4, we construct positive sample pairs with a prediction embedding and its corresponding ground-truth embedding and negative sample pairs with a prediction embedding and ground-truth embeddings at other spatial positions of the same image. We further utilize patch embeddings of images from the same minibatch to produce more negative Pred-GT pairs for contrastive learning.

The similarity score of the Pred-GT sample pair is calculated by dot product as $\hat{z}_i^T \bullet z_j$ where $\hat{z}_i^T, z_j \in \mathbb{R}^c$, i and j denote i-th and j-th patch $(i, j \in \{1, 2, \cdots, H \times W\})$. Hence, TriLFrame is to optimize the contrastive loss:

$$\mathscr{L} = -\sum_i \left[\log \frac{\exp\left(\hat{z}_i^T \bullet z_i\right)}{\sum_j \exp\left(\hat{z}_i^T \bullet z_j\right)}\right]. \tag{6}$$

The loss function in the above equation is typically the cross-entropy loss for distinguishing positive Pred-GT sample pairs from negative sample pairs. When training with minibatch, we define the following types of negative Pred-GT sample pairs to define the construction of negative samples:

(i) *Easy Negatives*. In the same minibatch, easy negatives are Pred-GT sample pairs from two images. Easy negatives are relatively easy to discriminate in general, but there exist similar patches from different images; for instance, image patches both contain a football

(ii) *Spatial Negatives*. In the same image, spatial negatives are constructed with predicted patch and ground-truth patch embeddings, where the two patches are at different spatial locations in the image, i.e., $(\hat{z}_i, z_j)$ pair with $i \neq j$

3.3. Transformer Encoder. We implement the conventional transformer architecture [28] as the transformer in TriLFrame except that the transformer decoder and the positional encoding are discarded. Although it is proven that positional embeddings make self-attention operation be aware of sequential information to some degree [44, 45], when using self-attention on image patches, the goal is to embed nonlocal correlations between patches of image, so it is not important to be aware of sequential information of patches.

As illustrated in Figure 5, the conventional transformer encoder operation gets a one-dimensional sequence of patch

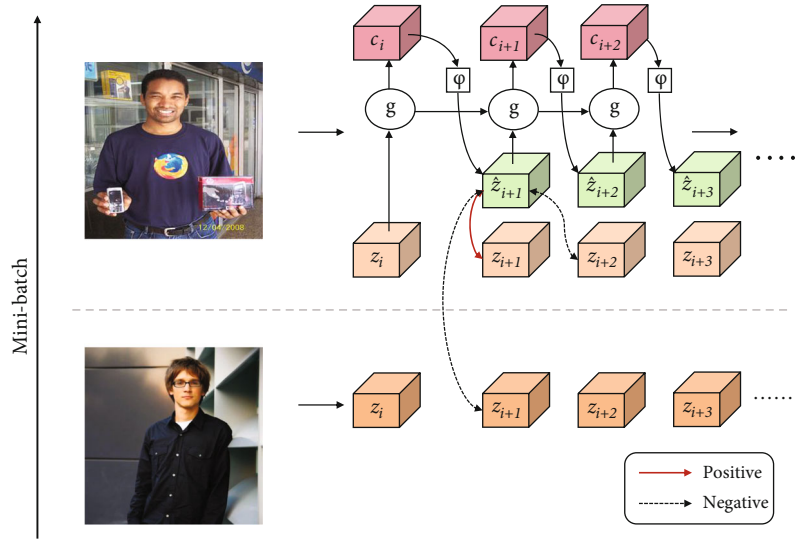


Figure 4: The construction of positive and negative samples of contrastive learning.

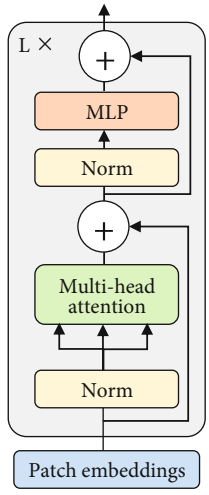


Figure 5: The architecture of transformer encoder applied in TriLFrame.

embeddings. To make the transformer applicable in three-dimensional latent embeddings of images, we break apart the image embedding $X \in \mathbb{R}^{C\times H\times W}$ along the spatial dimension $\mathbb{R}^{H\times W}$ to a series of one-dimensional patches $x_i \in \mathbb{R}^C$. Transformer encoder takes a total number of $H \times W$ patches as input. In our implementation, the transformer encoder is repeated L times, with a shared feature vector dimension C at all layers. Each layer has one multihead self-attention operation and one MLP operation, where the number of heads is H. After each transformer encoder, we get an output $Z \in \mathbb{R}^{C\times H\times W}$. Through the transformer encoder operation, each patch of image links to every other patch; thus, the nonlocal spatial dependencies are computed.

3.4. Image Processing Workflow. Due to the diversity of images generated by manufacturing systems, e.g., images of different resolution, images from different angles, panoramas or close shot, grey-scale image or RGB image, infrared images, and medical images (magnetic resonance imaging images, CT images), image data must be preprocessed for self-supervised learning afterwards. To help the model learn nonlocal semantics, we deploy the following frame-wise augmentation methods to every image in a minibatch, such as color jittering which includes random contrast, random brightness, random hue, random saturation, and random greyscale during self-supervised training. It is noted that, by introducing self-attention, in contrast to CPC, TriLFrame does not require image patches to be overlapped; this effectively avoids the network to perform feature extrapolation as the prediction.

In contrast to the one-off prediction of the latter patches presented in [6], we implement a successive predictive mechanism (i.e., latter patches are predicted in a progressive manner). As described in Equations (4) and (5), all previous context of the image (an aggregated context) is utilized to make the next inference. This successive prediction process ensures that the model makes use of every previous image patch when predicting the next patch embedding.

Batch normalization [59] (BN) is a conventional practice in deep neural networks; however, it is not adopted in CPC [6]. We argue that BN is necessary in TriLFrame, and it gives 2%-4% accuracy improvement in classification tasks we performed. It is difficult to train a hybrid CNN-transformer network without normalization either in self-supervised training stage or supervised fine-tuning stage. In this paper, BN is adopted for convolutional function and transformer encoder.

4. Experiments and Analysis

We show the experiment setups and the self-supervised training procedure in the following section, and then, we show the ablation study of TriLFrame and the evaluation of the model.

4.1. Experiment Setting

4.1.1. Network Architecture. A conventional ResNet [2] is implemented as the convolutional operation $f(.)$; ResNet consists of four residual blocks wrapping up with final channel dimension of 256. We use the output from the fourth residual block as the input to transformer encoder. In our experiments, ResNet18 is implemented. After ResNet encoder, the latent representation of an image is cut into a sequence of patches and then processed by transformer as [28] without the positional encoding module. Taking account of the number of the image patches, we set the number of attention heads and encoder layers to 2 and 1, respectively, in our experiments. This setting of transformer also forces encoder $f(.)$ to learn a better quality of semantic structural representation of image, i.e., in order to train a strong feature encoder, a weak self-attention operation transEnc$(.)$ and aggregation operation $g(.)$ is preferable. Thus, we apply a simple Convolutional Gated Recurrent Unit with the smallest kernel size (1, 1) as our aggregation operation. For inference, one simple MLP is applied in a progressive manner.

4.1.2. Self-Supervised Training. In our experiments, we use the ILSVRC ImageNet competition dataset [30] for self-supervised training. The ImageNet dataset has been used to evaluate unsupervised vision models by many works. Before encoder function $f(.)$, images are preprocessed for data augmentation; we implement random grey, random flip, random crop, and color jittering for each image before feeding to ResNet18. These augmentations help the network to avoid shortcuts, i.e., feature extrapolation, as discussed in Section 3.4.1. For self-supervision, we train TriLFrame end-to-end with Adam optimizer; we start with an initial learning rate of 10*e*-3 and weight decay rate of 10*e*-5. The learning rate is decayed by a linear function every 100 epoch and is settled at the rate of 10-5.

4.2. Evaluation Methods

4.2.1. Self-Supervised Learning Evaluation. The TriLFrame is first training with self-supervision on the ILSVRC ImageNet. Self-supervised training is initially evaluated by the validation Top 1 accuracy, i.e., the Top 1 accuracy of classifying the positive Pred-GT sample pairs from others in the validation set. Self-supervision with high validation accuracy tells that the model learns a good distribution of image embeddings. After self-supervised training, TriLFrame is further evaluated by downstream tasks, especially the image classification task on the ILSVRC ImageNet. TriLFrame is fine-tuned with a simple classification layer on the ILSVRC ImageNet in a supervised manner; after model converges, TriLFrame is then evaluated by the classification task on the ILSVRC ImageNet. We report all accuracy results as Top 1 accuracies.

4.2.2. Image Classification. Image classification is an important metric for evaluation on the self-supervised image learning approaches; thus, we take the image classification task to evaluate the TriLFrame.

After contrastive learning, the model should be able to encode the semantic structural representation of an input image from any source in manufacturing systems, and the image representation can then be used in classification task. The last aggregated context representation c is utilized to construct the image classification network as follows: at first stage, we first encode an image with the convolutional operation $f(.)$ to get the latent embedding X, which is subsequently broken into a sequence of image patches with no overlapping. The patch sequence is then sent to the self-attention operation transEnc$(.)$ to capture nonlocal dependencies, and finally, we use the aggregation function $g(.)$ to aggregate the whole sequence of patch embeddings into a context representation c which is a feature vector of the image. At second stage, the representation c is passed to a FC layer and a *Softmax* function to get the probabilities for image classification. The classification network is trained by Adam optimizer; we start with the learning rate of 10*e*-3 and a weight decay rate of 10*e*-3. Because TriLFrame is fine-tuned for image classification, so the initial learning rate of the convolutional operation $f(.)$, the self-attention operation transEnc$(.)$, and the aggregation operation $g(.)$ is set to 10^{-4}. At prediction stage, an image is preprocessed except for random crop. The final classification result is given by the *Softmax* probability.

4.3. Performance Analysis

4.3.1. Ablation Study. We conduct several ablation studies on the TriLFrame architecture, especially on the backbone encoder and transformer encoder module, to show the contribution of each module of TriLFrame and the effectiveness of deeper convolutional encoder. The ablation study is first conducted with ResNet18 as encoder, and the results are given in the upper part of Table 1. The baseline model is set up with random initialization and trained only in supervision for image classification on the ILSVRC ImageNet dataset. TriLFrame is pretrained with contrastive learning and fine-tuned with supervised learning on ILSVRC. We observe that Top 1 accuracy increases from 54.1% to 75.6%, after TriLFrame is pretrained with self-supervision. Also, it is obvious that after removing the transformer module transEnc(), the Top 1 accuracy incurs a significant drop: from 75.6% to 66.8%. This ablation study demonstrates that the TriLFrame framework is effective in capturing the semantic representation of image, and that the self-attention module plays an important role in the framework.

We also try to use deeper convolutional networks, e.g., ResNet50 and ResNet101. The results are given in the lower part of Table 1. It empirically shows that deeper convolutional encoders contribute to better self-supervised accuracy as well as better image classification accuracy. When adopting ResNet101 as backbone encoder, we observe that Top 1 accuracy of classification accuracy reaches 81.2%, which is a considerable increase compared with 75.6% by ResNet18 and 78.3% by ResNet50. This ablation study proves that deeper convolutional encoder plays a more effective role in TriLFrame architecture.

Table 1: The result of ablation studies on TriLFrame architecture. The model is pretrained with contrastive learning and then fine-tuned with supervised learning classification task on the ILSVRC ImageNet dataset. "Random Init." that represents random initialization is used for setting up. "Remove transEnc()" that represents the transformer of TriLFrame is removed; patch embeddings are aggregated without transformer encoder.

Encoder	Self-Sup. (ILSVRC) Setting	Top1 Acc.	Sup. (ILSVRC) Top1 Acc.
ResNet18	Random Init.	—	54.1
ResNet18	Remove transEnc()	68.8	66.8
ResNet18	TriLFrame	80.9	75.6
ResNet50	TriLFrame	83.7	78.3
ResNet101	TriLFrame	88.5	81.2

4.3.2. Self-Supervision Accuracy Compared with Classification Accuracy. We also conduct experiments to show the correlation of self-supervision and image classification tasks. During self-supervised training TriLFrame on the ILSVRC dataset, we perform several early stops at different self-supervision accuracy and fine-tune with supervision for image classification. Then, we demonstrate the relationship of self-supervision accuracy and image classification accuracy. For simplicity, the correlation experiments are conducted with ResNet18 as backbone encoder. The following Figure 6 shows our findings.

As shown in Figure 6, TriLFrame is trained using self-supervision on the ILSVRC dataset and is early-stopped at validation Top 1 accuracy of {54.8, 64.0, 71.4, 80.9}; for each early stop, the model is then fine-tuned with supervision for image classification. It is obvious that the performance of TriLFrame on downstream classification task is dependent on the self-supervision accuracy, i.e., TriLFrame of higher self-supervision accuracy yields higher accuracy of image classification. This correlation of self-supervision accuracy and classification accuracy shows that the image representation learnt in self-supervised training effectively capture semantic structural features which is generalized and can be used at downstream classification tasks.

4.4. Comparison with State-of-the-Art Methods. We show the comparison with state-of-the-art self-supervised methods using linear probing. Results of state-of-the-art methods are reported by the implementation in [39, 60, 61]. For fair comparison, all methods are pretrained with 224×224 image crop from the ILSVRC ImageNet dataset; data augmentation is applied accordingly. Note that we try ResNet18, ResNet50, and ResNet101 as backbone encoder in TriLFrame to show comparable results, which is different from other model settings. The results are given in Table 2. It shows that: first, the proposed TriLFrame framework outperforms the state-of-the-art method with 81.2% Top 1 accuracy, but it has a relatively large parameters of 485 M compared with other methods listed in the table. Second, when constructed with a shallow convolutional network ResNet18, TriLFrame is a quite light-weight architecture of only 75 M parameters, which is significantly less than other

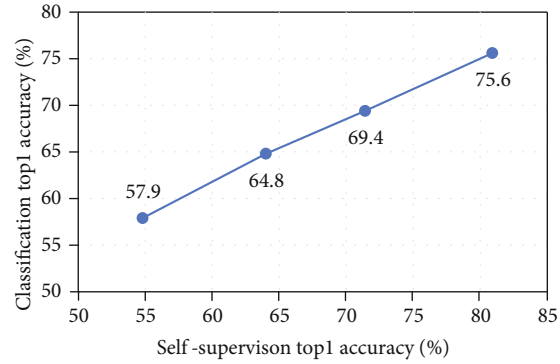


Figure 6: The correlation of self-supervision accuracy and image classification accuracy.

methods with close performance (CMC has a 94 M parameters but only 70.6% accuracy). This proves that the hybrid CNN-Transformer architecture which combines local convolutional operation and nonlocal self-attention operation is quite effective and efficient in capturing nonlocal semantic image features.

4.5. Transfer to Other Image Classification Tasks. Another important experiment to test whether self-supervised learning captures key semantic features is transfer learning. Model is first trained with self-supervision on the ImageNet; when the model settles, we follow the classification model setting as described in Section 4.2.2 except that the model will be fine-tuned end-to-end for transfer learning. Specifically, the parameters learned from self-supervision on ImageNet dataset are used to initialize the classification model, and then, the entire model will be trained with supervised learning on other image datasets. We follow the transfer learning settings and evaluation protocol in [39], and we use image datasets CIFAR [62], VOC2007 [63], Pets [64], and Flowers [65]. If features learned from self-supervision are generic and contain key semantic features, then they should be helpful in other the image datasets mentioned above. Please note that TriLFrame uses ResNet50 as encoder, which matches the conventional architecture setting. The transfer learning results are given in Table 3.

Although TriLFrame only surpasses other methods in one of the five classification tasks, it achieves competitive performances compared with state-of-the-art methods. We note that TriLFrame performs better with larger dataset; for example, TriLFrame achieves the highest accuracy on CIFAR100 and the second highest accuracy on CIFAR10; both datasets have more training samples than VOC2007, Pets, and Flowers. This characteristic of TriLFrame accords with ViT which also requires large amount of training data to get competitive performances.

4.6. Discussion. Through ablation study on the TriLFrame framework and comparison with SOTA models, we believe that the following factors contribute to the achievements of TriLFrame: First, the contrastive self-supervised training process forces the model to learn a strong semantic structural embedding of image through predicting

Table 2: The comparison with state-of-the-art self-supervised methods in ImageNet classification, evaluated by linear probing. "Params" and "Architecture" column shows the basic features of the corresponding method. "Top1 Acc." and "Top5 Acc." are reported by a linear classification on the ImageNet dataset, after models are pretrained with self-supervision.

Method	Params	Architecture	Top1 Acc. (%)	Top5 Acc. (%)
CMC	94 M	ResNet50 ×2	70.6	89.7
CPC v2	305 M	ResNet161	71.5	90.1
BYOL	375 M	ResNet50 ×4	78.6	94.2
SimCLR	375 M	ResNet50 ×4	76.5	93.2
MoCo v3	304 M	ViT-L/16	77.6	—
MoCo v3	304 M	ViT-BN-L/7	81.0	—
TriLFrame (ours)	75 M	ResNet18 + transformer	75.6	92.1
TriLFrame (ours)	265 M	ResNet50 + transformer	78.3	93.6
TriLFrame (ours)	485 M	ResNet101 + transformer	81.2	94.7

Table 3: Transfer learning results. Models are initialized with parameters from self-supervised learning on ImageNet dataset and are fine-tuned for classification task on relative datasets. A supervised training (on ImageNet) version "Supervised-IN" is also provided for comparison.

Method	Dataset Classes	CIFAR10 10	CIFAR100 100	VOC2007 20	Pets 37	Flowers 102
BYOL [39]		97.8	86.1	85.4	91.7	97.0
SimCLR [27]		97.7	85.9	84.1	89.2	97.0
Supervised-IN [27]		97.5	86.4	85.0	92.1	97.6
MoCo v3 [61]		98.9	90.5	—	93.2	97.7
TriLFrame (ours)		98.2	90.7	85.2	91.5	96.2

nondeterministic image patch embeddings given previous context knowledge of the image. Second, it is apparent that the hybrid CNN-Transformer architecture which combines local convolutional operation and nonlocal self-attention operation is quite effective and efficient in capturing nonlocal semantic image features. Third, we trust that the framework design of TriLFrame, which makes use of convolutional operation, self-attention operation, and aggregation operation, shows great performance and suits image learning tasks.

5. Conclusion

In this paper, we propose a novel self-supervised self-attention framework TriLFrame for image representation learning in manufacturing systems. TriLFrame combines the powerful local convolutional operation and the long-range nonlocal attention operation; TriLFrame learns image representation through contrasting predicted image patches and ground-truth image patches. We show that the proposed TriLFrame achieves state-of-the-art performances on image classification task on the ImageNet, with a Top 1 accuracy of 81.2%; TriLFrame also achieves competitive performance with a light-weight architecture of only 75 M parameters. When tested in transfer learning, TriLFrame is proven to be reliable in capturing semantic features for image classification tasks.

The work demonstrated shows that TriLFrame has a promising future in image-related tasks in manufacturing systems; it can be quickly transferred and deployed to applications such as anomaly detection, medical diagnosis, and road analysis. Nevertheless, TriLFrame requires extra supervision information (e.g., image labels and segmentation information) if it is to be deployed in manufacturing systems for a specific task, and the supervision information may require exquisite design and significant amount of effort. Therefore, we hope the proposed TriLFrame can be considered as a baseline or backbone framework when solving image-related tasks in manufacturing systems in the future.

Data Availability

The data that support the findings of this study are available on request from the corresponding author.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgments

This work was supported by the Education Department of Jiangxi Province of China (No. GJJ204912) and the Science and Technology Bureau of Ganzhou City of China (No. [2020]60).

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, U.S.A, 2016.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Conference and Workshop on Neural Information Processing Systems (NeurIPS)*, Lake Tahoe, U.S.A, 2012.

[4] B. Athiwaratkun, M. Finzi, P. Izmailov, and A. G. Wilson, Eds., "There are many consistent explanations of unlabeled data: why you should average," in *International Conference on Learning Representations (ICLR)*, New Orleans, U.S.A, 2019.

[5] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.

[6] A. Van Den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, https://arxiv.org/abs/1807.03748.

[7] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon et al.2019, https://arxiv.org/abs/1808.06670.

[8] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," 2019, https://arxiv.org/abs/1906.00910.

[9] X. Ji, J. F. Henriques, and A. Vedaldi, "Invariant information clustering for unsupervised image classification and segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea, 2019.

[10] L. Schmarje, M. Santarossa, S.-M. Schröder, and R. Koch, "A survey on semi-, self- and unsupervised techniques in image classification," 2020, https://arxiv.org/abs/2002.08721.

[11] G. Litjens, T. Kooi, B. E. Bejnordi et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, no. 9, pp. 60–88, 2017.

[12] Y. Taigman, M. Yang, M.'. A. Ranzato, and L. Wolf, "DeepFace: closing the gap to human-level performance in face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, U.S.A, 2014.

[13] J. Janai, F. Güney, A. Behl, and A. Geiger, "Computer vision for autonomous vehicles: problems, datasets and state of the art," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, U.S.A, 2017.

[14] H. Zhiyi, S. Haidong, Z. Xiang, Y. Yu, and C. Junsheng, "An intelligent fault diagnosis method for rotor-bearing system using small labeled infrared thermal images and enhanced CNN transferred from CAE," *Advanced Engineering Informatics*, vol. 46, article 101150, 2020.

[15] K. Yan, "Chiller fault detection and diagnosis with anomaly detective generative adversarial network," *Building and Environment*, vol. 201, article 107982, 2021.

[16] K. Yan, J. Su, J. Huang, and Y. Mo, "Chiller fault diagnosis based on VAE-enabled generative adversarial networks," *IEEE Transactions on Automation Science and Engineering*, vol. 19, no. 1, pp. 387–395, 2020.

[17] K. Yan, J. Huang, W. Shen, and Z. Ji, "Unsupervised learning for fault detection and diagnosis of air handling units," *Energy and Buildings*, vol. 210, article 109689, 2020.

[18] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[19] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, Eds., "Extracting and composing robust features with denoising autoencoders," in *ICML '08: Proceedings of the 25th international conference on Machine learning*, Helsinki, Finland, 2008.

[20] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2014, https://arxiv.org/abs/1312.6114.

[21] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 3, pp. 2672–2680, 2014.

[22] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2016, https://arxiv.org/abs/1511.06434.

[23] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "CVAE-GAN: fine-grained image generation through asymmetric training," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017.

[24] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9729–9738, 2020.

[25] O. J. Hénaff, "Data-efficient image recognition with contrastive predictive coding," in *Proceedings of the 37th International Conference on Machine Learning*, Long Beach, U.S.A, 2019.

[26] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, U.S.A, 2019.

[27] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning,*, 2020.

[28] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017.

[29] Y. Liu, J. Ma, Y. Xie et al., "Contrastive predictive coding with transformer for video representation learning," *Neurocomputing*, 2021.

[30] O. Russakovsky, J. Deng, H. Su et al., "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[31] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1422–1430, Santiago, Chile, 2015.

[32] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Computer Vision – ECCV 2016. ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9907 of Lecture Notes in Computer Science, pp. 649–666, Springer, Cham, 2016.

[33] C. Doersch and A. Zisserman, "Multi-task self-supervised visual learning," in *Proceedings ofthe IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2051–2060, Honolulu, U.S.A, 2017.

[34] M. U. Gutmann and A. Hyvarinen, *Noise Contrastive Estimation: A New Estimation Principle for Unnormalized Statistical Models*, AISTATS, 2010.

[35] H. Fang, S. Wang, M. Zhou, J. Ding, and P. Xie, "CERT: contrastive self-supervised learning for language understanding," 2005, https://arxiv.org/abs/2005.12766.

[36] J. M. Giorgi, O. Nitski, G. D. Bader, and B. Wang, "DeCLUTR: deep contrastive learning for unsupervised textual representations," 2006, https://arxiv.org/abs/2006.03659.

[37] T. Han, W. Xie, and A. Zisserman, "Video representation learning by dense predictive coding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Venice, Italy, 2020.

[38] T. Han, W. Xie, and A. Zisserman, "Memory-augmented dense predictive coding for video representation learning," in *European Conference on Computer Vision (ECCV)*, 2020.

[39] J.-B. Grill, F. Strub, F. Altché et al., "Bootstrap your own latent: a new approach to self-supervised learning," 2020, https://arxiv.org/abs/2006.07733.

[40] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[41] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 132–149, Munich, Germany, 2018.

[42] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," 2018, https://arxiv.org/abs/1810.04805.

[43] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, *Improving Language Understanding by Generative Pre-Training*, OpenAI, 2018.

[44] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision – ECCV 2020. ECCV 2020*, vol. 12346 of Lecture Notes in Computer Science, Springer, Cham.

[45] A. Dosovitskiy, L. Beyer, A. Kolesnikov et al., "An image is worth 16x16 words: transformers for image recognition at scale," 2021, https://arxiv.org/abs/2010.11929.

[46] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7794–7803, Salt Lake City, U.S.A, 2018.

[47] Y. Wang, X. Zhaoliang, X. Wang et al., "End-to-end video instance segmentation with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[48] M. Tan and V. L. Quoc, "EfficientNet: rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, U.S.A, 2019.

[49] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Z. Jiang, Q. Hou, and J. Feng, Eds., "DeepViT: towards deeper vision transformer," 2021, https://arxiv.org/abs/2103.11886.

[50] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, Eds., "Training data-efficient image transformers & distillation through attention," in *Proceedings of the 38th International Conference on Machine Learnin*, 2021.

[51] Q. Hou, Z. Jiang, L. Yuan, M.-M. Cheng, S. Yan, and J. Feng, Eds., "Vision permutator: a permutable MLP-like architecture for visual recognition," 2021, https://arxiv.org/abs/2106.12368.

[52] H. Touvron, P. Bojanowski, M. Caron et al., "ResMLP: feedforward networks for image classification with data-efficient training," 2021, https://arxiv.org/abs/2105.03404.

[53] I. Tolstikhin, N. Houlsby, A. Kolesnikov et al., "MLP-mixer: an all-MLP architecture for vision," 2021, https://arxiv.org/abs/2105.01601.

[54] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, Eds., "Early convolutions help transformers see better," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[55] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and V. L. Quoc, Eds., "Attention augmented convolutional networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea, 2019.

[56] C. Li, T. Tang, G. Wang et al., "BossNAS: exploring hybrid CNN-transformers with block-wisely self-supervised neural architecture search," https://arxiv.org/abs/2103.12424.

[57] J. Chen, T. Cai, W. He et al., "A blockchain-driven supply chain finance application for auto retail industry," *Entropy*, vol. 22, no. 1, p. 95, 2020.

[58] I. Sutskever, O. Vinyals, and V. L. Quoc, "Sequence to sequence learning with neural networks," in *Conference and Workshop on Neural Information Processing Systems (NeurIPS)*, Montréal, Canada, 2014.

[59] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, pp. 448–456, Lille, France, 2015.

[60] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," https://arxiv.org/abs/2003.04297.

[61] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," 2021, https://arxiv.org/abs/2104.02057.

[62] A. Krizhevsky, *Learning Multiple Layers of Features from Tiny Images*, Technical report, University of Toronto, 2009.

[63] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[64] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, "Cats and dogs," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012.

[65] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, Bhubaneswar, India, 2008.