Pattern Recognition 157 (2025) 110893

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Patient teacher can impart locality to improve lightweight vision transformer on small dataset

Jun Ling [a], Xuan Zhang [a,b,*], Fei Du [a], Linyu Li [c,d], Weiyi Shang [e], Chen Gao [g], Tong Li [f]

[a] School of Software, Yunnan University, Yunnan 650091, China
[b] Key Laboratory of Software Engineering of Yunnan Province, Yunnan 650091, China
[c] Key Laboratory of High-Confidence Software Technologies (MoU), Peking University, Beijing 100871, China
[d] School of Computer Science, Peking University, Beijing 100871, China
[e] Department of Electrical and Computer Engineering, University of Waterloo, Ontario, Canada
[f] School of Big Data, Yunnan Agricultural University, Yunnan, 650201, China
[g] School of Information Science and Engineering, Yunnan University, Yunnan 650091, China

ARTICLE INFO

Keywords:
Vision transformer
Knowledge distillation
Curriculum learning
Small dataset

ABSTRACT

Vision Transformer (ViT) has achieved unprecedented success in vision tasks with the assistance of abundant data. However, the lack of inductive bias in lightweight ViT makes learning locality challenging on small datasets, leading to poor performance. This limitation impedes the application of lightweight ViT in scenarios with limited datasets and computational power. Knowledge Distillation (KD) allows student models to benefit from the teacher model. However, in the progressive learning stage, traditional single-stage KD methods are usually suboptimal for delivering fixed knowledge to the growing student model. To address these issues, we propose a simple yet effective two-stage KD method called Curriculum Information Knowledge Distillation (CIKD) for the first time. Specifically, we incorporate a curriculum learning framework, progressing from easy to difficult, in the KD curriculum. At the first stage, *i.e.*, Attention Locality Imitation (ALI), the student model learns locality from the low-level semantic features of the teacher model through self-attention distillation. Afterward, at the second stage, *i.e.*, Logit Mimicking (LM), the student model learns label information and high-level semantic logit from the teacher model. Without bells and whistles, our approach achieves state-of-the-art results on 8 small-scale datasets with ViT-Tiny (5.0M). Our code and model weights are available at: https://github.com/newLLing/CIKD.

1. Introduction

Vision Transformer (ViT) [1] is now widely used in computer vision tasks, demonstrating unprecedented strong performance in situations with abundant data. For instance, in image classification tasks, ViT, which lacks inductive biases, is initially pre-trained on the extensive JFT-300M dataset. It is then fine-tuned on the Imagenet-1K dataset [2]. In this process, ViT set new state-of-the-art (SOTA) performance benchmarks. However, it is challenging for lightweight ViT to achieve optimal performance in practical applications when trained on small datasets. Many studies now attribute the reason for its poor performance to the lack of inductive bias in the ViT architecture [1,3]. Specifically, the locality, which is of great importance for understanding images, is hard to learn with a small dataset due to the high flexibility and the intrinsic globality of the self-attention mechanism in ViT.

Knowledge distillation (KD) allows lightweight ViT to benefit from high-performing ViT [4,5]. However, traditional KD only transfers fixed knowledge in the progressive learning process [4,6], which cannot maximize mutual information [7]. Recently, in the fields of Natural Language Processing and Computer Vision, two-stage KD methods have been proposed [8,9]. However, these methods are based on training using a masking mechanism, which requires an additional decoder to compute the knowledge distillation loss. These methods raise concerns as the artificially partitioned encoder and decoder may constrain the positions where semantic information can appear, potentially leading to the loss of low-level information. As a result, such methods generally exhibit suboptimal generalization capabilities [10], and they do not consider how to perform two-stage KD on small datasets.

As shown in Fig. 1, (a) ViT-Base pre-trained with the Dino method [11] has demonstrated an emphasis on local features of images on the

* Corresponding author at: School of Software, Yunnan University, Yunnan 650091, China.
E-mail addresses: cs.lingjun@gmail.com (J. Ling), zhxuan@ynu.edu.cn (X. Zhang), dufei@ynu.edu.cn (F. Du), xltx_youxiang@qq.com (L. Li), wshang@uwaterloo.ca (W. Shang), cg2@mail.yun.edu.cn (C. Gao), tli@ynu.edu.cn (T. Li).

https://doi.org/10.1016/j.patcog.2024.110893
Received 5 March 2024; Received in revised form 12 May 2024; Accepted 13 August 2024
Available online 16 August 2024

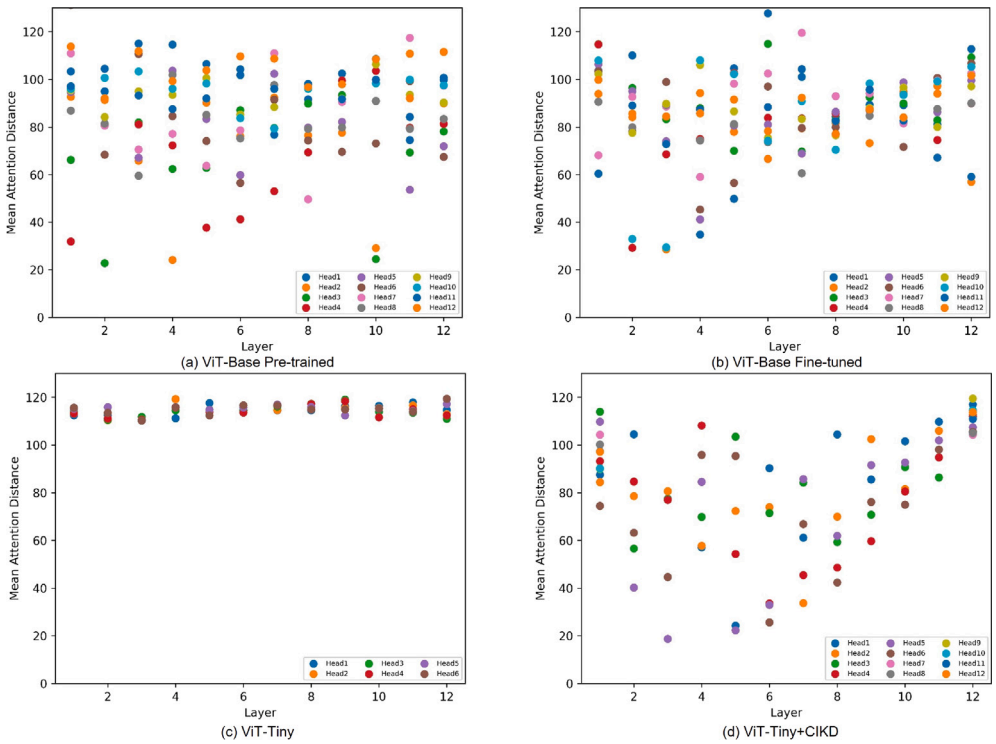

Fig. 1. Comparison of Mean Attention Distance (MAD) [1] of (a) ViT-Base Pre-trained, (b) ViT-Base Fine-tuned, (c) ViT-Tiny and (d) ViT-Tiny + CIKD.

Cifar100 dataset. Results from (b) ViT-Base Fine-tuned indicate that after fine-tuning, ViT learns to allocate attention more reasonably, with shallow blocks focusing more on local features and deep blocks prioritizing global features. However, (c) ViT-Tiny was trained directly on small datasets without pre-training. Its results indicate that it appears to have difficulty capturing local features and does not effectively learn hierarchical information from locality to globality. Both the shallow and deep layers of ViT-Tiny focus on global features, which is not conducive to its understanding of images.

We find that first using the Dino pre-training method, followed by fine-tuning, essentially teaches the model how to perceive objects in images during the pre-training stage and then uses fine-tuning to guide the model in recognizing what objects are present in images from a specific dataset. This approach shares a similarity with the principles of human education. In human education, teachers initially instruct students with simple courses, starting from the basics, allowing students to acquire fundamental skills. After students have mastered more straightforward courses, more complex ones are introduced to guide them in learning abstract and intricate concepts. We believe that the patient teacher model should possess patience, allowing the student model to learn progressively from easy to difficult tasks. Therefore, we develop a two-stage knowledge distillation for the student model called Curriculum Information Knowledge Distillation (CIKD) by combining knowledge distillation and curriculum learning. In the first stage, *i.e.*, Attention Locality Imitation (ALI) encourages the student model to focus only on the intrinsic features of the images and closely mimic the self-attention behavior (*i.e.*, scaled dot-product of query, key, and value) of the teacher model. This is akin to injecting local inductive biases into the student model, enhancing its feature extraction capability. This also enables the student model to effectively mimic the process of the teacher model transitioning from shallow-level capture of local features in images to deep-level capture of global features, ultimately enhancing feature extraction capability. The second stage, *i.e.*, Logit Mimicking (LM), involves constraining the student model's outputs to match those of the teacher model, enabling the student model to learn information corresponding to image labels. The learning process from the teacher model's low-level semantic features to high-level semantic logit and label follows a curriculum, progressing from easy to difficult, consistent with human learning paradigms. Furthermore, this two-stage knowledge distillation method completes knowledge transfer between the teacher and student [8,9], also known as maximizing mutual information [7]. To the best of our knowledge, we are the first to use total image inputs for a two-stage distillation method in ViT.

The simplicity of our approach lies in the fact that it only requires fine-tuning the ViT-Base pre-trained using the Dino method [11] on the target dataset once. Then, setting it to evaluate mode allows it to serve as the two-stage teacher model to guide the student model's learning process, and there is no need to modify particular structures for CIKD. The effectiveness of our proposed method is demonstrated by achieving impressive results, such as a TOP-1 Accuracy (%) of 89.8 on the CIFAR100 [12] dataset, 98.7 on the CIFAR10 [12] dataset, 96.8 on the Oxford Flowers [13] dataset, 79.7 on the FGVC-Aircraft [14] dataset, 93.7 on the CINIC10 [15] dataset, 93.5 on the Oxford-IIITPets [16] dataset, 88.5 on Chaoyang dataset [17] and 88.1 on the Stanford Cars [18] dataset using ViT-Tiny with CIKD. In particular, after using our method, CIKD, ViT-Tiny only needs to use existing pre-trained teachers for knowledge distillation without pre-training on large-scale data sets. Model parameters of ViT-Tiny are 15.5 times smaller than the teacher's model, but ViT-Tiny maintains the 97.8% accuracy of the teacher's model, which is 91.8% accuracy. For the Cifar10 dataset, the model parameters of ViT-Tiny are 15.5 times smaller compared to the teacher model, yet the accuracy surpasses that of the teacher model. Specifically, the Top-1 Accuracy of the student model reaches 98.7%, whereas the teacher model achieves 98.6%.

In summary, our main contributions are:

- We introduce a progressive knowledge distillation curriculum method called CIKD. This method simulates the human progressive learning process, with teachers imparting knowledge from easy to difficult, thereby enhancing the efficiency of knowledge distillation.
- The proposed Attention Locality Imitation (ALI) can facilitate the student model learning locality from the teacher model, which

can significantly improve student model generalization in the case of a small amount of data. Through the second stage, Logit Mimicking (LM), the student model acquires knowledge from the teacher model regarding label information, thereby maximizing model performance.

- The experimental results demonstrate that the lightweight ViT-Tiny trained using CIKD exhibits competitive performance across 8 small-scale datasets, comparable to the current state-of-the-art pre-training and fine-tuning methods. Furthermore, the effectiveness of our approach is further validated through visual analysis.

2. Related work

Vision Transformer: Convolutional Neural Networks [19,20] have long dominated the field of computer vision. Their inherent inductive biases, including locality and spatial invariance, were specifically designed for image recognition tasks. Locality, in particular, enables them to perform exceptionally well on small datasets. The Vision Transformer (ViT) model, based on the Self-Attention mechanism [1], has gradually started to replace Convolutional Neural Networks as the dominant architecture in computer vision when ample data is available [21,22]. However, ViT still struggles to perform well on small datasets.

ViT for Small Datasets: ViT demonstrates unsatisfactory performance when trained on small datasets such as Cifar100 [12]. One solution is to carefully design the architecture of ViT [23,24] to introduce inductive bias. Another approach is to use knowledge distillation methods to transfer inductive bias to ViT [25–27]. MCT-KD [28] proposed a momentum contrast transformer to enable ViT to be well-trained on small datasets. DeiT [3] explored for the first time the use of knowledge distillation for training ViT using a teacher model with a CNN architecture. This study primarily focuses on using knowledge distillation to introduce the inductive bias of locality on small datasets.

Curriculum Learning: Curriculum learning, originally introduced by prior research [29], is a method for training networks by organizing the sequence of learning tasks and incrementally increasing the learning difficulty. This training strategy has been widely adopted in computer vision [30] and natural language processing [31]. RCO [32] proposed using the teacher's intermediate state sequences as a curriculum to supervise the student model's learning at different stages. LFME [33] suggested using the teacher model as a measure of sample difficulty and organizing training samples from easy to hard, enabling the model to learn feature space progressively from simple to challenging samples. However, these methods often require complex design and computational processes. CEAD [34] allowed for a structured approach to teaching and learning within the Graph Transformers architecture, where teachers initially provide students with rigorous supervision and guidance and then gradually allow students to explore and innovate freely outside of the classroom. TC3KD [35] proposed a novel knowledge distillation method via teacher-student cooperative curriculum customization. In contrast, our proposed CIKD involves using knowledge distillation information as a curriculum, allowing the student model to first acquire locality in the ALI stage and then learn the teacher model's high-level semantic logit and label information in the LM stage to achieve better performance. Our approach is comparatively more straightforward.

Knowledge Distillation: The initial concept of knowledge distillation involved transferring "dark knowledge" from a large model to a smaller one. Subsequently, to improve distillation efficiency, a wide range of distillation methods emerged. These methods can mainly be categorized into two paradigms: logit-based distillation [36,37] and feature-based distillation [38,39]. TinyMIM [4] explored various distillation targets to transfer the success of large MIM-based pre-trained models to smaller ones. MiniLM [6] and MiniLMv2 [40] used knowledge distillation to compress the language model in the field of natural language processing. The mentioned research on the Multi-Head Self-Attention distillation in feature distillation has paid little attention to low-layer features because shallow features typically have smaller receptive fields and lack semantic content. LSFTN [41] adopted a student-aware teacher learning procedure before knowledge distillation. LG [42] used the locality of CNN to improve the performance of ViT. In CIKD, we utilize both shallow and deep features for imitation in the ALI stage because distilling this early self-attention behavior can guide the student model on how to form better attention maps initially [5]. This results in an improved feature extraction capability for the student model, enabling it to perform more efficiently in feature extraction on small datasets.

3. Curriculum information knowledge distillation

Our Curriculum Information Knowledge Distillation (CIKD) method primarily focuses on transferring the teacher model's self-attention behavior in the first stage. The student model learns the teacher's low-level semantic features, allowing the teacher model to impart locality, thereby enhancing the student model's feature extraction capability to focus correctly on objects within images. After going through the first stage, ALI, the student model can be seen as being imbued with a form of inductive bias towards locality. Combined with the second stage, LM, which imparts label information and the teacher model's high-level semantic logit, the student model gains a better understanding of the information related to object-label correspondence within images. As demonstrated in Table 9, reversing the two-stage knowledge distillation process, conducting the LM stage first, followed by the ALI stage, leads to poor performance. This further underscores the viability of our approach to knowledge distillation, starting with easier tasks before progressing to more challenging ones.

In this section, we introduce the first-stage distillation method, Attention Locality Imitation, in Section 3.1, followed by an explanation of the second-stage distillation method, Logit Mimicking, in Section 3.2. (For details, please refer to algorithm 1.) Finally, in Section 3.3, we provide a theoretical explanation of the feasibility of the two-stage distillation method.

3.1. Attention locality imitation

During the training process, the Attention Locality Imitation (ALI) involves the forward propagation of the teacher model, the forward propagation of the student model, and backpropagation. As shown on the right side of Fig. 3, the Vanilla Student only needs to mimic the self-attention behavior of two transformer blocks, specifically the first-level transformer block and the Learning Objective Block in the teacher model. The choice of the Learning Objective Block may vary depending on the scale of the dataset. During the forward propagation process, the self-attention behavior of the first-level transformer block and the Learning Objective Block in the teacher model is compared with the self-attention behavior of the first and last-level transformer blocks in the student model to calculate the Kullback–Leibler (KL) divergence, resulting in the ALI loss. In the first stage of ALI, label information is not utilized. The model's sole focus is on the image's features. It improves its feature extraction capabilities by mimicking the self-attention behavior of the teacher model. After the first stage, the teacher model essentially imparts local inductive bias to the student model.

3.1.1. Transformer block

A ViT consists of N sequentially stacked transformer blocks. For simplicity, let us assume that the input to the ith transformer block, denoted as $X_i \in \mathbb{R}^{P\times H}$, where $i \in [1, N]$, P represents the number of patches, and H is the dimension of hidden features. The output of each transformer block is used as the input of the next one. Each transformer block consists of a Multi-Head Self-Attention Layer (MHSA), a Fully

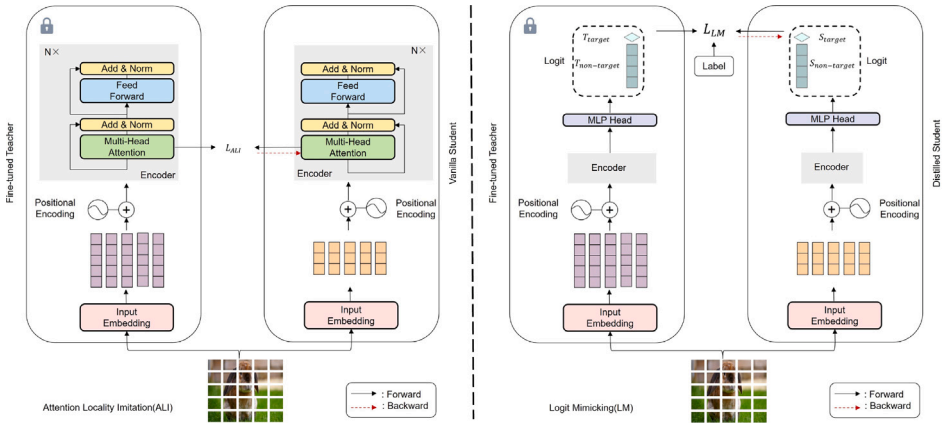

Fig. 2. Schematic diagram of Curriculum Information Knowledge Distillation (CIKD) proposed. In the Attention Locality Imitation (ALI) stage (left image), after undergoing patch transformation, the image is separately fed into the encoders of both the teacher model and the student model for feature extraction. In the Logit Mimicking (LM) stage (right image), the student model from the first stage continues to serve as the student model for the second stage.

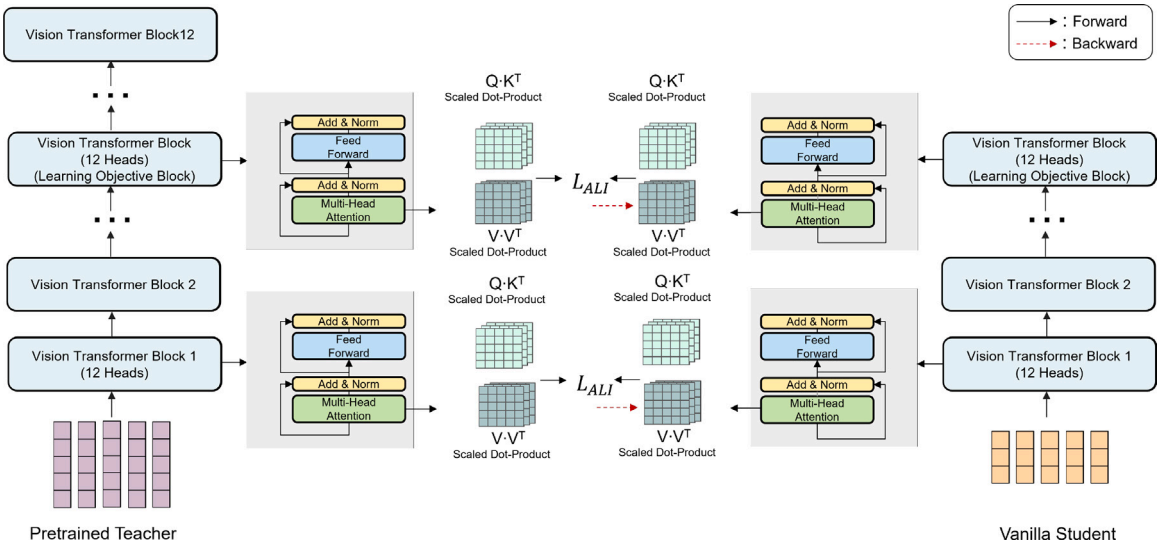

Fig. 3. Schematic representation of the proposed Attention Locality Imitation (ALI). Optimizing the student model (ViT-Tiny) to mimic the self-attention behavior of the teacher model (ViT-Base) using the original image converted into patches as input. We set the number of heads in the first layer and the Learning Objective Block (LOB) of the student model to be the same as the teacher model.

Connected Feed-Forward Network(FFN), and Layer Normalization(LN). The forward pass through the ith Transformer Block can be represented as:

$$\begin{aligned} T_i &= \text{MHSA}(\text{LN}(X_i)), \\ \dot{T}_i &= T_i + X_i, \\ \ddot{T}_i &= \text{FFN}(\text{LN}(\dot{T}_i)), \\ X_{i+1} &= \dot{T}_i + \ddot{T}_i, \end{aligned} \tag{1}$$

where MHSA() represents the Multi-Head Self-Attention layer, LN() represents the Layer Normalization, and FFN() represents the fully connected Feed-Forward Network.

3.1.2. Self-attention behavior

In the MHSA module, query, key, and value are the most fundamental and critical vectors. For the kth head in the ith Transformer Block, we can compute its Query and Key, as well as the Value's self-attention behavior. They are implemented as scaled dot-product:

$$SAB_{i,k}^{QK} = \text{Softmax}\left(\frac{Q_i^k (K_i^k)^T}{\sqrt{H/K}}\right), \tag{2}$$

$$SAB_{i,k}^{VV} = \text{Softmax}\left(\frac{V_i^k (V_i^k)^T}{\sqrt{H/K}}\right). \tag{3}$$

In the ith Transformer Block, SAB_i^{QK} is constructed by stacking K $SAB_{i,k}^{QK}$ modules, and similarly, SAB_i^{VV} is formed by stacking K $SAB_{i,k}^{VV}$ modules. The dimensions of SAB_i^{QK} and SAB_i^{VV} are denoted as (BS, K, N, N), where BS represents the batch size, K represents the number of heads in the Multi-Head Attention, and N represents the number of patches.

To ensure an accurate self-attention behavior, we match the number of heads in the Multi-head Self-Attention of the first and last layers of the student model with the number of heads in the Multi-head Self-Attention of the teacher model. To facilitate comprehensive imitation of the self-attention module, we transfer the self-attention behavior of $Query-Key^T$ and $Value-Value^T$ from the teacher model to the student model. This process ensures that the student model effectively learns to focus on local features from the teacher model.

For the student model to mimic the self-attention behavior of the teacher model, we set the number of attention heads in the first and last Transformer Blocks of the student model to be the same as that of the teacher model, allowing for the calculation of self-attention behavior. Our approach does not incur additional overhead. We define the number of teacher attention heads as K and calculate the KL-divergence between SAB_i^{QK} (see Eq. (2)) and SAB_i^{VV} (see Eq. (3)) of the teacher model and the student model as the loss function. This optimization

encourages the student model to imitate the self-attention behavior of the teacher model. We define $S^{QK}_{T-k,i}$ to represent the self-attention behavior of k heads of the $Query-Key^T$ of the teacher model's ith layer transformer block and $S^{QK}_{S-k,i}$ represent the self-attention behavior of k heads of the $Query-Key^T$ of the Student model's ith layer transformer block. The $Value-Value^T$ definitions of $S^{VV}_{T-k,i}$ and $S^{VV}_{S-k,i}$ are similar defined. We define the Learning Objective Block as LOB. Therefore:

$$L^{1-1}_{ALI} = \frac{1}{K}\sum_{k=1}^{K}\left(L_{KL}\left(SAB^{QK}_{T-k,1}, SAB^{QK}_{S-k,1}\right) + L_{KL}\left(SAB^{VV}_{T-k,1}, SAB^{VV}_{S-k,1}\right)\right), \tag{4}$$

$$L^{LOB-12}_{ALI} = \frac{1}{K}\sum_{k=1}^{K}\left(L_{KL}(SAB^{QK}_{T-k,LOB}, SAB^{QK}_{S-k,12}) + L_{KL}(SAB^{VV}_{T-k,LOB}, SAB^{VV}_{S-k,12})\right), \tag{5}$$

$$L_{ALI} = L^{1-1}_{ALI} + L^{LOB-12}_{ALI}. \tag{6}$$

During the entire training process, where the teacher model remains frozen, and the student model is trainable, we calculate the KL-divergence between the SAB of the first layer of the student model and the SAB of the first layer of the teacher model to obtain L^{1-1}_{ALI}. Further, we compute the KL-divergence between the SAB of the LOB layer of the student model and the SAB of the last layer of the teacher model to obtain L^{LOB-12}_{ALI}. Then, we add them together to obtain L_{ALI}.

In the ALI stage, the selection of the Learning Objective Block plays a crucial role. Therefore, we summarize the selection strategy of the Learning Target Block by extracting data proportionally from the same dataset and experimenting on different datasets (see Fig. 4). The principle is that the closer the block is to the input layer of the teacher's model, the smaller the amount of data; hence, it should be selected as the Learning Target Block. Subsequently, knowledge distillation should be conducted between the last layer of the selected block of the student and teacher models. Additionally, it is essential to distill knowledge from both the shallow block of the student model and the teacher model.

3.2. Logit mimicking

After undergoing the Attention Locality Imitation, the student model has gained locality. However, limited by the smaller model parameters and smaller dataset, after the first stage, the student model cannot achieve satisfactory performance even after further fine-tuning on the small dataset. To improve the performance further, we start the second stage of LM, which aims to quickly learn the correspondence between image features and their respective labels. For the LM stage, we continue to use the ViT-Base Fine-Tuned model as the teacher, but in contrast to the first stage of Attention Locality Imitation, we use the teacher model's logit outputs for distillation in this stage. The student model distilled in the first stage is referred to as the Distilled Student for the second stage (see Fig. 2). Inspired by prior research [37], the traditional distillation loss is decomposed into two parts, combined with the cross-entropy loss between the class probabilities from the Distilled Student and the label information as the loss function for the Distilled Student. We define the model's classification probability for C classes as $\mathbf{P} = [p_1, p_2, \ldots, p_i, \ldots, p_c] \in \mathbb{R}^{1\times C}$,where p_i is the probability of the ith class. Therefore:

$$p_i = \frac{\exp(z_i)}{\sum_{j=1}^{C}\exp(z_j)}, \tag{7}$$

where z_i and z_j represent the logit for the ith class and the jth class, respectively. The probability for the target class t can be represented as $\breve{p}_t$, and all other non-target classes can be represented as $\breve{p}_{\setminus t}$:

$$\breve{p}_t = \frac{\exp(\frac{z_t}{T})}{\sum_{j=1}^{C}\exp(\frac{z_j}{T})}, \tag{8}$$

$$\breve{p}_{\setminus t} = \frac{\sum_{k=1,k\neq t}^{C}\exp\left(\frac{z_k}{T}\right)}{\sum_{j=1}^{C}\exp\left(\frac{z_j}{T}\right)}. \tag{9}$$

We can calculate $\hat{p}_i = [\hat{p}_1, \ldots, \hat{p}_{t-1}, \hat{p}_{t+1}, \ldots, \hat{p}_C] \in \mathbb{R}^{1\times(C-1)}$ using a softmax function with temperature T as follows:

$$\hat{p}_i = \frac{\exp\left(\frac{z_i}{T}\right)}{\sum_{j=1,j\neq t}^{C}\exp\left(\frac{z_j}{T}\right)}. \tag{10}$$

The Target Class Knowledge Distillation (TCKD) [37] represents the similarity between the teacher's and student's binary probabilities of the target class:

$$TCKD = \breve{p}^{Te}_t \log\left(\frac{\breve{p}^{Te}_t}{\breve{p}^{St}_t}\right) + \breve{p}^{Te}_{\setminus t}\log\left(\frac{\breve{p}^{Te}_{\setminus t}}{\breve{p}^{St}_{\setminus t}}\right). \tag{11}$$

The Non-Target Class Knowledge Distillation (NCKD) [37] represents the similarity between the teacher's and student's probabilities among non-target classes:

$$NCKD = \breve{p}^{Te}_{\setminus t}\sum_{i=1,i\neq t}^{C}\hat{p}^{Te}_i \log\left(\frac{\hat{p}^{Te}_i}{\hat{p}^{St}_i}\right). \tag{12}$$

The knowledge distillation loss function for LM can be represented as:

$$L_{LM} = TCKD + \alpha NCKD, \tag{13}$$

we define the output probability of the student model as P, the label information as Y, the logit output of the student model as z^{St}, the logit output of the teacher model as z^{Te}, and the temperature as T. The loss function for the second stage is defined as:

$$L = L_{CE}(P, Y) + L_{LM}(z^{St}, z^{Te}, T), \tag{14}$$

where L_{CE} represents the cross-entropy loss function.

As indicated in Table 4, models utilizing single-stage distillation demonstrate lower performance. It is noteworthy that consolidating the two stages into a single stage may complicate model optimization and result in unsatisfactory performance. Due to our precise design of the two-stage knowledge distillation, incorporating LM allows us to achieve improved performance.

3.3. Analysis

Our proposed two-stage distillation method is more effective than single-stage methods. This can be observed from Fig. 6, where it is evident that two-stage distillation learns more locality. Additionally, Table 4 shows that the two-stage distillation achieves better performance. From a theoretical perspective, we can view two-stage distillation by considering mutual information as discussed in [7]. Our approach allows us to analyze how two-stage distillation facilitates substantial knowledge transfer. Knowledge distillation can be explained as the process of maximizing the mutual information (I) between the student model (F^S) and the teacher model (F^T). Representing the parameters of the student model as Θ^S and the teacher model parameters as Θ^T. The dataset used for feature-based distillation, which only utilizes the semantic information of images, can be represented as X^F. In contrast, the dataset used for logit-based distillation, which incorporates both the visual and label information of images, can be represented as X^L. Using single-stage knowledge distillation can be represented as:

$$\arg\max_{\Theta^S} \quad I_{\Theta^S,\Theta^T}(F^T, F^S|X^F), \tag{15}$$

Table 1
Summary of datasets used in the experiments.

Dataset	Image size	Train size	Test size	Classes
CIFAR100 [12]	32*32	50,000	10,000	100
CIFAR10 [12]	32*32	50,000	10,000	10
CINIC10 [15]	32*32	90,000	90,000	10
Oxford Flowers [13]	32*32	2040	6149	102
Oxford-IIITPets [16]	32*32	3680	3669	37
FGVC-Aircraft [14]	224*224	6667	3533	102
Stanford Cars [16]	224*224	8144	8041	196
Chaoyang [17]	512*512	3357	2139	4

$$\arg\max_{\Theta^S} \quad I_{\Theta^S,\Theta^T}(F^T, F^S|X^L), \tag{16}$$

where Eq. (15) represents the process of maximizing mutual information using Feature-based distillation, and Eq. (16) represents the process of maximizing mutual information using Logit-based distillation. Our proposed CIKD can be represented as:

$$\arg\max_{\Theta^S} \quad (I_{\Theta^S,\Theta^T}(F^T, F^S|X^F) + I_{\Theta^S,\Theta^T}(F^T, F^S|X^L) - I_{\Theta^S,\Theta^T}(F^T, F^S|(X^F, X^L))). \tag{17}$$

The mutual information defined in Eq. (17) is larger than that defined in Eqs. (15) and (16), indicating that our CIKD method can transfer more mutual information. Knowledge distillation primarily emphasizes transferring image feature knowledge from unlabeled datasets to the teacher model, enhancing its feature extraction capabilities. On the other hand, knowledge distillation on labeled datasets also imparts knowledge of feature extraction and labeling information. However, there are both similarities and disparities in the feature extraction knowledge between these two approaches. As illustrated in Fig. 6, "(a) ViT-Tiny + LM" and "(c) ViT-Tiny + ALI" exhibit distinct effects on the image, indicating differing focal points in feature extraction. As shown in Fig. 1, the student model and the teacher model have the closest CKA after using CIKD. This is because two-stage distillation completes knowledge transfer so that the mutual information between the student model and the teacher model is as large as possible.

4. Algorithm

To provide a more precise explanation of our approach, we illustrate the algorithmic process using pseudo-code in algorithm 1. In the first stage, ALI, given the fine-tuned teacher model, we optimize the student model through backward gradient propagation. This process involves utilizing the self-attention behavior of the first layer and the Learning Objective Block of the teacher model, along with the first and last layers of the student model, as inputs to the loss function. Moving to the second stage, we initialize the classified head of the distilled student model. Subsequently, we optimize the student model again through backward gradient propagation. In this stage, we use the logit output from the teacher model along with the logit output from the student model as inputs to the loss function.

Algorithm 1 The training process of CIKD

First Stage: The training process of the Attention Locality Imitation (ALI)
Input: Training set: X, the Learning Objective Block: LOB, teacher model: θ_T, epoch: N.
Output: Distilled student model: θ_{S_ALI}.
Initialize: Initialize Student model θ_S, $n = 1$, set the teacher model θ_T in evaluation mode.
for $n = 1$ to N **do**
  H_T = Embedded_Patches_Teacher(X) ▷ Convert image into token input.
  H_S = Embedded_Patches_Student(X) ▷ Convert image into token input.
  for $i = 1$ to LOB **do**
    if $i ==$ 1 or $i ==$ LOB **then**
      SAB_T = Transformer_Block(H_T) ▷ Obtain the Self-Attention Behavior of the first layer and the LOB layer of θ_T with *eq.* (2) and *eq.* (3).
      SAB_S = Transformer_Block(H_S) ▷ Obtain the Self-Attention Behavior of the first layer and the LOB layer of θ_S with *eq.* (2) and *eq.* (3).
    end if
    H_T = Transformer_Block(H_T) ▷ Forward propagate through the Transformer Block of the teacher model θ_T with *eq.* (1) .
    H_S = Transformer_Block(H_S) ▷ Forward propagate through the Transformer Block of the student model θ_S with *eq.* (1) .
    $i = i + 1$
  end for
  Obtain L_{ALI} by calculating *eq.* (4) to *eq.* (6).
  Update the student model θ_S.
  $N = N + 1$
end for
Output: Distilled student model: θ_{S_ALI}.

Second Stage: The training process of the Logit Mimicking (LM)
Input: Training set: X, label set: Y, teacher model: θ_T, distilled student model: θ_{S_ALI}, epoch: M, temperature: T.
Initialize: Initialize the classification's head of the distilled student model: θ_{S_ALI}, $m = 1$, set the teacher model θ_T in evaluation mode.
for $m = 1$ to M **do**
  $z^{Te} = \theta_T(X)$ ▷ Obtain the logit output of the teacher model.
  $z^{St} = \theta_{S\_ALI}(X)$ ▷ Obtain the logit output of the student model.
  Obtain the L_{LM} by calculating *eq.* (11) to eq. (14).
  Update the student model θ_S.
  $M = M + 1$
end for
Output: Student model: θ_{S_CIKD}.

5. Experiments

5.1. Setting

Datasets: We test the performance of our method on eight different datasets. The CIFAR-100 [12] dataset (Canadian Institute for Advanced Research, 100 classes) is a subset of the Tiny Images dataset and consists of 60 000 32 × 32 color images. The 100 classes in the CIFAR-100 are grouped into 20 superclasses. There are 600 images per class. The CIFAR-10 [12] dataset (Canadian Institute for Advanced Research, 10 classes) is a subset of the Tiny Images dataset and consists of 60 000 32 × 32 color images. CINIC-10 [15] is a dataset for image classification. It has a total of 270,000 images, 4.5 times that of CIFAR-10. It is constructed from two different sources: ImageNet and CIFAR-10. Specifically, it was compiled as a bridge between CIFAR-10 and ImageNet. Oxford Flower [13] is an image classification dataset consisting of 102 flower categories. The flowers chosen to be flowers commonly occur in the United Kingdom. Each class consists of between 40 and 258 images. The Oxford-IIIT Pet [16] Dataset is a 37-category pet dataset with roughly 200 images for each class. The images have large variations in scale, pose, and lighting. FGVC-Aircraft [14] contains 10,200 images of aircraft, with 100 images for each of 102 different aircraft model variants, most of which are airplanes. The Stanford Cars [16] dataset consists of 196 classes of cars with a total of 16,185 images taken from the rear. The data is divided into almost a 50–50 train/test split with 8144 training images and 8041 testing images. The Chaoyang [17] has 4021 training samples and 2139 test

samples for 4 classes of the medical image domain. The specific details of each dataset are shown in Table 1.

Model: Consistent with Deit [3], the teacher model utilizes the original ViT-Base architecture. Specifically, the teacher model utilizes the original ViT-Base architecture: 12 encoder layers, 768 embedding dimensions, an MLP ratio of 4, and 12 self-attention heads for all encoder layers. The teacher model is pre-trained on the Imagenet-1K dataset using the Dino method. The student model uses the original ViT-Tiny architecture: 12 encoder layers, 192 embedding dimensions, and an MLP ratio of 4. The number of self-attention heads for the first and last encoder layers is 12, while the rest are 6.

Implementation Details: On the Cifar100, Cifar10, CINIC10, and FGVC-Aircraft datasets during the Attention Locality Imitation stage, the student model uses a randomly initialized ViT-Tiny architecture. The teacher model, ViT-Base [3], is pre-trained on the Imagenet-1K dataset [2] and is further fine-tuned on each of these datasets. In the first stage, Attention Locality, with LOB set to 10, we trained on unlabeled datasets for 300 epochs. We used a batch size of 256 and a learning rate of 1×10^{-3}. We used a cosine decay schedule with a warm-up period of 5 epochs and the AdamW optimizer [43] with a weight decay of 0.05. We utilized data augmentation techniques, which involved random resizing and cropping, random horizontal flipping, and color jittering, and adjusted the image size to 224 × 224. The student model produced in the first stage is used as the student model for the second stage, while the teacher model continues to be the same as in the first stage. The labeled dataset is trained for 300 epochs using a batch size of 512 and a learning rate of 1×10^{-3}. The data augmentation used includes random resizing, random horizontal flipping, and TrivialAugmentWide. The image size is adjusted to 224 × 224. In Eq. (14), α is set to 4, and the temperature T is set to 1. In both stages, the teacher model ViT-Base is set to evaluation mode, and its gradients are not calculated. In other datasets, a learning rate of 8×10^{-3} is used in the ALI stage, the epoch is 1000, and the batch size is 64. In the LM stage, a learning rate of 2×10^{-3} is used, the epoch is 500, and the batch size is 512. Other parameters are consistent with the above parameters. All of the training devices are Nvidia 3090 GPUs. We use Pytorch tools, and our code is modified from timm [44].

5.2. Main results

As shown in Table 2, our method enables ViT-Tiny to achieve competitive results on Cifar100 and Cifar10. ViT-Tiny + CIKD achieves state-of-the-art performance on the Cifar100 dataset without pre-training and outperforms it by 2.8%. It achieves state-of-the-art performance on the Cifar10 dataset and outperforms pre-training the teacher model for fine-tuning the paradigm. Lightweight ViT uses a pre-trained teacher model without pre-training to obtain lightweight ViT that surpasses pre-training. It is worth mentioning that the performance of our ViT-Tiny + CIKD is better than the methods using the latest pre-training and fine-tuning paradigm. It should be noted that the parameters of the baseline model in Table 2 are the same as the model parameters we use, but our model parameters are rounded to one decimal place. Specifically, following the application of our CIKD method, ViT-Tiny requires no prior pre-training on extensive datasets but solely utilizes existing pre-trained teachers for knowledge distillation.

Our approach enabled ViT-Tiny to obtain competitive results on the Cifar100 dataset. Despite being 15.5 times smaller than the teacher model, it retains 97.8% of the accuracy exhibited by the teacher model. For the Cifar10 dataset, the model parameters are 15.5 times smaller compared to the teacher model, yet the accuracy surpasses that of the teacher model. Specifically, the Top-1 Accuracy of the student model reaches 98.7%, whereas the teacher model achieves 98.6%.

To verify the generality of our method, we conduct experiments on datasets from various domains and more minor scales. The results indicate that we consistently outperform the current state-of-the-art pre-training and fine-tuning paradigms. In Table 3, the dataset with

Table 2
Top-1 accuracy (%) on Cifar100, Cifar10. All models are pre-trained on ImageNet-1K, except ViT-Tiny+CIKD.

Method	#Param (M)	CIFAR10 [12]	CIFAR100 [12]
Feedforward networks			
ResMLP-S12 [45]	15.4	98.1	87.0
MLP-Mixer			
Mixer-B/16-SAM [46]	59.0	97.8	86.4
ConvMLP			
ConvMLP [47]	9.0	98.0	87.4
Vision transformer			
Teacher:ViT-Base [11]	85.5	98.6	91.8
MAE-Tiny [48]	6.0	–	78.9
D-MAE-lite [48]	6.0	–	85.0
DeiT-Tiny [3]	5.7	98.1	86.1
CRATE-T [49]	6.1	95.0	78.9
ViTAE-T [50]	4.8	97.3	85.0
Mini-DeiT-Ti [51]	3.0	97.5	83.8
DearKD-Tiny [52]	5.0	97.5	85.7
CSKD-Ti [53]	6.0	98.5	87.0
Student:ViT-Tiny+CIKD	5.5	**98.7**	**89.8**

"*" represents the accuracy obtained by training the ViT-Tiny model with the weights from the open-source ViT-Tiny model, following the fine-tuning method outlined in the Ref. [48] on the dataset. Only ViT-Tiny + CIKD is not pre-trained. The generality of our approach can be seen by conducting experiments on several small datasets of different types. We hope that our approach will advance the broader use of ViT for vision tasks, especially for small datasets.

As shown in Table 1, even though the Oxford Flowers dataset only has 2040 images and 102 categories, ViT-Tiny still achieves the same accuracy as MAE-Tiny-FT [48] with the help of our method CIKD. Where MAE-Tiny-FT represents pre-training and fine-tuning on the Imagenet dataset followed by a second fine-tuning on the Oxford Flowers dataset. In contrast to MAE-Tiny-FT, which costs a lot of computational power, we only need to use the existing pre-trained model ViT-Base to distill the knowledge on the Oxford Flowers dataset to achieve the same accuracy as the MAE-Tiny-FT. Our method still performs excellently on the Chaoyang dataset in the medical domain. Since Distilled MAE-lite and MAE-Tiny-FT do not have open-source weights, we are unable to fine-tune them on the CINIC10 dataset and the sunrise dataset. However, we also achieved better accuracy compared to other pre-training methods.

5.3. Selection strategy for the learning objective block

We conducted experiments on six datasets derived from the Cifar100 dataset, each with varying class proportions, and on two additional datasets with different selections of teacher model layers. The teacher model for the different class-proportioned Cifar100 datasets is trained using the respective datasets with varying class proportions.

To explore how the selection of the Learning Objective Block is carried out, we extracted a certain percentage of subsets from the CIFAR-100 dataset containing 50,000 images of the training set, as well as from other datasets, for the experiments. It is worth noting that we chose both the block of the first layer of the teacher model and the Learning Objective Block to distill knowledge with the block of the first layer of the student model and the block of the last layer. As shown in Fig. 4, choosing a block from an inappropriate layer of the teacher model as the Learning Objective Block can lead to suboptimal performance. This is because the features in layers closer to the input layer of the teacher model contain less semantic information, making it easier to learn with a smaller dataset. Conversely, features in layers farther away from the input layer of the teacher model contain more

Table 3
Top-1 accuracy (%) on Oxford Flowers, FGVC-Aircraft, Stanford Cars, Oxford Pets, CINIC10, Chaoyang. All models are pre-trained on ImageNet-1K, except ViT-Tiny+CIKD.

Method	#Param (M)	Oxford Flowers [13]	FGVC-Aircraft [14]	Stanford Cars [18]	Oxford Pets [16]	CINIC10* [15]	Chaoyang* [17]
Teacher: ViT-Base [11]	85.0	98.1	82.1	90.6	95.4	94.5	90.4
D-MAE-lite [48]	6.0	95.2	79.2	87.5	89.1	–	–
DeiT-Tiny [48]	6.0	96.4	73.5	85.6	93.1	92.4	87.3
MoCov3-Tiny [48]	6.0	94.8	73.7	83.9	87.8	91.1	84.3
MAE-Tiny [48]	6.0	85.8	64.6	78.8	76.5	90.5	83.7
MAE-Tiny-FT [48]	6.0	96.8	78.1	87.6	93.2	–	–
Student: ViT-Tiny+CIKD	5.5	**96.8**	**79.7**	**88.1**	**93.5**	**93.7**	**88.5**

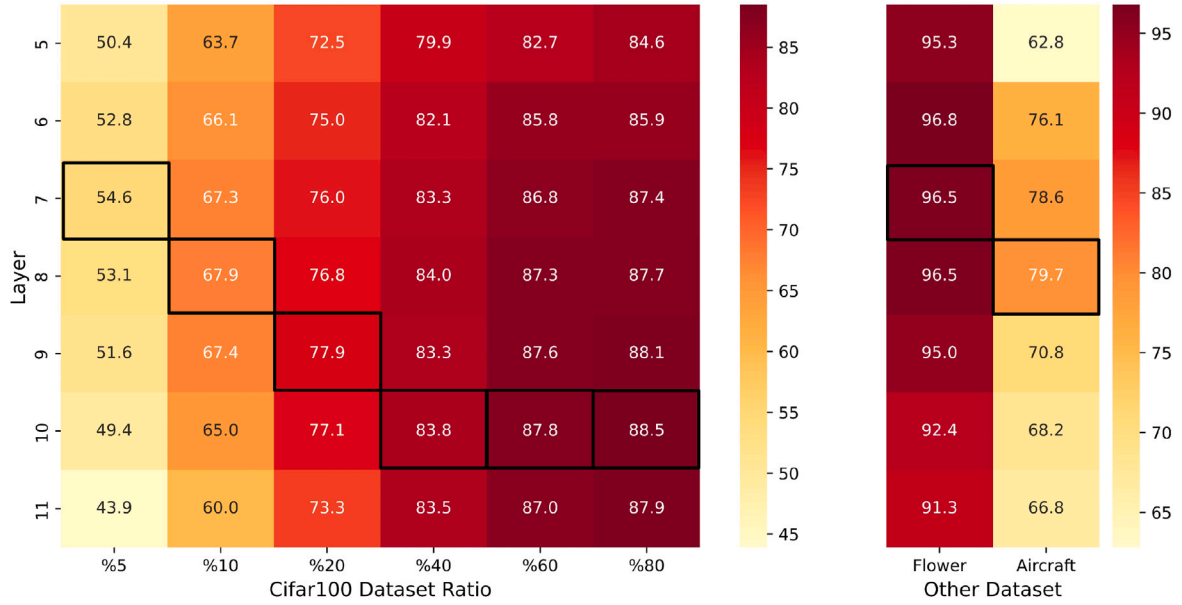

Fig. 4. The effect of selecting block of different layers of the teacher model as the Learning Objective Block on the same dataset extracted at different scales and on different datasets. Top-1 accuracy (%) is reported.

semantic information, hence making it more challenging to learn with a smaller dataset.

In Fig. 4, "Flower" represents the Oxford Flowers dataset, and "Aircraft" represents the FGVC-Aircraft dataset. Similar situations are observed on both datasets: as the amount of data decreases, we need to choose features from layers close to the input layer of the teacher model for knowledge distillation to achieve better results. There are 2040 training set images for Oxford Flowers and 6667 training set images for FGVC-Aircraft. It is worth noting that when the amount of data is small, choosing the Learning Objective Block for knowledge distillation away from the input layer can lead to more severe model performance degradation. Therefore, selecting a more appropriate layer for the first stage of ALI can achieve better results.

5.4. Visualization analysis

CKA Similarity. Knowledge distillation aims to transfer information from the teacher model to the student model, and the effectiveness of knowledge distillation is typically assessed by measuring the similarity between the teacher and student models. Therefore, we use the Centered Kernel Alignment (CKA) to analyze the similarity of representations between the student model and the teacher model:

$$CKA(K,L) = \frac{HSIC(K,L)}{\sqrt{HSIC(K,K)\cdot HSIC(L,L)}} \tag{18}$$

assuming that X and Y represent specific layer outputs of two feature vectors, the Gram matrices for these two layers are defined as $K = XX^T$ and $L = YY^T$. $HSIC$ stands for the Hilbert–Schmidt Independence Criterion [54].

Illustrated in Fig. 5, "ViT-Tiny + CIKD" corresponds to the model employing the Curriculum Information Knowledge Distillation (CIKD) method. In contrast, "ViT-Tiny + LM" is indicative of models undergoing knowledge distillation just through the Logit Mimicking (LM) stage. "ViT-Tiny + ALI" is associated with models that are distilled exclusively using the initial stage Attention Local Imitation (ALI) method. Lastly. "ViT-Tiny + (ALI + LM)" denotes simultaneous training with the initial stage Attention Local Imitation (ALI) method and the Logit Mimicking (LM) method . "ViT-Base FT" denotes the teacher model that has been fine-tuned. The incorporation of low-level semantic features in the small dataset enhances the resemblance between the student model and the teacher model. Conversely, focusing on high-level semantics during the Logit MImicking stage results in diminished similarity. Notably, the CIKD method manifests the most substantial similarity, indicating its effectiveness in aligning the student model closely with the teacher model's representational space. The gap in model capacity resulted in a decrease in CKA for both the student model and the teacher model in the final layer.

Mean Attention Distance. To intuitively observe whether CIKD enables the model to learn locality, we revealed the aggregation behavior of information in ViT's attention mechanism, which is computed through dot product operations from the compatibility between queries and keys. Therefore, we analyze the Mean Attention Distance of all tokens in different attention heads to assess the degree of aggregation of local and global information. The attention distance for the jth token in the hth head is calculated as follows:

$$D_{h,j} = \sum_i \text{softmax}(A_h)_{i,j} G_{i,j} \tag{19}$$

where $A_h \in \mathbb{R}^{l\times l}$ represents the attention map for the hth attention head, l is the number of tokens. $G_{i,j}$ is the Euclidean distance between the spatial positions of the ith and jth tokens.

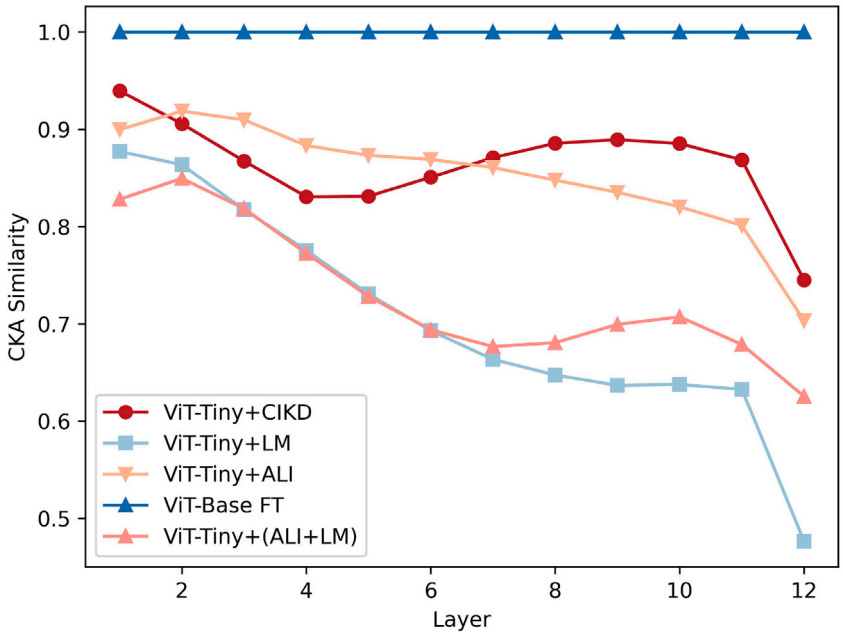

Fig. 5. CKA similarity between representations generated by ViT-Tiny with CIKD. "ViT-Tiny + CIKD" corresponds to the model employing the Curriculum Information Knowledge Distillation (CIKD) method. In contrast, "ViT-Tiny + LM" is indicative of models undergoing knowledge distillation just through the Logit Mimicking (LM) stage. "ViT-Tiny + ALI" is associated with models that are distilled exclusively using the initial stage Attention Local Imitation (ALI) method. Lastly. "ViT-Tiny + (ALI + LM)" denotes simultaneous training with the initial stage Attention Local Imitation (ALI) method and the Logit Mimicking (LM) method. "ViT-Base FT" denotes the teacher model that has been fine-tuned.

As shown in Fig. 6, it is evident from (a) and (b) that learning high-level semantic logit and simultaneously learning high-level semantic logit along with low-level semantic features make it challenging for ViT-Tiny to allocate attention effectively. However, in (c), it can be seen that after the ALI stage, ViT-Tiny has learned the concept of locality. By contrast, as observed in (d), after undergoing Curriculum Information Knowledge Distillation (CIKD), the model's attention allocation becomes more rational, *i.e.*, shallow blocks focus more on the local while deep blocks concentrate more on the global.

CAM. In Fig. 7, the Class Activation Mapping (CAM) visualization is conducted on an individual image from the Cifar100 dataset for an in-depth analysis of attention weight distribution. The visualization in (d) illustrates that, under the meticulous instruction of the teacher model, the student model exhibits a heightened and more targeted focus on the salient objects within the image. Contrastingly, in (c), the attention distribution is notably more refined and judicious compared to the initial states shown in (a) and (b). This highlights the effectiveness of the gradual guidance provided by the teacher model in enhancing the student model's capacity for rational attention allocation. In conclusion, the CAM visualization results show that the ViT-Tiny model can correctly characterize the objects in the images by our CIKD method, compared to the single-stage distillation method and the fusion of the two stages into a single-stage distillation method.

5.5. Ablation study

To validate the design choices in our method, we conduct experiments on the CIFAR100 dataset. In Table 4, we ablate our two-stage method so that we can see that the easy-to-hard knowledge distillation method we use is effective. In Table 5, we explore the generalizability of our approach using ViT-Small as the student model, where the teacher model ViT-Base has four times more parameters than the student model ViT-Small. However, the accuracy of the student model (91.4%) is indeed 99.5% of the accuracy of the teacher model (91.8%). In Table 6, we explore the effect of various pre-training methods on the CIKD of our method, from which we can see that the Dino pre-training method is the most appropriate. In Table 7, we ablate α of Eq. (13) in the second stage LM. In Table 8, we perform ablation experiments for knowledge distillation using only the Learning Target Block and not the shallow block. From the experimental results, we can see that good performance can only be obtained by using both shallow and deep learning target blocks. In Table 9, we perform a two-stage reversal to demonstrate the effectiveness of our knowledge distillation method designed from easy to difficult. In Table 10, we ablate the hyperparameter temperature (T) of the second-stage LM.

Design of two-stage distillation. We evaluate the effectiveness of our proposed Curriculum Information Knowledge Distillation (CIKD) through experiments with various configurations. Table 4 summarizes the results: "Student: ViT-Tiny + ALI" uses the Attention Locality Imitation (ALI) for 300 epochs followed by 300 epochs of fine-tuning. "Student: ViT-Tiny + LM" employs the Logit Mimicking (LM) for 600 epochs. "Student: ViT-Tiny + (ALI + LM)" denotes simultaneous training with ALI and LM for 600 epochs. "Student: ViT-Tiny + CIKD" begins with 300 epochs of ALI training and continues with 300 epochs of LM.

Through the experiences of two single-stage models "Student: ViT-Tiny + ALI" and "Student: ViT-Tiny + LM", it is evident that the performance of single-stage models is inferior to that of two-stage models. Moreover, even though the performance of "Student: ViT-Tiny + LM" is significantly lower than that of "Student: ViT-Tiny + ALI", it can be observed that the ALI played a crucial role in distilling knowledge in the two-stage process. However, it is precisely due to our proposed progressive knowledge distillation method, CIKD (from easy to difficult), that allows these two approaches to be effectively combined, resulting in the best performance. The simultaneous two-stage knowledge distillation approach denoted as "Student: ViT-Tiny + (ALI + LM)", exhibited poorer performance, reflecting the importance of the progressive two-stage distillation process from easy to difficult. Patient teachers who gradually impart knowledge from low-level feature information to high-level logit information can achieve higher performance, as shown in Table 4. It is observed that simultaneously imparting both types of knowledge directly leads to subpar performance (ViT + (ALI + LM)), accuracy of 66.3%.

Different teacher-student settings. As shown in Table 5, we explored the universality of our method and conducted experiments on other teacher-student model pairs. It can be seen from the results that our method has been effective for ViT-Base and ViT-Small. The

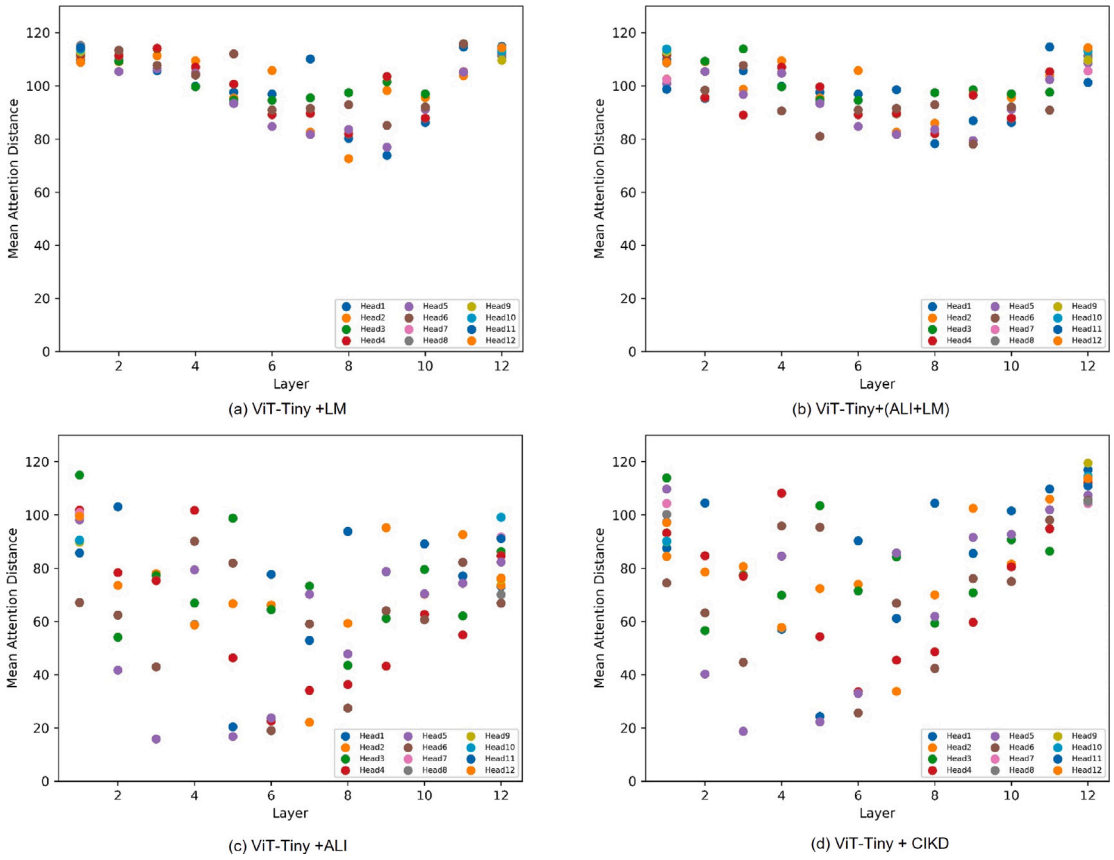

Fig. 6. Mean Attention Distance for ViT-Tiny at different stages. "(a) ViT-Tiny + LM" is indicative of models undergoing knowledge distillation just through the Logit Mimicking (LM) stage. "(b) ViT-Tiny + (ALI+ LM)" denotes simultaneous training with the initial stage Attention Local Imitation (ALI) method and the Logit Mimicking (LM) method. "(c) ViT-Tiny + ALI" is associated with models that are distilled exclusively using the initial stage Attention Local Imitation (ALI) method. Lastly. "(d) ViT-Tiny + CIKD" corresponds to the model employing the Curriculum Information Knowledge Distillation (CIKD) method. Mean Attention distance is computed for 128 example images by averaging the distance between the query pixel and all other pixels, weighted by the attention weight. Each dot shows the mean attention distance across images for one of 16 heads at one layer. Image width is 224 pixels [1].

Table 4
Results of ablation experiments for two-stage method.

Method	#Param (M)	Top-1 Acc (%)
Teacher: ViT-Base	85.5	91.8
ViT-Tiny + ALI	5.5	85.5
ViT-Tiny + LM	5.5	60.5
ViT-Tiny + (ALI + LM)	5.5	66.3
ViT-Tiny + CIKD	5.5	89.8

Table 5
Results of ablation experiments for different teacher-student settings.

Method	#Param (M)	Top-1 Acc (%)
Teacher: ViT-Base	85.0	91.8
ViT-Small + CIKD	21.5	91.4
ViT-Tiny + CIKD	5.5	89.8

hyperparameterization used is consistent with ViT-Base and ViT-Tiny, which also proves the effectiveness of our layer selection strategy. We use ViT-Small, which shrinks our parameters by a factor of almost 4 compared to the teacher model ViT-Base, but has 99.5% of the accuracy of the teacher model. It also demonstrates the generality of the selection strategy for the Learning Objective Block.

Different pre-training methods for teacher model. To explore which pre-training method is most suitable for the knowledge distillation of our method, we experiment with supervised pre-training methods, masked image modeling pre-training methods, and comparative learning pre-training methods and find that the Dino pre-training method is the most suitable.

Table 6
Results of ablation experiments for different teacher-student settings.

Method	Top-1 Acc (%) of ViT-Tiny
Deit [3]	88.4
MAE [55]	88.1
MocoV3 [56]	87.6
Dino [11]	89.8

Table 7
The results of the ablation study for α in Eq. (13).

α	1	2	3	4	5	6	7
Top-1 Acc (%)	89.4	89.7	89.7	89.8	89.8	89.6	89.4

α of Eq. (13). As shown in Table 7, to verify the impact of the hyperparameter α in the second stage of LM, we test various values of α distributed within the range [1, 7]. Setting α to extremely large or small values could hinder the model's performance. The best performance was obtained with values of α between 3 and 4, with larger or smaller values causing performance degradation.

Only the Learning Objective Block. To validate the effectiveness of the learning target block selection strategy, *i.e.*, selecting both the shallow block and the Learning Objective Block, we conduct experiments in which only the single-level Learning Objective Block is selected. The results are shown in Table 8. Distilling features from

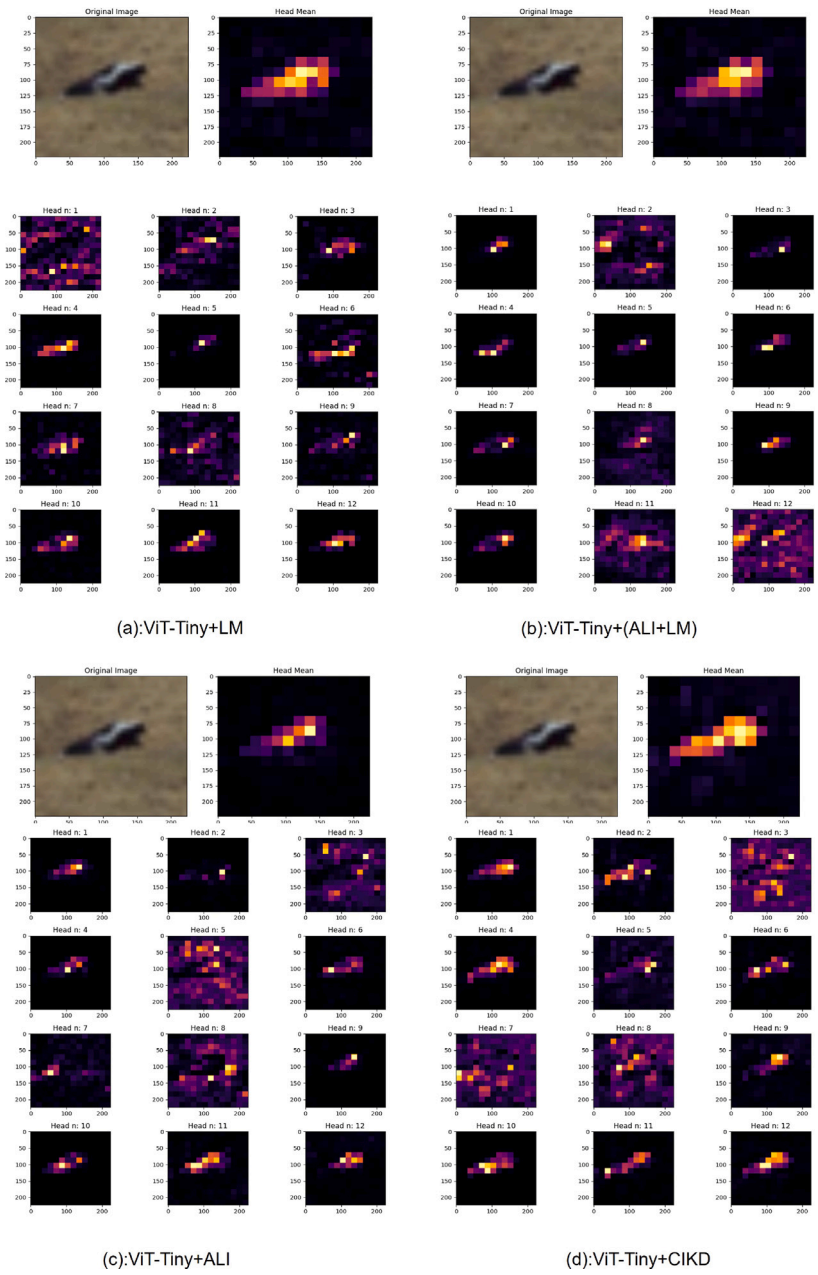

Fig. 7. Visualization of the last layer's Self-Attention in ViT-Tiny and each head using CAM. Brighter colors represent higher weights assigned by Self-Attention, while black indicates the background. "(a) ViT-Tiny + LM" is indicative of models undergoing knowledge distillation just through the Logit Mimicking (LM) stage. "(b) ViT-Tiny + (ALI + LM)" denotes simultaneous training with the initial stage Attention Local Imitation (ALI) method and the Logit Mimicking (LM) method. "(c) ViT-Tiny + ALI" is associated with models that are distilled exclusively using the initial stage Attention Local Imitation (ALI) method. Lastly. "(d) ViT-Tiny + CIKD" corresponds to the model employing the Curriculum Information Knowledge Distillation (CIKD) method. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 8
Results of the ablation study with only the Learning Objective Block.

Single teacher layer	7	8	9	10	11	12
Top-1 Acc (%)	86.8	87.2	88.1	88.3	88.6	88.2

a single layer of the teacher model leads to suboptimal performance compared to using features from two layers. This is due to the challenge of achieving robust convergence when optimizing deep network layers solely through backpropagation. This observation indirectly underscores the effectiveness of selecting two teacher model layers for distillation.

Reversal of two-stage distillation. To verify the rationality of our two-stage method and the effectiveness of the two-stage distillation from easy to difficult, our two-stage distillation method will be reversed, first performing Logit Mimicking (LM) distillation and then performing Attention Locality Imitation (ALI) distillation. As shown in Table 9, where ViT-Tiny + LM + ALI denotes LM distillation followed by ALI distillation, it can be observed from the results that our CIKD method yields better performance. The performance improvement of the two-stage approach relative to using only a single stage of ALI is not significantly enhanced. One possible reason is that the LM stage is performed first, resulting in the student model not learning much "dark knowledge" as in the CIKD, which undergoes the first stage of ALI followed by the second stage of LM, due to the advanced semantic properties of logits and labels.

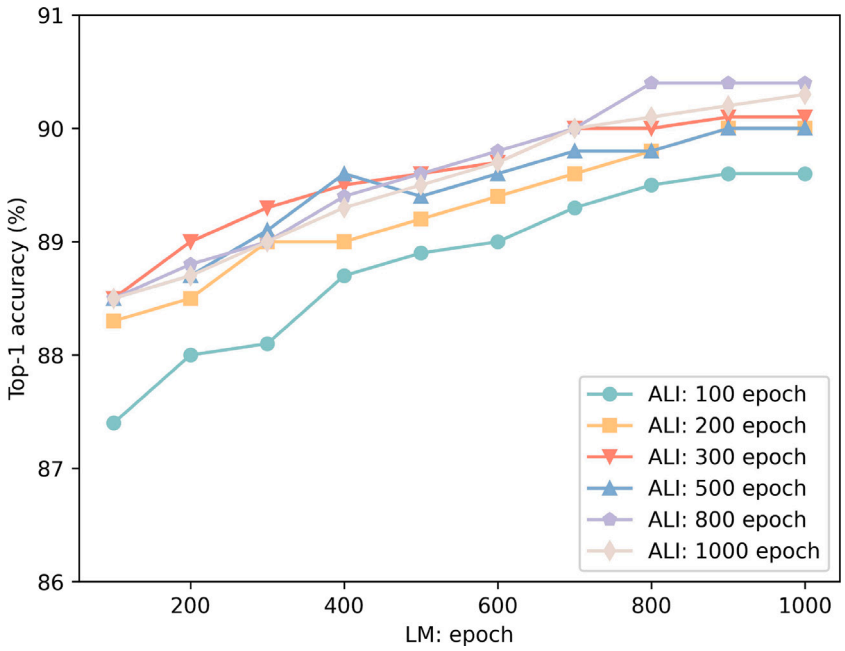

Fig. 8. Trade-off between training time and performance. Different line segments represent different training times in the first stage of ALI. The first-stage ALI with different training times is subjected to the second-stage LM training for 1000 epochs, and the performance changes of every 100 epochs are observed.

Table 9
Results of ablation experiments for reversal of two-stage distillation.

Method	#Param (M)	Top-1 Acc (%)
Teacher: ViT-Base	85.5	91.8
ViT-Tiny + LM + ALI	5.5	87.1
ViT-Tiny + CIKD	5.5	89.8

Table 10
Ablation results of Temperature (T).

Temperature (T)	1	2	3	4	5
Top-1 Acc (%)	89.8	89.3	89.2	89.1	89.1

Temperature (T). It can be seen from Table 10 that as the temperature (T) increases, the model accuracy gradually decreases. Therefore the temperature (T) is set to 1 by default in our experiments.

Trade-off between training time and performance. As shown in Fig. 8, we set different training times for the two stages to explore the impact of training time on model performance. Our approach is able to achieve state-of-the-art performance within a total of 600 epochs over two phases, with further increases in training duration only marginally improving performance.

6. Conclusions and future work

In this study, we have investigated a simple and efficient method that significantly improves the performance of ViT on small datasets. We have also introduced a selection strategy for the Learning Objective Block for teacher model on small datasets. When training ViT from scratch with limited data, it is challenging for the model to learn the local information in images. To address this issue, we proposed a knowledge distillation approach that combines curriculum learning. Our approach allows the student model to learn from easy to hard by first focusing on low-level semantic features for local information in the first stage and then incorporating high-level semantic logits and label information in the second stage. Extensive experiments have demonstrated the effectiveness and applicability of our method across 8 small-scale datasets, achieving competitive accuracy compared to the pre-training and fine-tuning paradigms. Furthermore, we validate the effectiveness of our approach through various visualization analyses. We believe that our method will advance the wider application of ViT in visual tasks, particularly in scenarios with small datasets.

Future Work: We only experimented on the classification task without evaluating dense-prediction downstream tasks, *i.e.*, object detection and segmentation. We leave this for further work. However, after applying our CIKD method, the model's feature extraction capability has been enhanced, and we believe it will also perform well in other tasks.

CRediT authorship contribution statement

Jun Ling: Writing – original draft, Validation, Software, Methodology. **Xuan Zhang:** Writing – review & editing, Supervision, Resources, Funding acquisition, Conceptualization. **Fei Du:** Formal analysis. **Linyu Li:** Data curation. **Weiyi Shang:** Writing – review & editing. **Chen Gao:** Visualization. **Tong Li:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

Science Foundation of Young and Middle-aged Academic and Technical Leaders of Yunnan under Grant No. 202205AC160040; Science Foundation of Yunnan Jinzhi Expert Workstation under Grant No. 202205AF150006; Major Project of Yunnan Natural Science Foundation under Grant No. 202302AE09002003; Knowledge-driven Smart Energy Science and Technology Innovation Team of Yunnan Provincial Department of Education; Science and Technology Project of Yunnan Power Grid Co., Ltd. under Grant No. YNKJXM20222254; Open Foundation of Yunnan Key Laboratory of Software Engineering under Grant No. 2023SE101;

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.

[3] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: International Conference on Machine Learning, PMLR, 2021, pp. 10347–10357.

[4] S. Ren, F. Wei, Z. Zhang, H. Hu, TinyMIM: An empirical study of distilling MIM pre-trained models, 2023, arXiv preprint arXiv:2301.01296.

[5] Z. Yang, Z. Li, A. Zeng, Z. Li, C. Yuan, Y. Li, Vitkd: Practical guidelines for vit feature knowledge distillation, 2022, arXiv preprint arXiv:2209.02432.

[6] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, Adv. Neural Inf. Process. Syst. 33 (2020) 5776–5788.

[7] S. Ahn, S.X. Hu, A. Damianou, N.D. Lawrence, Z. Dai, Variational information distillation for knowledge transfer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9163–9171.

[8] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, Q. Liu, Tinybert: Distilling bert for natural language understanding, 2019, arXiv preprint arXiv:1909.10351.

[9] W. Huang, Z. Peng, L. Dong, F. Wei, J. Jiao, Q. Ye, Generic-to-specific distillation of masked autoencoders, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15996–16005.

[10] Q. Han, Y. Cai, X. Zhang, RevColV2: Exploring disentangled representations in masked image modeling, 2023.

[11] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9650–9660.

[12] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images, 2009.

[13] M.-E. Nilsback, A. Zisserman, Automated flower classification over a large number of classes, in: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, 2008, http://dx.doi.org/10.1109/icvgip.2008.47.

[14] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, A. Vedaldi, Fine-grained visual classification of aircraft, 2013, arXiv preprint arXiv:1306.5151.

[15] L.N. Darlow, E.J. Crowley, A. Antoniou, A.J. Storkey, Cinic-10 is not imagenet or cifar-10, 2018, arXiv preprint arXiv:1810.03505.

[16] O.M. Parkhi, A. Vedaldi, A. Zisserman, C. Jawahar, Cats and dogs, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3498–3505.

[17] C. Zhu, W. Chen, T. Peng, Y. Wang, M. Jin, Hard sample aware noise robust learning for histopathology image classification, IEEE Trans. Med. Imaging (2022) 881–894, http://dx.doi.org/10.1109/tmi.2021.3125459.

[18] J. Krause, M. Stark, J. Deng, L. Fei-Fei, 3D object representations for fine-grained categorization, in: 2013 IEEE International Conference on Computer Vision Workshops, 2013, http://dx.doi.org/10.1109/iccvw.2013.77.

[19] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Commun. ACM (2017) 84–90, http://dx.doi.org/10.1145/3065386.

[20] X. Li, W. Wang, X. Hu, J. Yang, Selective kernel networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 510–519.

[21] Z. Peng, Z. Guo, W. Huang, Y. Wang, L. Xie, J. Jiao, Q. Tian, Q. Ye, Conformer: Local Features Coupling Global Representations for Recognition and Detection.

[22] X. Zhang, F. Liu, Z. Peng, Z. Guo, F. Wan, X. Ji, Q. Ye, Integral Migrating Pre-trained Transformer Encoder-decoders for Visual Object Detection.

[23] K. Yuan, S. Guo, Z. Liu, A. Zhou, F. Yu, W. Wu, Incorporating convolution designs into visual transformers, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, 2021, http://dx.doi.org/10.1109/iccv48922.2021.00062.

[24] Z. Lu, H. Xie, C. Liu, Y. Zhang, Bridging the gap between vision transformers and convolutional neural networks on small datasets, 2022.

[25] X. Chen, Q. Cao, Y. Zhong, J. Zhang, S. Gao, D. Tao, DearKD: Data-Efficient Early Knowledge Distillation for Vision Transformers.

[26] K. Wu, J. Zhang, H. Peng, M. Liu, B. Xiao, J. Fu, L. Yuan, M. Research, M. Cloud+ai, TinyViT: Fast Pretraining Distillation for Small Vision Transformers.

[27] J. Di, K. Han, Y. Wang, Y. Tang, J. Guo, C. Zhang, D. Tao, Efficient vision transformers via fine-grained manifold distillation., 2021, arXiv: Computer Vision and Pattern Recognition, arXiv: Computer Vision and Pattern Recognition.

[28] A. Dong, J. Liu, G. Zhang, Z. Wei, Y. Zhai, G. Lv, Momentum contrast transformer for COVID-19 diagnosis with knowledge distillation, Pattern Recognit. 143 (2023) 109732, http://dx.doi.org/10.1016/j.patcog.2023.109732, URL https://www.sciencedirect.com/science/article/pii/S0031320323004302.

[29] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in: Proceedings of the 26th Annual International Conference on Machine Learning, 2009, http://dx.doi.org/10.1145/1553374.1553380.

[30] S. Sinha, A. Garg, H. Larochelle, Curriculum by smoothing, Neural Inf. Process. Syst. Neural Inf. Process. Syst. (2020).

[31] Y. Tay, S. Wang, L. Tuan, J. Fu, M. Phan, X. Yuan, J. Rao, S. Hui, A. Zhang, Simple and Effective Curriculum Pointer-Generator Networks for Reading Comprehension over Long Narratives, Cornell University, 2019, arXiv, Cornell University - arXiv.

[32] X. Jin, B. Peng, Y. Wu, Y. Liu, J. Liu, D. Liang, J. Yan, X. Hu, Knowledge distillation via route constrained optimization, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV, 2019, http://dx.doi.org/10.1109/iccv.2019.00143.

[33] L. Xiang, G. Ding, J. Han, Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16, Springer, 2020, pp. 247–263.

[34] Y. Huang, J. Li, X. Chen, Y.-G. Fu, Training graph transformers via curriculum-enhanced attention distillation, in: The Twelfth International Conference on Learning Representations, 2023.

[35] C. Wang, K. Yang, S. Zhang, G. Huang, S. Song, TC3KD: Knowledge distillation via teacher-student cooperative curriculum customization, Neurocomputing 508 (2022) 284–292.

[36] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, 2015, arXiv preprint arXiv:1503.02531.

[37] B. Zhao, Q. Cui, R. Song, Y. Qiu, J. Liang, Decoupled knowledge distillation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11953–11962.

[38] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, J.Y. Choi, A comprehensive overhaul of feature distillation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1921–1930.

[39] B. Heo, M. Lee, S. Yun, J.Y. Choi, Knowledge transfer via distillation of activation boundaries formed by hidden neurons, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 3779–3787.

[40] W. Wang, H. Bao, S. Huang, L. Dong, F. Wei, Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers, 2020, arXiv preprint arXiv:2012.15828.

[41] D.Y. Park, M.-H. Cha, D. Kim, B. Han, et al., Learning student-friendly teacher networks for knowledge distillation, Adv. Neural Inf. Process. Syst. 34 (2021) 13292–13303.

[42] K. Li, R. Yu, Z. Wang, L. Yuan, G. Song, J. Chen, Locality guidance for improving vision transformers on tiny datasets, in: European Conference on Computer Vision, Springer, 2022, pp. 110–127.

[43] I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization, Cornell University, 2017, arXiv, Learning.

[44] R. Wightman, H. Touvron, H. Jégou, ResNet strikes back: An improved training procedure in timm, 2021, arXiv: Computer Vision and Pattern Recognition, arXiv: Computer Vision and Pattern Recognition.

[45] H. Touvron, P. Bojanowski, M. Caron, M. Cord, A. El-Nouby, E. Grave, G. Izacard, A. Joulin, G. Synnaeve, J. Verbeek, H. Jegou, ResMLP: Feedforward networks for image classification with data-efficient training, IEEE Trans. Pattern Anal. Mach. Intell. (2022) 1–9, http://dx.doi.org/10.1109/tpami.2022.3206148.

[46] X. Chen, C.-J. Hsieh, B. Gong, When Vision Transformers Outperform ResNets without Pre-training or Strong Data Augmentations, Cornell University, 2021, arXiv, Cornell University - arXiv.

[47] J. Li, A. Hassani, S. Walton, H. Shi, ConvMLP: Hierarchical Convolutional MLPs for Vision, Cornell University, 2021, arXiv, Cornell University - arXiv.

[48] S. Wang, J. Gao, Z. Li, J. Sun, W. Hu, A closer look at self-supervised lightweight vision transformers, 2022, arXiv preprint arXiv:2205.14443.

[49] Y. Yu, S. Buchanan, D. Pai, T. Chu, Z. Wu, S. Tong, B. Haeffele, Y. Ma, White-box transformers via sparse rate reduction, 2023.

[50] Q. Zhang, Y. Xu, J. Zhang, D. Tao, ViTAEv2: Vision Transformer Advanced by Exploring Inductive Bias for Image Recognition and Beyond.

[51] J. Zhang, H. Peng, K. Wu, M. Liu, B. Xiao, J. Fu, L. Yuan, Minivit: Compressing vision transformers with weight multiplexing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12145–12154.

[52] X. Chen, Q. Cao, Y. Zhong, J. Zhang, S. Gao, D. Tao, Dearkd: data-efficient early knowledge distillation for vision transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12052–12062.

[53] B. Zhao, R. Song, J. Liang, Cumulative Spatial Knowledge Distillation for Vision Transformers.

[54] A. Gretton, O. Bousquet, A. Smola, B. Schölkopf, Measuring statistical dependence with Hilbert-Schmidt norms, in: International Conference on Algorithmic Learning Theory, Springer, 2005, pp. 63–77.

[55] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, R. Girshick, Masked autoencoders are scalable vision learners, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, http://dx.doi.org/10.1109/cvpr52688.2022.01553.

[56] X. Chen, S. Xie, K. He, An Empirical Study of Training Self-Supervised Vision Transformers, Cornell University, 2021, arXiv, Cornell University - arXiv.

Jun Ling received bachelor's degree from Chengdu University's School of Computer Science in 2022. He is currently pursuing his master's degree at the School of Software, Yunnan University. His research interests include knowledge distillation, vision transformer, vision mamba, and curriculum learning.

Xuan Zhang received the B.S. and M.S. degrees in computer science, and the Ph.D. degree in system analysis and integration from Yunnan University, Kunming, China. She is a professor with the School of Software, Yunnan University, Kunming, China. She is author of 4 books and more than 120 articles. She has been principal investigator for more than 30 national, provincial, and private grants and contracts. She is the core scientist of Yunnan Key Laboratory of Software Engineering and Yunnan Software Engineering Academic Team. Her research interests include computer vision, knowledge graph, natural language processing, recommendation system, and blockchain.

Fei Du received the B.S. and M.S. degrees in software engineering, and the Ph.D. degree in computer science and technology from Yunnan University, Kunming, China. He is a lecturer with the School of Software, Yunnan University, Kunming, China. His research interests include deep learning, continual learning, imbalanced learning, and computer vision.

Linyu Li is currently pursuing a PhD degree with the School of Computer Science, Peking University, China. His research interests include knowledge graph completion and solving various knowledge graph problems using deep learning. He also serves as a reviewer for multiple journals.

Weiyi Shang is an Associate Professor at the University of Waterloo. His research interests include AIOps, big data software engineering, software log analytics and software performance engineering. He serves as a Steering committee member of the SPEC Research Group. He is ranked top worldwide SE research stars in a recent bibliometrics assessment of software engineering scholars. He is a recipient of various premium awards, including the SIGSOFT Distinguished paper award at ICSE 2013 and ICSE 2020, best paper award at WCRE 2011 and the Distinguished reviewer award for the Empirical Software Engineering journal. His research has been adopted by industrial collaborators (e.g., BlackBerry and Ericsson) to improve the quality and performance of their software systems that are used by millions of users worldwide. Contact him at wshang@uwaterloo.ca https://ece.uwaterloo.ca/~wshang/.

Chen Gao received the B.S. and M.S. degrees in School of Software from East China University of Science and Yunnan University, China, in 2017 and 2021, respectively. He is currently pursuing the Ph.D. degree with the School of Information Science and Engineering, Yunnan University. His research interests include knowledge graphs, information extraction, and few-shot learning.

Tong Li received his Ph.D. degree from De Montfort University, Leicester, UK, in 2007. He is currently a professor of Yunnan Agricultural University, China. His research interests are in software process, big data and concurrent process.