This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



# Mask-Guided Vision Transformer for Few-Shot Learning

Yuzhong Chen, Zhenxiang Xiao, Yi Pan, Lin Zhao, Haixing Dai, Zihao Wu, Changhe Li, Tuo Zhang, Changying Li, *Member, IEEE*, Dajiang Zhu, Tianming Liu, *Senior Member, IEEE*, and Xi Jiang

***Abstract*— Learning with little data is challenging but often inevitable in various application scenarios where the labeled data are limited and costly. Recently, few-shot learning (FSL) gained increasing attention because of its generalizability of prior knowledge to new tasks that contain only a few samples. However, for data-intensive models such as vision transformer (ViT), current fine-tuning-based FSL approaches are inefficient in knowledge generalization and, thus, degenerate the downstream task performances. In this article, we propose a novel mask-guided ViT (MG-ViT) to achieve an effective and efficient FSL on the ViT model. The key idea is to apply a mask on image patches to screen out the task-irrelevant ones and to guide the ViT focusing on task-relevant and discriminative patches during FSL. Particularly, MG-ViT only introduces an additional mask operation and a residual connection, enabling the inheritance of parameters from pretrained ViT without any other cost. To optimally select representative few-shot samples, we also include an active learning-based sample selection method to further improve the generalizability of MG-ViT-based FSL. We evaluate the proposed MG-ViT on classification, object detection, and segmentation tasks using gradient-weighted class activation mapping (Grad-CAM) to generate masks. The experimental results show that the MG-ViT model significantly improves the performance and efficiency compared with general fine-tuning-based ViT and ResNet models, providing novel insights and a concrete approach toward generalizing data-intensive and large-scale deep learning models for FSL.**

***Index Terms*— Domain adaptation, few-shot learning (FSL), mask, vision transformer (ViT).**

Manuscript received 12 June 2023; revised 4 January 2024 and 12 May 2024; accepted 19 June 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62276050, Grant 61976045, Grant 31971288, and Grant 62131009; and in part by Sichuan Science and Technology Program under Grant 2024NSFSC0655. *(Corresponding author: Xi Jiang.)*

Yuzhong Chen, Zhenxiang Xiao, Yi Pan, and Xi Jiang are with the Clinical Hospital of Chengdu Brain Science Institute, MOE Key Laboratory for Neuroinformation, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 611054, China (e-mail: chenyuzhong211@gmail.com; zhenxiang.up@gmail.com; dwaynepan5277@gmail.com; xijiang@uestc.edu.cn).

Lin Zhao, Haixing Dai, Zihao Wu, and Tianming Liu are with the Department of Computer Science, University of Georgia, Athens, GA 30602 USA (e-mail: lin.zhao@uga.edu; hd54134@uga.edu; zihao.wu1@uga.edu; tliu@cs.uga.edu).

Changhe Li and Tuo Zhang are with the School of Automation, Northwestern Polytechnical University, Xi'an 710071, China (e-mail: ChangheLi@mail.nwpu.edu.cn; tuozhang@nwpu.edu.cn).

Changying Li is with the Department of Agricultural and Biological Engineering, University of Florida, Gainesville, FL 32611 USA (e-mail: cli2@ufl.edu).

Dajiang Zhu is with the Department of Computer Science and Engineering, The University of Texas at Arlington, Arlington, TX 76019 USA (e-mail: dajiang.zhu@uta.edu).

Digital Object Identifier 10.1109/TNNLS.2024.3418527

## I. Introduction

DEEP neural networks (DNNs) have achieved great success in many computer vision tasks with a large amount of labeled data [1], [2]. However, learning with little data is often inevitable in a variety of application scenarios in which the labeled data are limited and costly. Recently, few-shot learning (FSL) has attained increasing interest in reconciling the demand and scarcity of large-scale labeled data in DNNs because of its generalizability of prior knowledge to new tasks given only a few samples. Existing literature has devoted extensive efforts to improving the effectiveness and efficiency of FSL from three aspects [3], data augmentation [4], [5], model design [6], and algorithm development [7], and correspondingly obtained promising results. By simply freezing the backbone and merely fine-tuning the last few layers, previous fine-tuning-based FSL approaches have achieved promising performances [8], [9], [10].

However, for data-intensive models with even more parameters, e.g., vision transformer (ViT) [11], current fine-tuning-based FSL approaches are inefficient in knowledge generalization [12]. For example, the large ViT model performs worse than ResNets when pretrained on a small dataset such as ImageNet, and the full benefit of larger models is observed only with large JFT-300M dataset [11]. One effective solution is transfer learning, which uses pretrained parameters from large-scale data. However, probably due to the nature of data-intensive models, such as ViT, which do not inherently encode the inductive biases that are useful for smaller datasets and, thus, require a large amount of labeled data to figure out the underlying modality-specific rules [11], [13], it results in unsatisfying performance when the labeled data are limited in FSL scenario. Therefore, figuring out how to efficiently generalize the domain knowledge of data-intensive models for FSL is still an open question and worth more effort to explore.

In this article, we propose a novel mask-guided ViT (MG-ViT) to efficiently and effectively adapt the domain knowledge of pretrained ViT to FSL tasks. The key idea of MG-ViT is to apply a mask on input image patches before the first transformer encoder layer to screen out the task-irrelevant patches and to guide ViT focusing on task-relevant and discriminative ones during FSL (see Fig. 1). The salience map is utilized to pinpoint patches that are highly relevant to the task and possess discriminative features, achieved by employing the gradient-weighted class activation mapping (Grad-CAM) [14] on images. In detail, the salience map is generated based on the

2162-237X © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: University of Georgia. Downloaded on September 30,2024 at 15:19:01 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



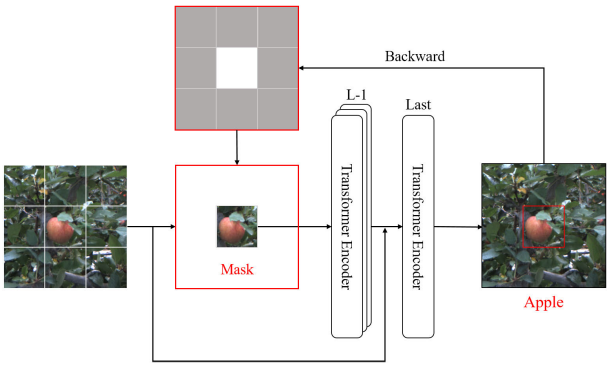



Fig. 1. Mask operation to select most task-relevant and discriminative image patches for MG-ViT model.

comparison of the target dataset with the most similar images from the source dataset. The proposed mask operation can take better advantage of the prior knowledge from the source domain and reduce the deviation between the source and target domains. Moreover, as visible patches that remain unmasked preserve task-relevant details but may lose essential global information, particularly positional information within the image, we introduce a residual connection between the initial and final encoder layers of ViT to maintain this global context. For FSL [9], MG-ViT adopts a two-stage training approach. The initial stage involves training exclusively on a base dataset using the vanilla ViT. Subsequently, in the second stage, the model undergoes fine-tuning on both novel and base datasets using the proposed MG-ViT. Notably, MG-ViT integrates the aforementioned mask operation and residual connection into ViT, seamlessly inheriting parameters from the pretrained ViT without necessitating additional resource-intensive pretraining steps.

To further enhance the overall generalizability of MG-ViT in FSL, we incorporate an active learning-driven sample selection method [15], [16]. The motivation stems from the fact that general FSL seeks to predict a substantial number of unlabeled samples by leveraging a restricted amount of labeled data [3]. The integration of active learning is envisaged to enhance the efficacy of model learning by annotating representative samples, consequently augmenting the model's predictive capability on other unlabeled samples [15], [17]. As part of this effort, we introduce a cluster-based sample selection method designed to judiciously identify representative few-shot samples.

We evaluate the MG-ViT model through comprehensive evaluations of image classification, object detection, and segmentation tasks. Extensive comparisons are conducted with the conventional fine-tuning-based ViT model and the widely employed ResNet model. The experimental findings unequivocally establish that the proposed MG-ViT model markedly enhances task performance and efficiency, surpassing the outcomes achieved by both the general fine-tuning-based ViT and ResNet models.

In general, the main novelties and contributions of our work are listed in the following.

1) We introduce a refined mask operation applied to image patches, offering guidance to the ViT in focusing on task-relevant and discriminative patches during FSL. In addition, a residual connection is incorporated to preserve global features from visible image patches while upholding the foundational structures of ViT. This innovative and artful approach contributes to the improvement of model effectiveness and efficiency, leveraging the inheritance of parameters from the pretrained ViT without incurring any additional costs.
2) We incorporate an active learning-based sample selection method to further enhance the efficiency of FSL within the proposed MG-ViT. This methodology strategically selects representative few-shot samples, thereby augmenting the overall efficiency of FSL within the MG-ViT framework. This represents a notable advancement in the optimization of FSL techniques.
3) We propose an effective and efficient FSL framework specifically tailored for data-intensive models (ViT in this study) and remarkably yields outstanding performance in image classification, object detection, and segmentation tasks. This accomplishment not only underscores the efficacy of the proposed approach but also provides novel insights into the generalization of data-intensive and large-scale deep learning models for FSL. This significant contribution advances the current understanding of FSL techniques within the context of sophisticated neural network architectures.

## II. Related Work

### A. Vision Transformer

ViT [11] is transferred from the structure of the transformer in natural language processing (NLP) tasks [18]. Recently, several refined ViT models such as DeiT [19], CeiT [20], local ViT [21], and NesT [22] are introduced with useful strategies, including knowledge distillation, depthwise convolution and tree-like structure, and achieve improved model performance and efficiency in the image classification task.

However, the data-intensive ViT is difficult to quickly adapt to the target domain with a small amount of labeled data. By using a distillation approach [16], smoothing the loss landscapes at convergence [23], or incorporating CNNs such as CCT [24] and NesT [22], ViT can reduce the demand for large sample data to some extent. Moreover, MAE [25] is proposed to mask most of the image patches and apply an unsupervised image reconstruction method to pretrain transformer to improve its generalizability for downstream tasks. BEiT [26] and BEiT v2 [27] propose a mask mechanism similar to MAE but reconstruct on latent codes of discrete VAE to better pretrain and solve the data-hungry issue. The MFGN [28] is a feature-augmentation approach based on the mask with a strong performance on FSL. However, these studies are still far from the requirement of fast adaption to the target domain for FSL.

### B. Few-Shot Learning

FSL is proposed for learning with only a few samples [3]. Metalearning, which is also known as learning-to-learn, is a crucial approach for FSL [29]. In recent studies for FSL,

Authorized licensed use limited to: University of Georgia. Downloaded on September 30,2024 at 15:19:01 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



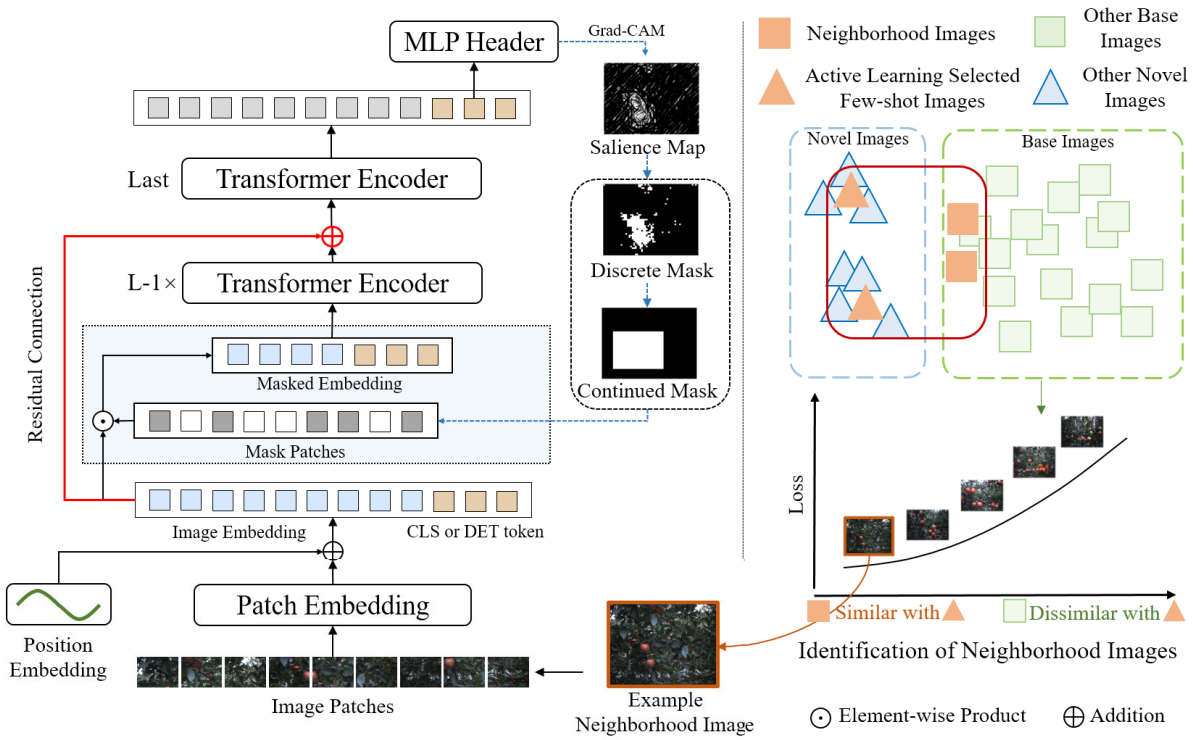



Fig. 2. Overall framework of MG-ViT for FSL. Left: model structure of MG-ViT. The proposed additional mask operation is highlighted with blue background, and the residual connection is highlighted with red line. The continued or discrete image patch mask of task-irrelevant regions is generated based on the salience map by Grad-CAM on the patch embedding layer. Right: orange triangles are representative few-shot samples selected from the novel dataset by active learning. Orange squares are identified neighborhood samples from the base dataset having lower loss values (i.e., higher similarity) with the representative few-shot samples. The identified neighborhood samples and representative few-shot samples highlighted in the red box are the training data during the joint fine-tuning stage. CLS: class token. DET: detect token.

Zhang et al. [7] proposed a novel absolute-relative learning paradigm to fully use the binary label and soft similarity information. Ma et al. [6] designed an inverted pyramid network (IPN) with global and local stages to learn the support–query relation and precise query-to-class similarity embedding. Yang et al. [4] calibrated the distribution of few sample classes by transferring statistics from the classes with sufficient examples subject to a Gaussian distribution of each feature representation. For transformer-based FSL, recent studies [30], [31], [32] merely fix the CNN-based feature extractor trained on the base class and apply the attention mechanisms to exploit the correlation between query and support sets to perform classification on FSL, which does not take full advantage of the powerful learning representation of ViT.

### C. Active Learning

Previous studies usually randomly select certain images from the dataset as training samples and have the model *learn from examples* [15]. However, by actively selecting a fixed number of training data, active learning is provably more powerful and with better generalizability [15], [17]. For example, Chitta et al. [16] designed an ensemble active learning method and achieved better performance on the test data with less training data. Therefore, active learning may improve the efficiency of FSL by labeling representative few-shot samples [3]. For example, Yan et al. [5] reported a better and more efficacy result on FSL using a graph convolution network (GCN)-based active learning data selection policy.

## III. Methods

The overall framework, as depicted in Fig. 2, includes two distinct parts: data augmentation from the base images and mask-guided model training. We first illustrate the definitions of our problem and ViT architecture in Section III-A. We next demonstrate the detailed structure of MG-ViT in Section III-B and the generation of image patch mask in Section III-C. We then introduce the identification of neighborhood image in Section III-D and active learning-based few-shot sample selection in Section III-E. Finally, we provide the overall training scheme in Section III-F.

### A. Problem Definition and ViT

*1) Problem Definition:* Given a base dataset $D_b = \{(x, y)\}$, where $y \in Y_b$, and a novel dataset $D_n = \{(x, y)\}$, where $y \in Y_n$, $Y_b \cup Y_n = Y$ and $Y_b \cap Y_n = \emptyset$, where $Y$ denotes the whole set of class. The base dataset is with a large amount of labeled data, while the novel dataset is with only a few labeled data. Our aim is to train a model with both base and novel datasets while achieving satisfying generalizability on the novel dataset for FSL. The common practice to evaluate the fast adaptation ability and generalizability of the FSL model is to build a $N$-way-$K$-shot task, where $N$ is the number of classes and $K$ is the number of labeled data in

Authorized licensed use limited to: University of Georgia. Downloaded on September 30,2024 at 15:19:01 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



the novel dataset. The model is trained with only $N \times K$ labeled data from the novel dataset. The performance of the FSL model is evaluated on the test split of the novel dataset.

*2) Vision Transformer:* ViT [11] receives a patch embedding sequence $x_{\text{PATCH}} \in \mathbb{R}^{N\times D}$, where $N = (H * W/P^2)$ is the number of patches, $D$ is the output dimension, and $(H, W)$ and $(P, P)$ are the resolutions of the image and patch, respectively. For different downstream tasks, $x_{\text{PATCH}}$ concatenates different tokens, e.g., $x_{\text{CLS}} \in \mathbb{R}^{1\times D}$ class token [11] for the image classification task and $x_{\text{DET}} \in \mathbb{R}^{100\times D}$ detect token [33] for the object detection task. To retain the position information of patches in the whole image, position embedding $\mathbf{P} \in \mathbb{R}^{(N+1+100)\times D}$ is also added to the concatenated inputs. Therefore, the input sequence $z_0$ of ViT with both class token and detect token is described as follows:

$$z_0 = [x_{\text{PATCH}}; x_{\text{DET}}; x_{\text{CLS}}] + \mathbf{P}. \tag{1}$$

The encoder layer of Transformers consists of one multihead self-attention (MSA) block and one multilayer perceptron (MLP) block. LayerNorm (LN) and residual connections are applied before and after every block, respectively. Therefore, the output embedding of the $l$th layer is

$$z'_l = \text{MSA}(\text{LN}(z_{l-1})) + z_{l-1}, \quad l = 1, \ldots, L \tag{2}$$

$$z_l = \text{MLP}\big(\text{LN}\big(z'_l\big)\big) + z'_l, \quad l = 1, \ldots, L. \tag{3}$$

For different downstream tasks, the task-relevant tokens are fed into the task-special MLP header for final prediction. For the image classification task

$$y_{\text{class}} = \text{MLP}(x_{\text{CLS}}). \tag{4}$$

For object detection task

$$y_{\text{class}} = \left[\text{MLP}_c\left(x^1_{\text{DET}}\right); \cdots; \text{MLP}_c\left(x^{100}_{\text{DET}}\right)\right] \tag{5}$$

$$y_{\text{bbox}} = \left[\text{MLP}_b\left(x^1_{\text{DET}}\right); \cdots; \text{MLP}_b\left(x^{100}_{\text{DET}}\right)\right]. \tag{6}$$

For segmentation task

$$y_{\text{seg}} = CNNs(x_{\text{PATCH}}). \tag{7}$$

### B. Mask-Guided ViT

We introduce the detailed structure of the proposed MG-ViT as follows. One of the core issues for FSL is to effectively and efficiently adapt the prior knowledge learned from the source domain (base dataset) to the target domain (novel dataset). Inspired by He et al. [25], we add an image patch mask (as detailed in Section III-C) to the patch embeddings before the first encoder layer of the transformer to screen out the task-irrelevant image patches and to guide ViT focusing on task-relevant and discriminative ones. Different from masking random patches in MAE [25], our design masks the task-irrelevant ones to focus on the task-relevant prior knowledge for better generalizability. Note that we only apply the mask operation on the base dataset but not on the novel dataset for two reasons. First, we want to make full use of the few-shot sample information in the novel dataset for better FSL. Second, it is difficult to identify the important features within only a few samples, which may lead to a noisy salience map. Therefore, the input of the first encoder layer $z_0^{\text{masked}}$ is

$$\begin{aligned} z_0^{\text{masked}} &= [x_{\text{PATCH}} \odot \text{Mask}; x_{\text{DET}}; x_{\text{CLS}}] \\ &\quad + [P_{\text{PATCH}} \odot \text{Mask}; P_{\text{DET}}; P_{\text{CLS}}] \end{aligned} \tag{8}$$

where $P_{\text{PATCH}}$, $P_{\text{DET}}$, and $P_{\text{CLS}}$ are the patch position embedding, detect position embedding, and class position embedding in $P$, respectively. $\odot$ is the elementwise production.

Moreover, we introduce a residual connection between the first and last encoder layers of ViT to retain the global features, especially the position information, of visible patches. Inspired by He et al. [25], [34], we add the embeddings of all image patches before the first encoder layer to the input of the last encoder layer, as shown in Fig. 2. The input of the last encoder layer $\hat{z}_{L-1}$ is

$$\begin{aligned} \hat{z}_{L-1} &= \left[\hat{x}^{L-1}_{\text{PATCH}}; x^{L-1}_{\text{DET}}; x^{L-1}_{\text{CLS}}\right] \\ \hat{x}^{L-1^i}_{\text{PATCH}} &= \begin{cases} x^{L-1^i}_{\text{PATCH}} + x^{0^i}_{\text{PATCH}}, & \text{if Mask}_i = 1 \\ x^{0^i}_{\text{PATCH}}, & \text{otherwise} \end{cases} \end{aligned} \tag{9}$$

where $\hat{x}^{L-1}_{\text{PATCH}} = \{x^{L-1^i}_{\text{PATCH}} | i = 1, \ldots, N\}$, $x^{L-1}_{\text{DET}}$, and $x^{L-1}_{\text{CLS}}$ are the image patch token, detect token, and class token of the last layer, respectively.

Compared to the vanilla ViT model, MG-ViT only introduces an additional mask operation and a residual connection, thus enabling the model to focus on the most task-relevant image patches and to inherit parameters from pretrained ViT without any other retraining. On the other hand, the computational costs of MLP (two-layer with hidden dimension $d$ and $d'$) and MSA in each ViT layer are $2Ndd'$ and $2N^2d + 4Nd$ (where $N$ is the patch number), respectively [35]. Therefore, the introduction of a mask reduces the computational costs in the masked layer with smaller $N$. We will also discuss the effects of mask areas and mask layers in Section V-E. In summary, MG-ViT can achieve fast domain adaption for FSL on ViT.

### C. Generation of Image Patch Mask

In order to take better advantage of the prior knowledge and to reduce the deviation between the source domain (base dataset) and the target domain (novel dataset), we generate the image patch mask based on the most task-relevant and discriminative patches from the salience map calculated by Grad-CAM [14]. Transformer treats graphs consisting of patches of nodes as complete graphs for processing, which requires the model to be able to accurately identify important nodes, so as to avoid the influence of background information on the model, e.g., shortcut learning [36]. Therefore, in order to enable the model to focus on important nodes faster when dealing with the task of fewer samples, we used the model interpretability-based techniques [37] to identify important nodes and restrict the model to process only these important nodes by adding a mask to the attention operation. Specifically, we first adopt Grad-CAM [14] to calculate the salience map of the identified neighborhood samples from the base dataset (as detailed in Section III-D). We then select top $k$ salient patches with the largest absolute sum values of the gradient of image

Authorized licensed use limited to: University of Georgia. Downloaded on September 30,2024 at 15:19:01 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



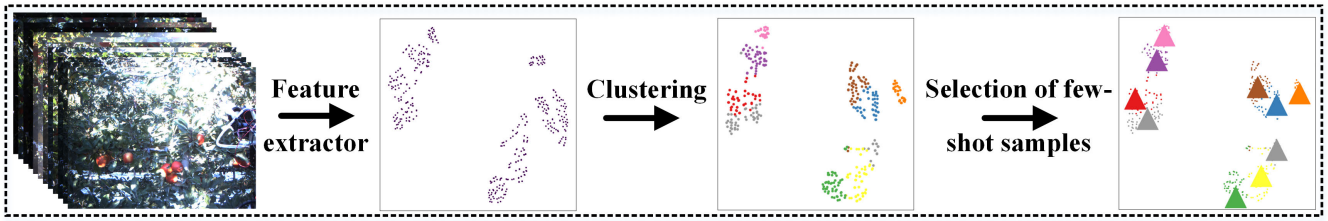



Fig. 3. Cluster-based active learning method for few-shot sample (triangles in the last subfigure) selection.

patch features after patch embedding in the salience map as the most task-relevant and discriminative ones. We finally perform a binary operation on the salience map to label those top $k$ salient patches as 1 and the remaining ones as 0. The generation of the image patch mask is written as follows:

$$g_i = \text{Sum}|\frac{\partial L(f(x), y)}{\partial x^i_{\text{PATCH}}}|, \quad i = 1, 2, \ldots, N \tag{10}$$

$$\text{Mask}^i = \begin{cases} 1, & \text{if } g_i \text{ in top}k\ G \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where $G \in \mathbb{R}^{\mathbb{N}} = \{g_1, g_2, \ldots, g_N\}$ is the salience map of patches $x_{\text{PATCH}} = \{x^1_{\text{PATCH}}, \ldots, x^N_{\text{PATCH}}\}$, top$k$ is the set of selected salient patches, and $\text{Mask} = \{\text{Mask}^1, \ldots, \text{Mask}^N\}$ is the binary mask. Besides the discrete mask generated from the Grad-CAM-based salience map, we also generate a continued one based on the center coordinates and designated length and width of the discrete mask in order to provide more contour information of target localization for object detection task (see Fig. 2).

### D. Identification of Neighborhood Samples From the Base Dataset

Identifying neighborhood samples [38] from the base dataset, which is similar to the few-shot samples in the novel dataset for joint fine-tuning, can improve the performance for FSL [4], [39], [40]. Inspired by Paul et al. [41] using the gradient normed and error $l_2$-norm scores to identify the important samples that are hard for model training with large loss, we measure the similarity as the negative loss value of the sample in the base dataset, with the model trained on the selected few-shot samples from the novel dataset (as detailed in Section III-E). The similarity of sample $(x_i, y_i) \in D_b$ with $D_n$ is written as follows:

$$\text{Sim}(x_i, D_n) = -L(f_{\hat{W}}(x_i), y_i) \tag{12}$$

$$\hat{W} = \arg\min_W L(f_W(x), y), \quad \forall (x, y) \in D_n \tag{13}$$

where $L(*)$ is the loss function, $f(*)$ represents the model's output of input $x$, $W$ is the set of parameters of the model, and $\hat{W}$ is the set of parameters of the model with least loss.

In this way, we effectively combine both the model characteristics and labeling information of the data without performing complex similarity calculations between the two datasets. The proposed method of loss-based image similarity is similar to the anomaly score in anomaly detection [42], which measures the distance of a data point from the center of a sphere. The lower loss allows the model to focus more on learning the representation of the novel dataset, which contributes to the performance of FSL.

### E. Active Learning-Based Few-Shot Sample Selection

To identify samples more conducive to model learning and further improve the generalizability of MG-ViT-based FSL, we introduce a cluster-based sample selection method to optimally select representative few-shot samples. Considering that the elaborate active learning methods may lead to overfitting problems, we design a simple and efficient method inspired by [39] and [43], as illustrated in Fig. 3. We first use a CNN-based model as a feature extractor to extract image features, then perform an unsupervised clustering to identify $k$ clusters as $k$-shot, and, finally, select the image with the highest node degree in each cluster as the few-shot sample. In this way, for each cluster divided by unsupervised clustering, we can get a representative sample of each cluster for model training, thus improving the model learning efficiency. In this study, we use the pretrained ResNet-101 [34] as the backbone, which contains 43 M parameters to extract features and is pretrained on ImageNet with labeled data for the classification task. We use the $k$-means to cluster image features and the Euclidean distance as the weight of adjacency.

### F. Overall Training Scheme

The overall training scheme of MG-ViT consists of two stages. The first stage is to train on the base dataset only with the vanilla ViT, and the second stage is to fine-tune the model on both the base and novel datasets with the proposed MG-ViT. We demonstrate the fine-tuning scheme at the second stage in Algorithm 1. Note that before we fine-tune MG-ViT with the novel dataset, we first initialize fine-tuning of the model on the base dataset for a few epochs since the change of computational flow of the model at the fine-tuning stage may lead to deviations in the input of the last encoder part.

## IV. Experiments

We evaluate our MG-ViT based on ViT-S and ViT-B [11] and conduct experiments on image classification, object detection, and segmentation tasks. We compare our methods with the general fine-tuning-based FSL methods [9]. We report the average accuracy (ACC) for the image classification task, the average precision (AP) score for the object detection task, and the mean Dice coefficient on the test split of the novel dataset.

Authorized licensed use limited to: University of Georgia. Downloaded on September 30,2024 at 15:19:01 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



TABLE I
AVERAGED ACCURACY FOR DIFFERENT NUMBERS OF SHOTS IN IMAGE CLASSIFICATION TASK AT SEVEN DIFFERENT DATASETS

| BACKBONE | AGRI-IMAGENET | | | FLOWER17 | | | MINIIMAGENET | | | COCO | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1-SHOT | 5-SHOT | 10-SHOT | 1-SHOT | 5-SHOT | 10-SHOT | 1-SHOT | 5-SHOT | 10-SHOT | 1-SHOT | 5-SHOT | 10-SHOT |
| RESNET101 [34] | 50.9 | 91.9 | 90.7 | 22.1 | 57.5 | 77.6 | 0.5 | 28.4 | 51.7 | 0.1 | 25.7 | 37.6 |
| SWIN-T [44] | 34.9 | 93.4 | 95.6 | 44.9 | 90.3 | 95.4 | **21.7** | **61.1** | 70.5 | 11.5 | 28.8 | 40.0 |
| VIT-S [11] | 62.8 | 86.1 | **96.9** | 31.2 | 81.9 | 87.0 | 15.4 | 54.7 | 63.0 | **13.2** | **36.0** | **45.3** |
| MG-VIT-S | **91.6** | **95.2** | **96.9** | **84.3** | **96.2** | **98.3** | 21.2 | 59.8 | **76.9** | 6.8 | 24.9 | 40.8 |
| VIT-B [11] | 69.1 | 87.5 | 91.4 | 64.9 | 93.4 | 96.2 | 16.3 | 45.5 | 62.0 | 11.6 | 29.0 | 34.6 |
| SWIN-B [44] | 69.2 | 93.3 | 97.2 | 67.1 | 94.1 | 96.5 | 42.0 | 68.2 | 76.3 | **15.0** | **40.8** | **47.7** |
| BEIT-B [26] | 62.4 | 94.7 | 96.1 | 63.8 | 91.3 | 95.7 | 9.8 | 54.4 | 71.2 | 12.8 | 29.7 | 37.2 |
| BEITV2-B [27] | 77.2 | 87.5 | 89.2 | 60.2 | 88.8 | 96.7 | 27.2 | 66.1 | 70.5 | 11.6 | 31.7 | 41.5 |
| MAE-B [25] | 85.0 | 96.5 | 97.6 | 63.4 | 96.9 | 97.5 | 26.4 | 67.7 | 84.3 | 13.7 | 34.1 | 44.4 |
| MG-VIT-B | **91.0** | **97.9** | **98.9** | **78.3** | **98.1** | **97.7** | **44.5** | **85.9** | **92.8** | 12.3 | 31.6 | 44.6 |

| BACKBONE | CIFAR-FS | | | CUB200 | | | STANFORDDOGS | | | AVERAGE | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1-SHOT | 5-SHOT | 10-SHOT | 1-SHOT | 5-SHOT | 10-SHOT | 1-SHOT | 5-SHOT | 10-SHOT | 1-SHOT | 5-SHOT | 10-SHOT |
| RESNET101 [34] | 0.2 | 8.9 | 32.0 | 5.5 | 47.1 | 66.0 | 0.8 | 32.2 | 45.3 | 11.4 | 41.7 | 57.3 |
| SWIN-T [44] | 11.2 | **38.0** | 47.9 | 11.3 | 50.2 | 67.7 | 15.9 | 44.9 | 53.1 | 21.6 | 58.1 | 67.2 |
| VIT-S [11] | 10.4 | 36.6 | 50.4 | 15.6 | 50.1 | 63.6 | 5.9 | 27.2 | 38.4 | 22.1 | 53.2 | 63.5 |
| MG-VIT-S | **14.9** | 34.3 | **58.6** | **23.2** | **70.3** | **84.7** | **22.2** | **57.2** | **80.4** | **37.7** | **62.6** | **76.6** |
| VIT-B [11] | 3.9 | 32.3 | 42.0 | 8.6 | 43.8 | 69.3 | 1.8 | 21.0 | 36.2 | 25.2 | 50.4 | 61.7 |
| SWIN-B [44] | 12.4 | 41.8 | 53.3 | 18.4 | 63.3 | 71.4 | 6.7 | 33.3 | 57.4 | 33.0 | 62.1 | 71.4 |
| BEIT-B [26] | 8.5 | 36.7 | 46.1 | 11.1 | 48.7 | 66.8 | 11.5 | 46.1 | 56.4 | 29.6 | 57.9 | 66.7 |
| BEITV2-B [27] | 12.2 | 34.5 | 46.5 | 18.4 | 73.7 | 80.7 | 10.2 | 52.2 | 74.7 | 32.8 | 65.1 | 75.1 |
| MAE-B [25] | 0.2 | 33.8 | 46.1 | 15.6 | 64.8 | 79.2 | 9.4 | 43.6 | 60.5 | 24.9 | 58.9 | 69.4 |
| MG-VIT-B | **24.4** | **46.9** | **61.9** | **19.6** | **76.9** | **86.8** | **31.0** | **74.4** | **87.2** | **43.0** | **73.1** | **81.4** |

**Algorithm 1** Fine-Tuning Scheme With MG-ViT

**Input:** base dataset $D_b$ and novel dataset $D_n$
**Initialize:** model pre-training on $D_b$
**for** epoch **in** initial fine-tuning step **do**
  $train\_one\_epoch(model, (D_b, Mask = None))$
**end for**
**while** training **do**
  **for** $(x, y)$ **in** $D_b$ **do**
    $Sim_{x_i} = -loss(f(x_i), y_i)$
  **end for**
  $D_b^{sub} = topk(Sim_{x_i})$
  $Mask_b = Grad - CAM(model, D_b^{sub})$ as Eq.10
  $train\_one\_epoch(model, (D_n, None) \cup (D_b^{sub}, Mask_b))$
**end while**

### A. Image Classification

*1) Dataset:* We carry out the image classification task based on Agri-ImageNet [45], Flower17 [46], miniImageNet [47], COCO [48], CIFAR-FS [49], CUB200 [50], and StanfordDogs [51] dataset. The Agri-ImageNet dataset contains three parent classes including fruit, weed, and vegetable. We randomly select the fruit (with nine subclasses) and weed (with eight subclasses) parent classes as the base dataset and the vegetable (with four subclasses) one as the novel dataset. The Flower17 dataset contains 17 kinds of flowers with 60 images for each class. We perform a four-way few-shot image classification task on Agri-ImageNet since there are four subclasses in the vegetable parent class and also in the Flower17 dataset. For the datasets miniImageNet, COCO, CIFAR-FS, CUB200, and StanfordDogs that have the class numbers 100, 80, 100, 200, and 120, respectively, we perform a 20-way few-shot image classification task. The base dataset is randomly split into training/test with 75%/25%. The remaining data in the novel dataset except for the actively selected few-shot samples is the test split of a novel dataset. For the training dataset, Rand-Augment [52], Random Erasing [53], and RandomResizeCrop to 224 × 224 are applied for data augmentation. For the test dataset, images are only resized and center-cropped to 224 × 224.

*2) Setting:* The ImageNet-1k pretrained model is first trained on the base dataset with the vanilla ViT. We adopt an AdamW optimizer with 100 epochs using a cosine decay learning rate scheduler and 10 epochs of linear warm-up. The batch size is set to 64, the initial learning rate is set to 0.0001, and the weight decay is set to 0.0001. Then, we fine-tune the model on the few-shot samples in the novel dataset together with the neighborhood samples from the base dataset with MG-ViT. We keep the same setting of regular training except for the initial learning rate of 0.001 and epochs to 30. The cross-entropy loss is adopted, and the label smoothing is set to 0.1. The top$k$ is set to 7 × 7 to select the salient patches for mask generation.

*3) Result:* We compare our MG-ViT with other vision models, including ResNet101, Swin, and ViT based on fine-tuning methods, and mask-based methods, including MAE, BEiT, and BEiT v2. As reported in Table I, we see that MG-ViT achieved the highest accuracy on most datasets and the average result in FSL tasks, demonstrating the superiority of the proposed MG-ViT in FSL. Fig. 4 displays visualizations of the patch

Authorized licensed use limited to: University of Georgia. Downloaded on September 30,2024 at 15:19:01 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



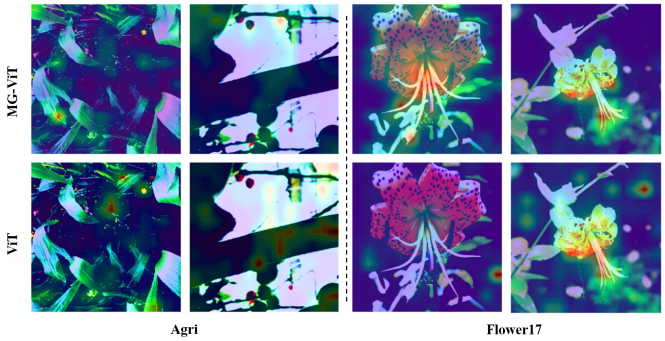



Fig. 4. Four representative patch salience maps generated by the proposed MG-ViT and ViT on Agri-ImageNet and Flower17 datasets.

salience maps, demonstrating that MG-ViT exhibits a higher precision in identifying task-relevant patches compared to fine-tuning-based ViT approaches.

### B. Object Detection

*1) Dataset:* We apply the object detection task on the apple class of the ACFR Orchard Fruit Dataset [54]. The dataset collected in 2016 and 2017 is used as the base and novel dataset, respectively. The base dataset is randomly split into training/validation/test with 64%/16%/20%. The remaining data in the novel dataset except for the few-shot samples are the test split of a novel dataset. For the training dataset, data augmentation includes resizing and cropping the input images so that the shortest side is between 432 and 720 pixels and the longest one is at most 960 pixels. For the test dataset, the images are only resized to 720 × 960 and normalized.

*2) Setting:* We follow Fang et al. [33] and use the provided ImageNet-1k pretrained ViT-S as the backbone. We continue pretraining the model on the Agri-ImageNet and Flower17 and keep the same setting as the training on the base dataset for better performance. Since the pretrained model is trained on images with a resolution of 224 × 224, while the ones in the apple dataset are with higher resolution, we adopt a bicubic interpolation [19] for position embedding to fit the apple images. AdamW optimizer with a cosine decay learning rate scheduler is employed during the training on the base dataset. Batch size, epochs, initial rate, and weight decay are set to 1, 200, 0.0001, and 0.0001, respectively. We use the same hyperparameters except for the initial rate and epochs set to 0.00001 and 400, respectively, during the fine-tuning of the novel dataset. We use the same loss function as in [55]. The top$k$ is set to 24 × 32 to select the salient patches for mask generation.

*3) Result:* We compare the proposed MG-ViT with the baseline method, a pure transformer structural object model of YOLOS-S [4] with fine-tuning [9]. As reported in Table II, we see that fine-tuning the full network (Ft-full) performs better than only fine-tuning the backbone (Ft-part) of the model, which is contrary to the conclusion of CNN-based few-shot object detection. By performing the mask operation, our MG-ViT outperforms the general fine-tuning-based method (ViT-S) by 2.7% from 47.9% to 50.6% in five-shot, 9.1% from 56.2% to 65.3% in ten-shot, and 1.1% from 74.9% to 76.0% in 30-shot. The patch salience maps of MG-ViT and YOLOS are further visualized and compared in Fig. 5 to illustrate that MG-ViT can locate the object (i.e., apple) more accurately compared to YOLOS.

TABLE II
AP SCORES FOR DIFFERENT NUMBERS OF SHOTS IN OBJECT DETECTION TASK BASED ON THE ACFR APPLE DATASET

| BACKBONE | METHOD | 5-SHOT | 10-SHOT | 30-SHOT |
|---|---|---|---|---|
| VIT-S [33] | FT-FULL | 48.2 | 63.2 | 72.9 |
| VIT-S [33] | FT-PART | 26.4 | 28.7 | 33.2 |
| MG-VIT-S | FT-FULL | **50.6** | **65.3** | **76.0** |

TABLE III
MEAN DICE COEFFICIENT FOR DIFFERENT NUMBERS OF SHOTS IN SEGMENTATION TASK BASED ON THE PASCAL VOC2012 DATASET

| BACKBONE | 5-SHOT | 10-SHOT | 30-SHOT |
|---|---|---|---|
| VIT-S [11] | 36.2 | 39.6 | 43.8 |
| VIT-B [11] | 38.1 | 45.8 | 46.6 |
| MG-VIT-S | 43.7 | 44.7 | 52.5 |
| MG-VIT-B | **45.9** | **51.3** | **54.7** |

### C. Segmentation

*1) Dataset:* We perform the segmentation task on the PASCAL visual object classes (VOCs) 2012 Dataset [56]. The images containing the categories "sheep," "sofa," "train," and "TV monitor" are used as novel datasets, and the rest of the remained images are the base dataset. In the novel dataset, all the data except for the selected few-shot images are used as a test set. The input images are all resized to 224 × 224 in both the training and test sets.

*2) Setting:* We follow Strudel et al. [57] and use the provided ImageNet-1k pretrained ViT-S and ViT-B as the backbone to extract the features from images. We simplified our segmentation process by replacing the decoder with two CNN layers and one interpolation layer for efficient prediction. The Adam optimizer with a cosine decay learning rate scheduler is employed during the training on the base dataset. Batch size, epochs, initial learning rate, and min learning rate are set to 128, 100, 0.0001, and 0.00001, respectively. We use the same hyperparameters except for batch size and epochs set to 4 and 30, respectively, during the fine-tuning of the novel dataset. We use the Dice loss function as in [58]. The top$k$ is set to 49 to select the salient patches for mask generation.

*3) Result:* We compare the proposed MG-ViT with the baseline, i.e., a pure ViT structural model. As reported in Table III, our method MG-ViT-S achieved mean Dice coefficient of 43.7%, 44.7%, and 52.5% for the five-, ten-, and 30-shot tasks, respectively, MG-ViT-B achieved mean Dice coefficient of 45.9%, 51.3%, and 54.7%, respectively, and ViT-S and ViT-B achieved mean Dice coefficient of 36.2%, 39.6%, 43.8%, and 38.1%, 45.8%, 46.6%, respectively. Intriguingly, our results suggest a relationship between the size of model parameters and the number of train data in enhancing segmentation. Increasing both parameters (from ViT-S/MG-ViT-S to ViT-B/MG-ViT-B) and data (five-shot, ten-shot, and 30-shot) progressively improves segmentation performance.

Authorized licensed use limited to: University of Georgia. Downloaded on September 30,2024 at 15:19:01 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



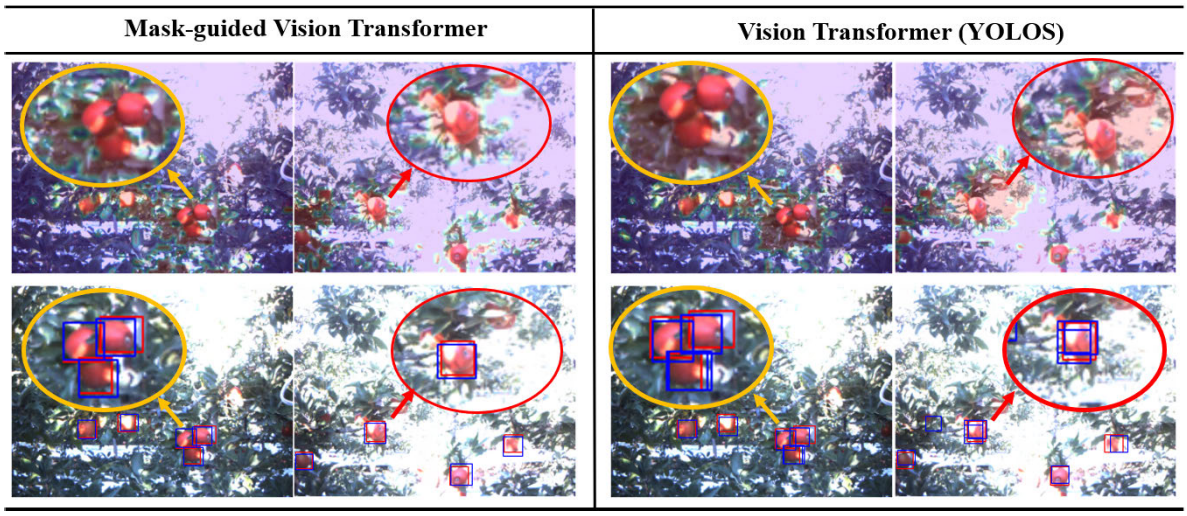



Fig. 5. Two representative patch salience maps (top) and predict *b*box (bottom, red: ground truth and blue: predicted) in MG-ViT and YOLOS, respectively. In each subfigure, yellow circle highlights the enlarged view of one example salient region.

TABLE IV
PERFORMANCES OF W/ AND W/O ACTIVE LEARNING-BASED FEW-SHOT SAMPLE SELECTION ON AGRI-IMAGENET IMAGE CLASSIFICATION AND ACFR APPLE OBJECT DETECTION WITH MG-VIT

| AGRI-IMAGENET | 1-SHOT | 5-SHOT | 10-SHOT |
|---|---|---|---|
| W/ | 85.6 | 98.1 | 98.5 |
| W/O | 84.6 | 96.5 | 96.6 |
| ACFR APPLE | 5-SHOT | 10-SHOT | 30-SHOT |
| W/ | 50.6 | 65.3 | 76.0 |
| W/O | 47.2 | 56.3 | 75.1 |

TABLE V
PERFORMANCES OF W/ AND W/O NEIGHBORHOOD SAMPLES' IDENTIFICATION ON AGRI-IMAGENET IMAGE CLASSIFICATION AND ACFR APPLE OBJECT DETECTION WITH VIT

| AGRI-IMAGENET | 1-SHOT | 5-SHOT | 10-SHOT |
|---|---|---|---|
| W/O | 77.0 | 90.4 | 96.7 |
| W/ | 82.1 | 97.8 | 97.9 |
| ACFR APPLE | 5-SHOT | 10-SHOT | 30-SHOT |
| W/O | 48.2 | 63.2 | 72.9 |
| W/ | 48.6 | 64.1 | 75.3 |

## V. ABLATION STUDY

In this section, we perform ablation studies to demonstrate the effectiveness of our model design including the active learning-based few-shot sample selection and neighborhood samples' identification. We also evaluate the influence of different mask shapes on different downstream tasks in our method.

### *A. Effect of Active Learning-Based Few-Shot Sample Selection*

We compare the effect of active learning-based few-shot sample selection with randomly selected few-shot samples in MG-ViT. As reported in Table IV, for the image classification task, active learning-based few-shot sample selection improves the accuracy by 1.0% from 84.6% to 85.6% in one-shot, 1.6% from 96.5% to 98.1% in five-shot, and 1.9% from 96.6% to 98.5% in ten-shot. Similarly, for object detection task, it improves the accuracy by 3.4% from 47.2% to 50.6% in five-shot, 9.0% from 56.3% to 65.3% in ten-shot, and 0.9% from 75.1% to 76.0% in 30-shot.

### *B. Effect of Neighborhood Samples' Identification*

We compare the effect of adopting identified neighborhood samples with randomly selected images from the base dataset in ViT. As reported in Table V, for image classification task, adopting identified neighborhood samples for joint fine-tuning with the novel dataset improves the accuracy by 5.1% from 77.0% to 82.1% in one-shot, 7.4% from 90.4% to 97.8% in five-shot, and 1.2% from 96.7% to 97.9% in ten-shot. Similarly, for the object detection task, it improves the accuracy by 0.4% from 48.2% to 48.6% in five-shot, 0.9% from 63.2% to 64.1% in ten-shot, and 2.4% from 72.9% to 75.3% in 30-shot.

### *C. Effect of Mask Strategy*

We compare the down-stream task performance between w/ and w/o the proposed mask strategy. For the model w/o masks, the identified neighborhood samples from the base dataset were only applied compared with ViT. As reported in Table VI, for the image classification task, adopting masks for joint fine-tuning improves the accuracy from 82.1% to 85.6% in one-shot, from 97.8% to 98.1% in five-shot, and from 97.9% to 98.5% in ten-shot. Similarly, for the object detection task, it improves the accuracy from 48.6% to 50.6% in five-shot, from 64.1% to 65.3% in ten-shot, and from 75.3% to 76.0% in 30-shot.

### *D. Effect of Discrete or Continued Mask*

We evaluate the effect of using a discrete or continued mask in MG-ViT. As reported in Table VII, the discrete mask

Authorized licensed use limited to: University of Georgia. Downloaded on September 30,2024 at 15:19:01 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



TABLE VI

AVERAGED ACC IN IMAGE CLASSIFICATION TASK AND AP IN OBJECT DETECTION TASK FOR DIFFERENT NUMBERS OF SHOTS W/ AND W/O MASK

| AGRI-IMAGENET | 1-SHOT | 5-SHOT | 10-SHOT |
|---|---|---|---|
| W/O MASK | 82.1 | 97.8 | 97.9 |
| W/ MASK | 85.6 | 98.1 | 98.5 |
| **ACFR APPLE** | **5-SHOT** | **10-SHOT** | **30-SHOT** |
| W/O MASK | 48.6 | 64.1 | 75.3 |
| W/ MASK | 50.6 | 65.3 | 76.0 |

TABLE VII

PERFORMANCES OF DISCRETE AND CONTINUED MASKS ON FIVE-SHOT IMAGE CLASSIFICATION AND TEN-SHOT OBJECT DETECTION WITH MG-VIT

| MASK SHAPE | DISCRETE | CONTINUED |
|---|---|---|
| 5-SHOT IMAGE CLASSIFICATION | **98.7** | 98.1 |
| 10-SHOT OBJECT DETECTION | 63.1 | **65.3** |

performs better for image classification, while the continued one performs better for object detection. This difference may derive from the fact that the continued mask could provide more contour information of target localization, which benefits object detection, while the discrete mask may provide more global semantic information for image classification tasks. We leave the exploration of different types of masks for different downstream tasks in FSL to future work.

### *E. Effect of Mask Areas and Mask Layers*

We evaluate the effect of different sizes of mask areas (i.e., number of patches) and different numbers of mask layers in MG-ViT.

The image classification task on Agri-ImageNet is selected as an example, and the five-shot accuracy is illustrated in Fig. 6. We see that the accuracy of the model shows an overall decrease tendency as the size of mask areas increases. To further investigate the relationship between mask area size and model accuracy, we conducted linear regression analyses on five- and ten-shot accuracies on the StanfordDogs dataset. The results, as visualized in Fig. 6, consistently revealed a trend of decreased accuracy as the size of mask areas increased. One possible reason is that increasing the size of mask areas could introduce certain task-irrelevant patches and, thus, decrease the accuracy of the model, which also validates the effectiveness of our model that focuses on task-relevant areas by mask for FSL. Moreover, although we use 11-layer mask layer in this study in order to minimize the computational cost and obtain superior performance compared to other methods, we see that the accuracy of the model is further increased as the number of mask layers decreases, indicating the effectiveness of our proposed mask strategy.

## VI. DISCUSSION

Our innovative MG-ViT strategically directs attention to the most task-relevant regions for optimal FSL. Our experiments

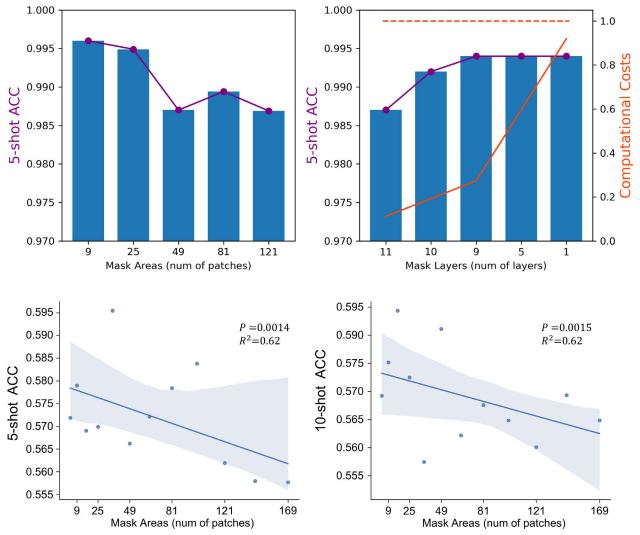



Fig. 6. Five-shot ACC (purple solid line) on Agri-ImageNet with different sizes of mask areas (top left) and the number of layers (top right). The orange dashed line shows the computational cost of vanilla ViT, while the orange solid line shows the computational cost of MG-ViT, which is evaluated in Section III-B (mask area = 9). The five-shot ACC (bottom left) and ten-shot ACC (bottom right) on StanfordDogs with different sizes of mask areas. $P$ and $R^2$ are key statistics arising from linear regression analysis.

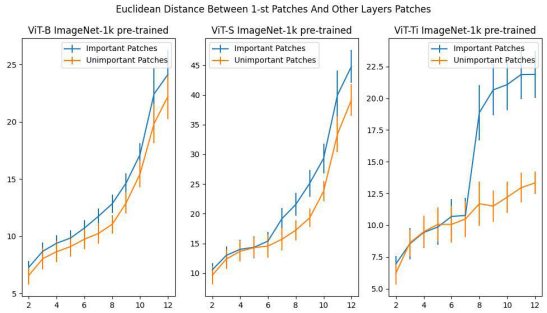



Fig. 7. Mean Euclidean distance between the first layer patches and other layers patches in ImageNet-1k pretrained ViT-B (left), ViT-S (middle), and ViT-Ti (right) on the ImageNet-1k validation dataset. The task-relevant (important) patches, compared to those task-irrelevant and less important ones, have higher similarity with the CLS token.

including image classification, object detection, and segmentation revealed superior performance of MG-ViT over the vanilla ViT. This improvement can be attributed to two key factors. First, MG-ViT mitigates redundant computations inherent in vanilla ViT. As depicted in Fig. 7, the model adeptly processes task-relevant and crucial patches while disregarding those that are task-irrelevant and less significant. Second, the mask-guided mechanism efficiently alleviates shortcut learning by eliminating unimportant patches within specific layers, as detailed in our findings.

## VII. CONCLUSION

We propose a novel MG-ViT to guide ViT more effectively and efficiently learn from the task-relevant prior knowledge for FSL. By simply adding an image patch mask operation and a residual connection to the vanilla ViT, MG-ViT significantly outperforms the general fine-tuning-based methods for FSL.

Authorized licensed use limited to: University of Georgia. Downloaded on September 30,2024 at 15:19:01 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



Our results are in agreement with ViT that learning the relevant patterns directly from data is sufficient, even beneficial. To further improve the efficiency of FSL, we also introduce an effective active learning-based few-shot sample selection method. Our two-stage fine-tuning-based framework could be widely applied to different downstream tasks, such as image classification, object detection, and segmentation in this study. In general, MG-ViT provides a concrete approach toward generalizing data-intensive and large-scale deep learning models for FSL. In the future, we can systematically explore the effectiveness of different types (e.g., size, shape, and data modality) of image patch masks for different downstream tasks for different downstream tasks such as brain disease identification [59] and brain cognitive assessment [60] based on medical images. Further combining the proposed mask operation with other models and modalities in other tasks such as NLP is another exciting direction of future work.

## REFERENCES


[1] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022.

[2] C. Chen, B. Wang, C. X. Lu, N. Trigoni, and A. Markham, "Deep learning for visual localization and mapping: A survey," *IEEE Trans. Neural Netw. Learn. Syst*, pp. 1–21, Sep. 2023.

[3] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surveys*, vol. 53, no. 3, pp. 1–34, Jun. 2020.

[4] S. Yang, L. Liu, and M. Xu, "Free lunch for few-shot learning: Distribution calibration," 2021, *arXiv:2101.06395*.

[5] S. Yan, S. Zhang, and X. He, "Budget-aware few-shot learning via graph convolutional network," 2022, *arXiv:2201.02304*.

[6] Y. Ma et al., "Few-shot visual learning with contextual memory and fine-grained calibration," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 811–817.

[7] H. Zhang, P. Koniusz, S. Jian, H. Li, and P. H. S. Torr, "Rethinking class relations: Absolute-relative supervised and unsupervised few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9427–9436.

[8] A. Nakamura and T. Harada, "Revisiting fine-tuning for few-shot learning," 2019, *arXiv:1910.00216*.

[9] X. Wang, T. E. Huang, T. Darrell, J. E. Gonzalez, and F. Yu, "Frustratingly simple few-shot object detection," 2020, *arXiv:2003.06957*.

[10] J. Cai and S. Mei Shen, "Cross-domain few-shot learning with meta fine-tuning," 2020, *arXiv:2005.10544*.

[11] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[12] Y. Liu et al., "A survey of visual transformers," *IEEE Trans. Neural Netw. Learn. Syst*, vol. 35, no. 6, pp. 7478–7498, Jun. 2024.

[13] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," 2021, *arXiv:2101.01169*.

[14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[15] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Mach. Learn.*, vol. 15, no. 2, pp. 201–221, 1994.

[16] K. Chitta, J. M. Álvarez, E. Haussmann, and C. Farabet, "Training data subset search with ensemble active learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 14741–14752, Sep. 2022.

[17] K. Konyushkova, R. Sznitman, and P. Fua, "Learning active learning from data," 2017, *arXiv:1703.03365*.

[18] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[19] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[20] K. Yuan, S. Guo, Z. Liu, A. Zhou, F. Yu, and W. Wu, "Incorporating convolution designs into visual transformers," 2021, *arXiv:2103.11816*.

[21] Y. Li, K. Zhang, J. Cao, R. Timofte, and L. Van Gool, "LocalViT: Bringing locality to vision transformers," 2021, *arXiv:2104.05707*.

[22] Z. Zhang, H. Zhang, L. Zhao, T. Chen, S. O. Arik, and T. Pfister, "Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding," 2021, *arXiv:2105.12723*.

[23] X. Chen, C.-J. Hsieh, and B. Gong, "When vision transformers outperform ResNets without pre-training or strong data augmentations," 2021, *arXiv:2106.01548*.

[24] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the big data paradigm with compact transformers," 2021, *arXiv:2104.05704*.

[25] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," 2021, *arXiv:2111.06377*.

[26] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," 2021, *arXiv:2106.08254*.

[27] Z. Peng, L. Dong, H. Bao, Q. Ye, and F. Wei, "BEiT v2: Masked image modeling with vector-quantized visual tokenizers," 2022, *arXiv:2208.06366*.

[28] Y. Yu, D. Zhang, and Z. Ji, "Masked feature generation network for few-shot learning," in *Proc. 31st Int. Joint Conf. Artif. Intell. (IJCAI)*, L. D. Raedt, Ed. Jul. 2022, pp. 3695–3701, doi: 10.24963/ijcai.2022/513.

[29] J. Vanschoren, "Meta-learning: A survey," 2018, *arXiv:1810.03548*.

[30] L. Liu, W. Hamilton, G. Long, J. Jiang, and H. Larochelle, "A universal representation transformer layer for few-shot image classification," 2020, *arXiv:2006.11702*.

[31] T. Gan, W. Li, Y. Lu, and Y. He, "Transformer-based few-shot learning for image classification," in *Proc. Int. Conf. Artif. Intell. Commun. Netw.* Xining, Ching: Springer, 2021, pp. 68–74.

[32] H. Chen, H. Li, Y. Li, and C. Chen, "Sparse spatial transformers for few-shot learning," 2021, *arXiv:2109.12932*.

[33] Y. Fang et al., "You only look at one sequence: Rethinking transformer in vision through object detection," 2021, *arXiv:2106.00666*.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, vol. 16, 2016, pp. 770–778.

[35] Y. Tang et al., "Patch slimming for efficient vision transformers," 2021, *arXiv:2106.02852*.

[36] R. Geirhos et al., "Shortcut learning in deep neural networks," *Nature Mach. Intell.*, vol. 2, no. 11, pp. 665–673, Nov. 2020.

[37] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021.

[38] Q. He, Z. Xie, Q. Hu, and C. Wu, "Neighborhood based sample and feature selection for SVM classification learning," *Neurocomputing*, vol. 74, no. 10, pp. 1585–1594, May 2011.

[39] W. Ge and Y. Yu, "Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1086–1095.

[40] O. Sbai, C. Couprie, and M. Aubry, "Impact of base dataset design on few-shot image classification," in *Proc. Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, 2020, pp. 597–613.

[41] M. Paul, S. Ganguli, and G. K. Dziugaite, "Deep learning on a data diet: Finding important examples early in training," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1–12.

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

[43] X. Cai, F. Nie, H. Huang, and F. Kamangar, "Heterogeneous image feature integration via multi-modal spectral clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1977–1984.

[44] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," 2021, *arXiv:2103.14030*.

[45] H. Dai et al., "Hierarchical semantic tree concept whitening for interpretable image classification," 2023, *arXiv:2307.04343*.

[46] M.-E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2006, pp. 1447–1454.

[47] O. Vinyals, C. Blundell, T. Lillicrap, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

[48] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Computer Vision–ECCV*. Zurich, Switzerland: Springer, 2014, pp. 740–755.


Authorized licensed use limited to: University of Georgia. Downloaded on September 30,2024 at 15:19:01 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



[49] L. Bertinetto, J. F. Henriques, P. H. S. Torr, and A. Vedaldi, "Meta-learning with differentiable closed-form solvers," 2018, *arXiv:1805.08136*.
[50] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 49–58.
[51] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. CVPR*, 2011, vol. 2, no. 1, pp. 1–2.
[52] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2020, pp. 702–703.
[53] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 13001–13008.
[54] S. Bargoti and J. P. Underwood, "Image segmentation for fruit detection and yield estimation in apple orchards," *J. Field Robot.*, vol. 34, no. 6, pp. 1039–1060, Sep. 2017.
[55] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.
[56] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.
[57] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7262–7272.
[58] F. Milletari, N. Navab, and S. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.
[59] L. Zhang, J. Qu, H. Ma, T. Chen, T. Liu, and D. Zhu, "Exploring Alzheimer's disease: A comprehensive brain connectome-based survey," *Psychoradiology*, vol. 4, Feb. 2024, Art. no. kkad033.
[60] J. Gao et al., "Prediction of cognitive scores by joint use of movie-watching fMRI connectivity and eye tracking via attention-CensNet," *Psychoradiology*, vol. 3, Mar. 2023, Art. no. kkad011.

**Yuzhong Chen** received the B.E. degree from the School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China, in 2020, where he is currently pursuing the master's degree under the supervision of Dr. Xi Jiang.

His current research interests include graph convolutional neural networks, vision transformers, and deep learning-based medical image analysis.

**Zhenxiang Xiao** received the Biomedical Engineering degree from the School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China, in 2021, where he is currently pursuing the master's degree under the supervision of Dr. Xi Jiang's.

His research interests include brain functional network analysis and graph convolutional neural networks.

**Yi Pan** is currently pursuing the bachelor's degree from Glasgow College, University of Electronic Science and Technology of China, Chengdu, China, conducting research under Dr. Xi Jiang and Dr. Tianming Liu's supervision.

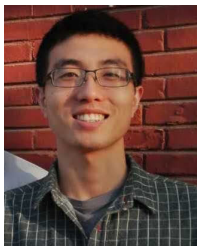
His research interests include brain-inspired artificial intelligence, multimodal biomedical artificial intelligence, and deep learning-based medical image analysis.

**Lin Zhao** received the B.E. degree from Northwestern Polytechnical University, Xi'an, China, in 2017. He is currently pursuing the Ph.D. degree with the School of Computing, University of Georgia, Athens, GA, USA, under the supervision of Prof. Tianming Liu.

His current research interests include deep learning and medical image analysis.

**Haixing Dai** received the B.E. degree from the School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, China, in 2016. He is currently pursuing the Ph.D. degree with the School of Computing, University of Georgia, Athens, GA, USA, under the supervision of Dr. Tianming Liu.

His current research interests include deep learning and medical image analysis.



**Zihao Wu** received the B.E. degree from the School of Microelectronics, Tianjin University, Tianjin, China, in 2017, and the master's degree from the Electrical Engineering and Computer Science Department, Vanderbilt University, Nashville, TN, USA, in 2020. He is currently pursuing the Ph.D. degree in computer science with the University of Georgia, Athens, GA, USA, under the supervision of Dr. Tianming Liu.

His current research interests include brain-inspired AI and deep learning-based medical image analysis.

**Changhe Li** received the master's degree in electronic information from Northwestern Polytechnical University, Xi'an, China, in 2023.

His research interests include deep learning, medical image processing, and other directions.

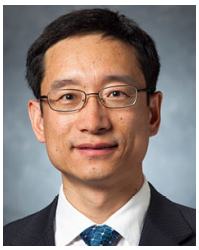
Authorized licensed use limited to: University of Georgia. Downloaded on September 30,2024 at 15:19:01 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



**Tuo Zhang** received the B.E. and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, China, in 2007 and 2015, respectively.

He is currently a Research Associate with Northwestern Polytechnical University. His research interests include machine learning and medical image analysis.



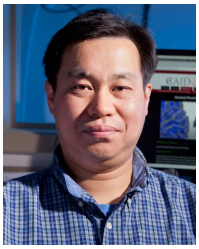
**Changying Li** (Member, IEEE) received the Ph.D. degree in agricultural and biological engineering from The Pennsylvania State University, University Park, PA, USA, in 2006.

He is a Professor and an IFAS AI Administrative Coordinator at the Department of Agricultural and Biological Engineering, University of Florida, Gainesville, FL, USA. He has more than 20 years of experience in developing innovative sensing and robotic technologies by leveraging AI for precision agriculture and automated phenotyping.

Dr. Li is a member of ASABE.

**Dajiang Zhu** received the Ph.D. degree in computer science at the University of Georgia, Athens, GA, USA, in 2014.

His current research interests include machine learning, neuroimaging, and computational neuroscience.



**Tianming Liu** (Senior Member, IEEE) is a Distinguished Research Professor of computer science at the University of Georgia, Athens, GA, USA. His research interests include brain imaging, computational neuroscience, and brain-inspired artificial intelligence. He has published more than 600 articles in these areas, his Google Scholar citations are more than 15 000, and his H-index is 63.

Dr. Liu is an Elected Fellow of American Institute for Medical and Biological Engineering. He is a recipient of the NIH Career Award and the NSF CAREER Award. He serves on the editorial boards of multiple journals, including *Medical Image Analysis*, IEEE TRANSACTIONS ON MEDICAL IMAGING, IEEE REVIEWS IN BIOMEDICAL ENGINEERING, IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, and IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS.

**Xi Jiang** received the B.E. degree in automation from Northwestern Polytechnical University, Xi'an, China, in 2009, and the Ph.D. degree in computer science from the University of Georgia, Athens, GA, USA, in 2016.

He is a Professor with the School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China. His research interests include machine learning/deep learning-based medical image analysis.

Dr. Jiang was a recipient of the Li Foundation Heritage Prize, USA, for "outstanding research and contributions in the interdisciplinary field of brain science" in 2019.

Authorized licensed use limited to: University of Georgia. Downloaded on September 30,2024 at 15:19:01 UTC from IEEE Xplore. Restrictions apply.