Class-specific Prompts in Vision Transformer for Continual Learning of New Diseases

Defeng Zhao[1,3], Zejun Ye[1,3], Wei-Shi Zheng[1,3], and Ruixuan Wang✉[1,2,3]

[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2]Department of Network Intelligence, Peng Cheng Laboratory, Shenzhen, China
[3]Key Laboratory of Machine Intelligence and Advanced Computing, MOE, China

***Abstract*—Current intelligent diagnosis systems are often trained to diagnose a small number of diseases and lack the ability of continually learning new disease knowledge. To have such continual learning ability, the deployed intelligent system needs to be continually updated based on training data of only new diseases, without accessing old data of previously learned diseaseue to privacy concerns and challenges in data sharing across medical centers. In this study, a novel and effective prompt learning strategy is proposed to help a pretrained and fixed Vision Transformer (ViT) continually learn new diseases. In particular, a set of unique prompts for each disease are effectively learned such that discriminative disease features can be well extracted from the fixed ViT under the instructions of prompts during feature extraction, even though the ViT feature extractor is pretrained in the natural image domain. Extensive empirical evaluations on two medical image datasets and one natural image dataset demonstrate the superior performance of the proposed method. The source code is available at https://github.com/zhaoedf/CSPrompt.**

***Index Terms*—Continual learning, Prompt learning, Disease diagnosis**

I. Introduction

Deep learning has been widely applied to intelligent diagnosis of various diseases [8]. It is desired for an intelligent system to diagnose all possible diseases associated with at least one body tissue, organ, or system. However, currently most intelligent medical systems can diagnose just one or a few diseases, although dozens or even hundreds of diseases may exist even for one tissue (e.g., skin) or organ. Considering that it is difficult to collect enough data of all diseases with limited time and resources, one possible solution is to enable an intelligent system to have the lifelong learning ability such that it can continually learn to diagnose more and more diseases over time, as human specialists do. Due to privacy concerns and challenges in data sharing across medical centers, when a deployed intelligent system later tries to continually learn new knowledge with data of new disease(s), old training data of the previously learned diseases are often not available. In this case, catastrophic forgetting of old disease knowledge will probably happen if the intelligent system is updated mainly based on only the data of new diseases [15], [21].

In order to alleviate the catastrophic forgetting issue, multiple strategies have been proposed recently. The regularization-based strategy tries to keep model parameters crucial to old knowledge unchanged when the model continually learns new knowledge [1], [7]. With more old knowledge learned and correspondingly more parameters kept unchanged, it will become more and more difficult for the model to learn new knowledge. Different from regularization-based strategy, the distillation-based strategy involves regularizing the model during the training of new tasks based on prior task knowledge, trying to preserve old knowledge in the updated model by keeping output, particularly at top layer(s) of the model, unchanged compared to the corresponding output of the old model [9], [10]. Since the output of the old model may not faithfully represent old knowledge if the input is only from classes of new knowledge, it turns out that storing a small amount of old data for each old class (i.e., rehearsal-based strategy) can significantly improve the performance of the distillation-based strategy in continual learning of new knowledge [5], [15]. When real data of old knowledge are not available, synthetic old data could be obtained with certain generative models, although continually synthesizing high-fidelity data of more and more old classes is a challenge [14], [16]. However, in numerous practical application scenarios, the enduring retention of training data would lead to violations of data privacy, while increasing the burden on memory cost. All the above strategies change model parameters and therefore old knowledge implicitly stored in model parameters probably will be gradually forgotten over continual learning of more and more new knowledge. Recent attempts have shown that including fixed old feature extractor(s) or extremely using the fixed single feature extractor in the updated model may better preserve old knowledge during continual learning of new knowledge [11], [20]–[22]. However, including more old feature extractors in the model over multiple times of continual learning would quickly expand the model, and simply using the same old feature extractor during model update could largely limit the capability of learning new knowledge.

Inspired by the recently developed prompt learning paradigm in natural language processing [13] and its initial applications in computer vision [19], we propose a novel and effective class-specific prompt learning strategy for class incremental learning under the condition that no data of previously learned diseases is preserved during learning of new diseases. Specifically, a set of visual instructions (i.e., prompts) are learned for each new disease (i.e., class) during continual

Corresponding author: wangruix5@mail.sysu.edu.cn

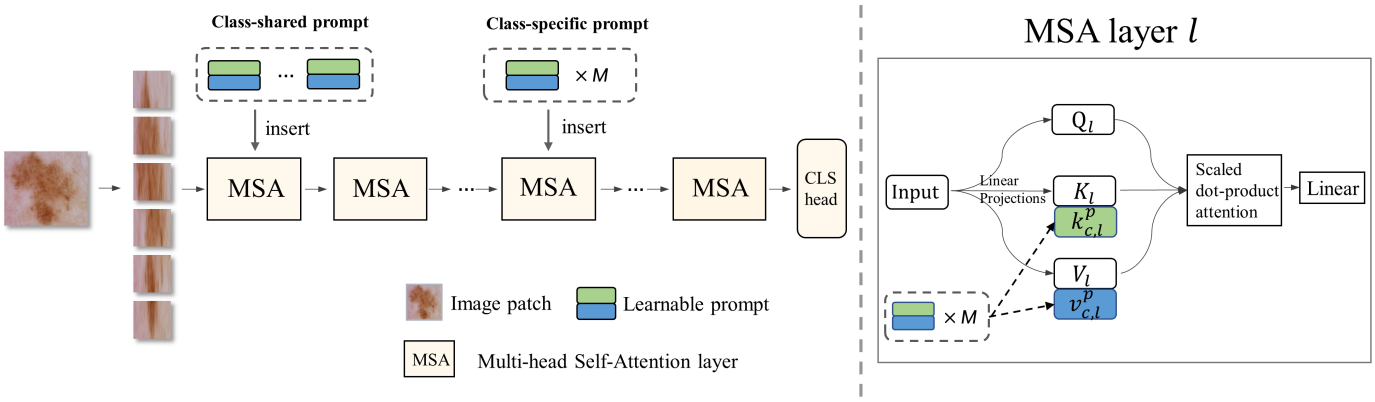

Fig. 1: The proposed prompt-based continual learning framework. Left: learnable class-specific prompts are included at certain intermediate self-attention layers, and learnable class-shared prompts are included at the first one or two self-attention layers. Both types of prompts will help the pretrained and fixed Vision Transformer extract discriminative features. Right: demonstrative interaction between patch tokens and a class-specific prompt (see Eq. 2 and description).

learning, and such prompts are used as additional keys and values at certain self-attention layers to help a pretrained and fixed Vision Transformer more effectively extract visual features of diseases from input medical images. A novel mismatched <input, prompt> design is proposed (Section II-B) to help learn effective class-specific prompts. Extensive evaluations on two medical image datasets and one natural image dataset support that the proposed method outperforms state-of-the-art methods in continual learning of new diseases.

II. Method

This study focuses on class-incremental learning (CIL) of new diseases, i.e., a classifier is continually updated over multiple rounds, with each round learning a certain number of new classes (diseases). Notably, data pertaining to any previously learned class is not preserved, and solely training data relevant to newly introduced classes are available to update the classifier at each round.

A. Prompt learning for CIL

The proposed CIL framework is built on any pretrained and fixed Vision transformer (ViT) backbone [4]. The pretrained ViT is used as a feature extractor, and its output is a feature vector representing the corresponding input image. An expandable classifier head is attached to the ViT, with ViT output as the input to the classifier head. The classifier head can be simply a fully connected layer with a following softmax operator. When a set of new classes are learned at a new learning round, correspondingly a set of new output neurons is added to the classifier head, and the new weight parameters linking the ViT output and the new output neurons will be learned. Since all learned classes share the same ViT feature extractor during CIL and no data of old classes are preserved during CIL, any change in the class-shared ViT parameters when learning new classes would be based on only the new classes' training data and therefore would probably only benefit the new classes but harm old ones. In order to alleviate catastrophic forgetting of old classes' knowledge during continual learning of new classes, the pretrained ViT is fixed across all the learning rounds in the proposed framework. However, in this case, the output from the pretrained and fixed ViT feature extractor may not be discriminative enough between different classes of data, particularly when the ViT is pretrained on the natural image domain (as in this study) and applied to the medical image domain.

In order to resolve this dilemma, a class-specific prompt learning strategy is proposed here (Figure 1). The basic idea is that learnable class-specific prompt(s) is included in certain self-attention layer(s) to instruct the ViT to extract more discriminative features from the input. Each prompt is associated with one specific class and consists of a number of learnable <$key, value$> vector pairs. Formally, denote by $\mathbf{p}_{c,l} = \{\mathbf{k}^p_{c,l,m}, \mathbf{v}^p_{c,l,m}\}_{m=1}^{M}$ the prompt for the c-th class at the l-th self-attention layer, where M is the number of keys $\{\mathbf{k}^p_{c,l,m}\}$ and values $\{\mathbf{v}^p_{c,l,m}\}$, and denote by $\{\mathbf{q}_{l,n}, \mathbf{k}_{l,n}, \mathbf{v}_{l,n}\}_{n=0}^{N}$ the original set of <$query, key, value$> triplet inputs to the l-th self-attention layer, where $n = 0$ corresponds to the special class token and all the others ($n = 1, \ldots, N$) correspond to the sequence of N image patch tokens. Note the triplets $\{\mathbf{q}_{l,n}, \mathbf{k}_{l,n}, \mathbf{v}_{l,n}\}_{n=0}^{N}$ are from the output of the previous self-attention block, while the prompt $\mathbf{p}_{c,l}$ implicitly represents characteristics of the c-th class at the semantic level associated with the l-th self-attention layer and is part of model parameters which will be learned during continual learning. In ViT, the conventional self-attention function is defined as

$$\mathbf{z}_{l,n} = \mathbf{V}_l \, \sigma\left(\frac{\mathbf{K}_l^{\mathrm{T}} \mathbf{q}_{l,n}}{\sqrt{d}}\right) \tag{1}$$

where $\mathbf{z}_{l,n}$ is the self-attention output for the n-th token at the l-th layer, and matrices $\mathbf{V}_l = [\mathbf{v}_{l,0}\ \mathbf{v}_{l,1}\ \ldots\ \mathbf{v}_{l,N}]$ and $\mathbf{K}_l = [\mathbf{k}_{l,0}\ \mathbf{k}_{l,1}\ \ldots\ \mathbf{k}_{l,N}]$ are respectively the collection of all value and key vectors of the $N+1$ input tokens at the l-th layer. $\sigma(\cdot)$ is a softmax-based normalization function to ensure that the degree of contributions of all $N+1$ values $\{\mathbf{v}_{l,n}\}_{n=0}^{N}$

to the n-th token are summed to 1, and d is the length of key or query vector. Compared to conventional self-attention, the learnable class-specific prompt $\mathbf{p}_{c,l}$ is used to modify the self-attention function as follows

$$\mathbf{z}_{c,l,n} = [\mathbf{V}_l \ \mathbf{V}^p_{c,l}] \ \sigma(\frac{[\mathbf{K}_l \ \mathbf{K}^p_{c,l}]^\top \mathbf{q}_{l,n}}{\sqrt{d}}) \tag{2}$$

where $\mathbf{z}_{c,l,n}$ is the self-attention output for the n-th token at the l-th layer when the prompt of the c-th class is used, and matrices $\mathbf{V}^p_{c,l} = [\mathbf{v}^p_{c,l,1} \ \mathbf{v}^p_{c,l,2} \ \cdots \ \mathbf{v}^p_{c,l,M}]$ and $\mathbf{K}^p_{c,l} = [\mathbf{k}^p_{c,l,1} \ \mathbf{k}^p_{c,l,2} \ \cdots \ \mathbf{k}^p_{c,l,M}]$ are respectively the collection of the M value and key vectors in the prompt $\mathbf{p}_{c,l}$ at the l-th layer. From Equation (2), it can be observed that the to-be-learned class-specific visual characteristics ($\{\mathbf{v}^p_{c,l,m}, m = 1, \ldots, M\}$) in the prompt $\mathbf{p}_{c,l}$ will be more or less embedded in each token output $\mathbf{z}_{c,l,n}, n = 0, \ldots, N$, depending on how similar between the query ($\mathbf{q}_{l,n}$) of each token and the keys ($\{\mathbf{k}^p_{c,l,m}, m = 1, \ldots, M\}$) of the prompt. Higher similarity between token query and prompt key(s) leads to more embedding of class-specific visual characteristics into the token output, which in turn would lead to more class-specific information in the output of the ViT feature extractor. On the contrary, if there is little similarity between the token query and prompt key(s), the class-specific prompt would not affect the token output and subsequently the output of the ViT feature extractor. As a result, if an input image contains visual features of a certain class, a prompt of this class at a certain self-attention layer would help the pretrained ViT generate more class-specific feature output, and such output would then help the classifier head more easily recognize the class of the input.

Besides class-specific prompts, class-shared prompt(s) particularly at lower (e.g., the 1st and 2nd) self-attention layer will also be investigated and empirically evaluated, considering its positive effect on continual learning in the previous study [18]. It is speculated that such prompt(s) at the lower layer(s) may help ViT extract class-shared low-level visual features which have been widely observed in the lower layers of convolutional neural networks [18]. Such class-shared prompt(s) could be particularly helpful when the ViT feature extractor is pretrained on the natural image domain and then applied to the medical image domain as in this study. For example, even if certain low-level features of medical images are crucial to discriminate between different medical classes and such features can be extracted by the first one or two layers in the ViT, they could be ignored at higher layers if the ViT considers such low-level features negligible for natural image classification. Therefore, class-shared prompt(s) associated with low-level features in the downstream medical domain may instruct the ViT to strengthen low-level medical image features at lower layer(s), which in turn would likely help the ViT pay more attention to (rather than ignore) such low-level medical features at subsequent layers. In this case, class-shared prompt(s) could improve the transfer learning ability of the pretrained and fixed ViT particularly for low-level features when the downstream task domain is different from the original task domain for ViT pretraining.

B. Optimization and inference

Suppose the t-th learning round in CIL contains C_t classes, and the CIL model will be updated to learn all the new C_t classes of knowledge. Let $\mathcal{C}_t = \{r_{t-1} + 1, \ldots, r_{t-1} + C_t\}$ denote the set of new class indices, where r_{t-1} is the total number of learned old classes before the t-th round, and $\mathcal{D}_c$ denote the training set for any class $c \in \mathcal{C}_t$. During model update, weight parameters connecting to the new output neurons in the classifier head, class-specific prompt(s) at certain self-attention layer(s) for each new class, and class-shared prompt(s) at the first one or two self-attention layers will be optimized. Additional effort needs to be made to effectively optimize class-specific prompts. For an input image $\mathbf{x}$ from any class $c \in \mathcal{C}_t$, the class-specific prompt(s) of class c can be used to form a *matched* <input, prompt> pair, and the class-specific prompt(s) of any different class c' can be used to form a *mismatched* <input, prompt> pair. If just matched <input, prompt> pairs are used during model update, there exists a risk of making the model lazy in the sense that the model may simply use class-specific prompt(s) to predict the class of input image, regardless of any visual information in the input. In order to make the model use both input and prompt information for prediction, we propose additionally using mismatched <input, prompt> pairs during model update. In particular, for any mismatched <input, prompt> pair, the model is enforced to correctly predict the class of the input as well. In this case, the mismatched class-specific prompt would not provide information about the class of input, and therefore the model has to learn to use visual information in the input for accurate prediction. Based on the above analysis, model update can be achieved by minimizing the expanded cross-entropy loss $\mathcal{L}$,

$$\mathcal{L} = \mathbb{E}_{\{c,c'\} \in \mathcal{C}_t, c \neq c', \mathbf{x} \in \mathcal{D}_c} \left[-\log \hat{y}_{c,c} - \log \hat{y}_{c,c'} \right], \tag{3}$$

where $\hat{y}_{c,c'}$ is the model output associated with class c when the class-specific prompt(s) are from class c'. Note that since there is no data of old classes involved in model update, and weight parameters of old classes in the classifier head and old class-specific prompts are fixed during model update, $\hat{y}_{c,c'}$ is obtained from the softmax operator over only the logits of the C_t new output neurons. In Eq. (3), the first cross-entropy loss term ($-\log \hat{y}_{c,c}$) corresponds to matched <input, prompt> pairs, while the second cross-entropy loss term ($-\log \hat{y}_{c,c'}$) corresponds to mismatched <input, prompt> pairs.

Once the model finishes the t-th round of continual learning, it can be used to predict any test input as one of all r_t (i.e., $r_{t-1} + C_t$) learned classes so far. Specifically, given any test input, the learned class-specific prompt(s) of every class $c \in \{1, \ldots, r_t\}$ is respectively paired with the input and the prediction probability of class c is collected from the output

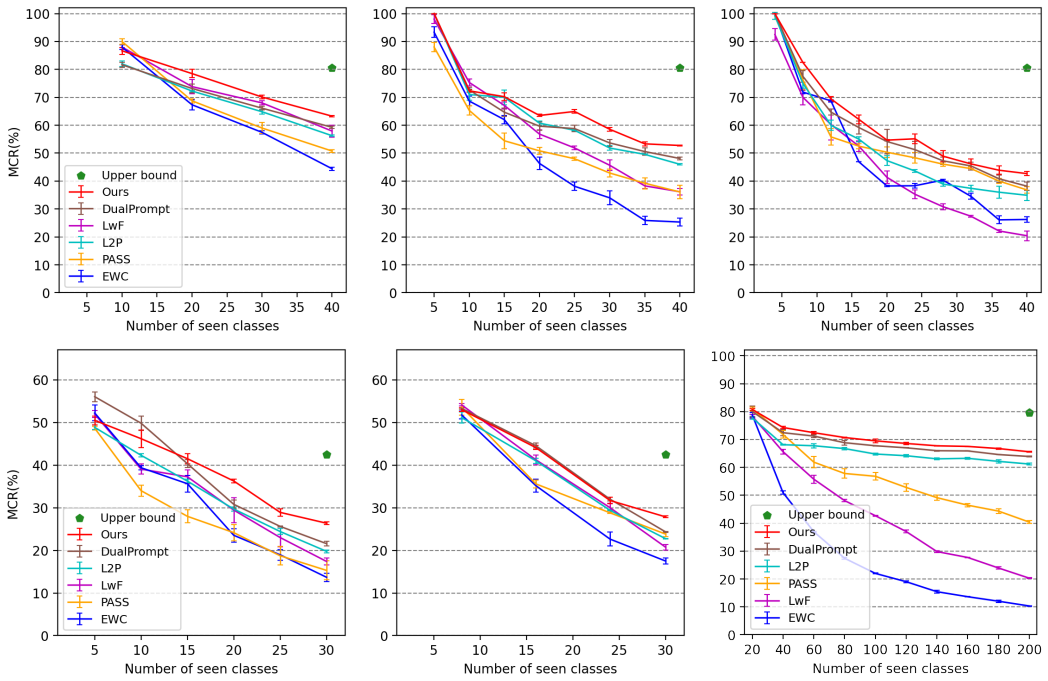

Fig. 2: Performance comparison between the proposed method and baselines on Skin40 (first row), TCGA30 (2nd row, first two), and ImageNet-R (2nd row, last). The X-axis in each figure represents the accumulated number of learned classes. The upper bound result (pentagon) is from the trained ViT with all classes of training data. All methods use a pretrained ViT-B/16.

of the unified classifier head. The class c^* with maximum prediction probability is selected as the predicted class, i.e.,

$$c^* = \max_{c \in \{1,\ldots,r_t\}} \hat{y}_{c,c}\,. \tag{4}$$

C. Comparison with relevant study

Our method is largely inspired by the recently developed DualPrompt learning for CIL [18], but with significant differences. First, DualPrompt proposes task-specific prompts (i.e., multiple classes learned at each round share the same prompts), while ours proposes class-specific prompts. Second, DualPrompt learns to match each input with one of the task-specific prompts based on a query function, while there is no such complicated learning and matching process in our method. Third, mismatched <input, prompt> pairs are novelly designed and applied to learn class-specific prompts in our method, which makes the optimization simpler and easier and leads to more effective prompts for CIL. Fourth, during inference, DualPrompt selects only one task's specific prompts for class prediction, while ours predicts based on each class' prompts and selects the optimal class. Overall, ours has a different learning and inference process and is simpler and more effective than DualPrompt.

III. Experiments

A. Experimental setup

The proposed CIL method was evaluated on two medical image datasets Skin40 [12], [22] and TCGA30, and one natural image dataset ImageNet-R [6] (Table I). The 40 data-balanced classes with relatively more number of images in the SD198 skin disease dataset [17] were selected to form the Skin40. TCGA30 is a set of histopathology image patches that were sampled from slides of 30 cancers in The Genome Cancer Atlas (TCGA). For each class, we randomly sampled training and test slides at the patient level. After that, we sampled 800 patches from each class of training slides and 200 patches from each class of test slides at a $10\times$ magnification (1.00 μm/per pixel). All training images in Skin40 and TCGA30 were randomly cropped with the scale range $[0.3, 1.0]$ and then resized to 224×224 pixels, followed by a random horizontal flipping. ImageNet-R [6] contains 200 classes of artistic renditions derived from the original ImageNet dataset [3]. Note that the classes in ImageNet-R are considered non-overlapped with those in ImageNet [18], [19]. A validation set was created by sampling 20% of the ImageNet-R training set for hyper-parameter selection (e.g., layers to include prompts). All training images in ImageNet-R were first resized to 256×256 pixels and then randomly cropped to 224×224 pixels, followed by a random horizontal flipping.

TABLE I: Statistics of three datasets. '37~345': the range of the number of images of each class. [50, 7016] represents the range of image height and width.

Datasets	#Classes	Training images per class	Test images per class	Image size
Skin40 [17]	40	50	10	[420, 1640]
TCGA30 [2]	30	800	200	256×256
ImageNet-R [6]	200	37~345	4~88	[50, 7016]

Skin40 and ImageNet-R were split into multiple (e.g., 8, 10) subsets of non-overlapped classes, with each subset corresponding to one learning round. These subsets were randomly ordered and then fixed for evaluation of the proposed method and baselines. The ViT feature extractor was pretrained on ImageNet [3] and its parameters were fixed in continual learning. One fully connected layer followed by the softmax was used as the classifier head. Following DualPrompt [18] and checked with the ImageNet-R validation set, class-shared prompts were included in the first two self-attention layers, and class-specific prompts were included in the 5-th and 7-th self-attention layers. By default, M was empirically set to 30 for each class-specific prompt and 5 for each class-shared prompt. When learning a set of new classes at each round, AdamW optimizer was adopted, with initial learning rate 0.01, weight decay coefficient 0.05, and batch size 64. The classifier was trained for up to 50 epochs, with consistent training convergence observed. Considering possible data imbalance among classes, mean class recall (MCR) over all learned classes so far after each round of continual learning was used as the performance measure. Note that MCR is equivalent to classification accuracy on a balanced dataset. For each experiment, three runs were performed and the mean and standard deviation of MCR were reported.

B. Quantitative evaluation

Effectiveness study: The effectiveness of our method was evaluated by comparing it with representative CIL methods, including LwF [10], EWC [7], PASS [23], L2P [19] and the current SOTA of prompt-based method DualPrompt [18]. For fair comparisons, no old data were used in these methods during CIL, and all methods used the same ViT-B/16 [4] backbone following previous work. A similar amount of effort was taken to tune each baseline. Figure 2 shows that our method outperforms all the baselines when the classifier continually learns either 4, 8, or 10 rounds. On Skin40, the primary medical image dataset, our method achieves significant performance gains. Specifically, under the 4-round, 8-round, and 10-round settings (i.e., first row, left to right), our method attained MCR of 63.25%, 52.75%, and 42.67% at the final round. Compared to the current strongest baseline DualPrompt, the improvements from our method are around 4% to 4.67%. This supports that our prompt learning approach can effectively mitigate catastrophic forgetting in continual learning of new disease categories. On another medical dataset TCGA30 and a natural image dataset ImageNet-R, our method similarly achieves superior performance, consistently surpassing all baseline methods. This further verifies the generalization capability and reliability of our method.

Ablation study: To check the effect of the key components in our method, an ablation study was performed. As shown in Table II, while solely adding class-shared prompts (second row) slightly improves the continual learning performance compared to the baseline (first row, only fine-tuning the classifier head with the fixed ViT feature extractor), additional inclusion of the class-specific prompts (third and fourth rows)

TABLE II: Ablation study on Skin40, with 5 classes learned in each round.

Class-shared prompts		✓		✓		✓
Class-specific prompts			✓	✓	✓	✓
Mismatched pairs					✓	✓
MCR (%)	47.50	48.15	37.75	38.00	49.25	**52.75**

unexpectedly downgrades the performance. This is probably because the model found a lazy way to predict class of training data, i.e., based on only the class-specific prompt information. In contrast, combining the mismatched <input, prompt> pairs with the class-specific prompts (fifth row) results in a better performance than the baseline, and additional inclusion of the class-shared prompt further improves the performance (last row). Note that class-specific prompts are involved in the mismatched <input, prompt> pairs. Therefore, the last two rows in Table II clearly support both class-specific prompts and mismatched <input, prompt> pairs are crucial in helping continual learning.

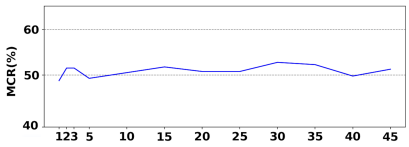

Fig. 3: Effect of the number (M) of keys and values in class-specific prompt on continual learning. Skin40 was used here, with 5 classes learned in each learning round.

TABLE III: Sensitivity of the insert location of the proposed class-specific prompt. The experiment was performed on Skin40 with 5 classes learned in each round. '{a,b}': class-specific prompt is inserted at a-th and b-th layers. The bold location $\mathbf{\{5,7\}}$ is determined by a grid search experiment on ImageNet-R validation set.

	Continuous layers						Non-continuous layers				Others	
Location	{3,4}	{4,5}	{5,6}	{6,7}	{7,8}	{8,9}	{3,5}	{4,6}	**{5,7}**	{6,8}	{8}	{8,9,10,11}
MCR (%)	49.75	51.00	51.25	50.50	50.75	52.00	49.50	51.00	**52.75**	50.25	52.25	50.25

Sensitivity study: In our method, each class-specific prompts consists of multiple (M) keys and values. By varying M from 1 to 40, the performance of our method changes in a relatively small range (Figure 3) and is always better than the best baseline, suggesting that our method is largely insensitive to the choice of element number in a prompt. Besides, another sensitivity study was performed to investigate the effect of the layer position of class-specific prompt on model performance. Regardless of prompt insertion between contiguous layers, non-contiguous layers, solitary layers, or multiple continuous layers, the variation in MCR fluctuated within a tight range of 49.50% to 52.75% (Table III), exhibiting stability of class-specific prompt in improving model performance. This indicates that our method is largely insensitive to prompt

TABLE IV: Comparison with Dualprompt on ImageNet-R and Skin40 with three different model sizes. The best performance is highlighted in **bold**.

	ImageNet-R			Skin40		
Backbone	ViT-S/16	ViT-B/16	ViT-L/16	ViT-S/16	ViT-B/16	ViT-L/16
DualPrompt [18]	57.25 ± 0.14	63.89 ± 0.20	$\mathbf{70.92 \pm 0.37}$	41.13 ± 2.4	48.08 ± 0.42	$\mathbf{47.30 \pm 0.80}$
Ours	$\mathbf{58.22 \pm 0.19}$	$\mathbf{65.58 \pm 0.21}$	68.36 ± 0.28	$\mathbf{44.65 \pm 0.92}$	$\mathbf{52.75 \pm 0.20}$	44.77 ± 0.68

insertion location. Furthermore, as can be seen from Table IV, our method consistently outperforms DualPrompt on both ViT-S/16 and ViT-B/16 backbones [4], further exhibiting its superiority. Our method falls slightly behind DualPrompt when using the larger ViT-L/16 model. One possible reason is that the prompts may not be optimal to guide feature learning in the very deep ViT-L/16 model with the prompts inserted in the first few layers.

IV. Conclusion

In this study, we propose a novel prompt learning strategy for class-incremental learning. Under the realistic condition of not storing any old data, our method outperforms state-of-the-art methods on multiple datasets. Prompt learning provides a new way to make use of well pretrained models for continual learning, and the extension of the proposed method to lesion detection and segmentation will be explored in future work. Pretraining the ViT using relevant medical image data may even further boost the performance, which will also be investigated in future work.

Acknowledgments. This work is supported in part by the Major Key Project of PCL (grant No. PCL2023AS7-1), the National Natural Science Foundation of China (grant No. 62071502), and Guangdong Excellent Youth Team Program (grant No. 2023B1515040025).

References

[1] A. Chaudhry, P. K. Dokania, T. Ajanthan, and P. H. Torr, "Riemannian walk for incremental learning: understanding forgetting and intransigence," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 532–547.

[2] R. J. Chen, C. Chen, Y. Li, T. Y. Chen, A. D. Trister, R. G. Krishnan, and F. Mahmood, "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 144–16 155.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: a large-scale hierarchical image database," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[5] A. Douillard, M. Cord, C. Ollion, T. Robert, and E. Valle, "Podnet: pooled outputs distillation for small-tasks incremental learning," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 86–102.

[6] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer, "The many faces of robustness: a critical analysis of out-of-distribution generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8320–8329.

[7] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabisa, C. Clopathb, D. Kumarana, and R. Hadsell, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[8] Y. Kumar, A. Koul, R. Singla, and M. F. Ijaz, "Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–28, 2022.

[9] K. Lee, K. Lee, J. Shin, and H. Lee, "Overcoming catastrophic forgetting with unlabeled data in the wild," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 312–321.

[10] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[11] Z. Li, C. Zhong, S. Liu, R. Wang, and W.-S. Zheng, "Preserving earlier knowledge in continual learning with the help of all previous feature extractors," *arXiv preprint arXiv:2104.13614*, 2021.

[12] Z. Li, C. Zhong, R. Wang, and W.-S. Zheng, "Continual learning of new diseases with dual distillation and ensemble strategy," in *Medical Image Computing and Computer Assisted Intervention*, 2020, pp. 169–178.

[13] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[14] O. Ostapenko, M. Puscas, T. Klein, P. Jahnichen, and M. Nabi, "Learning to remember: a synaptic plasticity driven framework for continual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 321–11 329.

[15] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "iCaRL: incremental Classifier and Representation Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2001–2010.

[16] H. Shin, J. K. Lee, J. Kim, and J. Kim, "Continual learning with deep generative replay," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[17] X. Sun, J. Yang, M. Sun, and K. Wang, "A benchmark for automatic visual classification of clinical skin disease images," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 206–222.

[18] Z. Wang, Z. Zhang, S. Ebrahimi, R. Sun, H. Zhang, C.-Y. Lee, X. Ren, G. Su, V. Perot, J. Dy, and T. Pfister, "Dualprompt: complementary prompting for rehearsal-free continual learning," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 631–648.

[19] Z. Wang, Z. Zhang, C.-Y. Lee, H. Zhang, R. Sun, X. Ren, G. Su, V. Perot, J. Dy, and T. Pfister, "Learning to prompt for continual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 139–149.

[20] Z. Xu, K. Chen, W.-S. Zheng, Z. Tan, X. Yang, and R. Wang, "Expert with outlier exposure for continual learning of new diseases," in *IEEE International Conference on Bioinformatics and Biomedicine*, 2022, pp. 1768–1772.

[21] S. Yan, J. Xie, and X. He, "Der: dynamically expandable representation for class incremental learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3014–3023.

[22] Y. Yang, Z. Cui, J. Xu, C. Zhong, W.-S. Zheng, and R. Wang, "Continual learning with bayesian model based on a fixed pre-trained feature extractor," *Visual Intelligence*, vol. 1, no. 1, p. 5, 2023.

[23] F. Zhu, X.-Y. Zhang, C. Wang, F. Yin, and C.-L. Liu, "Prototype augmentation and self-supervision for incremental learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5871–5880.