Vision Transformer for Small-Size Datasets

Seung Hoon Lee
Inha University
Incheon, South Korea
aanna0701@gmail.com

Seunghyun Lee
Inha University
Incheon, South Korea
lsh910703@gmail.com

Byung Cheol Song
Inha University
Incheon, South Korea
bcsong@inha.ac.kr

Abstract

Recently, the Vision Transformer (ViT), which applied the transformer structure to the image classification task, has outperformed convolutional neural networks. However, the high performance of the ViT results from pre-training using a large-size dataset such as JFT-300M, and its dependence on a large dataset is interpreted as due to low locality inductive bias. This paper proposes Shifted Patch Tokenization (SPT) and Locality Self-Attention (LSA), which effectively solve the lack of locality inductive bias and enable it to learn from scratch even on small-size datasets. Moreover, SPT and LSA are generic and effective add-on modules that are easily applicable to various ViTs. Experimental results show that when both SPT and LSA were applied to the ViTs, the performance improved by an average of 2.96% *in Tiny-ImageNet, which is a representative small-size dataset. Especially, Swin Transformer achieved an overwhelming performance improvement of* 4.08% *thanks to the proposed SPT and LSA.*

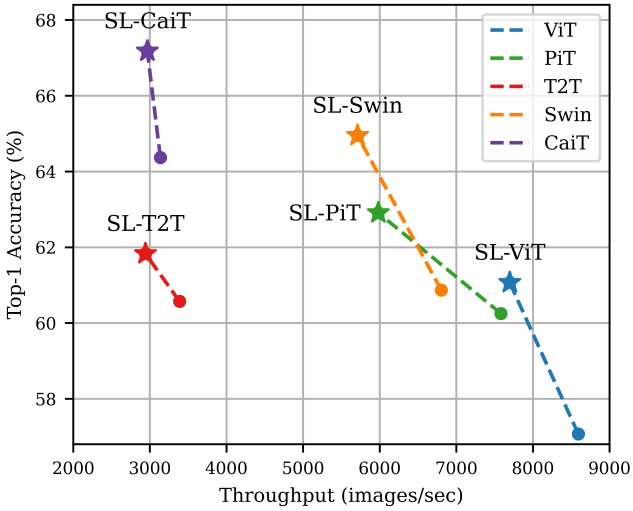


Figure 1. Effect of the proposed method on the overall performance when learning Tiny-ImageNet from scratch. Throughput refers to how many images can be processed per unit of time. The stars and dots indicate after and before the proposed method are applied, respectively.

1. INTRODUCTION

Convolutional neural networks (CNNs), which are effective in learning visual representations of image data, have been the main-stream in the field of computer vision (CV) [10, 14, 18, 30, 32, 37]. Meanwhile, in the field of Natural Language Processing (NLP), the so-called Transformer [35] based on self-attention mechanism has achieved tremendous success [5,8,20]. So, in the CV field, there have been attempts to combine the self-attention mechanism with CNNs [4,13,28,36,38,44]. These studies have succeeded in proving that the self-attention mechanism also works for the image domain. Recently, it was reported that Vision Transformer (ViT) [9], which applied a standard Transformer composed entirely of self-attention to image data, showed better performance than ResNet [10] and EfficientNet [32] in the image classification task. This made Transformer receive a lot of attention in the CV field.

ViT rarely uses convolutional filters, i.e., the core of CNNs. Convolutional filters were usually used only for their tokenization. Thus, ViT structurally lacks locality inductive bias than CNNs, and they require a too large amount of training data to obtain acceptable visual representation [26]. For example, just to learn a small-size dataset, ViT had to precede pre-training on a large-size dataset such as JFT-300M [29]. In order to alleviate the burden of pre-training, several ViTs which can learn a mid-size dataset such as ImageNet from scratch have been proposed. Such data-efficient ViTs tried to increase the locality inductive bias in terms of network architecture. For example, some adopted a hierarchical structure like CNNs to leverage various receptive fields [12, 24, 39], and the others tried to modify the self-attention mechanism itself [22,24,34,39,40]. However, learning from scratch on mid-size datasets still requires significant costs. Moreover, learning small-size datasets from

CNN认为信息具有空间局部性(Locality)可用滑动卷积共享权重降低参数空间

scratch is very challenging considering the trade-off between dataset capacity and performance. Therefore, we need to study ViT that can learn small-size datasets by sufficiently increasing the locality inductive bias.

Through observations, we found two problems that decrease locality inductive bias and limit the performance of the ViT. The first problem is poor tokenization. ViT divides a given image into non-overlapping patches of equal size, and linearly projects each patch to a visual token. Here, the same linear projection is applied to each patch. So, tokenization of the ViT has the permutation invariant property, which enables a good embedding of relations between patches [3]. On the other hand, non-overlapping patches allow visual tokens to have a relatively small receptive field. Usually, tokenization based on non-overlapping patches has a smaller receptive field than tokenization based on overlapping patches with the same down-sampling ratio. Small receptive fields cause ViT to tokenize with too few pixels. As a result, the spatial relationship with adjacent pixels is not sufficiently embedded in each visual token. The second problem is the poor attention mechanism. The feature dimension of image data is far greater than that of natural language and audio signal, so the number of embedded tokens is inevitably large. Thus, the distribution of attention scores of tokens becomes smooth. In other words, we face the problem that ViTs cannot attend locally to important visual tokens. The above two main problems cause highly redundant attentions that cannot focus on a target class. This redundant attention makes it easy for ViT to normally concentrate on the background and not capture the shape of the target class well (see Fig. 5).

This paper presents two solutions to effectively improve the locality inductive bias of ViT for learning small-size datasets from scratch. First, we propose Shifted Patch Tokenization (SPT) to further utilize spatial relations between neighboring pixels in the tokenization process. The idea of SPT was derived from Temporal Shift Module (TSM) [23]. TSM is effective temporal modeling which shifts some temporal channels of features. Inspired by this, we propose effective spatial modeling that tokenizes spatially shifted images together with the input image. SPT can give a wider receptive field to ViT than standard tokenization. This has the effect of increasing the locality inductive bias by embedding more spatial information in each visual token. Second, we propose Locality Self-Attention (LSA), which allows ViT to attend locally. LSA mitigates the smoothing phenomenon of attention score distribution by excluding self-tokens and by applying learnable temperature to the softmax function. LSA induces attention to work locally by forcing each token to focus more on tokens with large relation to itself. Note that the proposed SPT and LSA can be easily applied to various ViTs in the form of add-on modules without structural changes and can effectively improve performance (see Fig. 1 and Table 5).

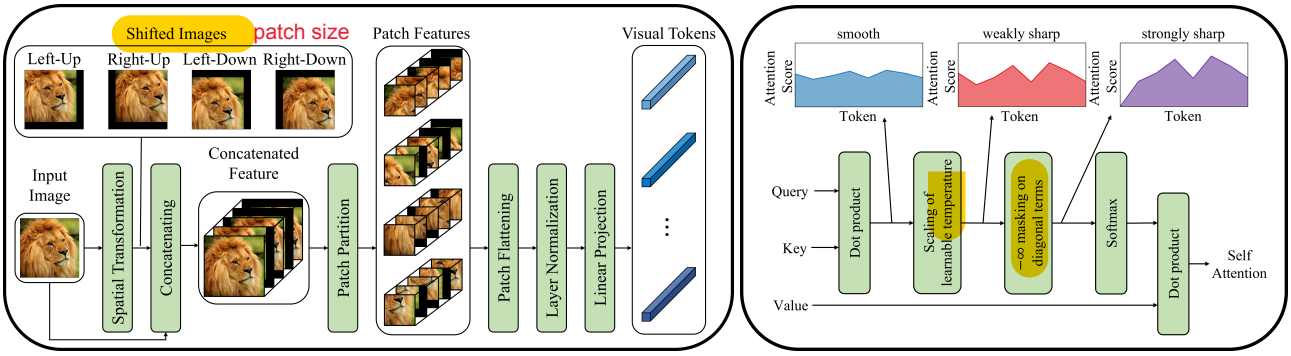
Our experiments show that the proposed method improves the performance of various ViTs both qualitatively and quantitatively. First, Fig. 5 illustrates that when SPT and LSA are applied to the ViTs, object shapes are better captured. From a quantitative aspect, SPT and LSA improve image classification performance. For example, in the experiment on Tiny-ImageNet, the classification accuracy is improved by an average of 2.96%, and a maximum of 4.08% (see Table 2). Also, SPT and LSA improve the performance of ViTs up to 1.06% in the mid-size dataset such as ImageNet (see Table 3). The main contribution points of this paper are as follows:

- To sufficiently embed spatial information between neighboring pixels, we propose new tokenization based on spatial feature shifting. The proposed tokenization can give a wider receptive field to visual tokens. This dramatically improves the performance of the ViTs.
- We propose a locality attention mechanism to solve or attenuate the smoothing problem of the attention score distribution. This mechanism significantly improves the performance of ViTs with only a small parameter increase and the addition of simple operations.

2. RELATED WORK

Recently, several data-efficient ViTs have been proposed to alleviate the dependence of ViT on large-size datasets. These ViTs can learn mid-size datasets from scratch. For example, DeiT [33] improved the efficiency of ViTs by employing data augmentations and regularizations and realized knowledge distillation by introducing the distillation token concept. T2T [41] used a tokenization method that flattened overlapping patches and applied a transformer. This makes it possible to learn local structure information around a token. PiT [12] produced various receptive fields through spatial dimension reduction based on the pooling structure of a convolutional layer. CvT [39] replaced both linear projection and multi-layer perceptron with convolutional layers. Also, like PiT, CvT generated various receptive fields only with a convolutional layer. Swin Transformer [24] presented an efficient hierarchical transformer that gradually reduces the number of tokens through patch merging while using attention calculated in non-overlapping local windows. CaiT [34] employed LayerScale, which converges well even in training ViTs with a large depth. In addition, the transformer layer of the CaiT is divided into a patch-attention layer and a class-attention layer, which is effective for class embedding.

However, the ViT for small-size datasets has not been reported yet. Therefore, this paper proposes tokenization using more spatial information and also a high-performance attention mechanism, which allows ViTs to effectively learn



(a) Shifted Patch Tokenization (b) Locality Self-Attention

Figure 2. Architectures of the proposed SPT and LSA.

small-size datasets from scratch.

3. PROPOSED METHOD

This section specifically describes two key ideas for increasing the locality inductive bias of ViTs: SPT and LSA. First, Fig. 2(a) depicts the concept of SPT. SPT spatially shifts an input image in several directions and concatenates them with the input image. Fig. 2(a) is an example of shifting in four diagonal directions. Next, patch partitioning is applied like standard ViTs. Then, for embedding into visual tokens, three processes are sequentially performed: patch flattening, layer normalization [2], and linear projection. As a result, SPT can embed more spatial information into visual tokens and increase the locality inductive bias of ViTs.

Fig. 2(b) explains the second idea, LSA. In general, a softmax function can control the smoothness of the output distribution through temperature scaling [11]. LSA primarily sharpens the distribution of attention scores by learning the temperature parameters of the softmax function. Additionally, the self-token relation is removed by applying the so-called diagonal masking, which forcibly suppresses the diagonal components of the similarity matrix computed by Query and Key. This masking relatively increases the attention scores between different tokens, making the distribution of attention scores sharper. As a result, LSA increases the locality inductive bias by making ViT's attention locally focused.

3.1. Preliminary

Before a detailed description of the proposed SPT and LSA, this section briefly reviews the tokenization and formulation of the self-attention mechanism of standard ViT [9].

Let $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ be an input image. Here, H, W, and C indicate the height, width, and channel of the image, respectively. First, ViT divides the input image into non-overlapping patches and flatten the patches to obtain a sequence of vectors. This process is formulated as Eq. 1:

$$\mathcal{P}(\mathbf{x}) = [\mathbf{x}_p^1; \mathbf{x}_p^2; \ldots; \mathbf{x}_p^N] \tag{1}$$

where $\mathbf{x}_p^i \in \mathbb{R}^{P^2\cdot C}$ represents the i-th flattened vector. P and $N = HW/P^2$ stand for the patch size and the number of patches, respectively.

Next, we obtain patch embeddings by linearly projecting each vector into the space of the hidden dimension of the transformer encoder. Each patch embedding corresponds to a visual token input to the transformer encoder, so this series of processes is called tokenization, i.e., $\mathcal{T}$. This is defined by:

$$\mathcal{T}(\mathbf{x}) = \mathcal{P}(\mathbf{x})\boldsymbol{E}_t \tag{2}$$

where $\boldsymbol{E}_t \in \mathbb{R}^{(P^2\cdot C)\times d}$ is the learnable linear projection for tokens, and d is the hidden dimension of the transformer encoder.

Note that the receptive fields of visual tokens in ViT are determined by tokenization. In the transformer encoder running after the tokenization step, the number of visual tokens does not change, so the receptive field cannot be adjusted there. and the tokenization (Eq. 2) of standard ViT is the same as the operation of the non-overlapping convolutional layer with the same size of kernel and stride. So, the receptive field size of visual tokens can be calculated by the following equation given in [1]:

RF(n-1)= (RF(n)-1) * s + k

$$r_{token} = r_{trans} \cdot j + (k - j) \tag{3}$$

where r_{token} and r_{trans} stand for the receptive field sizes of tokenization and transformer encoder, respectively. j and k are the stride and kernel size of the convolutional layer, respectively. As mentioned earlier, the receptive field is not adjusted in the transformer encoder, so $r_{trans} = 1$. Thus, r_{token} is the same as the kernel size. Here, the kernel size is the patch size of ViT. r_token = k = patch_size

At this time, let's investigate whether r_{token} is of sufficient size. For instance, we compare r_{token} with the receptive field size of the last feature of ResNet50 when training on the ImageNet dataset consisting of images of 224×224. The patch size of standard ViT is 16, so r_{token} of visual tokens is also 16. On the other hand, the receptive field size of the ResNet50 feature amounts to 483 [1]. As a result, the visual tokens of ViTs have a receptive field size that is about 30 times smaller than that of the ResNet50 feature. We interpret this small receptive field of tokenization as a major factor in the lack of local inductive bias. Therefore, Sec. 3.2 proposes the SPT to leverage rich spatial information by increasing the receptive field of tokenization.

Meanwhile, the self-attention mechanism of general ViTs operates as follows. First, a learnable linear projection is applied to each token to obtain Query, Key, and Value. Next, calculate the similarity matrix, that is, $\mathrm{R} \in \mathbb{R}^{(N+1)\times(N+1)}$, indicating the semantic relation between tokens through the dot product operation of Query and Key. The diagonal components of R represent self-token relations, and the off-diagonal components represent inter-token relations:

$$\mathrm{R}(\mathbf{x}) = \mathbf{x}\boldsymbol{E}_q(\mathbf{x}\boldsymbol{E}_k)^\intercal \tag{4}$$

Here, $\boldsymbol{E}_q \in \mathbb{R}^{d\times d_q}$, $\boldsymbol{E}_k \in \mathbb{R}^{d\times d_k}$ indicate learnable linear projections for Query and Key, respectively. And, d_q and d_k are the dimensions of Query and Key, respectively. Next, R is divided by the square root of the Key dimension, and then the softmax function is applied to obtain the attention score matrix. Finally, calculate the self-attention, defined by the dot product of the attention score matrix and Value, as in Eq. 5:

$$\mathrm{SA}(\mathbf{x}) = \mathrm{softmax}(\mathrm{R}/\sqrt{d_k})\mathbf{x}\boldsymbol{E}_v \tag{5}$$

where $\boldsymbol{E}_v \in \mathbb{R}^{d\times d_v}$ is a learnable linear projection of Value, and d_v is the Value dimension.

Eq. 5 was designed so that the attentions of tokens with large relations get large. However, due to the following two causes, attentions of standard ViT tend to be similar to each other regardless of relations. The first cause is as follows: Since Query ($\mathbf{x}\boldsymbol{E}_q$) and Key ($\mathbf{x}\boldsymbol{E}_k$) is linearly projected from the same input tokens, token vectors belonging to Query and Key tend to have similar sizes. Eq. 4 shows that R is the dot product of Query and Key. So, self-token relations which are dot products of similar vectors are usually larger than inter-token relations. Therefore, the softmax function of Eq. 5 gives relatively high scores to self-token relations and small scores to inter-token relations. The second cause is as follows: The reason why R is divided by $\sqrt{d_k}$ in Eq. 5 is to prevent the softmax function from having a small gradient. However, $\sqrt{d_k}$ can rather act as a high temperature of the softmax function and cause smoothing of the attention score distribution [11]. Our experiment proves

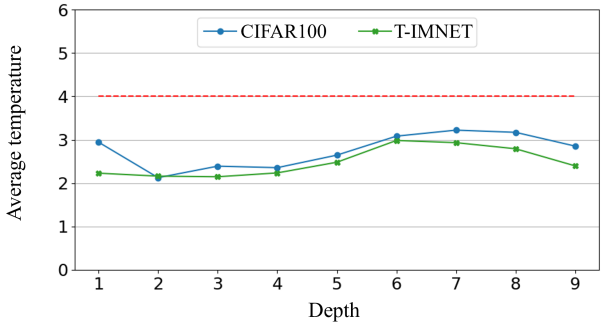


Figure 3. The learned temperature according to depth. Here, the red dashed line indicates the temperature of standard ViT.

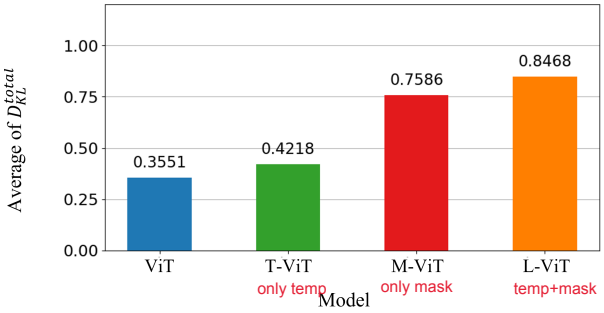


Figure 4. Kullback–Leibler Divergence (KLD) of attention score distributions. The average KLDs were measured on Tiny-ImageNet.

that the attention scores smoothed due to high temperature degrade the performance of ViT. For example, take a look at Table 1 that shows the top-1 accuracy of standard ViT on the small-size datasets, i.e., CIFAR100 and Tiny-ImageNet. Here, we can observe the best performance when the temperature of softmax is less than $\sqrt{d_k}$. Sec. 3.3 proposes the LSA for improving the performance of ViTs by solving the smoothing problem of the attention score distribution.

3.2. Shifted Patch Tokenization

This section first describes the overall formulation of SPT (Sec. 3.2.1) and applies the proposed SPT to the patch embedding layer and the pooling layer, i.e., two main tokenizations for ViTs (Sec. 3.2.2 and Sec. 3.2.3).

3.2.1 Formulation

First, each input image is spatially shifted by half the patch size in four diagonal directions, that is, left-up, right-up, left-down, and right-down. In this paper, this shifting strategy is named $\mathcal{S}$ for convenience, and the SPT of all experiments follows $\mathcal{S}$. Of course, various shifting strategies other than $\mathcal{S}$ are available, and they are dealt with in the

supplementary. Next, the shifted features are cropped to the same size as the input image and then concatenated with the input. Then, the concatenated features are divided into non-overlapping patches and the patches are flattened as in Eq. 1. Next, visual tokens are obtained through layer normalization (LN) and linear projection. The whole process is formulated as Eq. 6:

$$\mathrm{S}(\mathbf{x}) = \mathrm{LN}(\mathcal{P}([\mathbf{x}\,\mathbf{s}^1\,\mathbf{s}^2\,\ldots\,\mathbf{s}^{N_\mathcal{S}}]))\boldsymbol{E}_\mathcal{S} \tag{6}$$

Here, $\mathbf{s}^i \in \mathbb{R}^{H\times W\times C}$ represents the i-th shifted image according to $\mathcal{S}$ and $\boldsymbol{E}_\mathcal{S} \in \mathbb{R}^{(P^2\cdot C\cdot(N_s+1)\times d_\mathcal{S})}$ indicates a learnable linear projection. Also, $d_\mathcal{S}$ represents the hidden dimension of the transformer encoder, and $N_\mathcal{S}$ represents the number of images shifted by $\mathcal{S}$.

3.2.2 Patch Embedding Layer

This section describes how to use SPT as a patch embedding layer. We concatenate a class token to visual tokens and then add positional embedding. Here the class token is the token with representation information of the entire image, and the positional embedding gives positional information to the visual tokens. If a class token is not used, only positional embedding is added to the output of SPT. How to apply the SPT to the patch embedding layer is formulated as follows:

$$\mathrm{S}_{pe}(\mathbf{x}) = \begin{cases} [\mathbf{x}_{cls};\mathrm{S}(\mathbf{x})] + \boldsymbol{E}_{pos} & \text{if } \mathbf{x}_{cls} \text{ exist} \\ \mathrm{S}(\mathbf{x}) + \boldsymbol{E}_{pos} & \text{otherwise} \end{cases} \tag{7}$$

where $\mathbf{x}_{\boldsymbol{cls}} \in \mathbb{R}^{d_\mathcal{S}}$ is a class token and $\boldsymbol{E}_{pos} \in \mathbb{R}^{(N+1)\times d_\mathcal{S}}$ is the learnable positional embedding. Also, N is the number of embedded tokens in Eq. 6.

3.2.3 Pooling Layer

Tokenization is the process of embedding 3D-tensor features into 2D-matrix features. For example, it embeds $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ into $\mathbf{y} = \mathcal{T}(\mathbf{x}) \in \mathbb{R}^{N\times d}$. Since $N = HW/P^2$, the spatial size of the 3D feature is reduced by P^2 through the tokenization process. So, if tokenization is used as a pooling layer, the number of visual tokens can be reduced. Therefore, we propose to use SPT as a pooling layer as follows: First, class tokens and visual tokens are separated, and visual tokens in the form of 2D-matrix are reshaped into 3D-tensor with spatial structure, i.e., $\mathcal{R} : \mathbb{R}^{N\times d} \rightarrow \mathbb{R}^{(H/P)\times(W/P)\times d}$. Then, if the SPT of Eq. 6 is applied, new visual tokens with a reduced number of tokens are embedded. Finally, the linearly projected class token is connected with the embedded visual tokens. If there is no class token, only $\mathcal{R}$ is applied before the output of SPT. The whole process is formulated as Eq. 8:

$$\mathrm{S}_{pool}(\mathbf{y}) = \begin{cases} [\mathbf{x}_{cls}\boldsymbol{E}_{cls};\mathrm{S}(\mathcal{R}(\mathbf{y}))] & \text{if } \mathbf{x}_{cls} \text{ exist} \\ \mathrm{S}(\mathcal{R}(\mathbf{y})) & \text{otherwise} \end{cases} \tag{8}$$

Table 1. Top-1 accuracy (%) according to temperatures.

TEMPERATURE	TOP-1 ACCURACY (%)	
	CIFAR100	T-ImageNet
$\frac{1}{4}\sqrt{d_k}$	73.70	57.62
$\frac{1}{2}\sqrt{d_k}$	**74.54**	**57.65**
$\sqrt{d_k}$	73.81	57.07
$2\sqrt{d_k}$	72.77	56.98
$4\sqrt{d_k}$	71.55	56.43

where $\boldsymbol{E}_{cls} \in \mathbb{R}^{d\times d'_\mathcal{S}}$ is a learnable linear projection. In addition, $d'_\mathcal{S}$ is the hidden dimension of the next stage. As a result, SPT embeds rich spatial information into visual tokens by increasing the receptive field of tokenization as much as spatially shifted.

3.3. Locality Self-Attention Mechanism

This section describes the LSA. The core of LSA is the diagonal masking (Sec. 3.3.1) and the learnable temperature scaling (Sec. 3.3.2).

3.3.1 Diagonal Masking

Diagonal masking plays a role in giving larger scores to inter-token relations by fundamentally excluding self-token relations from the softmax operation. Specifically, diagonal masking forces $-\infty$ on diagonal components of R of Eq. 4. This makes ViT's attention more focused on other tokens rather than attending to its own tokens. The proposed diagonal masking is defined by:

$$\mathrm{R}^M_{i,j}(\mathbf{x}) = \begin{cases} \mathrm{R}_{i,j}(\mathbf{x}) & (i \neq j) \\ -\infty & (i = j) \end{cases} \tag{9}$$

where $\mathrm{R}^M_{i,j}$ indicates each component of the masked similarity matrix.

3.3.2 Learnable Temperature Scaling

The second technique for LSA is the learnable temperature scaling, which allows ViT to determine the softmax temperature by itself during the learning process. Fig. 3 shows the average learned temperature according to depth when the softmax temperature is used as the learnable parameter in Eq. 5. Note that the average learned temperature is lower than the constant temperature of standard ViT. In general, the low temperature of softmax sharpens the score distribution. Therefore, the learnable temperature scaling sharpens the distribution of attention scores. Based on Eq. 5, the LSA

Table 2. Top-1 accuracy comparison of different models on small-size datasets.

MODEL	THROUGHPUT (images/sec)	FLOPs (M)	PARAMS (M)	CIFAR10	CIFAR100	SVHN	T-ImageNet
ResNet 56	4295	506.2	0.9	95.70	76.36	97.73	58.77
ResNet 110	2143	1020.0	1.7	**96.37**	79.86	97.85	62.96
EfficientNet B0	4078	123.9	3.7	94.66	76.04	97.22	66.79
ViT	8593	189.8	2.8	93.58	73.81	97.82	57.07
SL-ViT	7697	199.2	2.9	94.53	76.92	97.79	61.07
T2T	3388	643.0	6.7	95.30	77.00	97.90	60.57
SL-T2T	2943	671.4	7.1	95.57	77.36	97.91	61.83
CaiT	3138	613.8	9.1	94.91	76.89	98.13	64.37
SL-CaiT	2967	623.3	9.2	95.81	**80.32**	**98.28**	**67.18**
PiT	7583	279.2	7.1	94.24	74.99	97.83	60.25
SL-PiT w/o S_{pool}	6632	280.4	7.1	94.96	77.08	97.94	60.31
SL-PiT w/ S_{pool}	5981	322.9	8.7	95.88	79.00	97.93	62.91
Swin	6804	242.3	7.1	94.46	76.87	97.72	60.87
SL-Swin w/o S_{pool}	6384	247.0	7.1	95.30	78.13	97.88	62.70
SL-Swin w/ S_{pool}	5711	284.9	10.2	95.93	79.99	97.92	64.95

with both diagonal masking and learnable temperature scaling applied is defined by:

$$\mathrm{L}(\mathbf{x}) = \mathrm{softmax}(\mathrm{R}^{\mathrm{M}}(\mathbf{x})/\tau)\mathbf{x}\boldsymbol{E}_v \tag{10}$$

where τ is the learnable temperature.

In other words, LSA solves the smoothing problem of the attention score distribution. Fig. 4 shows the depth-wise averages of total Kullback-Leibler divergence (D_{KL}^{total}) for all heads. Here, T and M mean that only learnable temperature scaling and diagonal masking is applied to ViTs, respectively, and L indicates that the entire LSA is applied to ViTs. The lower the average of D_{KL}^{total}, the flatter the attention score distribution. We can find that when LSA is fully applied, the average of D_{KL}^{total} is larger by about 0.5 than standard ViT, so LSA attenuates the smoothing phenomenon of the attention score distribution.

4. EXPERIMENT

This section verifies that the proposed method improves the performance of various ViTs through several experiments. Sec. 4.1 describes the settings of the following experiments. Sec. 4.2 quantitatively shows that the proposed method effectively improves various ViTs and reduces the gap with CNNs. Finally, Sec. 4.3 demonstrates that the ViTs are qualitatively enhanced by visualizing the attention scores of the final class token.

4.1. SETTING

4.1.1 Environment and Dataset

The proposed method was implemented in Pytorch [27]. In the small-size dataset experiment (Table 2), The details of throughput measurement are as follows: The inputs were Tiny-ImageNet, and the batch size was 128, and the GPU was RTX 2080 Ti.

For small-size dataset experiments, CIFAR-10, CIFAR-100 [17], Tiny-ImageNet [21], and SVHN [25] were employed and ImageNet [19] was employed for the mid-size dataset experiment.

4.1.2 Model Configurations

In the small dataset experiment, in the case of ViT, the depth was set to 9, the hidden dimension was set to 192, and the number of heads was set to 12. This configuration was determined experimentally. And in the ImageNet experiment, we used the ViT-Tiny suggested by DeiT [33]. In the case of PiT, T2T, Swin and CaiT, the configurations of PiT-XS, T2T-14, Swin-T and CaiT-XXS24 presented in the corresponding papers were adopted as they were, respectively. The performance of ViT improves as the number of tokens increases, but the computational cost increases quadratically. We were able to experimentally observe that it was effective when both the number of visual tokens in ViT without pooling and the number of tokens in the inter-

Table 3. Top-1 accuracy (%) of the proposed method on ImageNet dataset.

MODEL	TOP-1 ACCURACY (%)
ViT	69.95
SL-ViT	71.55 (+1.60)
PiT	75.58
SL-PiT	77.02 (+1.44)
Swin	79.95
SL-Swin	81.01 (+1.06)

Table 4. Effect of each component of LSA on performance.

MODEL	TOP-1 ACCURACY (%)	
	CIFAR100	T-ImageNet
ViT	73.81	57.07
T-ViT	74.35	57.95
M-ViT	74.34	58.29
L-ViT	**74.87**	**58.50**

Table 5. Effect of the proposed SPT (S) and LSA (L) on performance.

MODEL	TOP-1 ACCURACY (%)	
	CIFAR100	T-ImageNet
ViT	73.81	57.07
L-ViT	74.87	58.50
S-ViT	76.29	60.67
SL-ViT	**76.92**	**61.07**

mediate stage of ViT with pooling are 64, considering this trade-off. Accordingly, we modified the baseline models. In small-size dataset experiments, the patch size of the patch embedding layer was set to 8 and the patch size of ViTs using pooling layers such as Swin and PiT was set to 16. In the ImageNet dataset experiment, the patch size was set to be the same as that used in each paper. Also, the hidden dimension of MLP was set to twice that of the transformer in the small dataset experiment, and the configuration used in each paper was applied in the ImageNet experiment.

4.1.3 Training Regime

According to DeiT, various techniques are required to effectively train ViTs. Thus, we applied data augmentations such as CutMix [42], Mixup [43], Auto Augment [6], Repeated Augment [7] to all models. In addition, regularization techniques such as label smoothing [31], stochastic depth [15], and random erasing [45] were employed. Meanwhile, AdamW [16] was used as the optimizer. Weight decays were set to 0.05, batch size to 128 (however, 256 for ImageNet), and warm-up to 10 (however, 5 for ImageNet). All models were trained for 100 epochs, and cosine learning rate decay was used. In the small-size dataset experiments, the initial learning rate of ViT and CNNs was set to 0.003, and that of the remaining models was set to 0.001. On the other hand, in the ImageNet experiment, the initial learning rate was set to 0.00025 for all models.

4.2. QUANTITATIVE RESULT

4.2.1 Image Classification

This section presents the experimental results for small-size datasets and the ImageNet dataset. In the small-size dataset experiment, Throughput, FLOPs, and the number of parameters were measured in Tiny-ImageNet.

First, Table 2 shows the performance improvement when the proposed method was applied to ViTs. Here, SL indicates that both SPT and LSA were applied, and S_{pool} means the SPT applied to the pooling layer. In most cases, the proposed method effectively improved the performance of ViTs, especially in CIFAR100 and Tiny-ImageNet. For example, in CIFAR100, the performance of CaiT and PiT improved by $+3.43\%$ and $+4.01\%$ respectively and in Tiny-ImageNet, the performance of ViT and Swin improved up to $+4.00\%$ and $+4.08\%$ respectively. Also note that the performance was greatly improved only with the acceptable overhead of inference latency. In other words, the cost-effectiveness of the proposed method is remarkable. For example, for ViT, T2T, and CaiT, the proposed method causes only latency overhead of 1.12%, 1.15%, and 1.06% respectively. And it can be seen in the case of PiT and Swin that additional performance improvement can be obtained by replacing the pooling layer with S_{pool}. Therefore, we can find that spatial modeling provided by SPT is effective not only for patch embedding but also for the pooling layer. Also, This table shows that the proposed method effectively reduces the gap between ViT and CNN on small-size datasets. For example, SL-CaiT achieves the best performance over ResNet and EfficientNet on all datasets except CIFAR10. SL-Swin also offers better throughput while providing performance comparable to CNNs.

Table 3 shows performance when training a mid-size dataset ImageNet from scratch. In ViT, SPT was applied only to patch embeddings, and in PiT and Swin, SPT was applied to both patch embedding and pooling layers. We could observe that the proposed method is sufficiently ef-

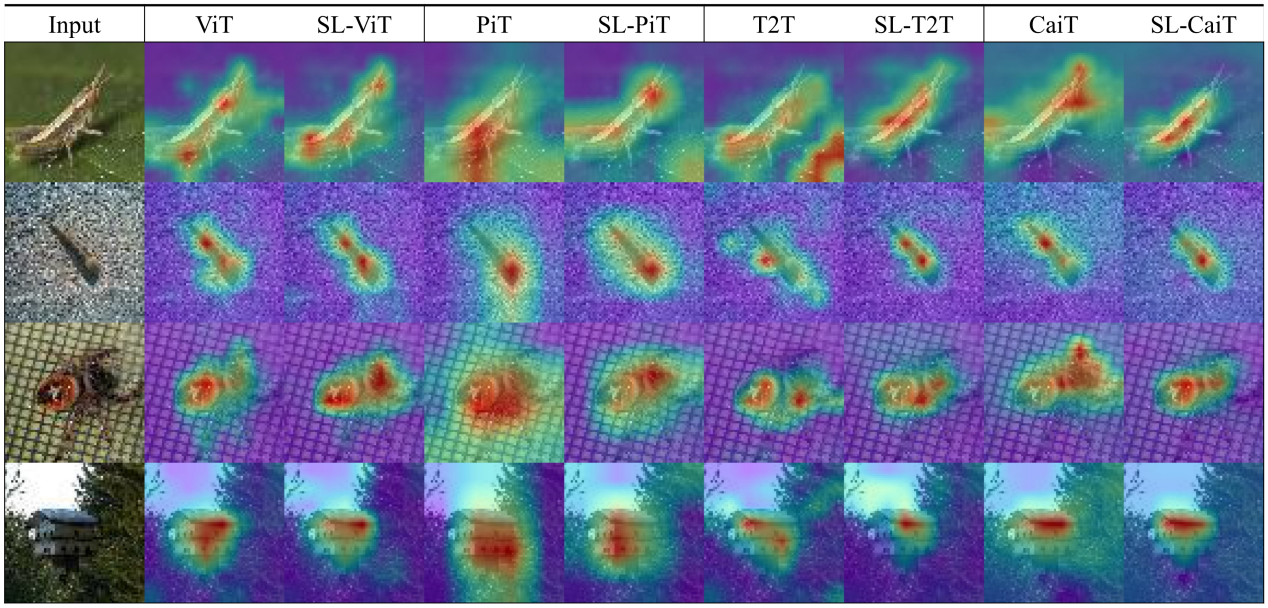


Figure 5. Visualization of attention scores of final class tokens.

fective for ImageNet. For example, the performance was improved by the proposed method as much as $+1.60\%$ for ViT, $+1.44\%$ for PiT, and $+1.06\%$ for Swin. As a result, we find that the proposed method noticeably improves the ViTs even on mid-size datasets.

4.2.2 Ablation Study

This section describes the ablation study on the proposed method. ViT was used for this experiment.

Elements of LSA Let's look at the effect of learnable temperature scaling and diagonal masking, two key elements of LSA, on overall performance. Table 4 shows that learnable temperature scaling and diagonal masking effectively resolves the smoothing phenomenon of attention score distribution (see Fig. 4). For example, learnable temperature scaling and diagonal masking in Tiny-ImageNet improved performance by $+0.88\%$ and $+1.22\%$, respectively. Considering that the LSA applied with both techniques shows a performance improvement of $+1.43\%$, we can claim that the contribution of each is sufficiently large and the two techniques produce a synergy.

SPT and LSA Table 5 shows that SPT and LSA can dramatically improve performance by increasing the locality inductive bias of ViT independently. In particular, in Tiny-ImageNet, SPT and LSA improved performance by $+1.43\%$ and $+3.60\%$, respectively. When both techniques were applied, the performance improvement was $+4.00\%$. This proves the competitiveness and synergy of the two key element technologies.

4.3. QUALITATIVE RESULT

Fig. 5 visualizes the attention scores of the final class token when SPT and LSA were applied to various ViTs. When the proposed method was applied, we can observe that the object shape is better captured as the attention, which was dispersed in the background, is concentrated on the target class. In particular, this phenomenon is evident in the CaiT of the first row, the T2T of the second row, the ViT of the third row, and the PiT of the last row. Therefore, we can find that the proposed method effectively increases the locality inductive bias and induces the attention of the ViTs to improve.

5. CONCLUSION

To train ViT on small-size datasets, this paper presents two novel techniques to increase the locality inductive bias of ViT. First, SPT embeds rich spatial information into visual tokens through specific transformation. Second, LSA induces ViT to attend locally through softmax with learnable parameters. The SPT and LSA can achieve significant performance improvement independently, and they are applicable to any ViTs. Therefore, this study proves that ViT learns small-size datasets from scratch and provides an opportunity for ViT to develop further.

References

[1] A. Araújo, W. Norris, and Jack Sim. Computing receptive fields of convolutional neural networks. 2019. 3, 4

[2] Jimmy Ba, J. Kiros, and Geoffrey E. Hinton. Layer normalization. *ArXiv*, abs/1607.06450, 2016. 3

[3] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018. 2

[4] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V. Le. Attention augmented convolutional networks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3285–3294, 2019. 1

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, J. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. Henighan, R. Child, A. Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020. 1

[6] Ekin D. Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 113–123, 2019. 7

[7] E. D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3008–3017, 2020. 7

[8] J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 1

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1, 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[11] Yu-Lin He, Xiaoliang Zhang, W. Ao, and Joshua Zhexue Huang. Determining the optimal temperature parameter for softmax function in reinforcement learning. *Appl. Soft Comput.*, 70:80–85, 2018. 3, 4

[12] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. pages 11936–11945, October 2021. 1, 2

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 1

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 1

[15] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. Deep networks with stochastic depth. In *ECCV*, 2016. 7

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015. 7

[17] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[18] A. Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90, 2012. 1

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 6

[20] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942, 2020. 1

[21] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 6

[22] Yawei Li, K. Zhang, Jie Cao, R. Timofte, and L. Gool. Localvit: Bringing locality to vision transformers. *ArXiv*, abs/2104.05707, 2021. 1

[23] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7082–7092, 2019. 2

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, October 2021. 1, 2

[25] Yuval Netzer, T. Wang, A. Coates, A. Bissacco, Bo Wu, and A. Ng. Reading digits in natural images with unsupervised feature learning. 2011. 6

[26] Behnam Neyshabur. Towards learning convolutions from scratch. *arXiv preprint arXiv:2007.13657*, 2020. 1

[27] Adam Paszke, S. Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, N. Gimelshein, L. Antiga, Alban Desmaison, Andreas Köpf, E. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 6

[28] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck

transformers for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16519–16529, 2021. 1

[29] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 843–852, 2017. 1

[30] Christian Szegedy, W. Liu, Y. Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, D. Erhan, V. Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. 1

[31] Christian Szegedy, V. Vanhoucke, S. Ioffe, Jonathon Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. 7

[32] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *ArXiv*, abs/1905.11946, 2019. 1

[33] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. pages 10347–10357. PMLR, 2021. 2, 6

[34] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Herve Jegou. Going deeper with image transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 32–42, October 2021. 1, 2

[35] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017. 1

[36] Fei Wang, Mengqing Jiang, Chen Qian, S. Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6450–6458, 2017. 1

[37] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, D. Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1

[38] X. Wang, Ross B. Girshick, A. Gupta, and Kaiming He. Non-local neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 1

[39] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22–31, October 2021. 1, 2

[40] Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. Co-scale conv-attentional image transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9981–9990, October 2021. 1

[41] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis E.H. Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 558–567, October 2021. 2

[42] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision (ICCV)*, 2019. 7

[43] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. 7

[44] Han Zhang, I. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, 2019. 1

[45] Zhun Zhong, L. Zheng, Guoliang Kang, Shaozi Li, and Y. Yang. Random erasing data augmentation. In *AAAI*, 2020. 7

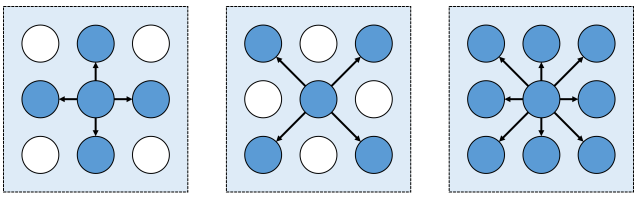
Supplementary

This section investigates the various shifting strategies that SPT can employ. Specifically, we explored the shift direction and shift intensity (shift ratio), which have the most impact on performance.

We examined the following three shift directions. The first is the 4 cardinal directions consisting of up, down, left and right directions (Fig. 1(a)). The second is 4 diagonal directions including up-left, up-right, down-left and down-right (Fig. 1(b)). The last is the 8 cardinal directions including all the preceding directions (Fig. 1(c)). Table 1 shows top-1 accuracy in small-size datasets such as CIFAR-10, CIFAR-100, SVHN, and Tiny-ImageNet for each shift direction. This experiment adopted a model applying SPT to standard ViT. 4 cardinal directions showed the best performance in CIFAR-10 and SVHN. On the other hand, 4 diagonal directions and 8 cardinal directions provided the best performance in CIFAR-100 and Tiny-ImageNet, respectively. This shows that the shift direction is somewhat dependent on the characteristics of datasets. For example, in CIFAR-10 or CIFAR-100, the target class tends to be in the center of the image, whereas other datasets do not. The location of the target class has some degree of correlation with the shift direction, and the correlation can affect the performance. However, since the performance difference was experimentally marginal, in this paper, the shift direction in the experiment was fixed to 4 diagonal directions.

Table 1. Top-1 Accuracy (%) of Various Shift Directions.

DIRECTIONS	TOP-1 ACCURACY (%)			
	CIFAR10	CIFAR100	SVHN	T-ImageNet
4 Cardinal	**94.44**	76.29	**97.87**	60.35
4 Diagonal	94.33	76.29	97.86	**60.67**
8 Cardinal	94.41	**76.40**	97.81	60.57



(a) (b) (c)

Figure 1. Various Shift Directions.

Next, we look at various shift ratios. The degree of image shifting in SPT is defined as follows: $\text{SHIFT} = P \times r_{shift}$, where P represents the patch size, and r_{shift} represents the shift ratio. Table 2 shows the performance according to shift ratio for CIFAR-100, Tiny-ImageNet, and ImageNet. In this experiment, a model with SPT applied to standard ViT was used, and 4 diagonal directions were adopted. In CIFAR-100 and ImageNet, a ratio of 0.5 was the best, and in Tiny-ImageNet, a ratio of 0.25 was the best. This experimental result shows that the optimal shift ratio also depends on the datasets. Since the relatively most reasonable shift ratio is 0.5 according to our experiment, all the experiments in this paper fixed the shift ratio to 0.5. Note that more various shifting strategies will be available in addition to the methods considered here. The exploration of optimal shifting strategy according to datasets remains as a further work.

Table 2. Top-1 Accuracy (%) of Various Raitos.

SHIFT RATIO	TOP-1 ACCURACY (%)		
	CIFAR100	T-ImageNet	ImageNet
0.125	-	60.63	-
0.25	76.24	**61.01**	70.65
0.5	**76.29**	60.78	**70.83**
0.75	75.73	60.18	70.57
1.00	74.63	59.35	-