SP-ViT: Learning 2D Spatial Priors for Vision Transformers

Yuxuan Zhou[1*]
zhouyuxuanyx@gmail.com
Wangmeng Xiang[2*]
cswxiang@comp.polyu.edu.hk
Chao Li[4]
lllcho.lc@alibaba-inc.com
Biao Wang[4]
wb.wangbiao@alibaba-inc.com
Xihan Wei[4]
xihan.wxh@alibaba-inc.com
Lei Zhang[2]
cslzhang@comp.polyu.edu.hk
Margret Keuper[1,3]
margret.keuper@uni-siegen.de
Xiansheng Hua[4]
xiansheng.hxs@alibaba-inc.com

[1] University of Mannheim
[2] The Hong Kong Polytechnic University
[3] University of Siegen,
Max Planck Institute for Informatics,
Saarland Informatics Campus
[4] Alibaba Group

*

Abstract

Recently, transformers have shown great potential in image classification and established state-of-the-art results on the ImageNet benchmark. However, compared to CNNs, transformers converge slowly and are prone to overfitting in low-data regimes due to the lack of spatial inductive biases. Such spatial inductive biases can be especially beneficial since the 2D structure of an input image is not well preserved in transformers. In this work, we present Spatial Prior–enhanced Self-Attention (SP-SA), a novel variant of vanilla Self-Attention (SA) tailored for vision transformers. Spatial Priors (SPs) are our proposed family of inductive biases that highlight certain groups of spatial relations. Specifically, the attention score is calculated with emphasis on certain kinds of spatial relations at each head, and such learned spatial foci can be complementary to each other. Based on SP-SA we propose the SP-ViT family, which consistently outperforms other ViT models with similar GFlops or parameters. Our largest model SP-ViT-L achieves 86.3% Top-1 accuracy with a reduction in the number of parameters by almost 50% compared to previous state-of-the-art model (150M for SP-ViT-L↑384 vs 271M for CaiT-M-36↑384) among all ImageNet-1K models trained on 224×224 and fine-tuned on 384×384 resolution w/o extra data. Code can be found at https://github.com/ZhouYuxuanYX/SP-ViT.

*Work done during internship at Alibaba Group.

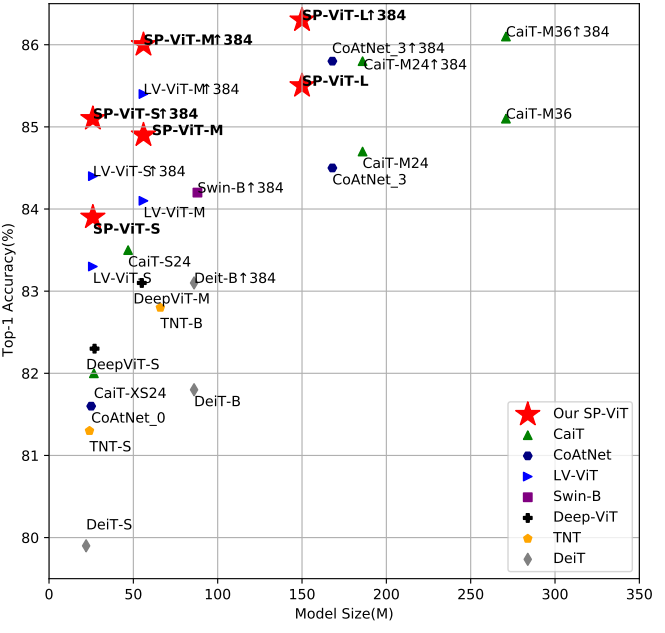
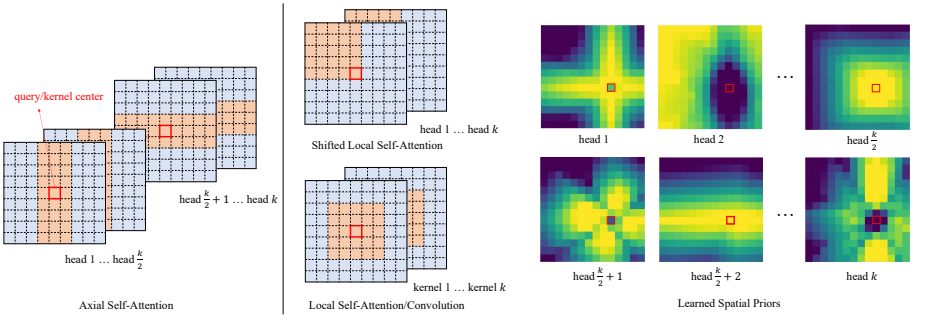


Figure 1: ImageNet-1K top-1 accuracy of our proposed SP-ViT and state-of-the-art ViTs. The models shown are all trained on 224×224 resolution, $\uparrow$ denotes that models are fine-tuned on a higher resolution. Note that we exclude models pretrained on extra data or larger resolution than 224×224 for a fair comparison.

1 Introduction

Transformers [21] have recently achieved exciting results in image classification [3, 5, 8, 9, 12, 14, 16, 19, 20, 27], after dominating in natural language processing (NLP) [1, 7, 15]. At the heart of transformer lies the so-called self-attention mechanism, which captures the content relations between all pairs of input tokens and focuses on related pairs selectively. Self-attention is more flexible in comparison to convolution, which is hard-coded to capture local dependencies exclusively. This can possibly equip transformer models with larger capacity and greater potential for computer vision tasks. As reported in recent works, transformers outperform Convolutional Neural Networks (CNNs), when pretrained on large dataset [9], facilitated with knowledge distillation [19] or pseudo labels [14] from pretrained CNNs.

Nevertheless, CNNs generalize better and converge faster than Vision Transformers (ViT). This suggests that certain types of inductive biases employed in convolution can still be beneficial to vision tasks. Not surprisingly, many recent studies [5, 6, 8, 11, 16, 19, 23, 26, 27] propose to incorporate convolutional inductive biases into ViTs in different ways. The effectiveness of convolution relies on the fact that neighboring pixels of natural images are highly correlated, but there may exist other highly correlated contents outside the local receptive field of a convolutional filter. Therefore, we propose to make use of a variety of inductive biases simultaneously, just as humans do, e.g., if we see a part of a horizontal object, we



(a) Hand-crafted convolutional inductive biases (b) Proprosed SP-SA

Figure 2: (a) Convolutional inductive biases proposed for ViTs: axial self-attention in CSWin-Transformer [8] and shifted local self-attention in Swin-Transformer [16]. (b) Our Spatial Priors (SPs) are learned by our model automatically. The learned SPs assign different scores for different spatial relations. Given a certain SP, attention is forced to be within high-score regions. Our SP-SA handles different types of spatial relations in a complementary manner, e.g., SPs which focus on local and non-local relations are both learned.

naturally look along its direction instead of restricting our sight within a local area.

In this work, we introduce a novel family of inductive biases named *Spatial Priors* (SPs) into ViTs via an extension of vanilla self-attention (SA), called *Spatial Prior–enhanced Self-Attention* (SP-SA). SP-SA highlights a certain group of 2D spatial relations at each attention head based on the relative position of key and query patches. Since the construction and validation of appropriate spatial priors are extremely laborious, we introduce the idea of *learnable spatial priors*. More specifically, we only impose the weak prior knowledge to the model that different relative distances should be treated differently. Yet we do not force the model to favor any kind of spatial relation, e.g., neither local nor non-local. Effective spatial priors (SPs) are supposed to be discovered by the model itself in the training stage. For this purpose, SPs are represented by a family of mathematical functions which map the relative coordinates to abstracted scores, called *spatial relation functions*. To search for desirable spatial relation functions, we parameterize these functions by neural networks and optimize them jointly with ViTs. Thereby, the model can learn spatial priors similar to the ones induced in convolutions, as well as spatial relationships over larger distances. Examples for learned SPs are shown in Fig. 2(b). Diverse complementary patterns are presented in different attention heads, so that different types of spatial relations are handled individually.

As a matter of fact, convolutional inductive biases can be seen as a special kind of spatial priors: they first divide coordinate spatial relations into two categories, i.e., ones focusing on the local neighborhood and ones focusing on non-local regions. Then they learn priors of the local neighborhoods and ignore the non-local relations. For comparison, some of the existing approaches to combine such convolutional biases with ViTs are illustrated in Fig. 2(a).

In summary, we make the following contributions:

- We propose a family of inductive biases for ViTs that focus on different types of spatial relations, called Spatial Priors (SP). SPs generalize convolutional inductive biases to both local and non-local correlations. Parameterized with neural networks, SPs are automatically learned during training, w/o preference for any hard-coded region.

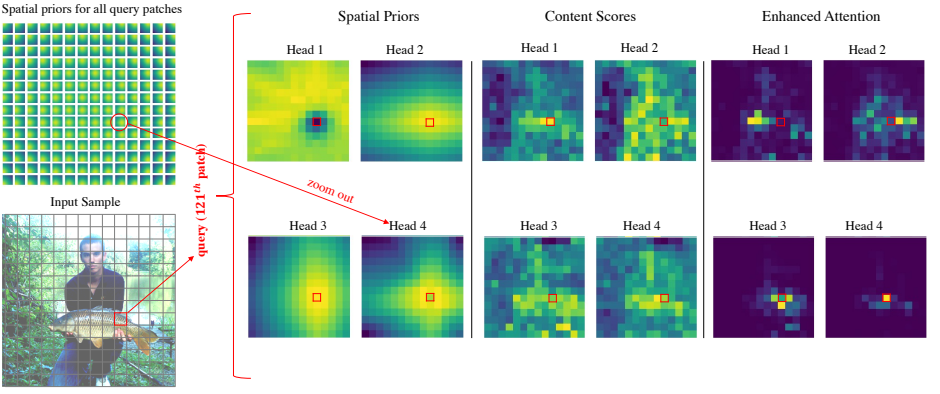


Figure 3: Visualization of the learned 2D SPs, content scores and the enhanced attention. The input image is shown in the bottom-left and the query patch is marked in red. Different SPs are learned, including horizontal and vertical (head 2 and 3), non-local (head 1), as well as cross-shaped (head 4). The attention scores at each head are obtained within the context of a certain type of spatial relations. The original attention is distracted by background objects, whereas our Spatial Priors help the model to focus on the object of interest.

- We propose SP-SA, a novel self-attention variant that automatically learns beneficial spatial inductive biases. Built on SP-SA, we construct a ViT variant called SP-ViT. SP-ViTs establish state-of-the-art results on the ImageNet Benchmark w/o extra data.

- Our SPs are compatible with various input sizes, as they are derived from relative coordinates . SP-ViTs also demonstrate improved classification performance over the baseline model when fine-tuned on higher resolution.

2 Related Work

Vision Transformers Recently, Dosovitskiy et al. [9] showed that purely attention-based transformers can achieve state-of-the-art performance in image classification, when pre-trained on large-scale datasets. Since then, a vast amount of efforts have been made to improve ViTs. Some works [10, 14] find it effective to add additional losses or regularization terms, while others propose new patch embedding blocks [12] or scale-up methods [20, 29]. [8, 16, 22, 28] propose to utilize multi-scale information, where local attention are is adopted to reduce the overall computation. It is noteworthy that an cross-shaped 2D structure, similar to the design in CSWin-Transformer [8], is also learned by our model.

Inductive Biases for ViTs ViTs' performance degrades rapidly with a reduced amount of training data. To alleviate this issue, many studies focus on emphasizing local correlations by introducing a convolutional inductive bias into ViTs, either by restricting SA to local windows [8, 16, 17], combining vanilla transformers with implicit or explicit convolutional operations [4, 5, 11, 23, 26, 27], knowledge distillation [19], or convolutional initialization [6]. Our work also incorporates inductive biases into ViTs, but they are not locally restricted and are automatically learned by the model. Indeed, as shown in Fig. 2(b), patterns that focus solely on local or remote regions are both present in the learned SPs.

Relative Spatial Information Transformers are by their very nature permutation invariant, thus extra spatial information is often supplied to better handle ordered input data. Besides the common absolute positional embedding, the relative spatial information is also considered in Swin-Transformer [16] by an trainable bias term called relative positional bias. ConViT [6] also introduces a function based on coordinates relative to force the attention to be within a local region. In comparison to ViTs, using relative positional information is more common in NLP transformers. The relative positional embedding [18] is built on the distances between tokens and has been improved in XL-Net [25] and DEBERTA [13]. It can be extended to 2D for ViTs with little effort, and is proved to be effective in [24]. The essential difference of our method is the focus on various learned spatial relations at each head, which proves to be beneficial in Sec. 4.3.

3 Method

3.1 Spatial Prior-enhanced Self-Attention

Motivated by the observation that certain inductive biases on spatial relations can be beneficial to transformers, we propose an extension of self-attention enhanced by a combination of learned 2D Spatial Priors (SPs), called Spatial Prior–enhanced Self-Attention (SP-SA). Each SP $\Omega \in \mathbf{R}^{N\times N}$ forms a specific spatial context for computing attention scores $\mathbf{A} \in \mathbf{R}^{N\times N}$, and it is derived from coordinate spatial relations between input tokens, i.e. relative positions between the key and query patches for ViTs. Thus an SP has exactly the same form of attention scores and we can simply integrate it in the equation of vanilla SA [21] by multiplicative interaction:

$$A_{ij} = \frac{\exp(e_{ij} \cdot \Omega_{ij})}{\sum_{k=1}^{n} \exp(e_{ik} \cdot \Omega_{ik})}, \tag{1}$$

with

$$e_{ij} = \frac{(\vec{x}_i^{\top} \mathbf{W}^{Q})(\vec{x}_j^{\top} \mathbf{W}^{K})^{\top}}{\sqrt{d_z}}, \tag{2}$$

where $\vec{x}_i$ and $\vec{x}_j$ are the i^{th} and j^{th} input tokens.

3.1.1 Learnable 2D Spatial Priors

Taking query patch i as the reference point, we can obtain a relative coordinate $\vec{r}_{ij} \in \mathbf{R}^2$ for image patch j. Then we employ a shared mapping f_p for all query and key patch pairs, named spatial relation function:

$$\Omega_{ij} = f_p(\vec{r}_{ij}), \tag{3}$$

the outputs together form the so-called 2D SP Matrix Ω.

To enable the model to learn desirable inductive biases automatically, we employ Multilayer Perceptron (MLP) to parameterize the mapping from 2D relative coordinates to Ω. Thereby, we allow Ω to learn a weighting for the attention scores for query $\vec{x}_i$ and key $\vec{x}_j$ which depends solely on their relative coordinates and is applied in a non-linear way, i.e. before the softmax. We extend SP-SA to its multi-head version by adding a unique network to each head. This design follows the same motivation as multi-head self-attention and assumes that a combination of different SPs should boost the performance.

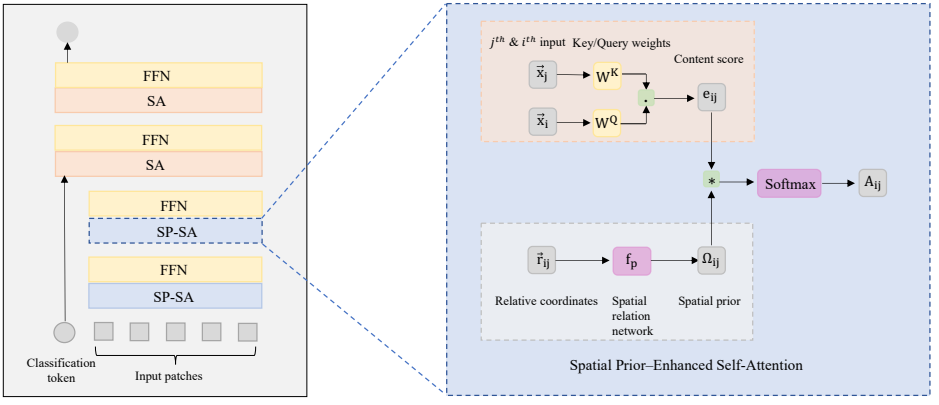


Figure 4: The schema of SP-ViT. SP-SA can be used as a drop-in replacement for the vanilla SA layer at a range of depths. Because the classification token does not have a valid 2D relative coordinate, it is simply concatenated with the hidden representation after the last SP-SA layer. FFN: feedforward network (2 linear layers separated by a GeLU activation).

3.2 Relation to Other Methods

In the following, we discuss the relation of SP-SA to the most related work.

Relation to Local Windows The square and cross-shaped windows used in [8, 16] can be seen as a special form of our proposed spatial relation functions in practice:

$$f_p(\vec{r}_{ij}) = \begin{cases} 1, & \text{if } \|(\vec{r}_{ij} - \vec{\Delta}) \odot (a,b)\|^\infty <= 1 \\ & \text{or } \|(\vec{r}_{ij} - \vec{\Delta}) \odot (b,a)\|^\infty <= 1, \\ 0, & \text{else} \end{cases} \tag{4}$$

where $\vec{\Delta}$, a and b control the shift, window width and height respectively. If $a = b$, it generates a square window, otherwise it results in a cross-shaped window. Both works only adopt some hard-coded patterns for the whole network, while our method proposes to benefit from a variety of beneficial 2D structures.

Relation to PSA The Positional Self-Attention (PSA) proposed in [6] can also be regarded as a manually designed family of spatial relation functions:

$$f_p(\vec{r}_{ij}) = \alpha(\|(\Delta_x, \Delta_y)\|^2 - \|\vec{r}_{ij} - (\Delta_x, \Delta_y)\|^2), \tag{5}$$

where the parameters Δ_x and Δ_y are specially initialized to approximate convolution effect.

Note that their main contribution is the so-called local/convolutional initialization, which restricts the number of heads to the square of integer numbers, and the initial values of both α and $\vec{\Delta}$ require extra hyperparameter tuning. In order to compare with their method, we adopt a ViT baseline with 9 heads for ablation analysis as in [6].

Relation to Relative Positional Embeddings Shaw et al. [18] introduce the so-called 1D Relative Positional Embedding (RPE) for transformers to take relative distances into account:

$$A_{ij} = \frac{\exp(e_{ij} + (\vec{x}_i \mathbf{W}^Q) T_{r_{ij}})}{\sum_{k=1}^{n} \exp(e_{ik} + (\vec{x}_i \mathbf{W}^Q) T_{r_{ij}})}, \tag{6}$$

where T is a learnable embedding table from which the RPE is taken. Then it interacts multiplicatively with the query. If extended to 2D, it is equivalent to applying a linear transformation to one-hot representations of relative distances. For one-hot representations, the magnitude of distances is neglected, while this is not the case for relative coordinates.

The main difference of our approach to all previous methods is the combination of complementary spatial priors at each layer. As shown in Table 6, performance drops form 83.6% to 82.1% with the same spatial priors per layer. In addition, we can see in Figure 2(b) that different spatial foci (local and non-local) are learned for each layer. With such spatial foci, our model is less distracted by noises w.r.t. a certain context. For example, as discussed in Sec. 4.1, our SP-ViT shows a significantly diminished class activation in background regions compared to DeiT. We also confirmed experimentally that our SP-SA outperforms these methods, see Tab. 3.

4 Experiments

We first provide an experimental evaluation of the proposed SP-SA in the context of image classification on the ImageNet-1k dataset and show that SP-ViTs achieve state-of-the-art results for training without extra data. Further, we provide an extensive ablation study to analyze the impact of all proposed model details.

Table 1: Comparing to state-of-the-art models trained on ImageNet-1k 224×224 resolution. Models are by default trained and tested on 224×224 resolution if not specified. $\uparrow$ plus size denotes the model is trained on 224×224 resolution then fine-tuned and tested on size $\times$ size resolution. The performance of LV-ViT-L trained on 224×224 resolution is not available in [14]. And LV-ViT-L trained on 288×288 resolution has a lower accuracy of 85.3%.

Network	Top-1 (%)	Parameters	FLOPs
DeiT-S [19]	79.9	22M	4.6B
CaiT-XS-24 [20]	82.0	27M	5.4B
LV-ViT-S [14]	83.3	26M	6.6B
Our SP-ViT-S	**83.9**	26M	6.6B
DeiT-B [19]	81.8	86M	17.5B
Swin-B [16]	83.3	88M	15.4B
CaiT-S-24 [20]	83.5	47M	9.4B
LV-ViT-M [14]	84.1	56M	12.7B
Our SP-ViT-M	**84.9**	56M	12.7B
CaiT-M-24 [20]	84.7	186M	36.0B
Our SP-ViT-L	**85.5**	150M	34.7B
LV-ViT-S$\uparrow$384 [14]	84.4	26M	22.2B
SP-ViT-S$\uparrow$384	**85.1**	26M	22.2B
CaiT-S-24$\uparrow$384 [20]	85.1	47M	32.2B
LV-ViT-M$\uparrow$384 [14]	85.4	56M	42.2B
Our SP-ViT-M$\uparrow$384	**86.0**	56M	42.2B
CaiT-M-24$\uparrow$384 [20]	85.8	186M	116.1B
Our SP-ViT-L$\uparrow$384	**86.3**	150M	110.6B

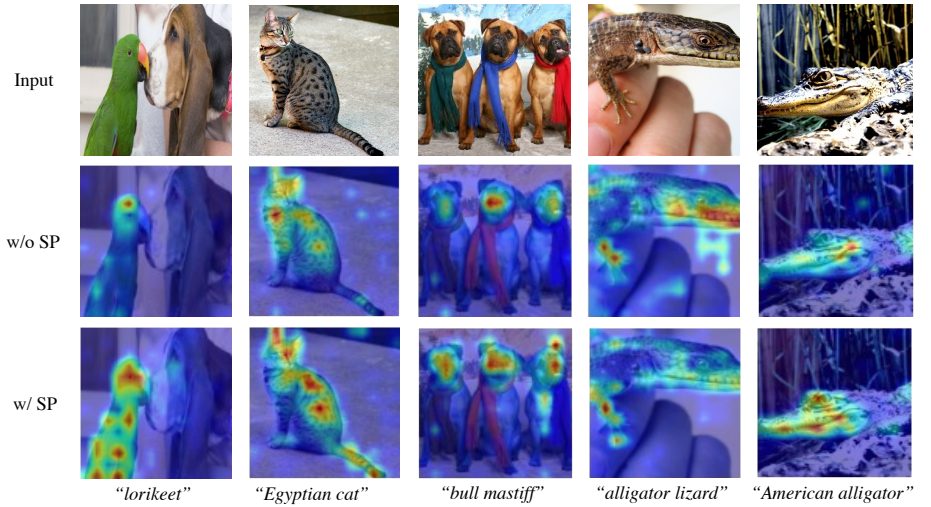


Figure 5: Visualization using Transformer Explainability [2]. The second row are results of DeiT baseline w/o SP layers. The Last row are results of SP-ViT. Our SP-ViT generate results with more focus on areas of interests and less distraction from background.

4.1 Image Classification on ImageNet-1K

Settings All models for ImageNet-1K classification are trained on a single machine node with 8 Tesla V100 GPUs. Our code is based on DeiT [19]. To obtain our SP-ViT, we replace the vanilla SA layers of the baseline with SP-SA till the last 2 layers and follow the training settings in [14] (with Token Labeling). We keep the vanilla SA in the last 2 layers, based on the ablation analysis conducted on a fraction of ImageNet, please refer to the Appendix for more details. When fine-tuning on higher resolution (indicated by ↑384 in Tab. 1), we set batch size to 512, learning rate to 5e-6, weight decay to 1e-8 and we fine-tune the model for 30 epochs.

Comparing to State-of-the-Art Models We compare our SP-ViT (based on LV-ViT) with other recent ViTs in Tab. 1. Within all groups of comparable model sizes, SP-ViT outperforms competing models. Our best result of 86.3% is achieved with SP-ViT-L↑384. It outperforms all previous models with about only 150M parameters as compared to 271M parameters of the second best CaiT-M-36↑384. Also note that our smaller SP-ViT-M↑384 already achieves 86.0% accuracy, on par with CaiT-M-36↑384 while reducing parameters from 271M to 56M (by a factor of about 4.8).

Qualitative results We present visualizations of target class activation maps using the recent technique [2] in Figure 5 to showcase the behavior of SP-ViT. While the DeiT model only shows class activations on small parts of the target class regions, for example on the head of the "Lorikeet", the fur of the "Egyptian cat" or the jaw of the "American alligator", the proposed SP-ViT model shows class activations on wider target class regions. Thereby, it follows well class specific image regions such as the pointy ears as well as the tail of the "Egyptian cat", and the dogs' ears in the "Bull mastiff" class. The "Alligator lizard" example as well as the "American alligator" further show a significantly diminished class activation

in background regions compared to DeiT. In summary, we make two observations: 1) The results generated by SP-ViT focus more on areas of target class objects comparing to DeiT. In “Lorikeet", “Bull mastiff", “Egyptian cat" and “American alligator", SP-ViT’s activation maps clearly have a better coverage of target class; 2) The distraction by background is better suppressed, e.g. in “Alligator lizard", resulting in a cleaner activation map.

4.2 Semantic Segmentation

Following [14] and [16], we utilize UperNet as our base framework and our SP-Vit trained ImageNet1K as the backbone to perform semantic segmentation on ADE20K. We adopt the same training setup as [13] and [15] and obtain 49.8 mIoU, which improves the result of LV-ViT-S by 1.2 mIoU. This shows that our proposed SPs benefit downstream tasks as well.

Table 2: Performance of our proposed SP-ViT in the downstream semantic segmentation task. SP-ViT improves over its baseline on both single-scale (SS) and multi-scale (MS) setups on the validation set.

Method	mIoU (SS)	P.Acc. (SS)	mIoU (MS)	P.Acc. (MS)
LV-ViT-S	47.9	82.6	48.6	83.1
SP-ViT-S	49.0	83.0	49.8	83.4

4.3 Ablation Analysis

For ablation, we employ a small DeiT model as the baseline with 12 layers, 9 heads and 432 embedded dimensions. The choice of head numbers is simply for a fair comparison with other methods, because Positional Self-Attention (PSA) introduced by d’Ascoli et al. [6] requires such specific numbers (square of integer numbers) of heads. Due to limited available computation resources, we train all model variants on the first 100 classes of ImageNet-1K called ImageNet-100 for 300 epochs, following the setup in [6]. In this section, we simply take the accuracy at the last epoch for all models. This should be a fair comparison, since we adopt the same hyperparameters for different models without tuning. For all experiments in this section, we train the models on 4 NVIDIA P100 GPUs and adopt a batch size of 256. The rest of settings are kept the same as DeiT’s w/o knowledge distillation in [19].

Table 3: Comparing to SA with Relative Positional Bias [16], Positional SA[6], SA with the 2D extension of Relative Positional Embedding (RPE) [18] as well as a more advanced version proposed in DEBERTA [13] on ImageNet-100.

Method	Top-1 acc (%)
2D RPE [18]	79.9
Improved 2D RPE [13]	82.8
Relative Positional Bias [16]	81.3
Positional Self-Attention [6]	82.5
SP-SA Additive	83.5
SP-SA	**83.6**

Comparing to Related Approaches SP-SA Additive is obtained by replacing the multiplication in Eq. (1) with a summation. It is more comparable to other methods which also employ additive interaction between spatial information and content scores. As shown in Tab. 3, our SP-SA has much higher Top-1 accuracy than all previous methods. The advantage of our SP-SA can be largely credited to the combination of different spatial foci at each head, see Sec. 4.3. As opposed to our method, the Relative Positional Bias [16] directly adds a univariate bias term to the content score before applying softmax, and the bias term is taken from a parameter table based on the relative coordinates. Adding such a bias term is a straightforward idea to include relative spatial information, but it is neither based on the idea of nor capable of learning complex 2D spatial priors, as reflected in Tab. 3.

We have also compared SP-SA to Positional Self-Attention [6] with hand-crafted spatial relation function. Our method delivers better performance, which shows that the effort in such a manual design process can be saved by our learnable SP.

Table 4: The effect of unique Spatial Priors (SPs) per head. This setting performs best.

SP-SA	Top-1 (%)
shared SP	82.6
unique SPs per layer	82.1
unique SPs per layer&head (default)	**83.6**

Single vs Multiple Spatial Priors To validate the benefit of combining various learned SPs, we compare SP-SA to two variants: one only adopting a single SP for each layer, the other learning the same SP for the whole network. As shown in Tab. 4, a shared SP for the whole network provides better results than a single SP for each layer. However, the proposed setting with a unique SP per layer&head performs best, providing evidence of the benefit of combining different SPs.

5 Conclusions and Discussions

In this paper, we introduce a variant of self-attention (SA) named Spatial Prior-enhanced Self-Attention (SP-SA) to facilitate vision transformers with automatically learned spatial priors. Based on the SP-SA, we further proposed SP-ViT and experimentally demonstrate the effectiveness of our method. Our proposed SP-ViTs establish state-of-the-art results for models trained on ImageNet-1K only. For example, SP-ViT-M achieves a 0.8% higher accuracy comparing to the previous state-of-the-art LV-ViT-M. We hope that our powerful SP-SA can stimulate more studies on designing appropriate inductive biases for ViTs.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[2] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791, June 2021.

[3] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 357–366, 2021.

[4] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Conditional positional encodings for vision transformers. *arXiv preprint arXiv:2102.10882*, 2021.

[5] Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *arXiv preprint arXiv:2106.04803*, 2021.

[6] Stéphane d'Ascoli, Hugo Touvron, Matthew Leavitt, Ari Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *ICML 2021: 38th International Conference on Machine Learning*, 2021.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2018.

[8] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. *arXiv preprint arXiv:2107.00652*, 2021.

[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 *ICLR 2021: The Ninth International Conference on Learning Representations*, 2021.

[10] Chengyue Gong, Dilin Wang, Meng Li, Vikas Chandra, and Qiang Liu. Improve vision transformers training by suppressing over-smoothing. *arXiv: Computer Vision and Pattern Recognition*, 2021.

[11] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. *arXiv preprint arXiv:2104.01136*, 2021.

[12] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *arXiv preprint arXiv:2103.00112*, 2021.

[13] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *ICLR 2021: The Ninth International Conference on Learning Representations*, 2021.

[14] Zihang Jiang, Qibin Hou, Li Yuan, Daquan Zhou, Yujun Shi, Xiaojie Jin, Anran Wang, and Jiashi Feng. All tokens matter: Token labeling for training better vision transformers. *arXiv preprint arXiv:2104.10858*, 2021.

[15] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[16] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[17] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. *Advances in Neural Information Processing Systems*, 32, 2019.

[18] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 464–468, 2018.

[19] Hugo Touvron, Matthieu Cord, Douze Matthijs, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *ICML 2021: 38th International Conference on Machine Learning*, 2021.

[20] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 32–42, 2021.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017.

[22] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021.

[23] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. *arXiv preprint arXiv:2103.15808*, 2021.

[24] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10033–10041, 2021.

[25] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32, pages 5753–5763, 2019.

[26] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating convolution designs into visual transformers. *arXiv preprint arXiv:2103.11816*, 2021.

[27] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Francis E. H. Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021.

[28] Zizhao Zhang, Han Zhang, Long Zhao, Ting Chen, and Tomas Pfister. Aggregating nested transformers. *arXiv preprint arXiv:2105.12723*, 2021.

[29] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*, 2021.