

Maximizing the Position Embedding for Vision Transformers with Global Average Pooling

Wonjun Lee^{1,2}, Bumsub Ham¹, Suhyun Kim^{2*}

¹Yonsei University, Republic of Korea

²Korea Institute of Science and Technology, Republic of Korea
{velpegor, bumsub.ham}@yonsei.ac.kr, dr.suhyun.kim@gmail.com

Abstract

In vision transformers, position embedding (PE) plays a crucial role in capturing the order of tokens. However, in vision transformer structures, there is a limitation in the expressiveness of PE due to the structure where position embedding is simply added to the token embedding. A layer-wise method that delivers PE to each layer and applies independent Layer Normalizations for token embedding and PE has been adopted to overcome this limitation. In this paper, we identify the conflicting result that occurs in a layer-wise structure when using the global average pooling (GAP) method instead of the class token. To overcome this problem, we propose MPVG, which maximizes the effectiveness of PE in a layer-wise structure with GAP. Specifically, we identify that PE counterbalances token embedding values at each layer in a layer-wise structure. Furthermore, we recognize that the counterbalancing role of PE is insufficient in the layer-wise structure, and we address this by maximizing the effectiveness of PE through MPVG. Through experiments, we demonstrate that PE performs a counterbalancing role and that maintaining this counterbalancing directionality significantly impacts vision transformers. As a result, the experimental results show that MPVG outperforms existing methods across vision transformers on various tasks.

Introduction

Recently, vision transformers have become essential architecture in the field of computer vision due to their superior performance, surpassing CNNs in various tasks such as image classification, object detection, and semantic segmentation. This superiority has led to extensive research into numerous elements of vision transformer architecture, starting with ViT (Dosovitskiy et al. 2020).

Among the research on vision transformers, image representation methods for class prediction have been studied. In ViT, the class token is used to perform image representation, and the output of this token is then used to make class predictions via Multi-Layer Perceptron (MLP) (Dosovitskiy et al. 2020). However, in several vision transformers, global average pooling (GAP) has been preferred over the class token method due to its translation-invariant characteristics and superior performance (Chu et al. 2021b). As a result, the GAP

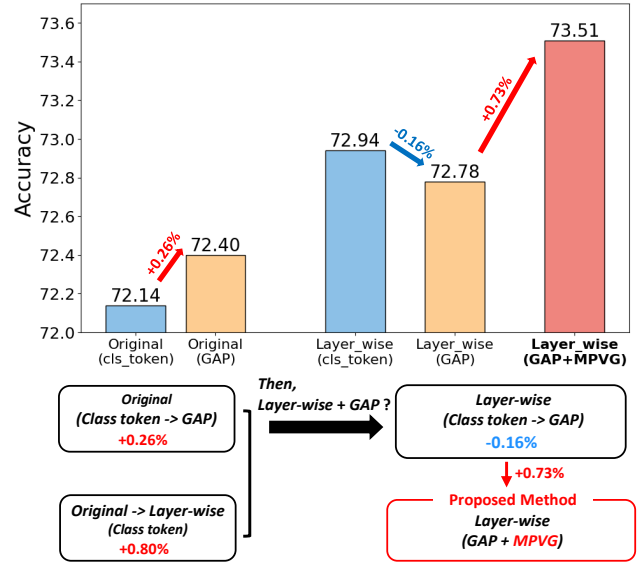


Figure 1: The conflicting result between the GAP method and the Layer-wise method. In DeiT-Ti, using the GAP method and the Layer-wise method separately results in performance improvements, but combining these two methods leads to a decrease in performance. As a result, MPVG resolves this phenomenon between the GAP and Layer-wise structure, maximizing the effect of PE.

method has been widely adopted in vision transformers (Liu et al. 2021; Chu et al. 2021a; Chang et al. 2023; Zhai et al. 2022).

Another research topic in vision transformers is position embedding (PE). PE plays a crucial role in providing positional information of tokens in the vision transformer, as the self-attention mechanism has an inherent deficiency in capturing the ordering of input tokens (Wu et al. 2021; Xu et al. 2024). In the original vision transformer, the expressiveness of the PE is limited due to its structure, where PE is simply added to the token embedding before being input into the first layer. To address this problem, each layer has independent Layer Normalizations (LNs) for the token embedding and PE, with PE being gradually delivered across

*Corresponding author.

all the layers (Yu et al. 2023). We refer to this method as "Layer-wise". The Layer-wise structure resolves the existing limitations and enhances the expressiveness of PE.

Interestingly, as shown in Fig 1, we observed results that differed from our expectations between the class token and GAP methods with PE delivered in the Layer-wise structure. On image classification, the GAP method demonstrates superior performance compared to the class token method (Chu et al. 2021b). Additionally, the Layer-wise structure also improves the performance of vision transformers by enhancing the expressiveness of PE (Yu et al. 2023). However, we observed a conflicting result where performance decreased when the GAP and Layer-wise structure were applied together. Therefore, to overcome the conflicting results, we propose a method to maximize the effectiveness of PE in the GAP approach.

We observe that PE exhibits distinct characteristics at each layer in the Layer-wise structure, which are not seen in the original vision transformer. As shown in Fig 2, we find that PE tends to counterbalance the values of token embedding passing through the layers in the Layer-wise structure. Additionally, we observe that this tendency becomes more pronounced as the layers deepen. We also discover that in the Layer-wise structure, while the token embedding values maintain the counterbalanced effect by PE as they progress through the layers, as shown in Fig 2-(b), the directional balance of the token embedding is not adequately compensated by PE after passing through the last layer, even though it is still maintained. Through this observation, we establish two hypotheses: (1) in the Layer-wise structure, PE initially provides position information, but as the layers deepen, PE plays a role in counterbalancing the values of token embedding; (2) after the last layer, it is beneficial for vision transformers to maintain the directional balance by counterbalancing the token embedding values with PE.

To validate these hypotheses, we simply add PE to the Layer Normalization (LN) that exists outside the layers and before the MLP head. We call this LN as "Last LN". We refer to the method that uses an improved Layer-wise structure, different from the conventional Layer-wise structure, and does not deliver PE to the Last LN as PVG. Additionally, we refer to the method that maximizes the role of PE by delivering it to the Last LN as MPVG. By comparing PVG and MPVG, we demonstrate that MPVG effectively maximizes PE and that maintaining the counterbalancing directionality of PE is beneficial for vision transformers. Our experiments validate our hypothesis and demonstrate that MPVG outperforms other methods. The results demonstrate that MPVG consistently performs well for vision transformers.

In this paper, our contributions are as follows:

1. We propose a simple yet effective method called MPVG, which maximizes the effect of PE in the GAP method. We show that MPVG leads to better performance in vision transformers.
2. We provide an analysis of the phenomenon observed in PE when using the Layer-wise structure and offer insights into the role of PE.
3. Through experiments, we verify that MPVG is generally

effective for vision transformers on various tasks such as image classification, semantic segmentation, and object detection.

Related Work

Vision Transformers

The vision transformer design is adapted from Transformer (Vaswani et al. 2017), which was designed for natural language processing (NLP). This adaptation makes it suitable for computer vision tasks such as image classification (Dosovitskiy et al. 2020; Touvron et al. 2021; Liu et al. 2021), object detection (Carion et al. 2020; Zhu et al. 2020), and semantic segmentation (Zheng et al. 2021; Wang et al. 2021; Strudel et al. 2021).

Class Token & Global Average Pooling ViT (Dosovitskiy et al. 2020) conducts ablation studies comparing the class token and GAP. Additionally, there are other studies on the use of GAP and class tokens in vision transformers (Raghu et al. 2021; Chu et al. 2021b). Studies such as CeiT (Yuan et al. 2021a) and T2T-ViT (Yuan et al. 2021b) use class token, while others like Swin Transformer (Liu et al. 2021) and CPVT (Chu et al. 2021b) adapt GAP. CPVT achieves performance improvements by using GAP instead of the class token. Although the class token is not inherently translation-invariant, it can become so through training. By adopting GAP, which is inherently translation-invariant, better improvements in image classification tasks are achieved (Chu et al. 2021b). Furthermore, GAP results in even less computational complexity because it eliminates the need to compute the attention interaction between the class token and the image patches.

Position Embeddings in Vision Transformers

Absolute Position Embedding In the transformer, absolute position embedding is generated through a sinusoidal function and added to the input token embedding (Vaswani et al. 2017). The sinusoidal functions are designed to give the position embedding locally consistent similarity, which helps vision transformers focus more effectively on tokens that are close to each other in the input sequence. This local consistency enhances the model's ability to capture spatial relationships and patterns (Vaswani et al. 2017).

Besides sinusoidal positional embedding, position embedding can also be learnable. Learnable position embedding is created through training parameters, which are initialized with a fixed-dimensional tensor and updated along with the model's parameters during training. Recently, many models have adopted absolute position embedding due to their effectiveness in encoding positional information (Dosovitskiy et al. 2020; Touvron et al. 2021; Liu et al. 2021).

Relative Position Embedding In addition to absolute position embedding, there is also relative position embedding (Shaw, Uszkoreit, and Vaswani 2018). Relative PE encodes the relative position information between tokens. The first to propose relative PE in computer vision was (Shaw, Uszkoreit, and Vaswani 2018). Furthermore, (Bello et al.

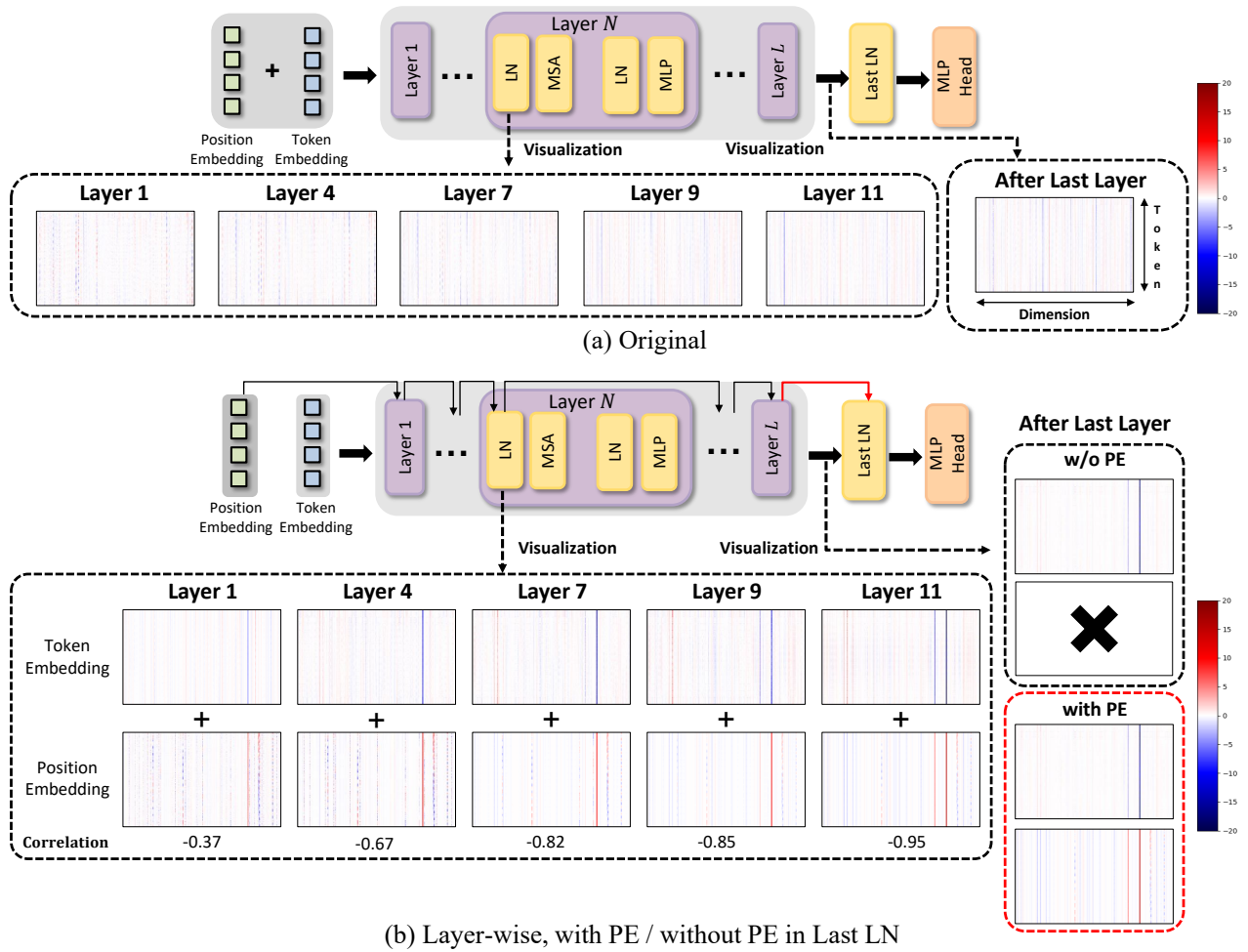


Figure 2: The heatmaps depict the characteristics of each layer in both the original structure and the Layer-wise structure with the GAP method. For the Layer-wise structure, the heatmaps illustrate cases both with and without PE in the Last LN. For each heatmap based on DeiT-Ti, the x-axis represents the dimension of DeiT-Ti (256), and the y-axis represents the number of tokens (196). In both (a) and the top row (token embedding) of (b), the heatmaps represent the average value of token embedding in each layer, while the bottom row of (b) shows the heatmap of PE. The correlation in (b) refers to the correlation coefficient between token embedding and position embedding.

2019) proposed a 2-D relative position encoding for image classification that showed superior performance compared to traditional 2-D sinusoidal embedding. This relative encoding captures spatial relationships between tokens more effectively. In related research, iRPE (Wu et al. 2021) improves relative PE by incorporating query interactions and relative distance modeling in self-attention. RoPE (Heo et al. 2024) introduces flexible sequence lengths, decaying inter-token dependency, and relative position encoding in linear self-attention.

Methodology

In this section, we first explain the absolute position embedding and then provide a detailed overview of the Layer-wise structure (Yu et al. 2023). Next, we introduce PVG, an improved Layer-wise structure, along with MPVG, which

effectively leverages the characteristics of PE in the Layer-wise structure.

Preliminary: Absolute Position Embedding

The method of absolute position embedding used in vision transformers is as follows. As shown in Fig 3-(a), PE is added to the token embedding before they are input into the layer. This can be expressed as follows:

$$x_0 = [x_{cls}; p^1; p^2; \dots p^N;] + pos, \quad (1)$$

where p and pos represent the patch and position embedding, respectively. N represents the number of patches, calculated as HW/P^2 , where H and W are the height and width of the image, and $P \times P$ is the resolution of each patch. The combined token embedding and PE, denoted as x , can be expressed in a layer as follows:

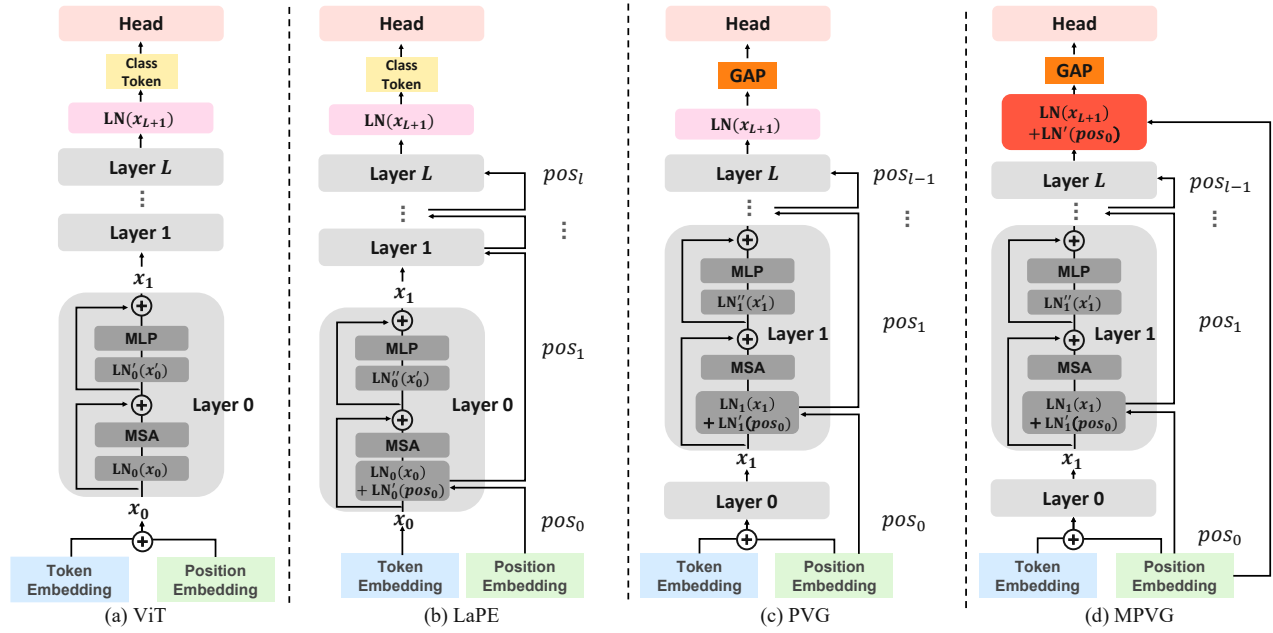


Figure 3: The overview of the various methods. (a) ViT. (b) LaPE (Yu et al. 2023). (c) PVG, an improved Layer-wise structure. Specifically, we adopt a structure where the token embedding and PE are added before entering layer 0 and a hierarchical structure for delivering PE, excluding layer 0. (d) MPVG. The main difference from PVG is whether the initial PE is delivered to the Last LN.

$$x'_l = \text{MSA}(\text{LN}_l(x_l)) + x_l \quad (l = 0 \dots L), \quad (2)$$

$$x_{l+1} = \text{MLP}(\text{LN}'_l(x'_l)) + x'_l \quad (l = 0 \dots L), \quad (3)$$

$$y = \text{LN}(x_{L+1}) \quad (4)$$

where LN, LN', and LN'' represent different Layer Normalizations, Multi-head Self-Attention is denoted as MSA, and Multi-Layer Perceptron is denoted as MLP. x_{L+1} refers to the value after passing through the last layer L .

Preliminary: Layer-wise Structure

LaPE (Yu et al. 2023) points out problems with the joining method that position embedding and token embedding in the vision transformers. As shown in Eq. (1), when PE is added to the token embedding before the first layer, and the same LN is applied to both the token embedding and PE as in Eq. (2), they share the same affine parameters in LN. This method limits the expressiveness of PE. Therefore, the Layer-wise structure is used to resolve these problems. This can be expressed as follows:

$$x_0 = [x_{\text{cls}}; p^1; p^2; \dots p^N], \quad (5)$$

$$x'_l = \text{MSA}(\text{LN}_l(x_l) + \text{LN}'_l(pos_l)) + x_l \quad (6)$$

Eq. (1) is modified to Eq. (5) and Eq. (2) to Eq. (6). In Eq. (6), the Layer-wise structure uses independent LN for token embedding(x) and PE. PE is delivered in each layer as follows :

$$\begin{cases} pos_0 = pos \\ pos_l = \text{LN}'_{l-1}(pos_{l-1}) \quad (l = 1 \dots L) \end{cases} \quad (7)$$

Maximizing the Position Embedding with GAP

In this section, we propose two methods, MPVG and PVG, to validate our hypothesis. In Fig 2-(b), we observed that, in Layer-wise structure, the effect of PE in counterbalancing the values of token embedding(x) becomes more pronounced as the layers deepen, as evidenced by the correlation between the two. However, in Layer-wise structure, although the directionality of the token embedding is maintained outside the layer, there is no PE to counterbalance that value. Therefore, we validate our hypothesis by comparing MPVG, which delivers PE to the Last LN, with PVG, which does not.

We remove the class token as we adapt the Global Average Pooling (GAP) method. Although we use the Layer-wise structure, we modify specific details. Specifically, we combine two structural approaches: (1) adding token embedding and PE before inputting the layer. (2) delivering PE to each layer except the 0th layer. We call this method as PVG. In PVG method, as shown in Figure 3-(c), is as follows:

$$x_0 = [p^1; p^2; \dots p^N] + pos, \quad (8)$$

$$x'_l = \begin{cases} \text{MSA}(\text{LN}_0(x_0)) + x_0 & \text{if } l = 0 \\ \text{MSA}(\text{LN}_l(x_l) + \text{LN}'_l(pos_{l-1})) + x_l & \text{if } 1 \leq l \leq L \end{cases} \quad (9)$$

$$\begin{cases} pos_0 = pos \\ pos_l = \text{LN}'_l(pos_{l-1}) \quad (l = 1 \dots L) \end{cases} \quad (10)$$

The subsequent process is the same as in Eq. (3) and Eq. (4). In MPVG, we modify Eq. (4) as follows after going through the process of PVG:

$$y = \text{LN}(x_{L+1}) + \text{LN}'(pos_0) \quad (11)$$

To verify whether maintaining the counterbalance effect of PE is beneficial, we deliver PE to the Last LN in PVG, as shown in Eq. (11). We refer to this method as MPVG. In the next section, we verify the superiority of MPVG by comparing the two methods. Also, we show that MPVG outperforms previous approaches through experiments across various vision transformers and datasets.

Experiment

Training Settings All experiments are conducted on an RTX 4090 with 4 GPUs using AdamW optimizer (Loshchilov and Hutter 2019), while DeiT-B is trained on an RTX 4090 with 8 GPUs.

Image Classification

We evaluate the performance of our methods on ImageNet-1K (Deng et al. 2009) and CIFAR-100 (Krizhevsky, Hinton et al. 2009). On ImageNet-1K, we conduct experiments with DeiT (Touvron et al. 2021), Swin (Liu et al. 2021), CeiT (Yuan et al. 2021a), and T2T-ViT (Yuan et al. 2021b). In the case of Swin, due to its staged architecture that generates hierarchical representations with the same feature map resolution as convolutional networks, both PVG and MPVG exceptionally include layer 0. All vision transformers are trained on 224×224 resolution images for 300 epochs, except T2T-ViT-7, which is trained for 310 epochs.

On CIFAR-100, we conduct experiments using ViT-Lite (Hassani et al. 2022) and T2T-ViT-7 (Yuan et al. 2021b). ViT-Lite was trained for 310 epochs on 32×32 resolution images with a batch size of 128. In the case of T2T-ViT-7, we transfer our pretrained T2T-ViT to downstream datasets such as CIFAR-100 and finetune the pretrained T2T-ViT-7 for 60 epochs with a batch size of 128.

As shown in Table 1, For MPVG, the performance on DeiT-Ti improved from 72.14% to 73.51%, representing an increase of approximately 1.37%. For DeiT-S, the performance improved from 79.81% to 80.61%, an increase of approximately 0.80%. Additionally, there were performance improvements of 0.57% in DeiT-B, 0.27% in Swin-Ti, 0.58% in CeiT, and 0.52% in T2T-ViT. Overall, MPVG outperforms the existing methods in all cases. Moreover, we confirm that MPVG consistently demonstrates superior performance compared to PVG across various vision transformers.

As shown in Table 2, MVPG achieves overall performance improvements on CIFAR-100. Specifically, MPVG improves the performance of ViT-Lite by 1.97%, from 74.90% to 76.87%, and enhances the performance of T2T-ViT-7 by 0.29% over the default. Additionally, MPVG

Model	Method	#Params (M)	Top-1 Acc (%)
DeiT-Ti (Touvron et al. 2021)	Default	5.717	72.14
	LaPE	5.721	72.94
	PVG	5.721	73.17
	MPVG	5.721	73.51
DeiT-S (Touvron et al. 2021)	Default	22.050	79.81
	LaPE	22.059	80.39
	PVG	22.058	80.38
	MPVG	22.059	80.61
DeiT-B (Touvron et al. 2021)	Default	86.567	81.85
	LaPE	86.586	82.15
	PVG	86.583	82.21
	MPVG	86.584	82.42
Swin-Ti (Liu et al. 2021)	Default	28.589	81.37
	LaPE	28.599	81.48
	PVG	28.598	81.52
	MPVG	28.599	81.64
CeiT-Ti (Yuan et al. 2021a)	Default	6.356	76.62
	LaPE	6.361	76.89
	PVG	6.361	77.14
	MPVG	6.361	77.20
T2T-ViT-7 (Yuan et al. 2021b)	Default	4.310	71.76
	LaPE	4.313	72.01
	PVG	4.312	71.91
	MPVG	4.313	72.28

Table 1: Top-1 accuracy comparison with various methods, using DeiT-T, DeiT-S, DeiT-B, Swin-Ti, CeiT-Ti, T2T-ViT-7 on ImageNet-1K.

Model	Method	#Param (M)	Top-1 Acc (%)
ViT-Lite (Hassani et al. 2022)	Default	3.740	74.90
	LaPE	3.744	75.52
	PVG	3.742	76.67
	MPVG	3.743	76.87
T2T-ViT-7 (Yuan et al. 2021b)	Default	4.078	83.22
	LaPE	4.082	83.41
	PVG	4.081	83.39
	MPVG	4.081	83.51

Table 2: Top-1 accuracy comparison with various methods, using ViT-Lite and T2T-ViT-7 on CIFAR-100. In the case of T2T-ViT, the results are based on fine-tuning the pretrained model on the downstream dataset, CIFAR-100.

shows a 0.2% and 0.12% improvement over PVG for ViT-Lite and T2T-ViT-7, respectively. Overall, MPVG outperforms existing methods across all cases on CIFAR-100.

Object Detection

On object detection, we evaluate our methods on COCO 2017 (Lin et al. 2014). To demonstrate the effectiveness of

Model	Pre-trained	Method	AP ^{box} / AP ^{mask}
ViT-Adapter-Ti	DeiT-Ti	Default	45.9 / 41.0
		LaPE	46.2 / 41.2
		PVG	46.1 / 41.2
		MPVG	46.5 / 41.4

Table 3: Performance comparison of Object Detection on COCO2017. For comparison, DeiT-Ti model pretrained on ImageNet-1K with each method is used.

Model	Pre-trained	Method	mIoU
ViT-Adapter-Ti	DeiT-Ti	Default	40.55
		LaPE	41.42
		PVG	41.07
		MPVG	41.69

Table 4: Performance comparison of Semantic Segmentation on ADE20K. For comparison, DeiT-Ti model pretrained on ImageNet-1K with each method is used.

our method on object detection tasks, we select the ViT-Adapter-Ti (Chen et al. 2022) model based on Mask R-CNN (He et al. 2017) in MMDetection framework (Chen et al. 2019). Additionally, we use the default settings and train it for 36 epochs using the 3x+MS(multi-scale training) schedule. As shown in Table 3, MPVG achieves improvements of +0.6 in box AP and +0.5 in mask AP compared to the default setting. MPVG, in particular, demonstrates superior performance with an increase of +0.5 in box AP and +0.4 in mask AP over PVG.

Semantic Segmentation

On semantic segmentation, we evaluate our methods on ADE20K (Zhou et al. 2019). We select the ViT-Adapter-Ti (Chen et al. 2022) model based on UperNet (Xiao et al. 2018) in MMsegmentation framework (Contributors 2020) and train it using the default settings. As shown in Table 4, MPVG achieves an improvement of +1.14 mIoU compared to the default. Furthermore, MPVG outperforms PVG, achieving a performance improvement of +0.62 mIoU.

Analysis

Through experiments on image classification, object detection, and semantic segmentation, we demonstrate the effectiveness of MPVG. In all tasks, MPVG not only outperforms the baseline but also achieves the best performance among all methods. This validates our hypothesis and proves that our method is an effective approach to maximizing PE in the GAP method. Fig 4 shows that in Layer-wise structure, token embedding and position embedding exhibit increasingly opposing directions as the layers deepen. This suggests that PE not only provides positional information in the initial layers but also may play a counterbalancing role that becomes more pronounced in deeper layers. To further explore this, we compare PVG and MPVG to confirm that PE has a counterbalancing effect. This comparison proves that maintaining the counterbalancing role of PE impacts the per-

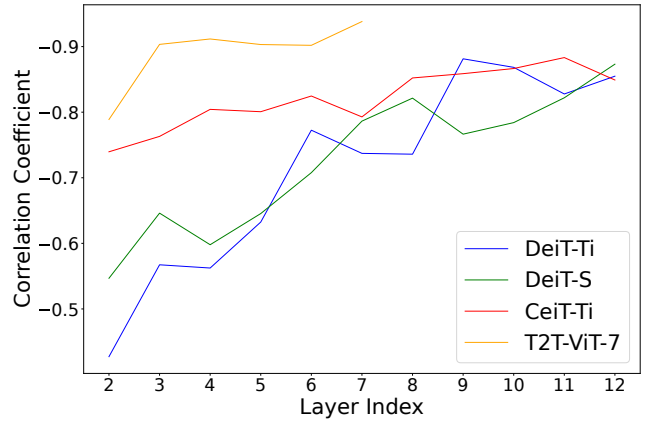


Figure 4: Correlation coefficient between token embedding and position embedding in Layer-wise. Each token embedding and position embedding is based on the values after applying LN. DeiT-Ti, DeiT-S, and CeiT-Ti each have a total of 12 layers, but T2T-ViT-7 has 7 layers.

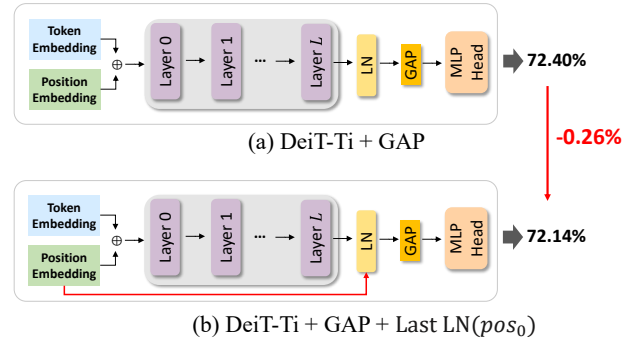


Figure 5: Comparison of two methods on DeiT-Ti. (a) Structure with only GAP applied, showing 72.40% performance; and (b) Structure with GAP and position embedding added to the Last LN in a non-Layer-wise structure, also showing 72.14% performance.

formance of vision transformers.

In conclusion, several key points can be identified: (1) In the initial layers, PE primarily provides positional information, enabling the model to understand the spatial relationships between tokens. However, as the layers deepen, PE plays a role in counterbalancing the token embedding. (2) This counterbalancing effect of PE has a significant impact on the performance of vision transformers. Therefore, MPVG demonstrates that maintaining this direction is beneficial for vision transformers and proves that PE can perform additional roles to sustain this effect.

Effect of PE in Last LN

We conduct additional experiments to validate our hypothesis. Specifically, we aim to confirm that adding PE to the Last LN effectively maintains the counterbalancing role of PE in Layer-wise structure. We compare the method using only GAP with the method that adds PE to the Last LN in

Model	PE Method	Last LN	Top-1 Accuracy (%)
DeiT-Ti	MPVG	pos_{11}	73.30
		pos_8	73.38
		pos_5	73.39
		pos_0	73.51

Table 5: Comparison of the value of PE added to the Last LN in MPVG. pos_0 refers to the initial position embedding, and pos_{11} represents the position embedding after applying LN in the last layer. pos_N ($=LN'_N(pos_{N-1})$) indicates the PE input for the $(N + 1)$ th layer.

a non-Layer-wise structure while using GAP. We perform these experiments with DeiT-Ti on ImageNet-1K.

In Fig 5, we compare (a), where only GAP is applied, with (b), where PE is delivered to the Last LN in a non-Layer-wise structure with GAP. Fig 5-(a) shows a 72.40% performance, while Fig 5-(b) shows a decreased performance of 72.14%. This indicates that adding PE to the Last LN is only effective in Layer-wise structure where PE is delivered to each layer. In other words, these experimental results prove that in Layer-wise structure, the token embedding contains values that are counterbalanced by PE. Specifically, in Fig 5-(b) structure, the token embedding that passes through the layers does not contain the directional values, which is counterbalanced by PE. Thus, adding PE to the Last LN not only has no effect but actually leads to a decrease in performance. As a result, as shown in Fig 2-(b), in Layer-wise structure, the token embedding progresses while retaining values that are meant to be counterbalanced by PE, but after passing through the final layer, this directional value is not adequately compensated by PE. This proves that PE is necessary to perform this additional counterbalancing role in the Last LN.

Ablation Study

The impact of PE values delivered to the Last LN We conduct experiments to investigate the impact of varying the PE values passed to the Last LN in MPVG. In Table 5, pos_N represents the value of $LN'_N(pos_{N-1})$. Since MPVG does not deliver PE to layer 0, N ranges from 1 to $L - 1$, where L is the number of layers. Experiments show that MPVG consistently outperforms PVG, which achieves a performance of 73.17%, regardless of the PE values passed to the Last LN. This suggests that delivering PE in the Last LN has a significant positive impact on the performance of vision transformers. Furthermore, it demonstrates that maintaining the role of PE in the Last LN is generally effective. MPVG adopts pos_0 , which shows the best performance by comparing various PE values delivered to the Last LN.

Structural differences in MPVG We also experiment by varying the architecture structure in MPVG. Table 6 presents the ablations for the differences in architecture within MPVG. The experiments are conducted on DeiT-Ti using the ImageNet-1K. Through this experiment, we adopt an improved Layer-wise structure that differs from the conventional Layer-wise structure.

Model	Layer 0	Structure Hierarchical	($x+PE$)	Top-1 Acc (%)
MPVG	X	✓	✓	73.31
	✓	X	✓	73.48
	✓	✓	X	73.28
	✓	✓	✓	73.51

Table 6: Structural Differences in MPVG. "Layer 0" denotes whether layer 0 is included when delivering PE to layers. "Hierarchical" denotes whether pos_l is $LN'_l(pos_{l-1})$ or $LN'_l(pos_0)$. "($x+PE$)" denotes whether PE is added to the token embedding(x) before entering layer 0 or not.

Specifically, we conduct comparative experiments on three structural differences: (1) Our methods exclude layer 0 when delivering PE. Through our experiments, we find that delivering PE to layer 0, which was previously included, is not only unnecessary but also improves the performance of vision transformers when excluded. (2) We add PE to the token embedding before it enters layer 0. Unlike LaPE where token embedding x and PE are separated before entering the first layer, our methods add PE to x before entering layer 0. This structure does not limit the expressiveness of PE because independent LN is applied to both the token embedding and PE in each layer, and PE is delivered in a Layer-wise structure. Moreover, the ($x+PE$) structure boosts performance by approximately 0.23%. (3) We observe that the performance is similar between hierarchical and non-hierarchical structures. However, in non-hierarchical structures, performance often declines in small or large vision transformers due to overfitting (Yu et al. 2023). Through Table 6, we demonstrate that our methods represent the optimal structure in the Layer-wise structure.

Conclusion

We reveal that position embedding can play additional roles in vision transformers using the GAP method. Specifically, in a Layer-wise structure, PE has a counterbalancing effect on the values of token embedding, and maintaining this directional balance by PE is beneficial for vision transformers. Based on these observations, we propose a simple yet effective method, MPVG. MPVG utilizes the characteristics of PE observed in the Layer-wise structure to maximize the PE. Through extensive experiments, we demonstrate that MPVG is generally effective on vision transformers, outperforming previous methods. However, MPVG has a potential limitation in that it is incompatible with the class token method. Through these limitations, we will further explore the broader applicability of MPVG and the effects of PE's counterbalancing as part of our future work. In this paper, we demonstrate that MPVG effectively addresses the issues arising in GAP and Layer-wise structures, providing a significantly more meaningful approach. Through this, we look forward to MPVG offering a broader perspective on position embedding.

Acknowledgments

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2024-RS-2023-00258649, 80%) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation) and was partly supported by the IITP grant funded by the Korea government (MSIT) (No.RS-2022-00143524, Development of Fundamental Technology and Integrated Solution for Next-Generation Automatic Artificial Intelligence System) and (No.RS2023-00225630, Development of Artificial Intelligence for Text-based 3D Movie Generation).

References

- Bello, I.; Zoph, B.; Vaswani, A.; Shlens, J.; and Le, Q. V. 2019. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3286–3295.
- Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.
- Chang, S.; Wang, P.; Lin, M.; Wang, F.; Zhang, D. J.; Jin, R.; and Shou, M. Z. 2023. Making vision transformers efficient from a token sparsification view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6195–6205.
- Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; Zhang, Z.; Cheng, D.; Zhu, C.; Cheng, T.; Zhao, Q.; Li, B.; Lu, X.; Zhu, R.; Wu, Y.; Dai, J.; Wang, J.; Shi, J.; Ouyang, W.; Loy, C. C.; and Lin, D. 2019. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv preprint arXiv:1906.07155*.
- Chen, Z.; Duan, Y.; Wang, W.; He, J.; Lu, T.; Dai, J.; and Qiao, Y. 2022. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*.
- Chu, X.; Tian, Z.; Wang, Y.; Zhang, B.; Ren, H.; Wei, X.; Xia, H.; and Shen, C. 2021a. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in neural information processing systems*, 34: 9355–9366.
- Chu, X.; Tian, Z.; Zhang, B.; Wang, X.; and Shen, C. 2021b. Conditional positional encodings for vision transformers. *arXiv preprint arXiv:2102.10882*.
- Contributors, M. 2020. MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark. <https://github.com/open-mmlab/mms Segmentation>.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Hassani, A.; Walton, S.; Shah, N.; Abuduweili, A.; Li, J.; and Shi, H. 2022. Escaping the Big Data Paradigm with Compact Transformers. *arXiv:2104.05704*.
- He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.
- Heo, B.; Park, S.; Han, D.; and Yun, S. 2024. Rotary position embedding for vision transformer. *arXiv preprint arXiv:2403.13298*.
- Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.
- Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, 740–755. Springer.
- Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.
- Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
- Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; and Dosovitskiy, A. 2021. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34: 12116–12128.
- Shaw, P.; Uszkoreit, J.; and Vaswani, A. 2018. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*.
- Strudel, R.; Garcia, R.; Laptev, I.; and Schmid, C. 2021. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7262–7272.
- Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357. PMLR.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, Y.; Xu, Z.; Wang, X.; Shen, C.; Cheng, B.; Shen, H.; and Xia, H. 2021. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8741–8750.
- Wu, K.; Peng, H.; Chen, M.; Fu, J.; and Chao, H. 2021. Rethinking and improving relative position encoding for vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10033–10041.
- Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; and Sun, J. 2018. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, 418–434.

- Xu, H.; Xiang, L.; Ye, H.; Yao, D.; Chu, P.; and Li, B. 2024. Permutation Equivariance of Transformers and Its Applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5987–5996.
- Yu, R.; Wang, Z.; Wang, Y.; Li, K.; Liu, C.; Duan, H.; Ji, X.; and Chen, J. 2023. LaPE: Layer-adaptive position embedding for vision transformers with independent layer normalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5886–5896.
- Yuan, K.; Guo, S.; Liu, Z.; Zhou, A.; Yu, F.; and Wu, W. 2021a. Incorporating convolution designs into visual transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 579–588.
- Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.-H.; Tay, F. E.; Feng, J.; and Yan, S. 2021b. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, 558–567.
- Zhai, X.; Kolesnikov, A.; Houlsby, N.; and Beyer, L. 2022. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12104–12113.
- Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P. H.; et al. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6881–6890.
- Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; and Torralba, A. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127: 302–321.
- Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.