JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Masked Jigsaw Puzzle : A Versatile Position Embedding for Vision Transformers

Bin Ren[1,2*], Yahui Liu[2*], Yue Song[2], Wei Bi[3], Rita Cucchiara[4], Nicu Sebe[2], Wei Wang[5†]
([*]:Equal Contribution. [†]Corresponding author)

Tag: THU-AM-371
Paper ID: 1672

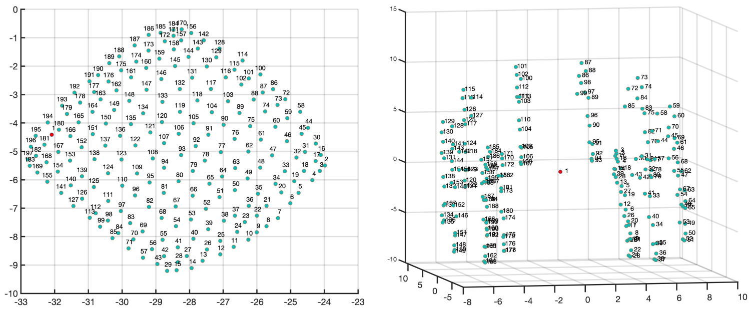
Code: https://github.com/yhlleo/MJP
E-Mail: bin.ren@unitn.it; yahui.cvrs@gmail.com

[1]University of Pisa, Italy. [2]University of Trento, Italy. [3]Tencent AI Lab , China.
[4]University of Modena and Reggio Emilia, Italy. [5]Beijing Jiaotong University, China

Masked Jigsaw Puzzle : A Versatile Position Embedding for Vision Transformers

Bin Ren[1,2*], Yahui Liu[2*], Yue Song[2], Wei Bi[3], Rita Cucchiara[4], Nicu Sebe[2], Wei Wang [5†]
(*:Equal Contribution. †:Corresponding author)
[1]University of Pisa, Italy. [2]University of Trento, Italy. [3]Tencent AI Lab , China.
[4]University of Modena and Reggio Emilia, Italy. [5]Beijing Jiaotong University, China

What do Position Embeddings (PEs) learn in ViTs?



(a) UMAP (b) PCA

Low-dimensional projection of PEs from DeiT-S[1]

- The 2D spatial relationship of image patches
- The spatial relationship learned in the high-dimensional space still manifests in the low-dimensional space

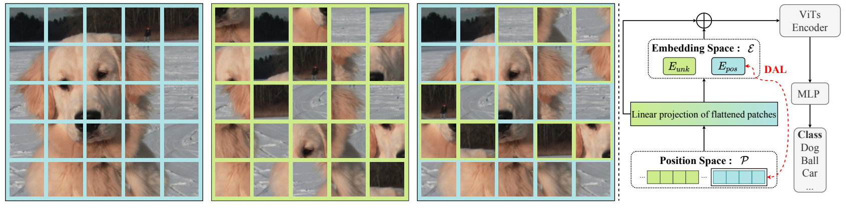
What do the spatial relation bring for vision tasks?

- **Accuracy Increase** for Classification Problem [2]
- **Privacy Leakage** under Gradient Attack [3]
- **Consistency Drop** when facing transformed input [4]

How to alleviate the conflict in PEs?

- Removing PEs? No!
- Training ViTs with all image patches naively shuffled ? No!
- The proposed Masked Jigsaw Puzzle (**MJP**) PEs? **Yes!**

Simple yet Effective technique: MJP



(a) $\mathbf{x}$ (b) $\tilde{\mathbf{x}}$ ($\gamma = 1$) (c) $\tilde{\mathbf{x}}$ ($\gamma = 0.4$) (d) Overview of our MJP

- Step1: Block-wise Random Selection
- Step2: Jigsaw Puzzle Shuffling
- Step3: An **un-known** PEs to the shuffled patches
- Step4: Dense Absolute Localization (**DAL**) for the unshuffled patches

Experiments:

- **Regular ImageNet-1K Training**

Method	Param.	Top-1 Acc. ↑	Diff. Norm. ↓	Consistency ↑
ResNet-50 [17]	25	79.3	11.77	51.5
ResNet-50 + MJP	25	79.4	7.11	69.3
DeiT-S [40]	22	79.8	16.21	64.3
DeiT-S + MJP	22	80.5	8.96	82.9
Swin-T [31]	29	81.3	15.49	41.5
Swin-T + MJP	29	81.3	12.36	66.9

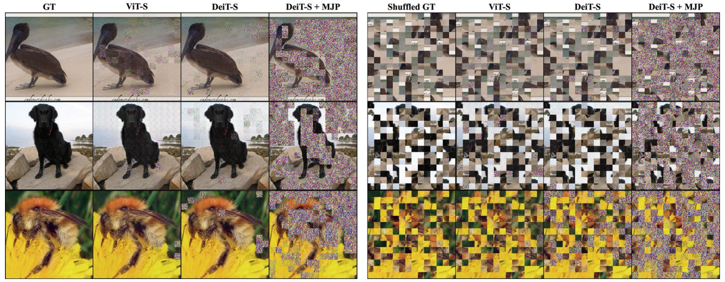
- **Robustness on Challenging Sets**

Method	ImageNet-C	ImageNet-A		ImageNet-O
	mCE ↓	Acc ↑	AURRA ↑	AUPR ↑
DeiT-S	54.6	19.2	25.1	20.9
DeiT-S + MJP	**51.6**	**21.6**	**29.8**	**22.6**

- **Privacy Preservation**

	Model	Set.	Acc. ↑	MSE ↑	FFT_{2D} ↑	PSNR ↓	SSIM ↓	LPIPS ↑
(1)	ViT-S [8]	a	78.1	.0278	.0039	19.27	.5203	.3623
	DeiT-S [40]		79.8	.0350	.0057	18.94	.5182	.3767
	DeiT-S (w/o PEs)		77.5	.0379	.0082	20.22	.5912	.2692
	DeiT-S+MJP		**80.5**	**.1055**	**.0166**	**11.52**	**.4053**	**.6545**
(2)	ViT-S [8]	b	18.7	.0327	.0016	18.44	.6065	.2836
	DeiT-S [40]		36.0	.0391	.0024	17.60	.5991	.3355
	DeiT-S (w/o PEs)		**77.5**	.0379	.0025	20.25	.6655	.2370
	DeiT-S+MJP		62.9	**.1043**	**.0059**	**11.66**	**.4493**	**.6519**
(3)	DeiT-S+MJP (w/o)	a	40.6	.1043	.0059	11.66	.4493	.6519
(4)	DeiT-S+MJP	c	62.9	.1706	.0338	8.07	.0875	.8945

(a) $\phi(\nabla\mathcal{M}(\mathbf{x}), \mathbf{x})$ (b) $\phi(\nabla\mathcal{M}(\tilde{\mathbf{x}}), \tilde{\mathbf{x}})$ (c) $\phi(\nabla\mathcal{M}(\tilde{\mathbf{x}}), \mathbf{x})$



Summary/Conclusion

- The concrete 2D spatial relation of image patches learned in the high-dimensional position embedding is **visually** demonstrated
- PEs bring conflict among accuracy, privacy, and consistency (i.e., position-insensitive property, robustness) in vision task
- The proposed MJP is able for preserving the consistency versus maintaining the accuracy
- **Models and Code are publicly available:** https://github.com/yhlleo/MJP

References:
[1] Touvron, Hugo, et al. "Training data-efficient image transformers & distillation through attention." *ICML*2021.
[2] Dosovitskiy, Alexey, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." *ICLR*2020.
[3] Lu, Jiahao, et al. "April: Finding the achilles' heel on privacy for vision transformers.." *CVPR*2022.
[4] Xie, Qizhe, et al. "Unsupervised data augmentation for consistency training." *NeurIPS*2020.

Motivations

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

- **Position Embeddings (PEs) in Vision Transformers (ViTs)**

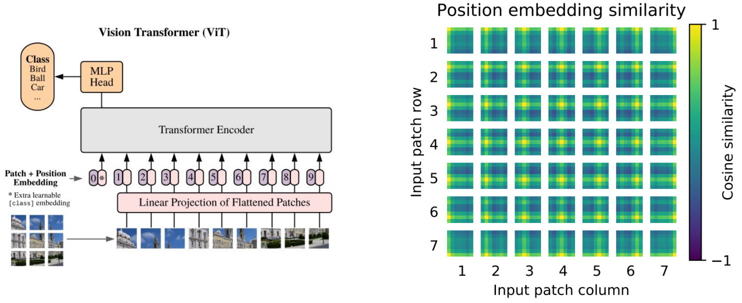
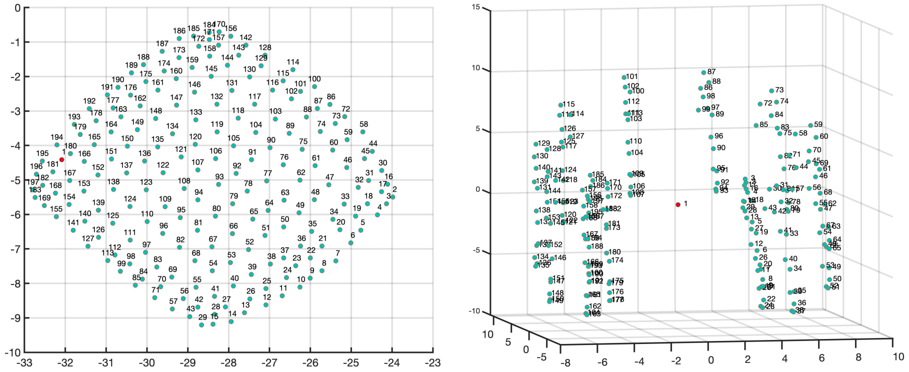


Figure 1: Vision Transformer (ViT) and the similarity of PEsof ViT-L/32[1].

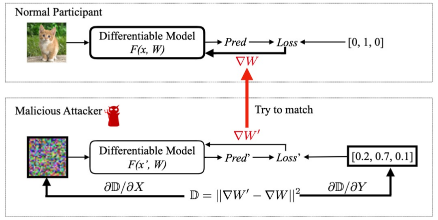
- **Q1: What do Position Embeddings (PEs) learn in ViTs?**



(a) UMAP (b) PCA

Figure 3. Low-dimensional projection of position embeddings from DeiT-S[2]. (a) The 2D UMAP projection, (b) The 3D PCA projection.

- **Q2: What do the PEs bring for vision tasks?**
  - **Accuracy Increase** for Classification Tasks[1]
  - **Privacy Leakage** under Gradient Attack in Federated Learning (FL)[3]
  - **Consistency Drop** when facing transformed input data[4]



$$\mathbb{E}_{\mathbf{x}}[\mathbb{1}\{\arg\max P(\mathcal{F}(\mathbf{x})) = \arg\max P(\mathcal{F}(\tilde{\mathbf{x}}))\}].$$

Gradient Attack Consistency

- **Observations**
  - The <u>2D spatial relationship</u> of image patches
  - The spatial relation learned in the high-dimensional space still <u>manifests</u> in the low-dimensional space
  - The learned spatial relation from PEs brings conflict among accuracy, privacy, and consistency

- **Goal:** Alleviate the conflict by improving the consistency (robustness, position-insensitive, safety) of ViTs without hurting the regular performance (i.e., accuracy)

Masked Jigsaw Puzzle (MJP)

[1] Dosovitskiy, Alexey, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." *ICLR*2020.
[2] Touvron, Hugo, et al. "Training data-efficient image transformers & distillation through attention." *ICML*2021.
[3] Lu, Jiahao, et al. "April: Finding the achilles' heel on privacy for vision transformers.." *CVPR*2022.
[4] Xie, Qizhe, et al. "Unsupervised data augmentation for consistency training." *NeurIPS*2020.

Introduction

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

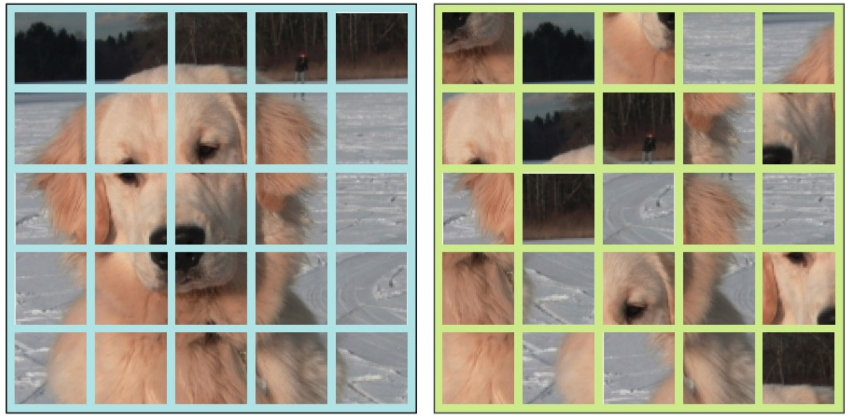
- **Alleviating the Conflict Brought by PEs**

➢ Removing PEs in the ViTs ? NO!

Method	Top-1 Acc. ↑	Consistency ↑
A: DeiT-S [40]	79.8	64.3
B: A - PEs	77.5 (-2.3)	100.0

The Consistency is evaluated by: $\mathbb{E}_{\mathbf{x}}[\mathbb{1}\{\arg\max P(\mathcal{F}(\mathbf{x})) = \arg\max P(\mathcal{F}(\tilde{\mathbf{x}}))\}]$.

➢ Training ViTs with all the image patches naively shuffled ? NO!

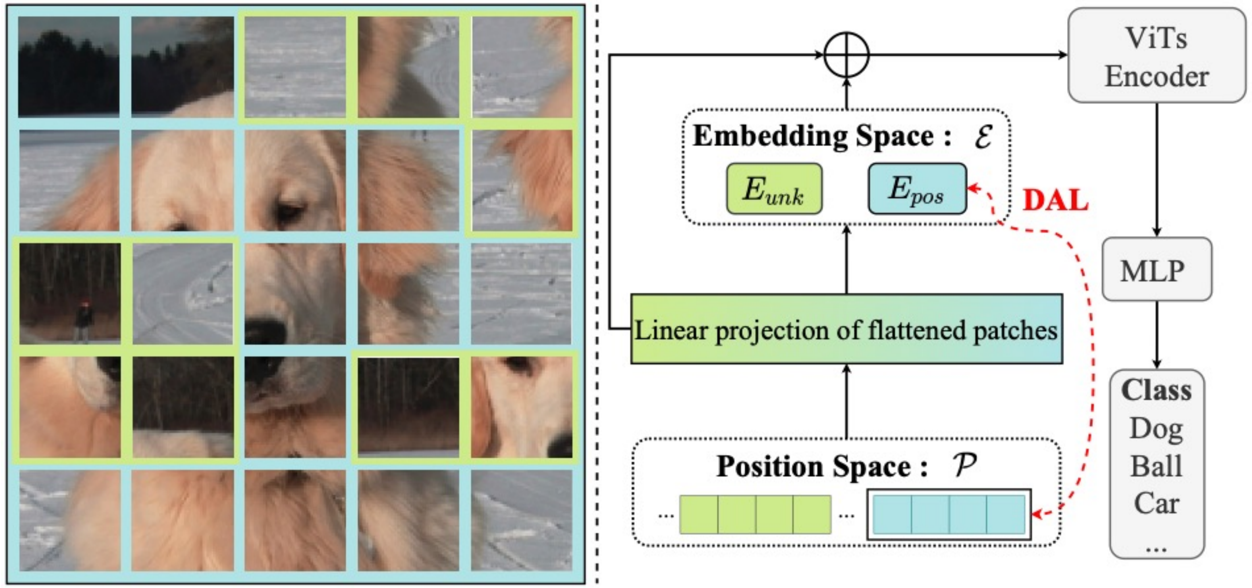


(a) $\mathbf{x}$ (b) $\tilde{\mathbf{x}}\ (\gamma = 1)$

Accuracy Marginally Drop

- **The proposed Masked Jigsaw Puzzle (MJP) PEs** YES!

Adjusting Input Data & PEs



(c) $\tilde{\mathbf{x}}\ (\gamma = 0.4)$ (d) Overview of our MJP

[40] Touvron, Hugo, et al. "Training data-efficient image transformers & distillation through attention." *ICML*2021.

The proposed MJP

- **Adjust The Input Data**

Algorithm 1 Block-wise Random Jigsaw Puzzle Shuffle

Input: Input image: $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$;
Shuffle Ratio: γ;
Patch Size: P

Output: Shuffled image patches: $\widetilde{\boldsymbol{x}}_p$

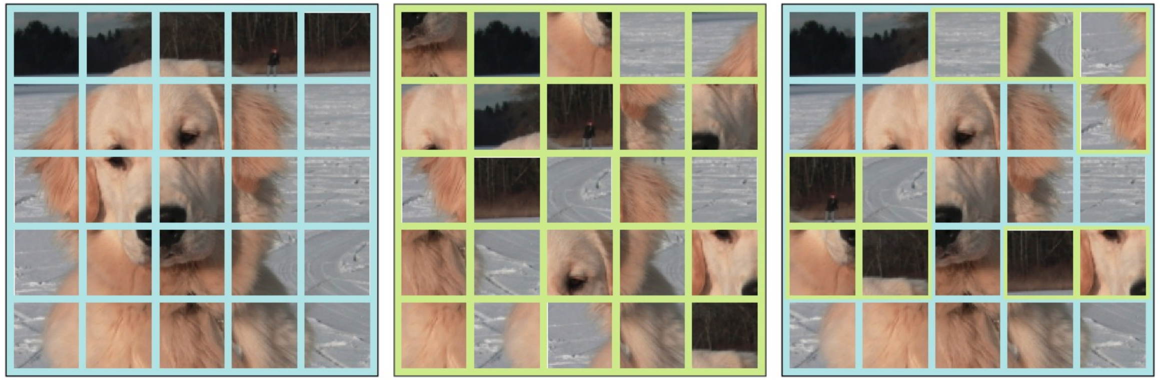
1: $\mathbf{x}_p \in \mathbb{R}^{N \times (P^2 \cdot C)} \longleftarrow Patchlize(\mathbf{x}, P)$

2: $\mathbf{m} \in \mathbb{R}^{\frac{H}{P} \times \frac{W}{P}} \longleftarrow BinaryInitialize(\mathbf{x}_p, 0)$

3: $\widetilde{\mathbf{m}} \in \mathbb{R}^{\frac{H}{P} \times \frac{W}{P}} \longleftarrow BlockwiseMask(\mathbf{m}, \gamma)$ [28]

4: $\widetilde{\boldsymbol{x}}_p \in \mathbb{R}^{N \times (P^2 \cdot C)} \longleftarrow JigsawPuzzle(\mathbf{x}_p, \widetilde{\mathbf{m}})$

5: **return** $\widetilde{\boldsymbol{x}}_p$

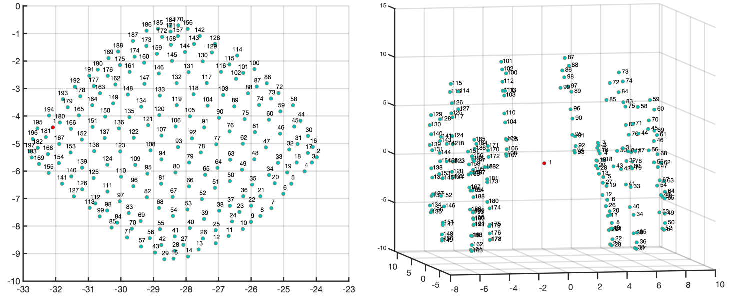


(a) $\mathbf{x}$ (b) $\tilde{\mathbf{x}}$ $(\gamma = 1)$ (c) $\tilde{\mathbf{x}}$ $(\gamma = 0.4)$

[28] Bao, Hangbo, et al. "BEiT: BERT Pre-Training of Image Transformers." *ICLR*2019.

The proposed MJP

- **Adjust The PEs**



(a) UMAP (b) PCA

➢ The spatial relation learned in the high-dimensional space still manifests in the low-dimensional space

➢ PEs capture the absolute position of the input patches, to some extent, the position information could be reconstructed via a reversed mapping function:

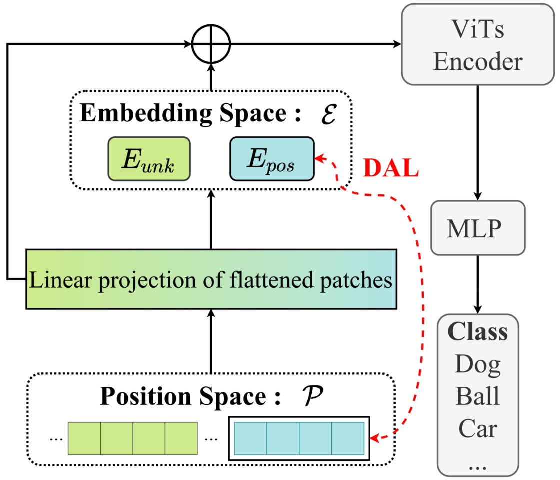
$$g(\cdot) : \mathcal{E} \rightarrow \mathcal{P}$$

➢ Specifically, given $(\widetilde{i}, \widetilde{j})^T$ is the predicted patch position, and $\mathbf{E}_{\text{pos}}^{i,j}$ is the position embedding of the patch (i, j) in the K × K grid.

$$(\widetilde{i}, \widetilde{j})^T = g(\mathbf{E}_{\text{pos}}^{i,j}),$$

Constructing the dense absolute localization (DAL) loss

$$\mathcal{L}_{\text{DAL}} = \mathbb{E}_{\mathbf{E}_{\text{pos}}^{i,j}, 1 \leq i,j \leq K}[\|(i,j)^T - (\widetilde{i}, \widetilde{j})^T\|_1],$$



➢ The final MJP

Algorithm 2 The pipeline of the proposed MJP.

Input: Input image: $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$;
Shuffle Ratio: γ; Patch Size: P

1: $\widetilde{\boldsymbol{x}}_p \leftarrow Alg.\ 1(\mathbf{x}, P, \gamma)$ // 1st & 2nd procedures
2: $\mathbf{E}_{\text{unk}}(\widetilde{\boldsymbol{x}}_p)$ // 3rd procedure
3: $\mathbf{DAL}(\mathbf{x} - \mathbf{x} \cap \widetilde{\boldsymbol{x}}_p)$ // 4th procedure, only for **training**

$$\mathbf{z}_0 = [\mathbf{x}_{\text{CLS}}; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}, \cdots, \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{\text{pos}},$$

$$\boxed{\widetilde{\mathbf{E}}_{\text{pos}}^i = \begin{cases} \mathbf{E}_{\text{pos}}^i, & \text{if } \widetilde{\mathbf{m}}_i = 0 \\ \mathbf{E}_{\text{unk}}, & \text{if } \widetilde{\mathbf{m}}_i = 1 \end{cases}}$$

$$\widetilde{\mathbf{z}}_0 = [\mathbf{x}_{\text{CLS}}; \widetilde{\mathbf{x}}_p^1\mathbf{E}; \widetilde{\mathbf{x}}_p^2\mathbf{E}, \cdots, \widetilde{\mathbf{x}}_p^N\mathbf{E}] + \widetilde{\mathbf{E}}_{\text{pos}}.$$

Experimental Results

- **Regular ImageNet-1K Training**

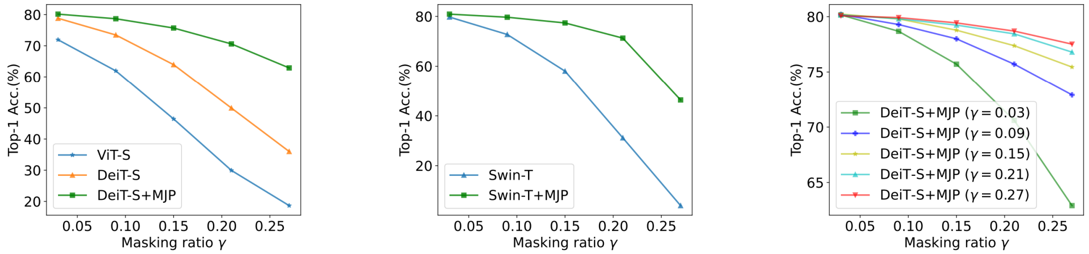
➢ Comparisons of different backbones on ImageNet-1K classification. Note that the image size here are all set to 224x224.

Method	Param.	Top-1 Acc. ↑	Diff. Norm. ↓	Consistency ↑
ResNet-50 [17]	25	79.3	11.77	51.5
ResNet-50 + MJP	25	79.4	7.11	69.3
DeiT-S [40]	22	79.8	16.21	64.3
DeiT-S + MJP	22	80.5	8.96	82.9
Swin-T [31]	29	81.3	15.49	41.5
Swin-T + MJP	29	81.3	12.36	66.9

➢ Ablation study on the proposed MJP trained with different mask ratio

Metric	Masking Ratio					
	0	0.03	0.09	0.15	0.21	0.27
Top-1 Acc.	80.0	80.5	80.3	80.4	80.2	80.3
Diff. Norm.	16.56	8.96	6.36	5.23	4.39	3.97
Consistency	64.0	82.9	88.1	90.5	92.3	93.1

➢ Ablation on the mask ratio during inference



(a) (b) (c)

(a) comparisons among ViT-S, DeiT-S and our method (trained with γ = 0.03); (b) comparisons between Swin-T and our method (trained with γ = 0.03); (c) comparisons of our method on DeiT-S trained with different γ.

➢ Ablation study on the variants of the proposed MJP

Method	Top-1 Acc. ↑	Consistency ↑
A: DeiT-S [40]	79.8	64.3
B: A - PEs	77.5 (-2.3)	100.0
C: A + SPP [39]	74.9 (-4.9)	74.8
D: A + DAL (NLN)	80.0 (+0.2)	64.0
E: A + JP	79.2 (-0.6)	73.8
F: A + JP + IDX	79.9 (+0.1)	79.6
G: A + JP + UNK	80.1 (+0.3)	83.8
H: A + JP + UNK + DAL (PCA)	79.9 (+0.1)	83.4
I: A + JP + UNK + DAL (LN)	80.0 (+0.2)	83.8
J: A + JP + UNK + DAL (NLN)	80.5 (+0.7)	82.9

[17] He, Kaiming, et al. "Deep residual learning for image recognition." *CVPR*2016.
[31] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." *CVPR*2021.
[39] Tolstikhin, Ilya O., et al. "Mlp-mixer: An all-mlp architecture for vision." *NeurIPS*2021.
[40] Touvron, Hugo, et al. "Training data-efficient image transformers & distillation through attention." *ICML*2021.

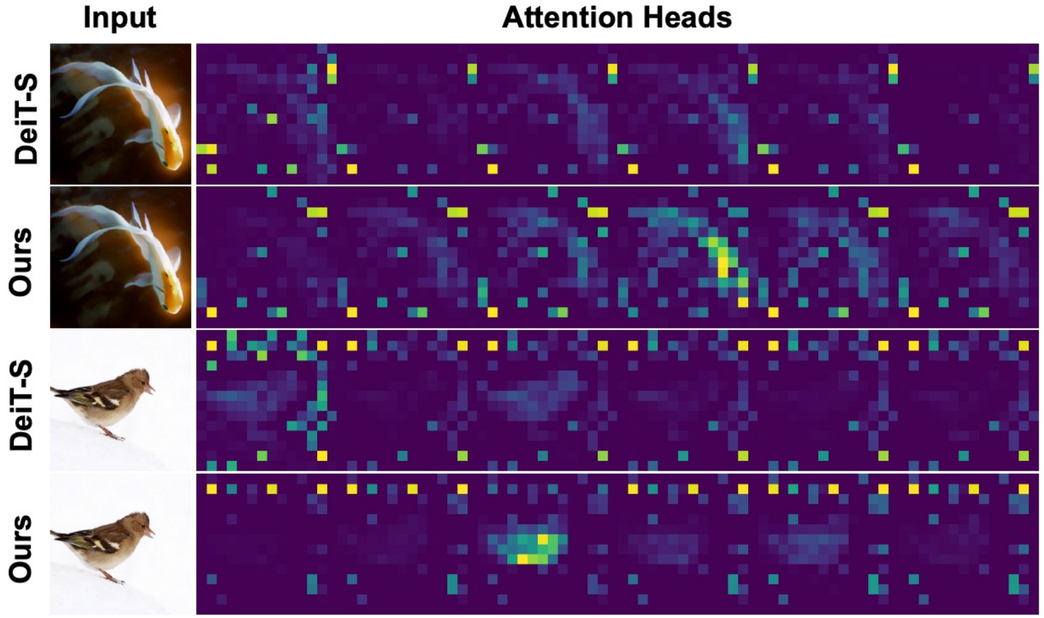
Experimental Results

Robustness on Challenging Sets

Robustness to common corruptions and adversarial examples

Method	ImageNet-C	ImageNet-A		ImageNet-O
	mCE ↓	Acc ↑	AURRA ↑	AUPR ↑
DeiT-S	54.6	19.2	25.1	20.9
DeiT-S + MJP	**51.6**	**21.6**	**29.8**	**22.6**

➢ The visualization Maps of the last self-attention in DeiT-S



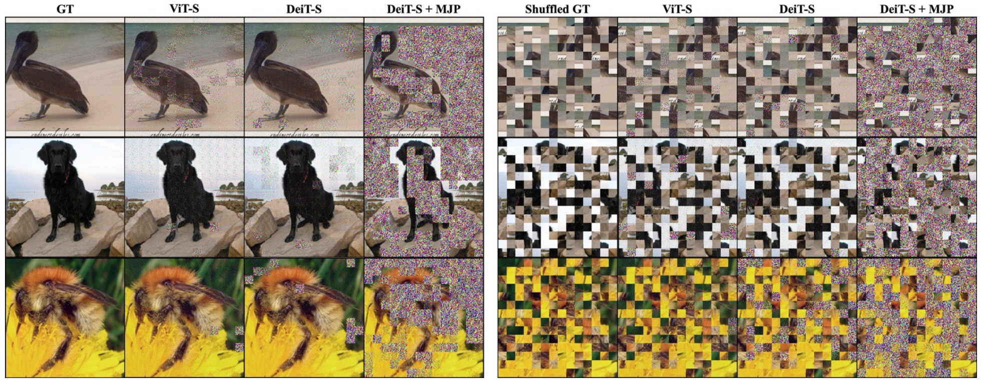
The underlying reason might be that MJP enforces the ViTs aware of both local and global context features, and it helps ViTs to get rid of some unnecessary sample-specific local features during the training.

Privacy Preservation

Comparisons on gradient leakage by analytic attack [32] with ImageNet-1K validation set, where we test (1) ViT-S, DeiT-S and our model in the setting (a); (2) ViT-S, DeiT-S and our model in the setting (b) (i.e., MJP with γ = 0.27); (3) ablation on without (w/o) using Eunk in setting (a); and (4) Our model in setting (c).

	Model	Set.	Acc. ↑	MSE ↑	FFT_{2D} ↑	PSNR ↓	SSIM ↓	LPIPS ↑
(1)	ViT-S [8]	a	78.1	.0278	.0039	19.27	.5203	.3623
	DeiT-S [40]		79.8	.0350	.0057	18.94	.5182	.3767
	DeiT-S (w/o PEs)		77.5	.0379	.0082	20.22	.5912	.2692
	DeiT-S+MJP		**80.5**	**.1055**	**.0166**	**11.52**	**.4053**	**.6545**
(2)	ViT-S [8]	b	18.7	.0327	.0016	18.44	.6065	.2836
	DeiT-S [40]		36.0	.0391	.0024	17.60	.5991	.3355
	DeiT-S (w/o PEs)		**77.5**	.0379	.0025	20.25	.6655	.2370
	DeiT-S+MJP		62.9	**.1043**	**.0059**	**11.66**	**.4493**	**.6519**
(3)	DeiT-S+MJP (w/o)	a	40.6	.1043	.0059	11.66	.4493	.6519
(4)	DeiT-S+MJP	c	62.9	.1706	.0338	8.07	.0875	.8945

(a) $\phi(\nabla\mathcal{M}(\mathbf{x}),\mathbf{x})$ (b) $\phi(\nabla\mathcal{M}(\tilde{\mathbf{x}}),\tilde{\mathbf{x}})$ (c) $\phi(\nabla\mathcal{M}(\tilde{\mathbf{x}}),\mathbf{x})$



Visual comparisons on image recovery with gradient updates [32]. Our proposed DeiT-S+MJP model significantly outperforms the original ViT-S [8] and DeiT-S [40] models

[8] Dosovitskiy, Alexey, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." *ICLR*2020.

Summary

- We for the first time **visually** demonstrate that PEs can explicitly learn the 2D spatial relationship from the input patch sequences;
- We experimentally verified that PEs bring conflict among accuracy, privacy, consistency (i.e., position-insensitive property, robustness) in vision task;
- A versatile Position embedding method, **MJP**, is proposed, for preserving the consistency versus maintaining the accuracy;
- MJP can improve the privacy preservation capacity of ViTs under typical gradient attacks by a large margin, which may pilot a new direction for privacy preservation.