Rotary Position Embedding for Vision Transformer
–Supplementary Material–

Byeongho Heo Song Park Dongyoon Han Sangdoo Yun

NAVER AI Lab

Appendix

A Experiments (cont'd)

We demonstrated the performance of 2D RoPE with performance graphs through various input resolutions in the main paper. This appendix provides additional ablation studies and the entire performance numbers for the multi-resolution experiments. We measured the performance of ViTs and Swin Transformers with default position embedding (APE or RPB), 2D RoPE variants (RoPE-Axial and RoPE-Mixed), and 2D RoPE variants with default position embedding (RoPE-Axial and RoPE-Mixed + APE or RPB). Note that figures in the paper do not include small resolutions such as 96×96 to improve the visualization.

A.1 Impacts of learnable frequencies

Studies on applying RoPE to ViT [1–3] have not considered the learnable frequencies. However, a comparison with RoPE-Axial + learnable frequencies can be an interesting ablation study by revealing the contribution of learnable frequencies on RoPE-Mixed. Table A.1 and Table A.2 show learnable RoPE-Axial performance compared to fixed RoPE-Axial and RoPE-Mixed. The learnable frequencies improve RoPE-Axial on high-resolution (384, 512) but are ineffective on other resolutions. These results imply that the effects of RoPE-Mixed originate from frequency mixing rather than frequency learning, as we claimed in the paper.

Table A.1: RoPE-Axial with learnable frequencies for ViT-S.

Resolution	144	192	224	256	320	384	512
Axial	73.6	79.2	80.9	81.7	81.5	80.0	76.1
Axial+learn	73.5 (-0.1)	79.1 (-0.1)	80.7 (-0.2)	81.5 (-0.2)	81.8 (+0.3)	81.3 (+1.3)	77.8 (+1.7)
Mixed	74.2 (+0.6)	79.6 (+0.4)	80.9 (0.0)	81.8 (+0.1)	82.2 (+0.7)	81.8 (+1.8)	79.1 (+3.0)

Table A.2: RoPE-Axial with learnable frequencies for ViT-B.

Resolution	144	192	224	256	320	384	512
Axial	78.9	82.8	83.6	84.2	84.3	83.9	82.0
Axial+learn	78.9 (0.0)	82.7 (-0.1)	83.6 (0.0)	84.2 (0.0)	84.4 (+0.1)	83.9 0.0	82.3 (+0.3)
Mixed	79.4 (+0.5)	82.8 (0.0)	83.8 (+0.2)	84.3 (+0.1)	84.7 (+0.4)	84.4 (+0.5)	82.9 (+0.9)

A.2 Multi-resolution classification – ViT

Table A.3, A.4, and A.5 report the total numbers of multi-resolution classification, which is illustrated in Figure 4. 2D RoPE variants outperform APE in the smallest resolution 96×96 with significant gap.

Table A.3: Multi-resolution performance of ViT-S. Table reports the performance of 2D RoPE variants corresponding to the first graph in Figure 4.

	Test resolution							
Position embeds	96×96	144×144	192×192	224×224	256×256	320×320	384×384	512×512
APE	35.4	73.6	79.1	80.4	80.9	80.6	79.4	75.4
RoPE-Axial	55.9	73.6	79.2	80.9	81.7	81.5	80.0	76.1
RoPE-Mixed	55.7	74.2	79.6	80.9	81.8	82.2	81.8	79.1
APE + RoPE-Axial	58.4	74.2	79.2	80.7	81.6	81.9	81.2	75.3
APE + RoPE-Mixed	58.5	74.4	79.5	80.9	81.8	82.1	81.7	78.5

Table A.4: Multi-resolution performance of ViT-B. Table reports the performance of 2D RoPE variants corresponding to the second graph in Figure 4.

	Test resolution							
Position embeds	96×96	144×144	192×192	224×224	256×256	320×320	384×384	512×512
APE	57.6	79.1	82.7	83.4	83.8	83.5	82.8	80.5
RoPE-Axial	66.9	78.9	82.8	83.6	84.2	84.3	83.9	82.0
RoPE-Mixed	68.1	79.4	82.8	83.8	84.3	84.7	84.4	82.9
APE + RoPE-Axial	68.9	79.3	82.8	83.7	84.2	84.4	83.8	81.4
APE + RoPE-Mixed	70.2	79.7	83.0	83.8	84.4	84.6	84.3	82.4

Table A.5: Multi-resolution performance of ViT-L. Table reports the performance of 2D RoPE variants corresponding to the third graph in Figure 4.

	Test resolution							
Position embeds	96×96	144×144	192×192	224×224	256×256	320×320	384×384	512×512
APE	61.5	80.8	83.8	84.6	84.9	84.7	84.2	82.2
RoPE-Axial	71.0	80.9	83.9	84.7	85.1	85.3	85.1	84.0
RoPE-Mixed	71.7	81.1	83.9	84.8	85.4	85.7	85.6	84.7
APE + RoPE-Axial	72.4	81.1	84.0	84.7	85.2	85.3	85.1	83.8
APE + RoPE-Mixed	73.2	81.3	84.0	84.9	85.3	85.6	85.5	84.4

A.3 Multi-resolution classification – Swin Transformer

Table A.6, A.7, and A.8 show the total numbers of multi-resolution classification of Swin Transformer with 2D RoPE variants corresponding to Figure 5 of paper. Similar to ViT cases, 2D RoPE variants significantly outperform RPB in small resolutions: 96×96 and 128×128.

Table A.6: Multi-resolution performance of Swin-T. Table reports Swin-T performance with 2D RoPE variants corresponding to the first graph in Figure 5.

	Test resolution							
Position embeds	96×96	128×128	160×160	192×192	224×224	256×256	320×320	384×384
RPB	39.6	67.4	77.8	79.8	81.2	80.9	80.0	78.9
RoPE-Axial	47.6	69.5	77.6	80.0	81.3	81.6	81.0	79.2
RoPE-Mixed	53.6	71.9	78.4	80.2	81.4	81.7	80.8	79.5
RPB + RoPE-Axial	43.8	67.9	77.8	80.2	81.5	81.6	80.9	79.1
RPB + RoPE-Mixed	50.9	69.4	78.1	80.3	81.5	81.8	80.7	78.5

Table A.7: Multi-resolution performance of Swin-S. Table reports Swin-S performance with 2D RoPE variants corresponding to the second graph in Figure 5.

	Test resolution							
Position embeds	96×96	128×128	160×160	192×192	224×224	256×256	320×320	384×384
RPB	47.0	72.7	80.2	81.8	82.9	82.8	82.2	81.0
RoPE-Axial	44.0	72.0	79.9	82.0	83.1	83.3	83.0	80.9
RoPE-Mixed	55.7	75.5	80.5	82.3	83.0	83.3	82.9	81.4
RPB + RoPE-Axial	55.4	74.7	80.8	82.4	83.2	83.3	82.8	81.3
RPB + RoPE-Mixed	57.4	75.2	80.8	82.5	83.3	83.4	82.8	81.1

Table A.8: Multi-resolution performance of Swin-B. Table reports Swin-B performance with 2D RoPE variants corresponding to the third graph in Figure 5.

	Test resolution							
Position embeds	96×96	128×128	160×160	192×192	224×224	256×256	320×320	384×384
RPB	48.8	73.3	80.9	82.3	83.3	83.1	82.3	81.2
RoPE-Axial	52.7	74.3	80.8	82.6	83.6	83.7	83.2	81.8
RoPE-Mixed	61.5	76.7	81.4	82.9	83.7	83.8	83.3	82.1
RPB + RoPE-Axial	55.3	74.4	81.3	82.8	83.6	83.8	83.1	81.5
RPB + RoPE-Mixed	62.2	76.3	81.4	82.8	83.6	83.7	83.1	81.4

A.4 2D RoPE with APE or RPB

In Figure 6 of the paper, we report the performance improvement of RoPE-Mixed compared to base position embeddings: APE or RPB. We provide numbers for Figure 6 in Table A.9 and A.10. Note that each number means performance improvement (%p.) compared to base position embeddings (APE or RPB).

Table A.9: Performance improvement compared to APE in ViT-B. Table shows the improvement over APE, which is shown in the left graph of Figure 6.

	Test resolution								
Position embeds	112×112	128×128	160×160	192×192	224×224	256×256	320×320	384×384	512×512
APE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
RoPE-Mixed	1.1	0.5	0.2	0.1	0.4	0.5	1.2	1.6	2.4
RoPE-Mixed + APE	2.2	1.1	0.4	0.3	0.4	0.6	1.1	1.5	1.9

Table A.10: Performance improvement compared to RPB in Swin-B. Table shows the improvement over RPB, which is shown in the right graph of Figure 6.

	Test resolution						
Position embeds	128×128	160×160	192×192	224×224	256×256	320×320	384×384
RPB	0	0	0	0	0	0	0
RoPE-Mixed	3.4	0.5	0.6	0.4	0.7	1.0	0.9
RoPE-Mixed + RPB	3.0	0.5	0.5	0.3	0.6	0.8	0.2

References

1. Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva-02: A visual representation for neon genesis. arXiv preprint arXiv:2303.11331 (2023)
2. Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., Hoiem, D., Kembhavi, A.: Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. arXiv preprint arXiv:2312.17172 (2023)
3. Lu, Z., Wang, Z., Huang, D., Wu, C., Liu, X., Ouyang, W., Bai, L.: Fit: Flexible vision transformer for diffusion model. arXiv preprint arXiv:2402.12376 (2024)