



Improving efficiency in ViTs

- Transformers are state-of-the-art in most image and video tasks.
- ViTs are computationally expensive.
- Compute and memory grows quadratically (N x N). 14x14 for image classification to $128 \times 128 = 16$ K for image denoising.
- How to improve the efficiency of vision transformers?

CKA analysis

- attention between High correlation matrices layers. across
- blocks of MSA • Same for output (features Z).

	1	2	3	4	5	6	7	8	9	10	11	12
1	1		0.86		0.76	0.72	0.53	0.63	0.63	0.56	0.5	0.44
2		1	0.91	0.9			0.66	0.74	0.73	0.64	0.51	0.41
3	0.86	0.91	1	0.96	0.92	0.89	0.74			0.71	0.55	0.45
4		0.9	0.96	1	0.96	0.94	0.78			0.74	0.56	0.44
5	8.76		0.92	0.96	1	0.95	018			0.75	0.55	0.43
6	0.72		0.89	0.94	0.95	1	0.85			0.77	0.55	0.44
7	0.53	0.66	0.74	0.78	0.8	9.85	1	0.79	0.81		0.52	0.39
8	0.63	0.74					0.79	1		0.79	0.57	0.42
9	0.63	0.73	0.79						1		0.6	0.45
10	0.56	0,64	0.71	0,74	0.75	0.77		0.79	0.82	1	0.72	0.56
11	0.5	0.51	0.55	0.56	0.55	0.55	0.52	0.57	0.6	0.72	1	0.92
12	0.44	0.41	0.45	0.44	0.43	0.44	0.39	0.42	0.45	0.56	0.92	1

(a) CKA of A^[CLS]

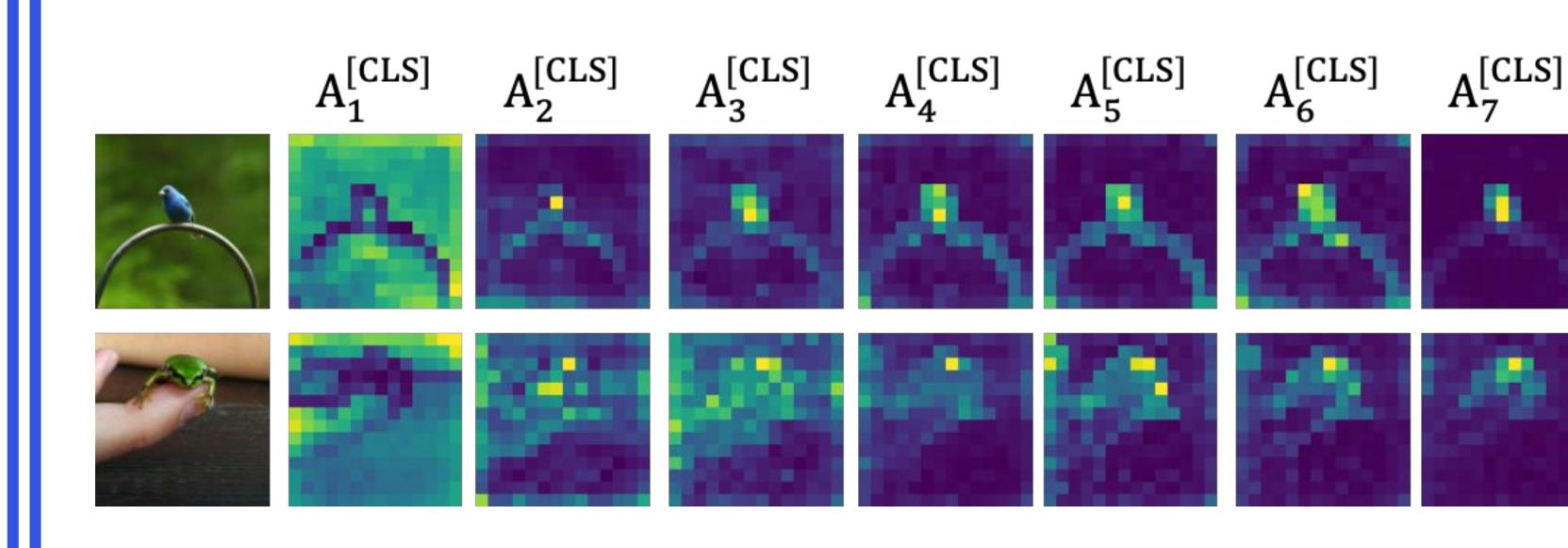
1	2	3	4	5	6	7	8	9	10	11	12
1	0.32	0.29	0.31	0.25	0.25	0,12	0.13	0.2	0.078	0.072	0.073
0.32	1	0.85	0.8			0.62	0.57	0.35	0.46	0.3	0.41
0.29	0.85	1	0.87	0.74		0.67	0.59	0.35	0.48	0.27	0.44
0.31	0.8	0.87	1	0.81	0.82	0.68	0.63	0.39	0.51	0.32	0.44
0.25	0.75		0.81	1	0.8	0.71		0.49	0.6	0.37	0.52
0.25			0.82	0.8	1	0.7	0.66	0.44	0.52	0.34	0.44
0.12	0.62	0.67		0.71	0.7	1	0.83	0.38	0.72	0.47	0.66
0.13	0.57	0.59	0.63	0.74	0.66	0.83	1	0.63	0.78	0.49	0.65
0.2	0.35	0.35	0.39	0.49	0.44	0.38	0.63	1	0.48	0.41	0.38
0.078	0.46	0.48	0.51	0.6	0.52	0.77	0.78	0.48	1	0.5	0.69
0.072	0.3	0.27	0.32	0.37	0.34	0.47	0.49	0.41	0.5	1	0.42
0.073	0.41	0.44	0.44	0.52	0.44	0.66	0.65	0.38	0.69	0.42	1

(b) CKA of Z^[MSA]

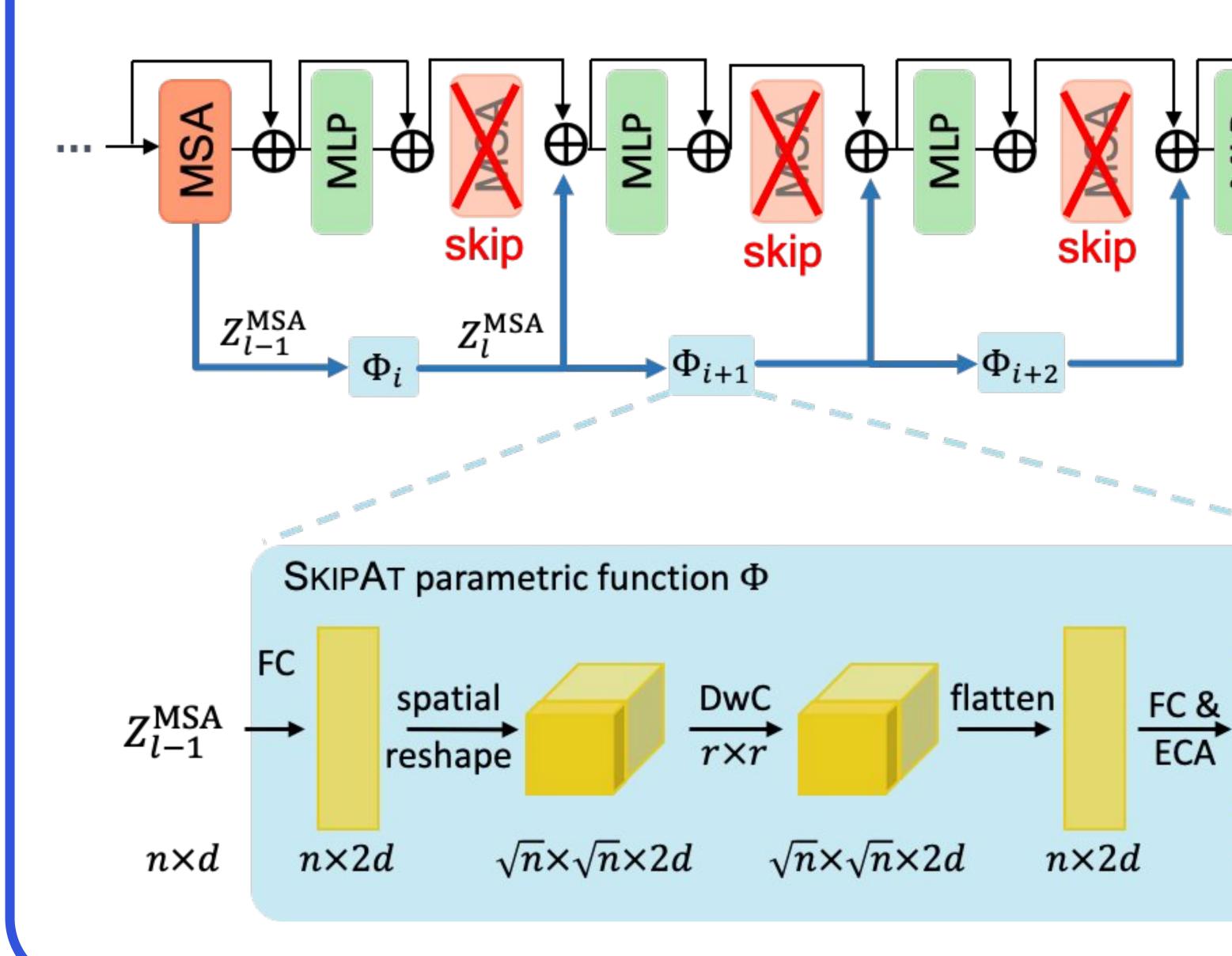
Skip-Attention: Improving Vision Transformers by Paying Less Attention

ashanka Venkataramanan*, Amir Ghodrati*, Yuki M. Asano, Fatih Porikli, Amirhossein Habibian

We leverage the high correlation across MSA blocks to improve efficiency of ViTs



Approximating MSA with SkipAT parametric function



 $Z_l \leftarrow Z_l^{\text{MSA}} + Z_{l-1},$ $Z_l \leftarrow \mathrm{MLP}(Z_l) + Z_l.$

 $A_{12}^{[CLS]}$

 $A_{11}^{[CLS]}$

 $A_{o}^{[CLS]}$

 $A_{8}^{[CLS]}$

 $Z_l \leftarrow \Phi(Z_{l-1}^{\text{MSA}}) + Z_{l-1}$ $Z_l \leftarrow \text{MLP}(Z_l) + Z_l$

 $n \times d$

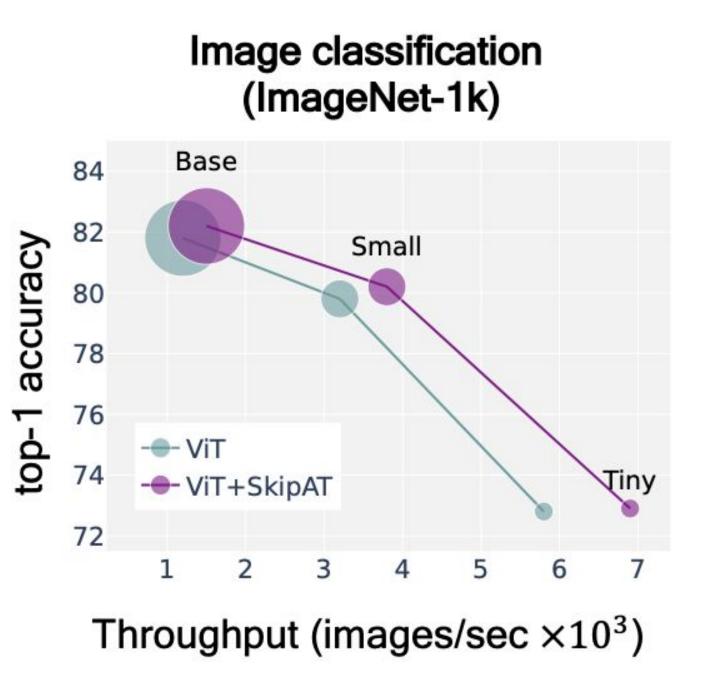
 $\rightarrow Z_l^{MSA}$

 $n \times d$



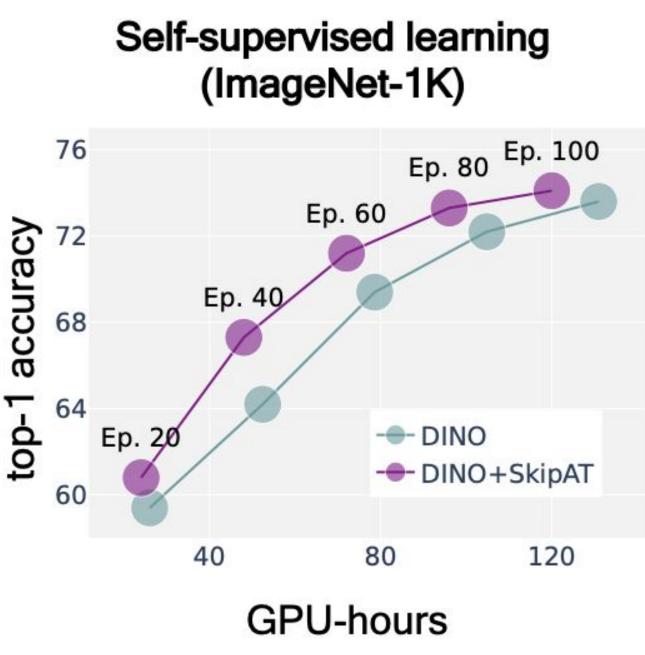


Image classification

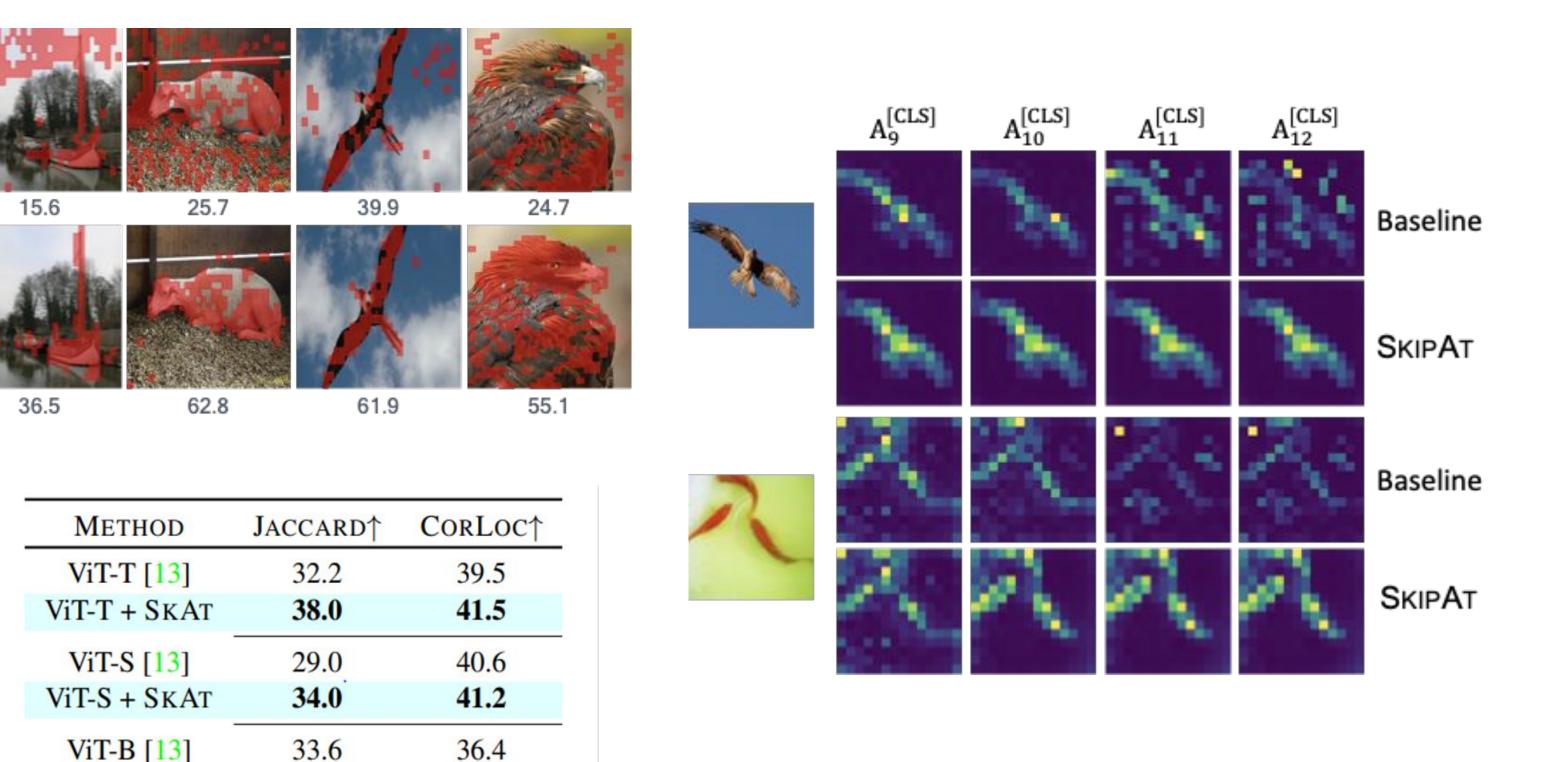


36.8

ViT-B + SKAT



Unsupervised Object Discovery



Additional Downstream Tasks

