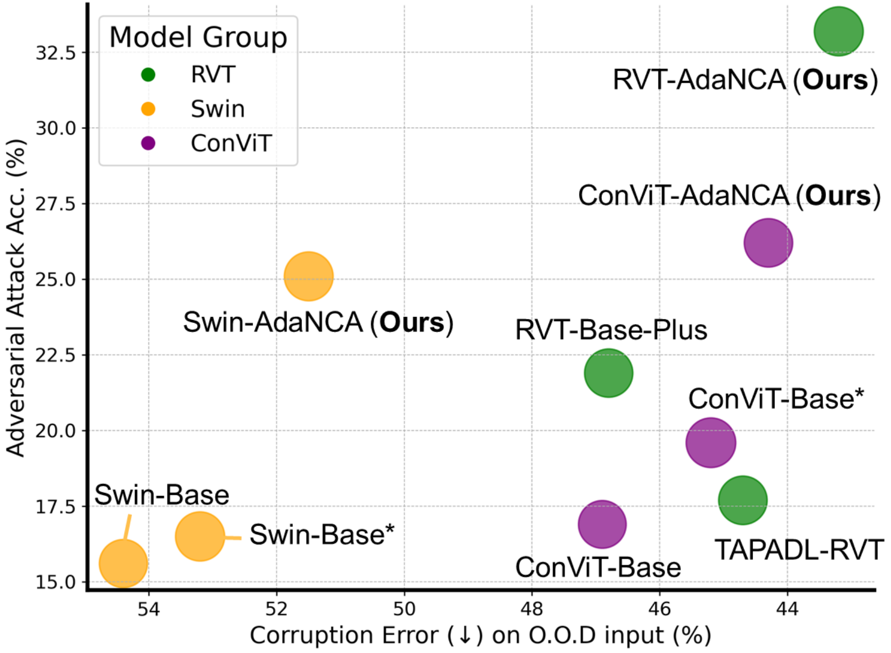
AdaNCA: Neural Cellular Automata as Adaptors for More Robust Vision Transformer

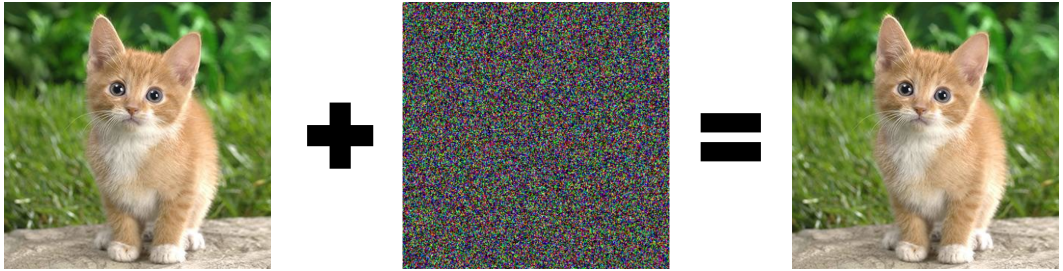
Yitao Xu
Tong Zhang
Sabine Süsstrunk



Background

- **Vision Transformer (ViT) is vulnerable against:**
  - Adversarial samples

Szegedy et al. ICLR 2014



Cat　　Adversarial noise　　Dog

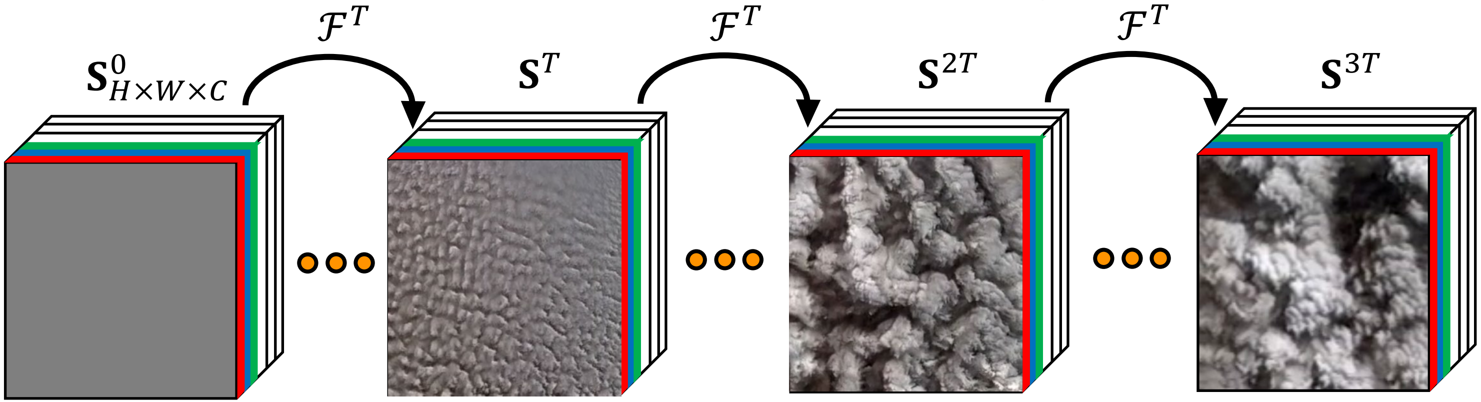
  - Out-of-distribution inputs

Hendrycks et al. ICLR 2018

Cat　　Dog　　Bird　　Fish　　Basketball

Background

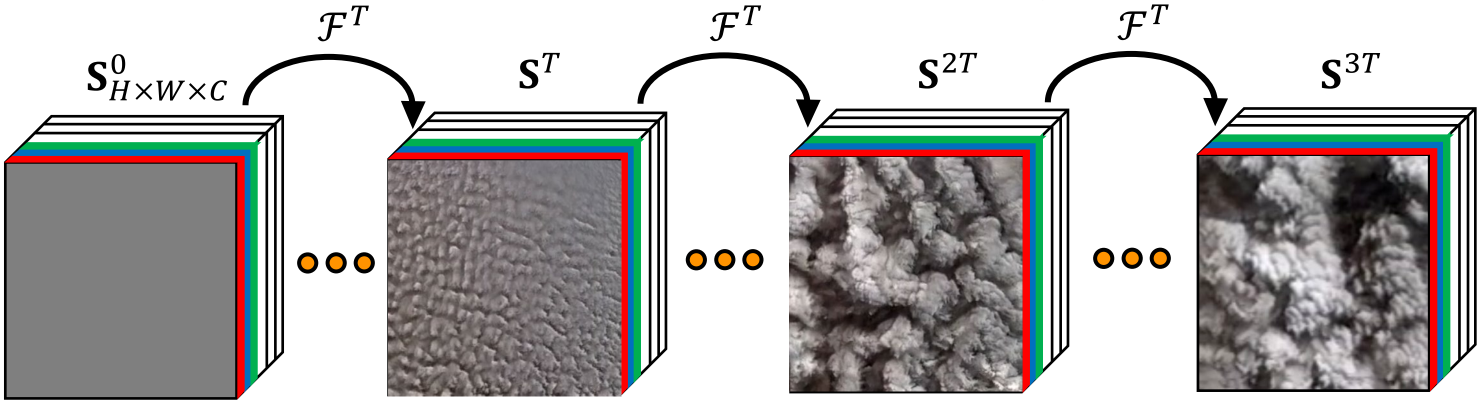
- **Neural Cellular Automata (NCA) is robust against perturbations**

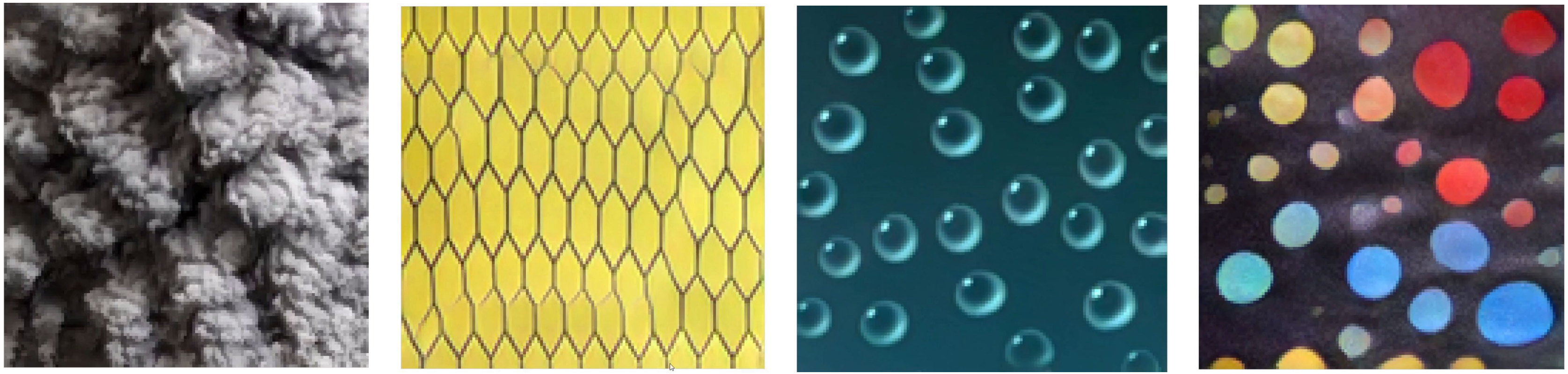


Pajouheshgar & Xu et al. CVPR 2023

Background

- **Neural Cellular Automata (NCA) is robust against perturbations**

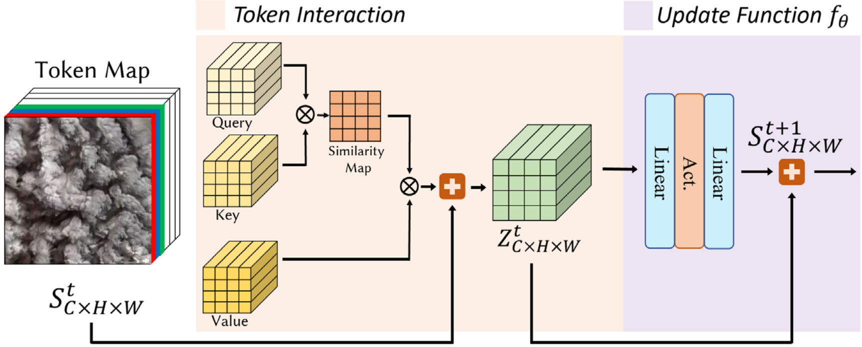




Pajouheshgar & Xu et al. CVPR 2023

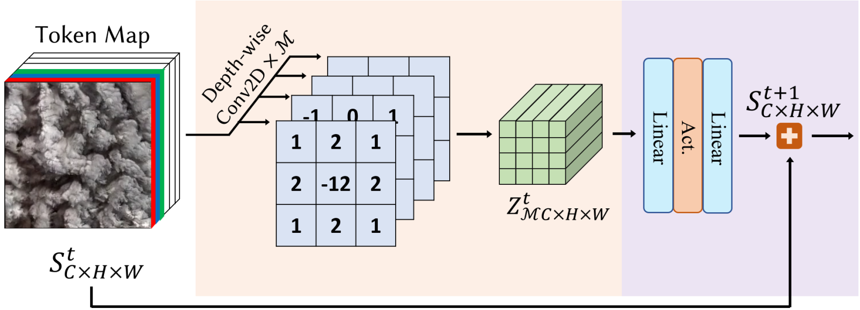
Background

- **ViT and NCA are similar in token interaction learning**



$$\mathbf{X}_{attn} = \sigma\left(\frac{\mathbf{QK}^\top}{\sqrt{C}}\right)\mathbf{V}$$

$$\mathbf{X}_{out} = f_\theta(\mathbf{X}_{attn})$$

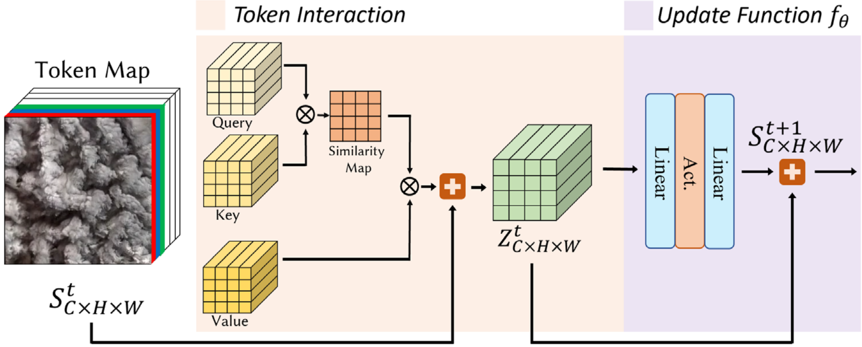


$$\mathbf{S}_{\mathcal{I}} = (\mathbf{S} \circledast [\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_{\mathcal{M}}])_{\oplus}$$

$$\mathbf{S}_{out} = f_\theta(\mathbf{S}_{\mathcal{I}})$$

Background

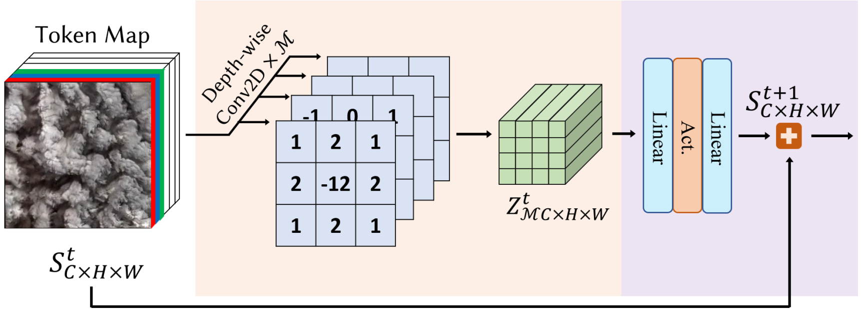
- **ViT and NCA are similar in token interaction learning**



$$\mathbf{X}_{attn} = \sigma\left(\frac{\mathbf{QK}^\top}{\sqrt{C}}\right)\mathbf{V}$$

Vulnerable

$$\mathbf{X}_{out} = f_\theta(\mathbf{X}_{attn})$$



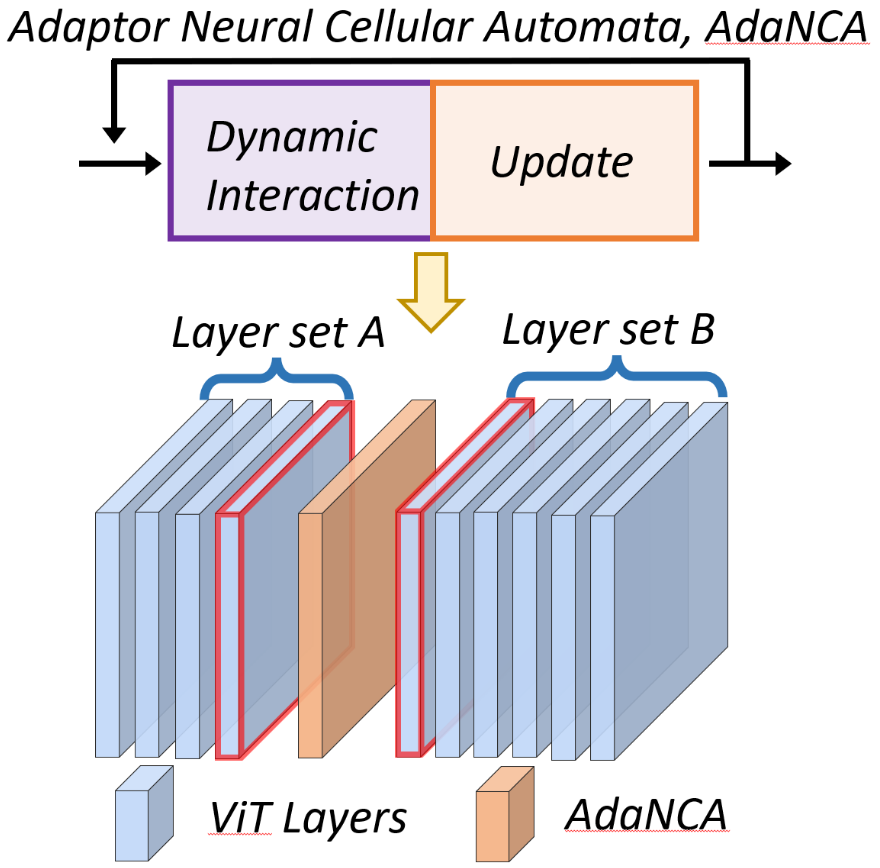
$$\mathbf{S}_{\mathcal{I}} = (\mathbf{S} \circledast [\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_{\mathcal{M}}])_\oplus$$

Robust

$$\mathbf{S}_{out} = f_\theta(\mathbf{S}_{\mathcal{I}})$$

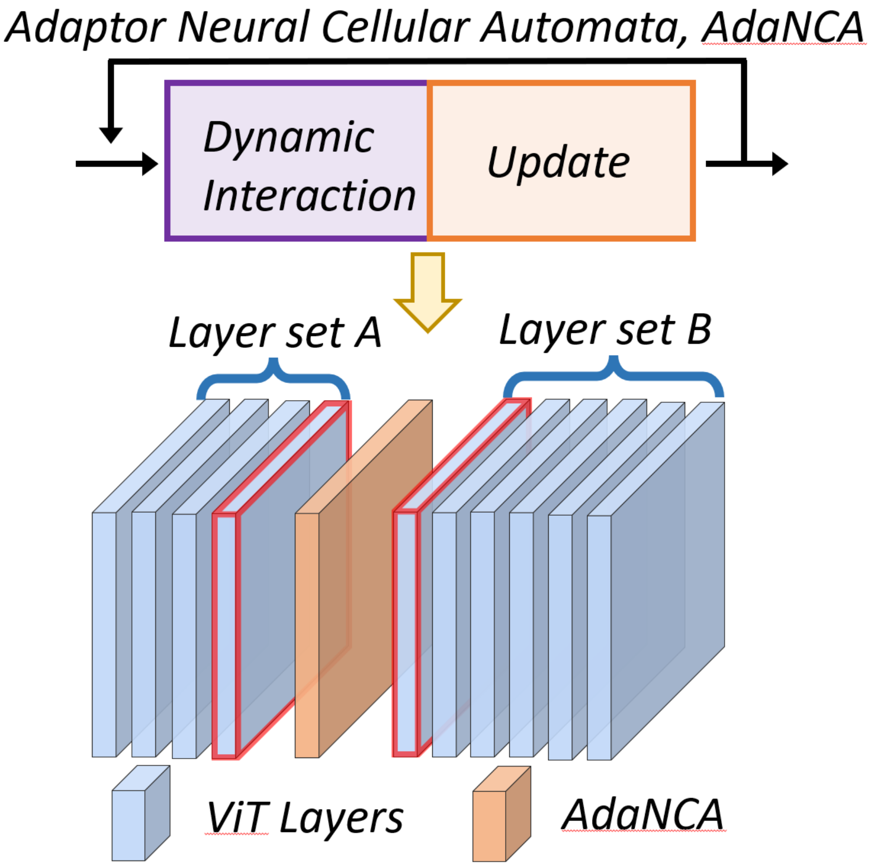
Can NCA improve the robustness of ViT?

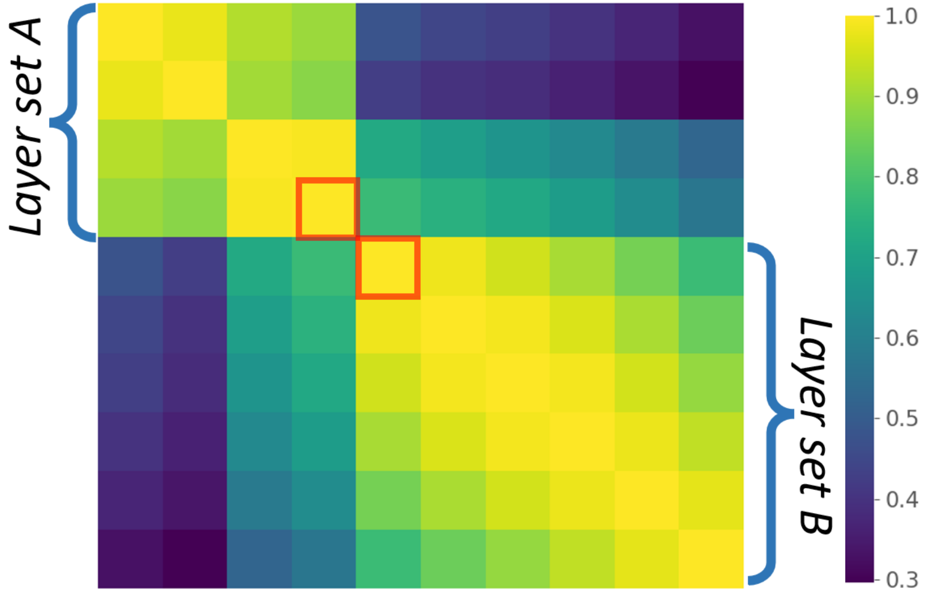
Framework



Adaptor Neural Cellular Automata, AdaNCA
Dynamic Interaction
Update
Layer set A
Layer set B
ViT Layers
AdaNCA

Framework

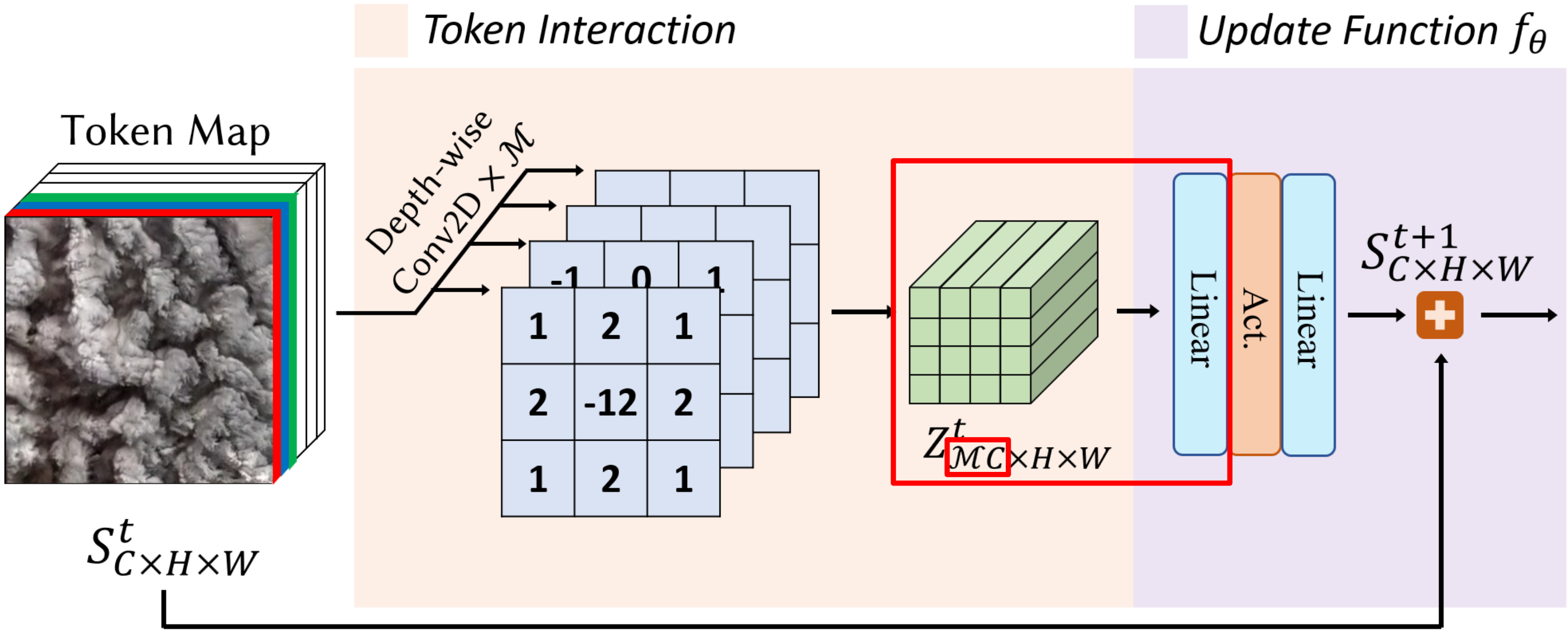




Pair-wise Layer Similarity

Problem with Vanilla NCA

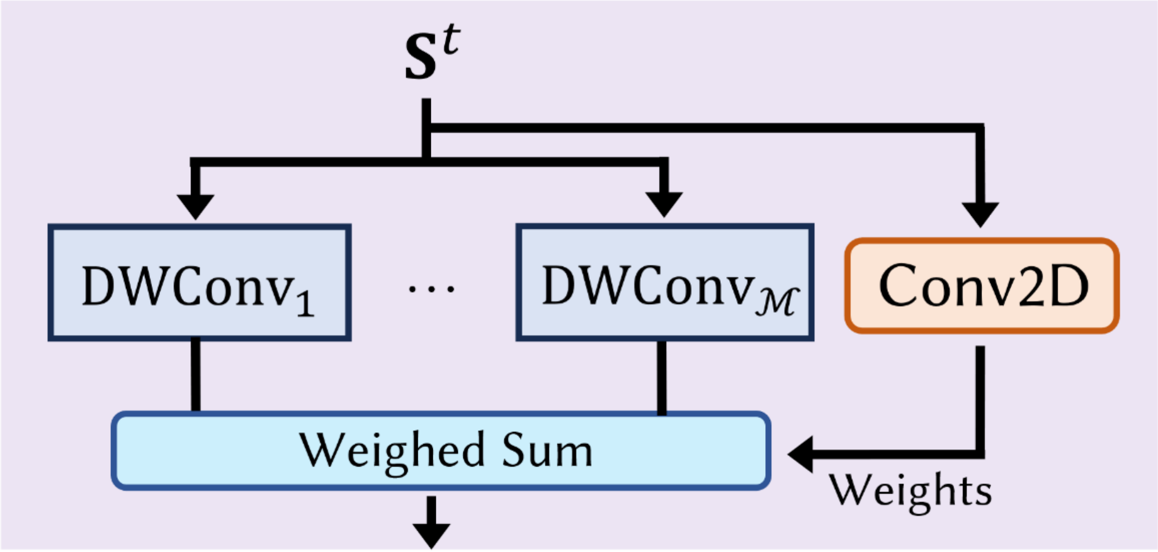
- **Too computationally intensive**



Struggle on scaling up

Dynamic Interaction

- **Reduce computational cost**

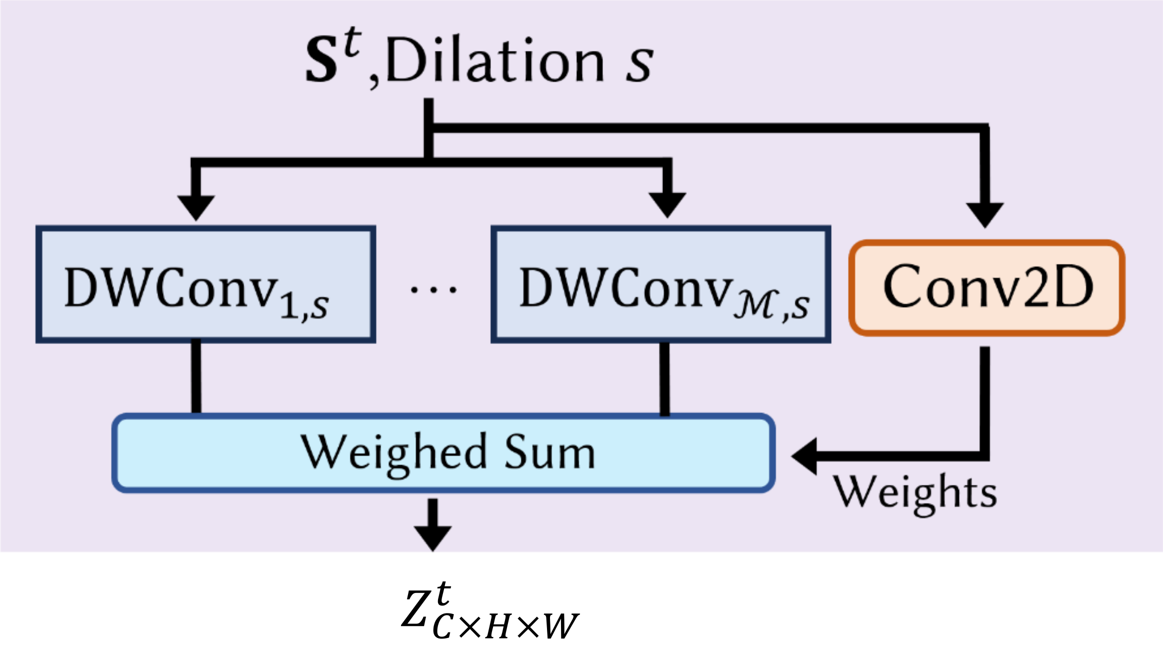


$Z^t_{C\times H\times W}$

FLOPS: $\mathcal{M}\text{HWC}^2 \rightarrow HWC^2$

Multi-scale Dynamic Interaction

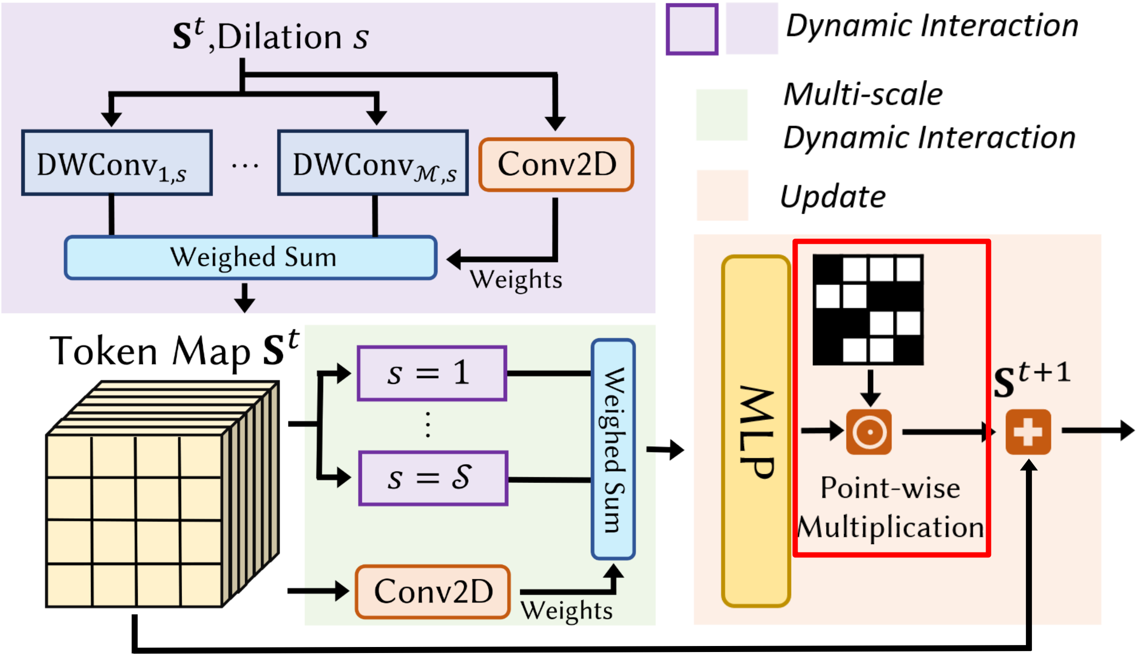
- **Improve model capacity**



FLOPS: $\mathcal{M}\mathrm{HWC}^2 \rightarrow HWC^2$

Multi-scale Dynamic Interaction

- **Improve model capacity**



Insert Positions of AdaNCA

- **Different insert positions result in different performance**

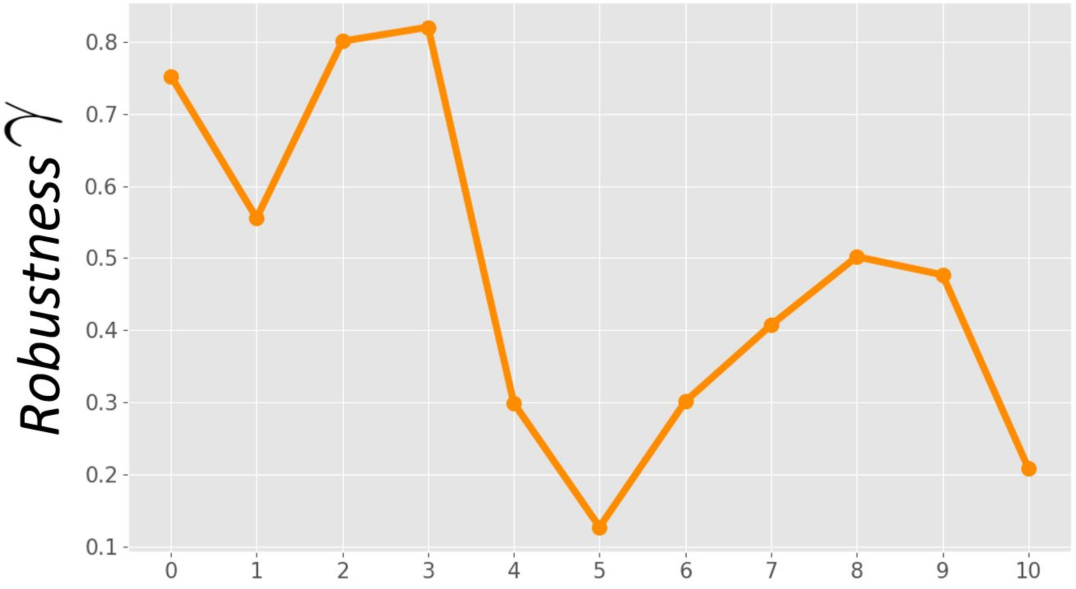
$$\alpha \; = \text{Clean Accuracy}$$

$$\alpha' = \text{Accuracy under Adv.}$$

$$\beta \; = \frac{\alpha'}{\alpha}$$

$$\gamma \; = \frac{\beta_{AdaNCA} - \beta_{Base}}{\beta_{Base}}$$

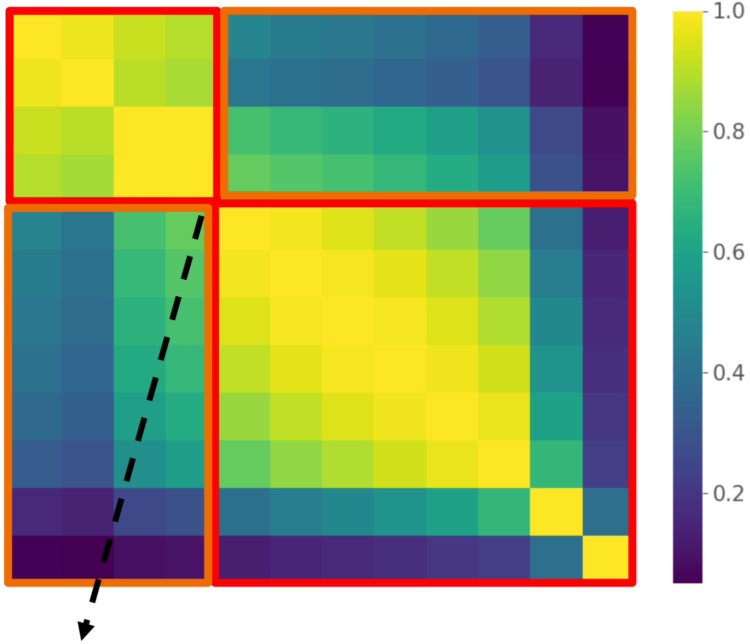
$Base$: Baseline ViT
$AdaNCA$: AdaNCA-Enhanced ViT



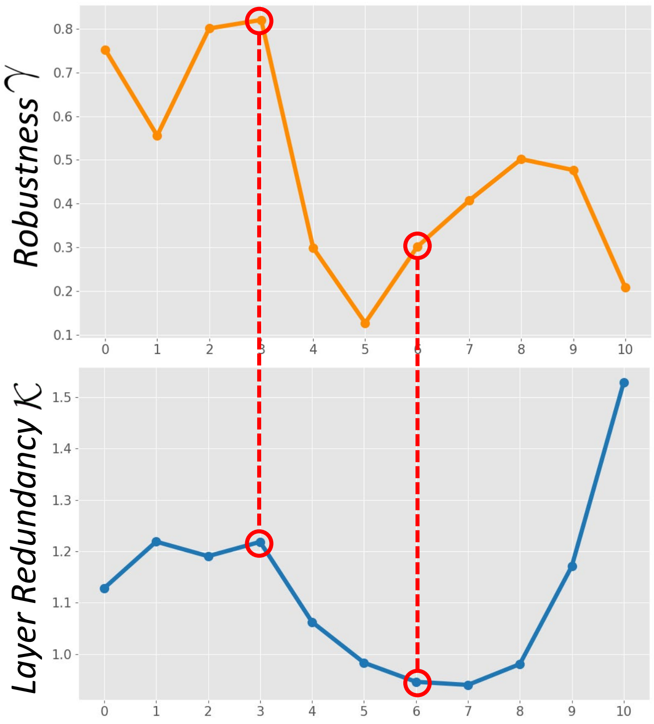
Swin-tiny (12 layers, 11 insert points)

Insert Positions of AdaNCA

- **Different insert positions result in different performance**

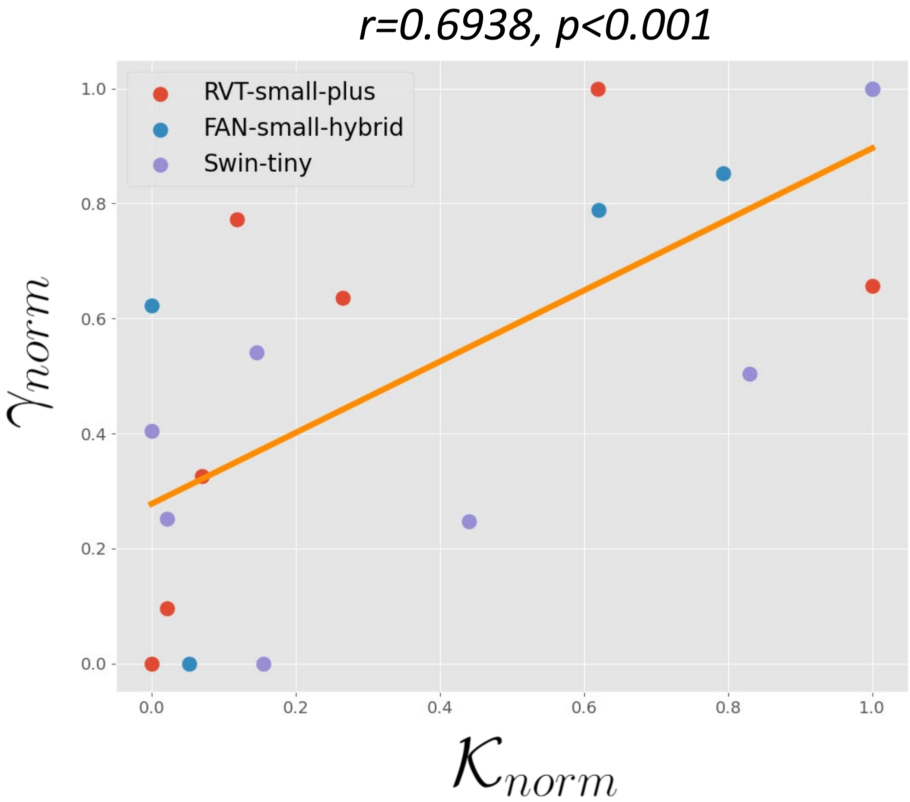


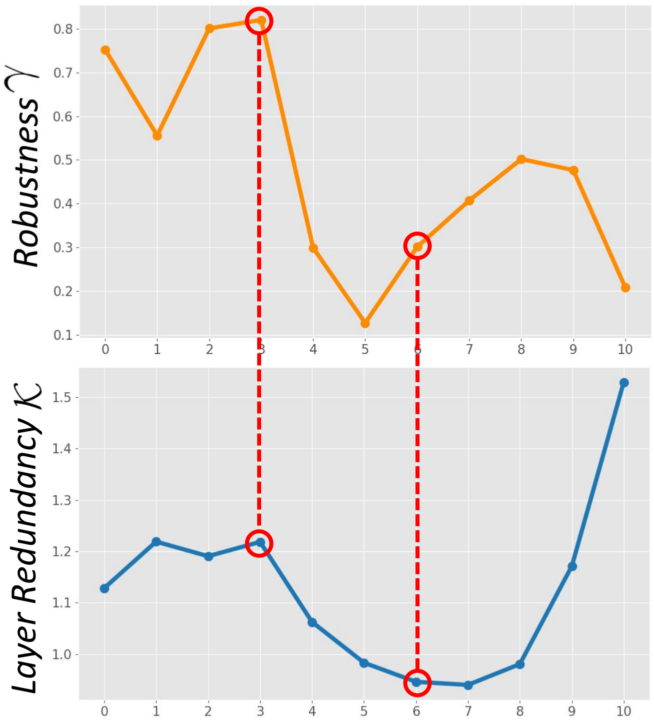
$$\mathcal{K}(3)=Sim(\boxed{})-Sim(\boxed{})$$



Insert Positions of AdaNCA

- **Different insert positions result in different performance**





Main Results

Model	Params (M)	FLOPS (G)	ImageNet Clean Acc.	Adversarial Inputs PGD [6]	CW [7]	APGD-DLR [8]	APGD-CE [8]	IM-A [9]	OOD inputs IM-C (↓) [10]	IM-R [11]	IM-SK [12]
RVT-B [1]	**88.5**	**17.7**	82.7	29.9	21.5	21.9	31.4	28.5	46.8	48.7	36.0
TAPADL-RVT [2]	89.4	17.9	83.1	27.6	19.3	17.7	26.8	**32.7**	44.7	50.2	38.6
RVT-B-AdaNCA	91.0	19.0	**83.3**	**36.7**	**30.2**	**33.2**	**36.2**	31.9	**43.2**	**51.7**	**39.0**
FAN-B [3]	**50.4**	**11.7**	83.9	15.0	7.6	10.4	13.1	39.6	46.1	52.7	40.8
TAPADL-FAN [2]	50.7	11.8	**84.3**	18.6	9.2	13.5	16.9	42.3	**43.7**	**54.6**	40.7
FAN-B-AdaNCA	51.7	12.4	84.1	**20.3**	**10.6**	**14.1**	**19.1**	**42.9**	44.7	53.4	**41.0**
Swin-B [4]	**87.8**	**15.4**	83.4	21.3	13.4	15.6	23.1	35.8	54.3	46.6	32.4
Swin-B*	94.1	16.7	83.3	22.8	14.6	15.9	23.8	35.2	53.2	46.9	33.7
Swin-B-AdaNCA	90.7	16.3	**83.7**	**24.1**	**20.5**	**25.1**	**24.8**	**36.0**	**51.5**	**48.2**	**35.5**
ConViT-B [5]	**86.5**	**17.7**	**82.4**	21.2	8.9	16.9	20.3	29.0	46.9	48.4	35.7
ConViT-B*	93.6	19.2	82.7	24.1	10.0	20.5	23.9	30.1	45.2	49.9	37.8
ConViT-B-AdaNCA	89.0	19.0	**83.2**	**29.2**	**20.1**	**26.3**	**28.4**	**33.0**	**44.3**	**51.1**	**39.1**

[1] Mao et al. CVPR 2022
[2] Guo et al. ICCV 2023
[3] Zhou et al. ICML 2022
[4] Liu et al. ICCV 2021
[5] D'Ascoli et al. ICML 2021

[6] Madry et al. ICLR 2018
[7] Carlini et al. IEEE SP 2017
[8] Croce et al. ICML 2020
[9] Djolonga et al. CVPR 2021

[10] Hendrycks et al. ICLR 2018
[11] Hendryck et al. ICCV 2021
[12] Wang et al. NeurIPS 2019

Main Results

Model	Params (M)	FLOPS (G)	**ImageNet** Clean Acc.	**Adversarial Inputs** PGD [6]	CW [7]	APGD-DLR [8]	APGD-CE [8]	IM-A [9]	**OOD inputs** IM-C (↓) [10]	IM-R [11]	IM-SK [12]
RVT-B [1]	**88.5**	**17.7**	82.7	29.9	21.5	21.9	31.4	28.5	46.8	48.7	36.0
TAPADL-RVT [2]	89.4	17.9	83.1	27.6	19.3	17.7	26.8	**32.7**	44.7	50.2	38.6
RVT-B-AdaNCA	91.0	19.0	**83.3**	**36.7**	**30.2**	**33.2**	**36.2**	31.9	**43.2**	**51.7**	**39.0**
FAN-B [3]	**50.4**	**11.7**	83.9	15.0	7.6	10.4	13.1	39.6	46.1	52.7	40.8
TAPADL-FAN [2]	50.7	11.8	**84.3**	18.6	9.2	13.5	16.9	42.3	**43.7**	**54.6**	40.7
FAN-B-AdaNCA	51.7	12.4	84.1	**20.3**	**10.6**	**14.1**	**19.1**	**42.9**	44.7	53.4	**41.0**
Swin-B [4]	**87.8**	**15.4**	83.4	21.3	13.4	15.6	23.1	35.8	54.3	46.6	32.4
Swin-B*	94.1	16.7	83.3	22.8	14.6	15.9	23.8	35.2	53.2	46.9	33.7
Swin-B-AdaNCA	90.7	16.3	**83.7**	**24.1**	**20.5**	**25.1**	**24.8**	**36.0**	**51.5**	**48.2**	**35.5**
ConViT-B [5]	**86.5**	**17.7**	**82.4**	21.2	8.9	16.9	20.3	29.0	46.9	48.4	35.7
ConViT-B*	93.6	19.2	82.7	24.1	10.0	20.5	23.9	30.1	45.2	49.9	37.8
ConViT-B-AdaNCA	89.0	19.0	**83.2**	**29.2**	**20.1**	**26.3**	**28.4**	**33.0**	**44.3**	**51.1**	**39.1**

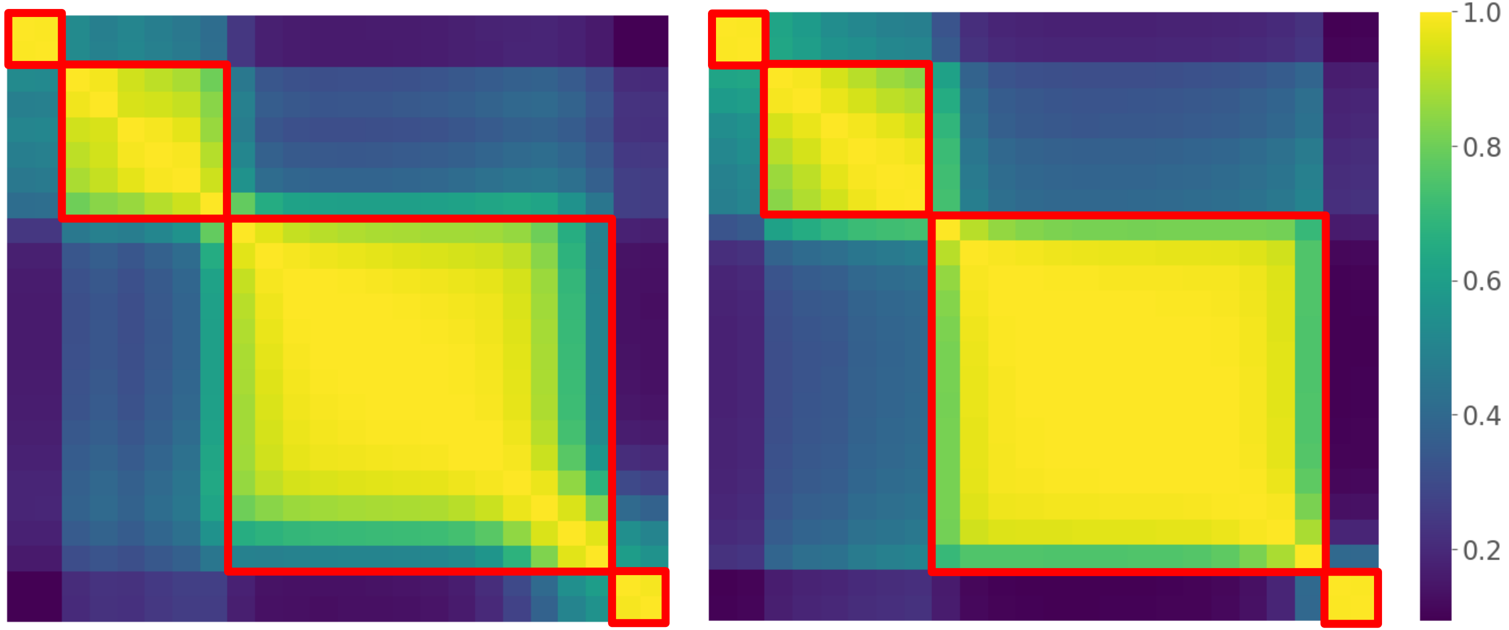
[1] Mao et al. CVPR 2022
[2] Guo et al. ICCV 2023
[3] Zhou et al. ICML 2022
[4] Liu et al. ICCV 2021
[5] D'Ascoli et al. ICML 2021

[6] Madry et al. ICLR 2018
[7] Carlini et al. IEEE SP 2017
[8] Croce et al. ICML 2020
[9] Djolonga et al. CVPR 2021

[10] Hendrycks et al. ICLR 2018
[11] Hendryck et al. ICCV 2021
[12] Wang et al. NeurIPS 2019

Layer Similarity Structure

- **AdaNCA increases the network redundancy**



Swin-Base

κ_{mean}=0.47

Swin-Base-AdaNCA

κ_{mean}=0.51

Ablation Studies

Exp. Type	Recur	StocU	RandS	DynIn	Params (M)	FLOPS (G)	Accuracy (↑)	Attack Failure Rate (↓)
Baseline	✗	✗	✗	✗	27.59	4.5	86.56	12.29
Ablation	✗	✓	✗	✓	28.97	4.7	**87.36**	19.04
	✓	✗	✓	✓	27.94	4.7	86.92	19.56
	✓	✓	✗	✓	27.94	4.7	87.12	19.34
	✓	✓	✓	✗	27.93	4.7	86.72	21.98
Ours	✓	✓	✓	✓	27.94	4.7	87.18	**22.35**

✗ Recur: Unrolling the recurrent structure into different single-step NCA

✗ StocU: Remove stochastic update

✗ RandS: AdaNCA evolves for a fixed time step

✗ DynIn: Replace Dynamic Interaction with a simple mean aggregation

Thank you for listening!