SAPIENZA
UNIVERSITÀ DI ROMA

From One to Many: Virtual Subnetworks for Vision Transformer Ensembles

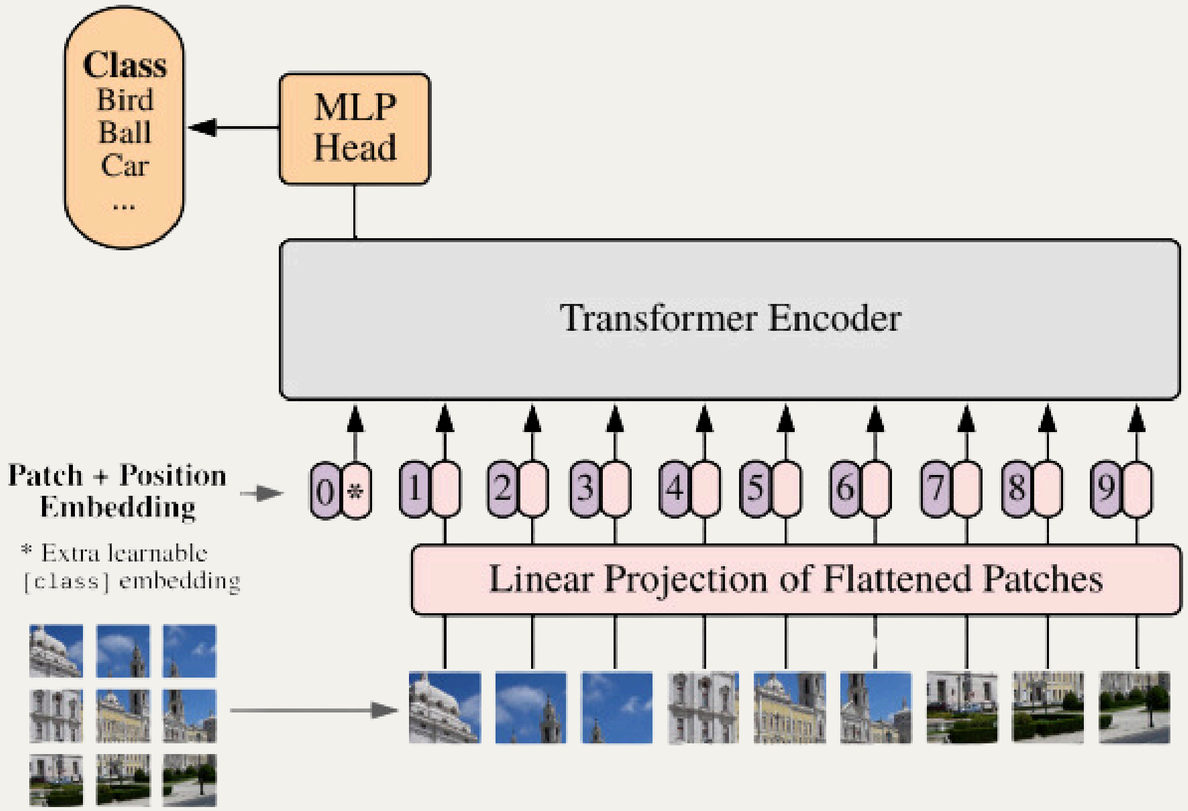
Alessandro Ferrante

Engineering in
Computer Science

Professor Simone Scardapane

Sapienza University of Rome **2025**

Vision Transformers



PRO

- Global context awareness
- State of the art performance (>CNN)
- Flexible to many inputs

CONS

- Data-hungry
- Computational costs
- Vulnerable to noise and attacks

Extending ViT

Custom ViT

Additional tokens

Vanilla ViT

→ **Virtual subnetworks**

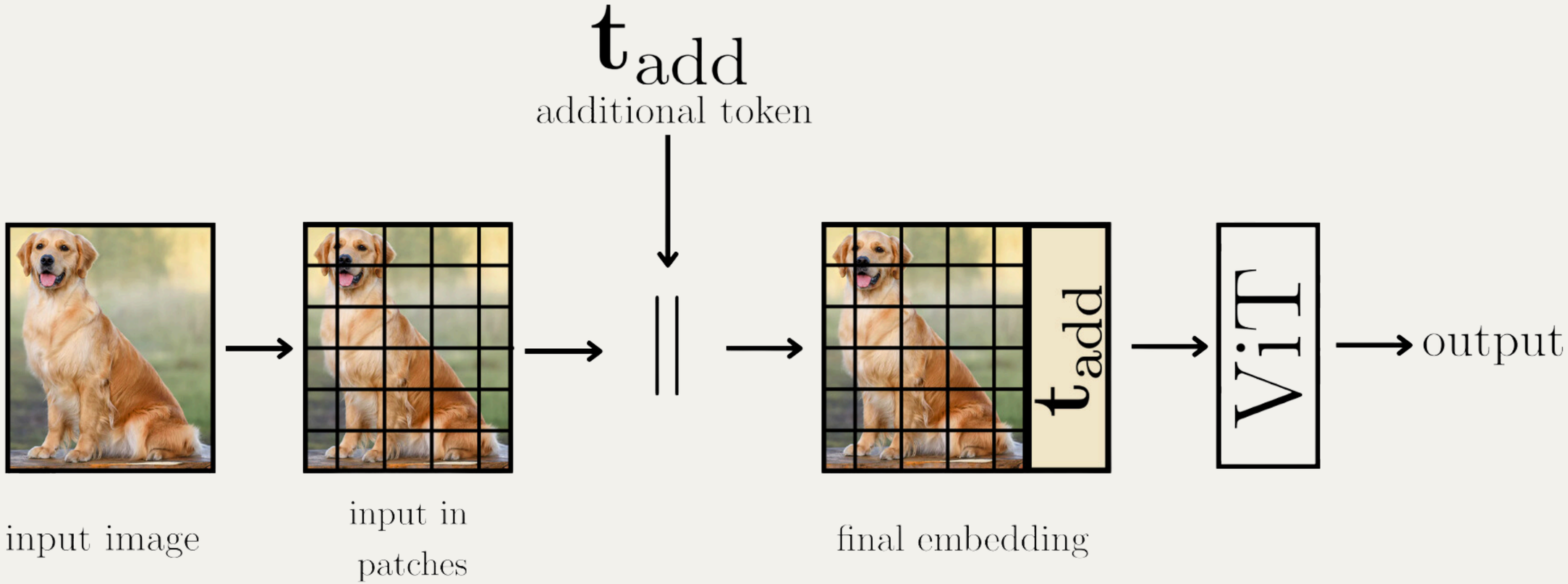
 **Ensembling predictions**

Only **1** training

Recycling data

Robustness

Virtual Subnetworks



Ensembling

Goals

Maintain accuracy

Increase confidence

Improve generalization

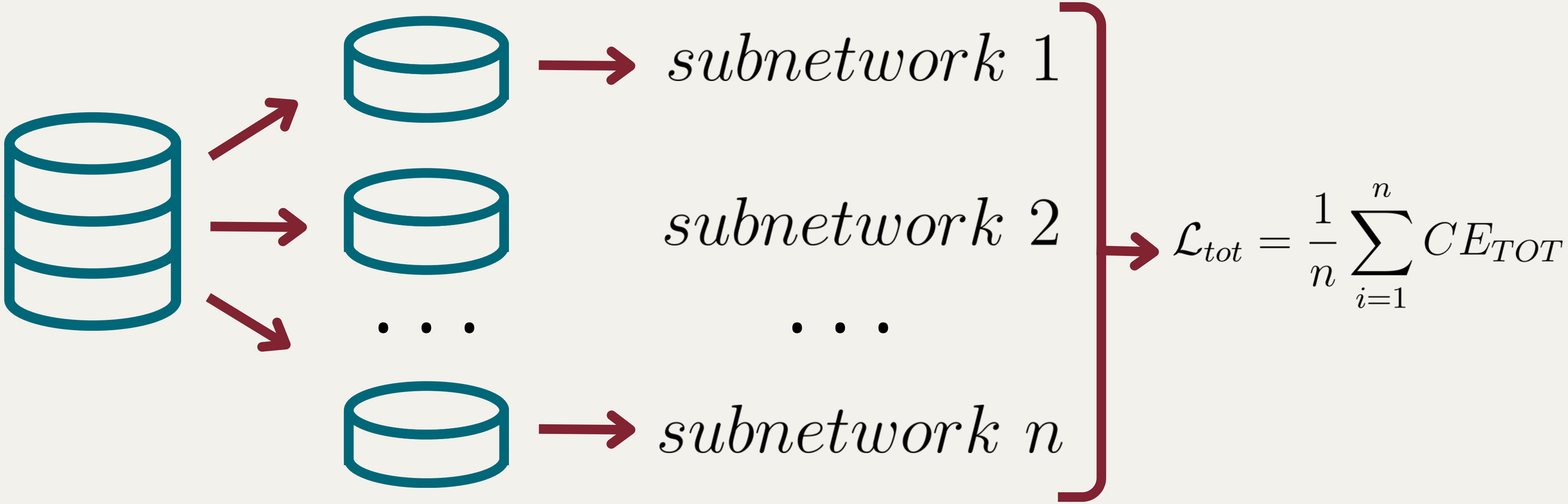
Ensemble by averaging

$$Ensemble\ Logits = \frac{1}{n}\sum_{1}^{n} Logits\ list$$

$$KL\ Divergence = \frac{1}{n}\sum_{i=1}^{n} \mathcal{D}_{\mathrm{KL}_i}(P_{Logits_i} \| P_{Ensemble\ Logits})$$

$$\mathcal{L} = \mathcal{L}_{CE_{\text{ensemble}}} - \lambda_{\mathrm{KL}} \cdot \mathcal{D}_{\mathrm{KL}}$$

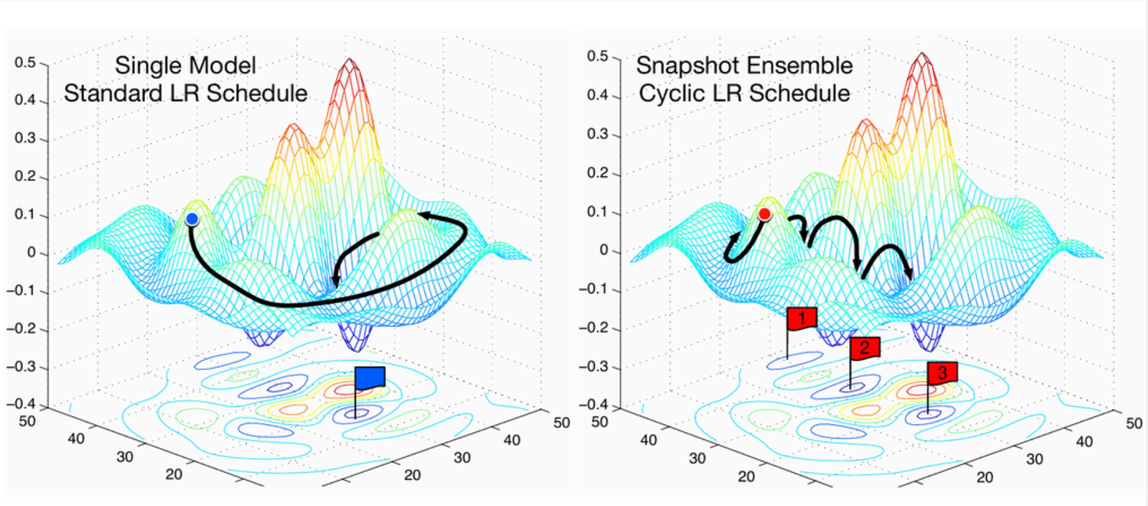
Ensemble by data partitioning

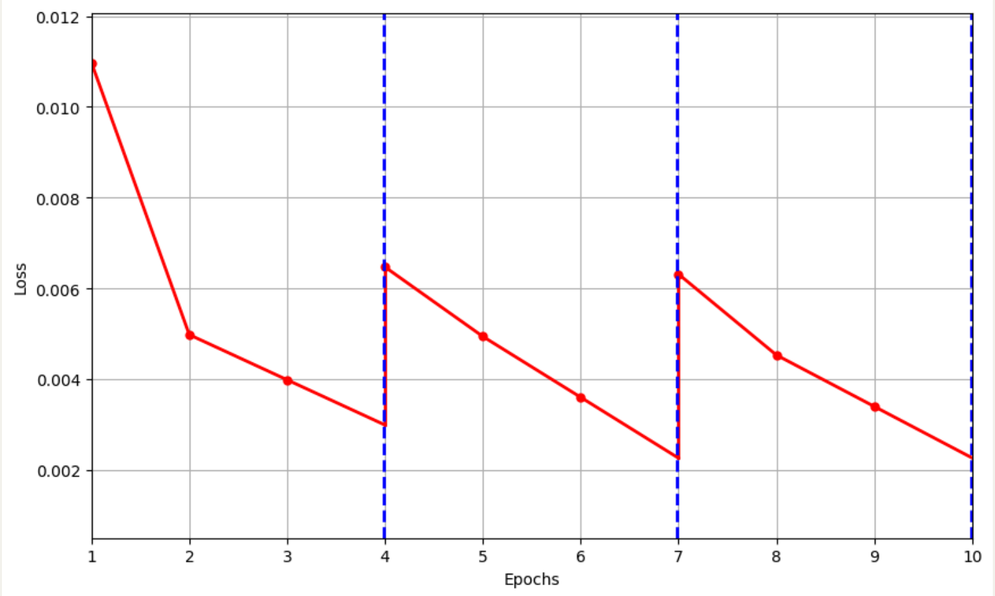


Original trainset

n partitions

Snapshot Ensemble



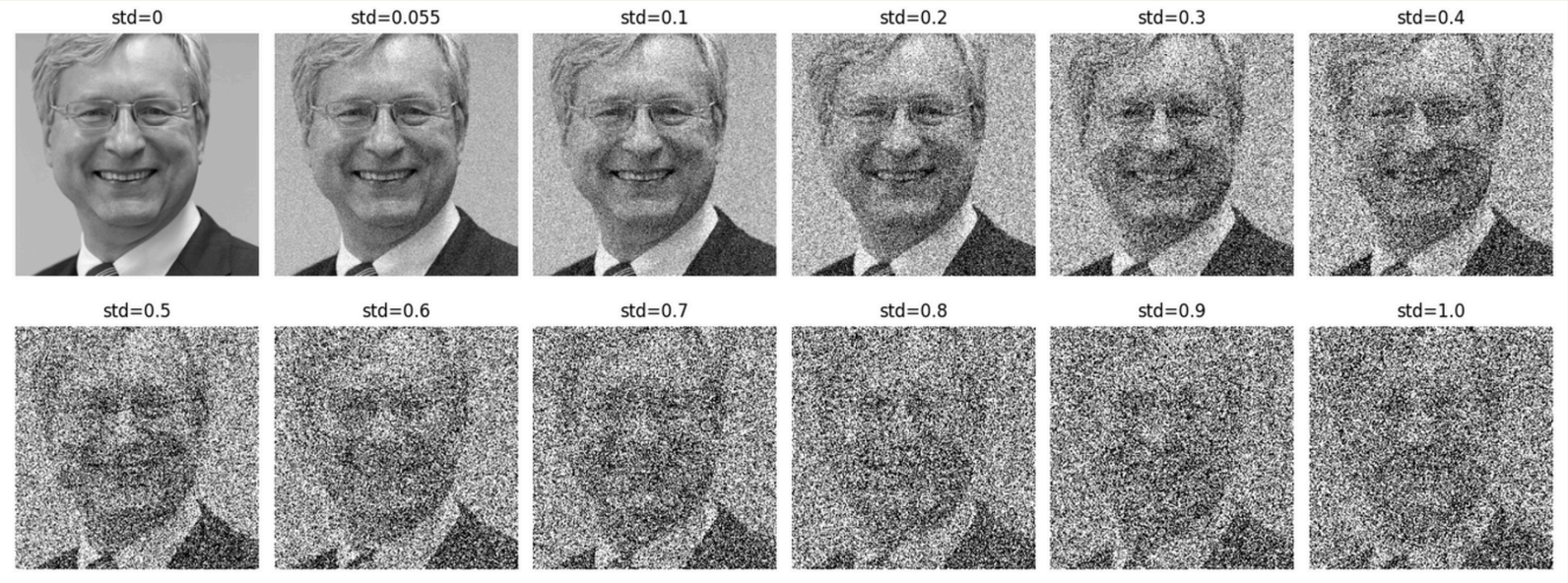


1. Still a **single** training
2. Up to **m** snapshots
3. Even less computation

Preliminary results

		Accuracy		ECE	
Standard	ViT Vanilla	0.944	**+4.5%**	0.042	**-66.7%**
	2x Vanilla	0.948	**+4.0%**	0.040	**-65.0%**
Custom ViT	**3 Subnetworks Averaging**	**0.986**		0.014	
	5 Subnetworks Partitioning	**0.986**		0.015	
	9 Subnetworks Snapshot	0.957		**0.013**	

Gaussian noised input



Accuracy

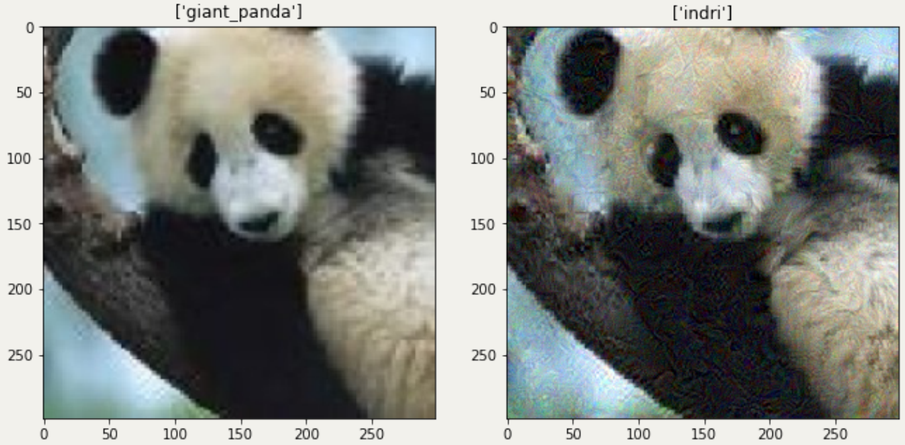
	std=0.1	std=1.0
3 Subnetworks Averaging	**+16.4%**	**+28.3%**
5 Subnetworks Partitioning	**+12.6%**	**+25.8%**
9 Subnetworks Snapshot	**+13.2%**	**+25.8%**

ECE

	std=0.1	std=1.0
3 Subnetworks Averaging	**-55.7%**	**-56.8%**
5 Subnetworks Partitioning	**-37.4%**	**-33.8%**
9 Subnetworks Snapshot	**-40.7%**	**-41.5%**

Adversarial attacks

FGSM - PGD

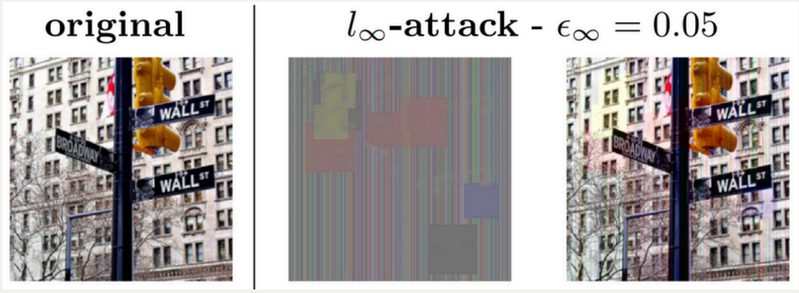


White box

CW

Black box

Square attack

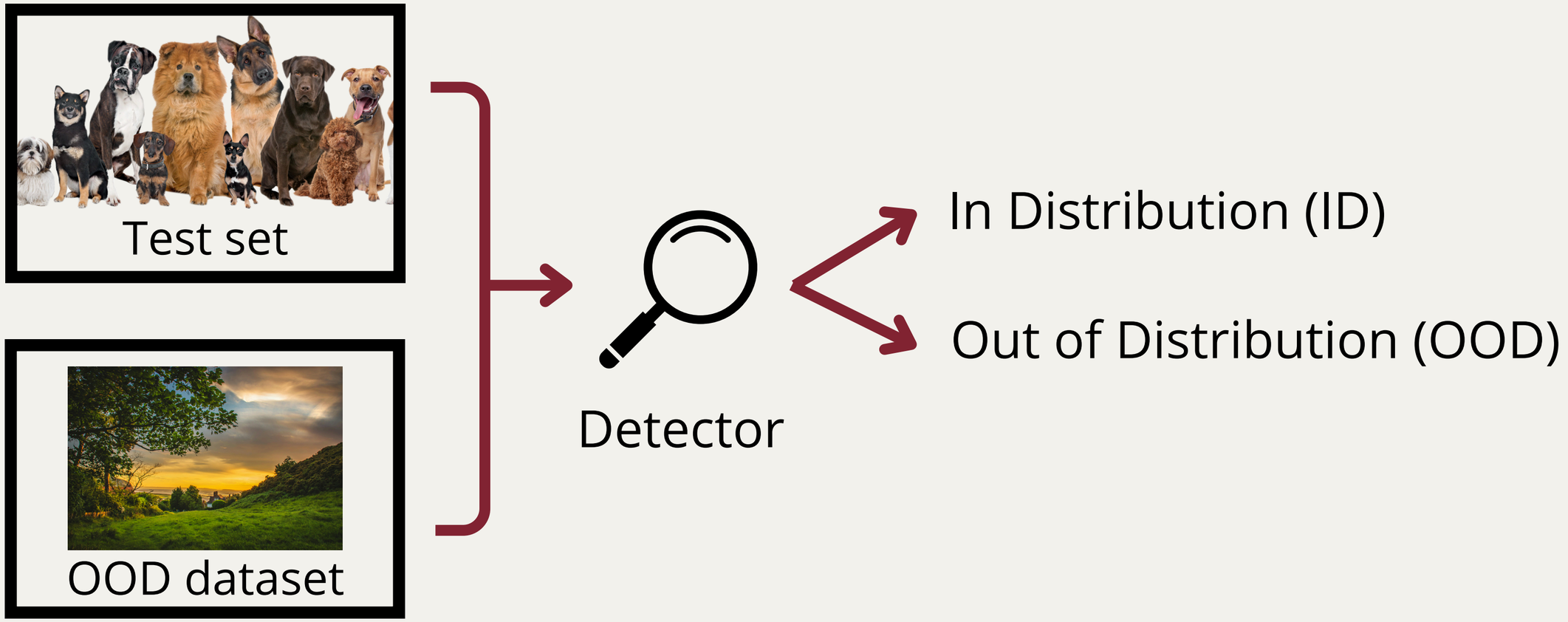


Adversarial attacks (2)

Accuracy and ECE comparison:

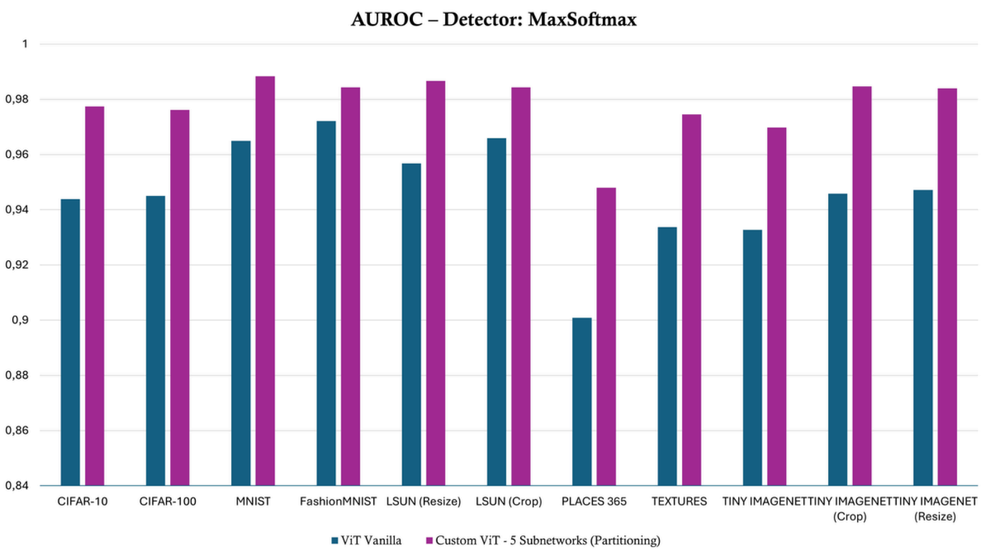
	FGSM		PGD		CW		Square		
ViT Vanilla	0.62	0.29	0.68	0.22	0.68	0.21	0.75	0.18	Standard
2 Vanilla	0.63	0.30	0.70	0.26	0.70	0.20	0.75	0.18	
2 Subnetworks Average	0.79	**0.16**	**0.78**	**0.16**	0.78	**0.14**	0.88	0.09	**Custom ViT**
3 Subnetworks Partitioning	**0.80**	0.18	**0.78**	0.20	**0.81**	**0.14**	**0.89**	**0.08**	
9 Subnetworks Snapshot	0.62	0.32	0.71	0.25	0.72	0.20	0.78	0.16	
	+28%	**-46%**	**+13%**	**-33%**	**+14%**	**-32%**	**+16%**	**-56%**	

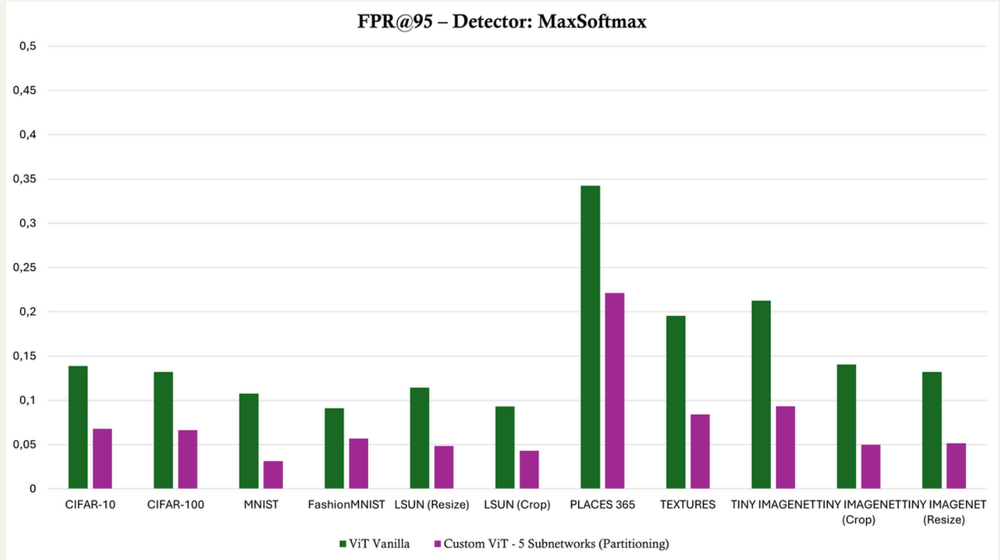
OOD analysis



OOD analysis (2)

Detector: MaxSoftmax



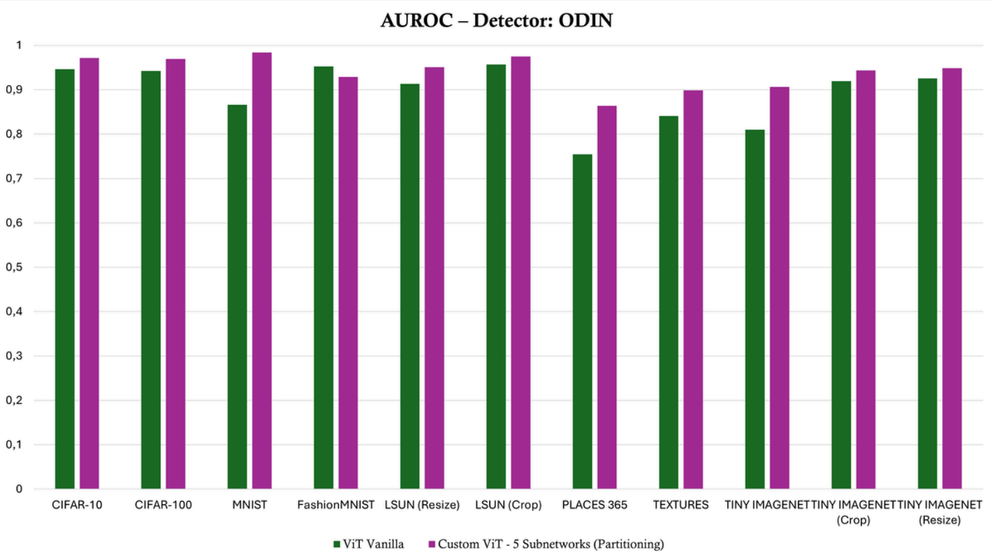


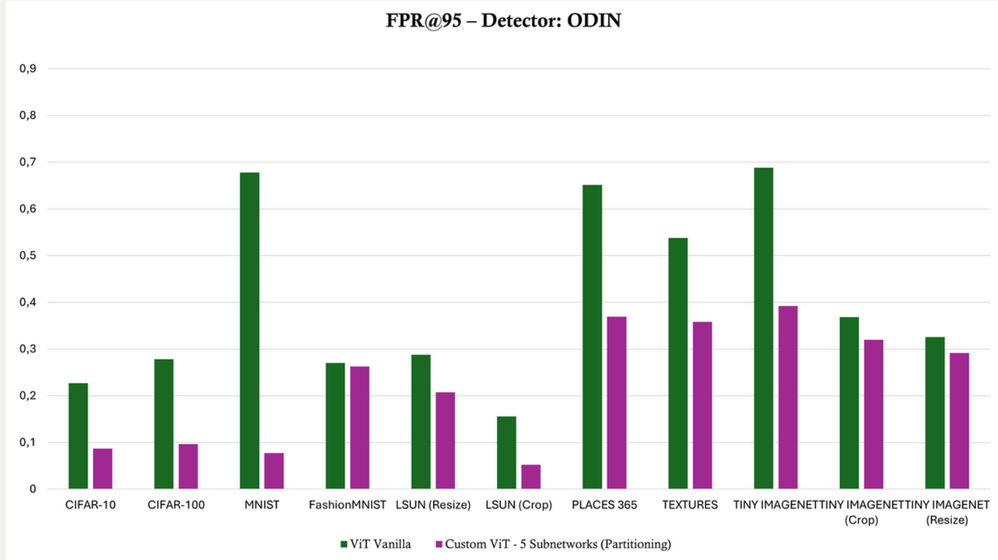
Average increment: +3.36%

Average reduction: -52.11%

OOD analysis (3)

Detector: ODIN





Average increment: +5.23%

Average reduction: -17.76%

Conclusions

The proposed models outperform vanilla ViTs in both accuracy and confidence.

Ensembling boosts robustness against noise, attacks, and OOD tasks.

Custom ViT flexibility allows ad hoc solutions for specific tasks.