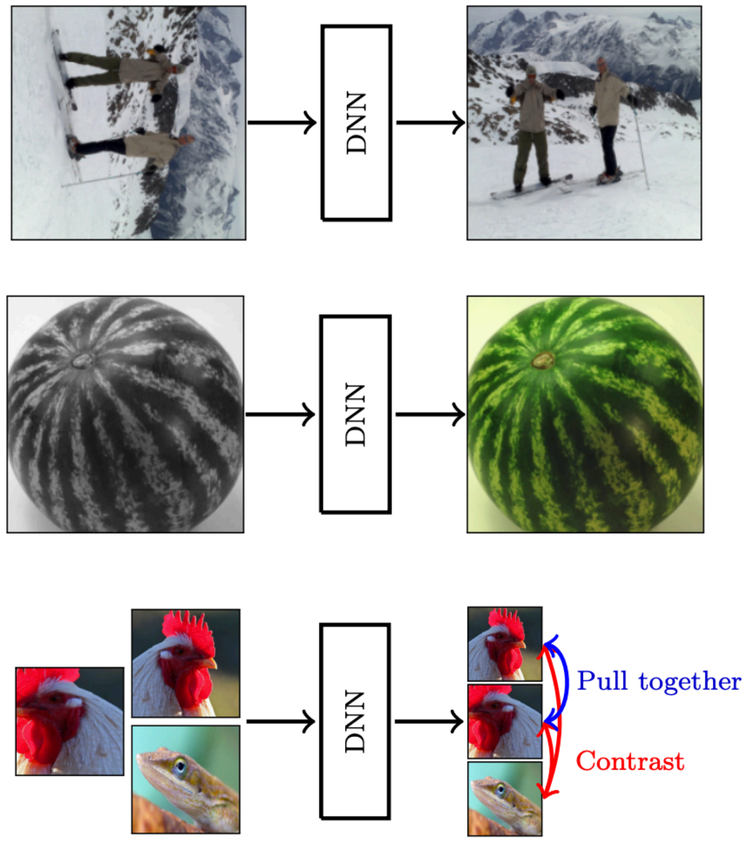
UNIVERSITEIT VAN AMSTERDAM *

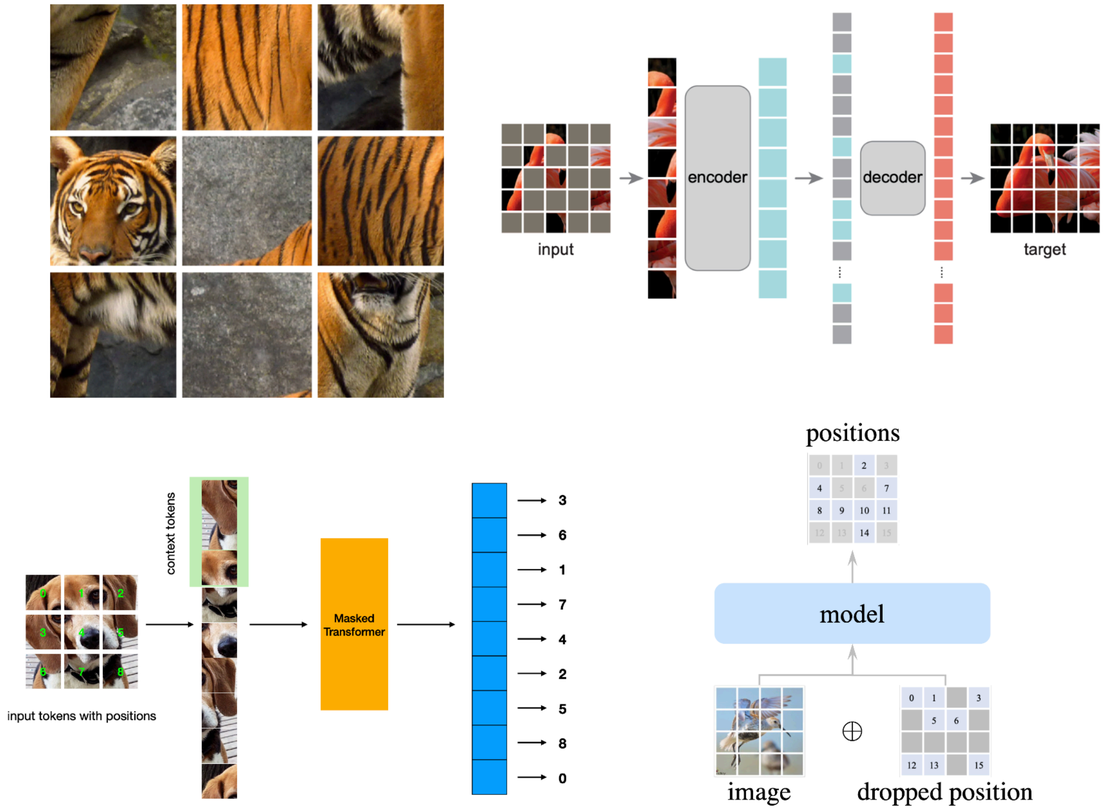
How *PART*s assemble into wholes: Learning the relative composition of images

Melika Ayoughi*, Samira Abnar, Chen Huang, Chris Sandino, Sayeri Lala, Eeshan Gunesh Dhekane, Dan Busbridge, Shuangfei Zhai, Vimal Thilak, Josh Susskind, Pascal Mettes*, Paul Groth*, Hanlin Goh

Self-supervised Learning: Global versus local



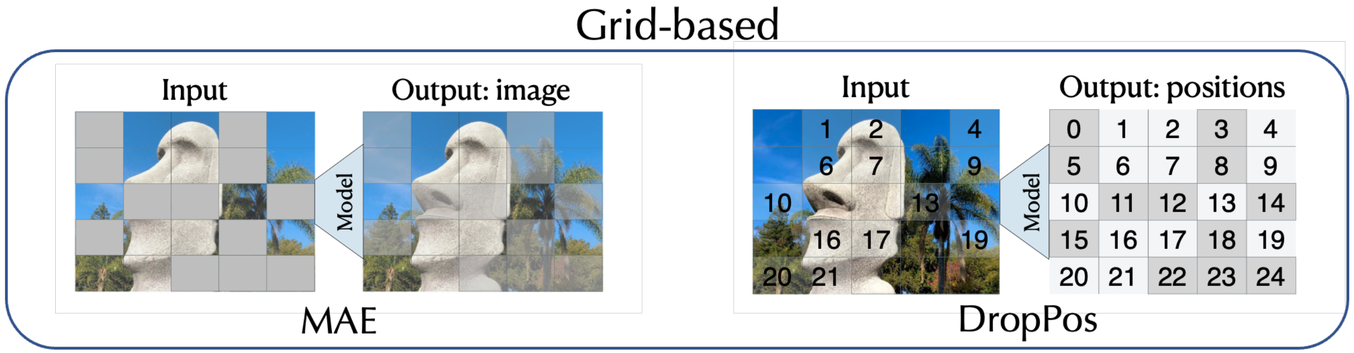
Global visual invariances

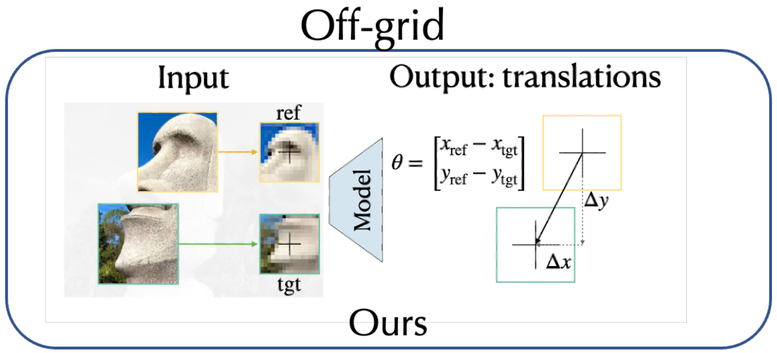


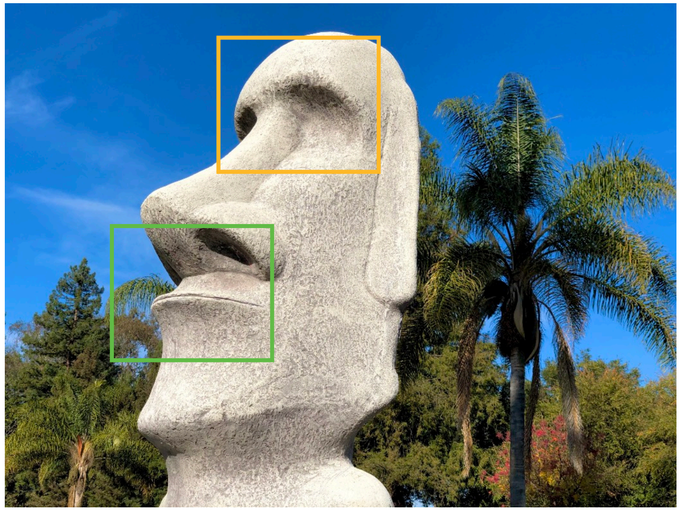
Local spatial structure

U. Ozbulak, H. J. Lee, B. Boga, E. T. Anzaku, H. Park, A. Van Messem, W. De Neve, and J. Vankerschaver. “Know your self-supervised learning: A survey on image-based generative and discriminative training”
Wang, Haochen, et al. "DropPos: Pre-Training Vision Transformers by Reconstructing Dropped Positions." NeurIPS 2023.

Grid Structure?

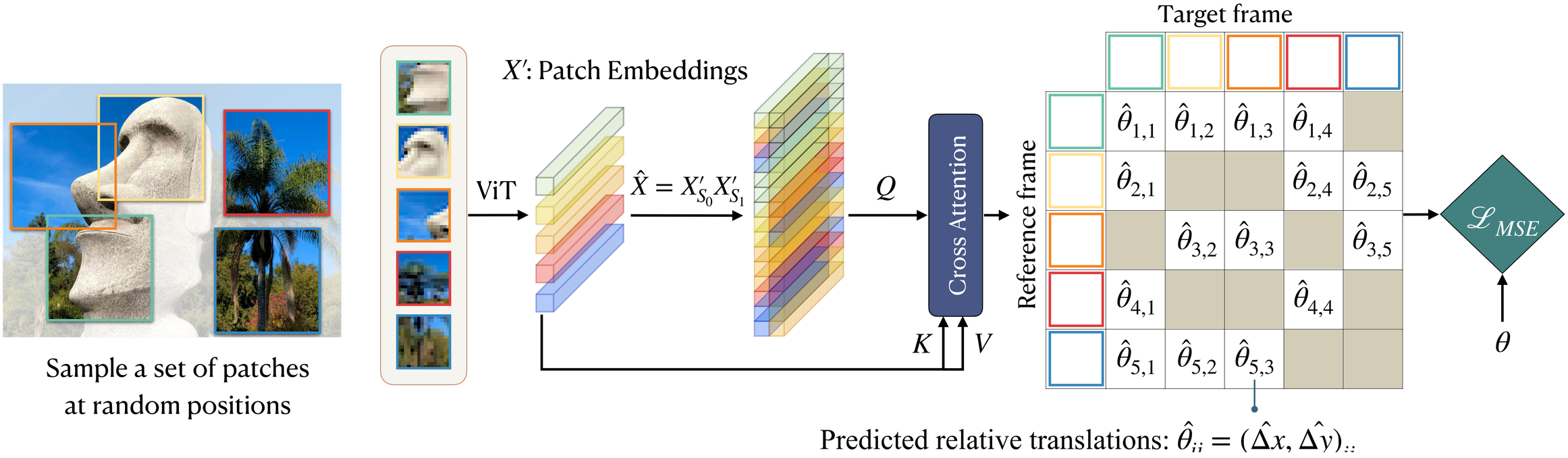






Current self-supervised learning approaches (i) rely on a fixed grid and (ii) focus on absolute pretext tasks

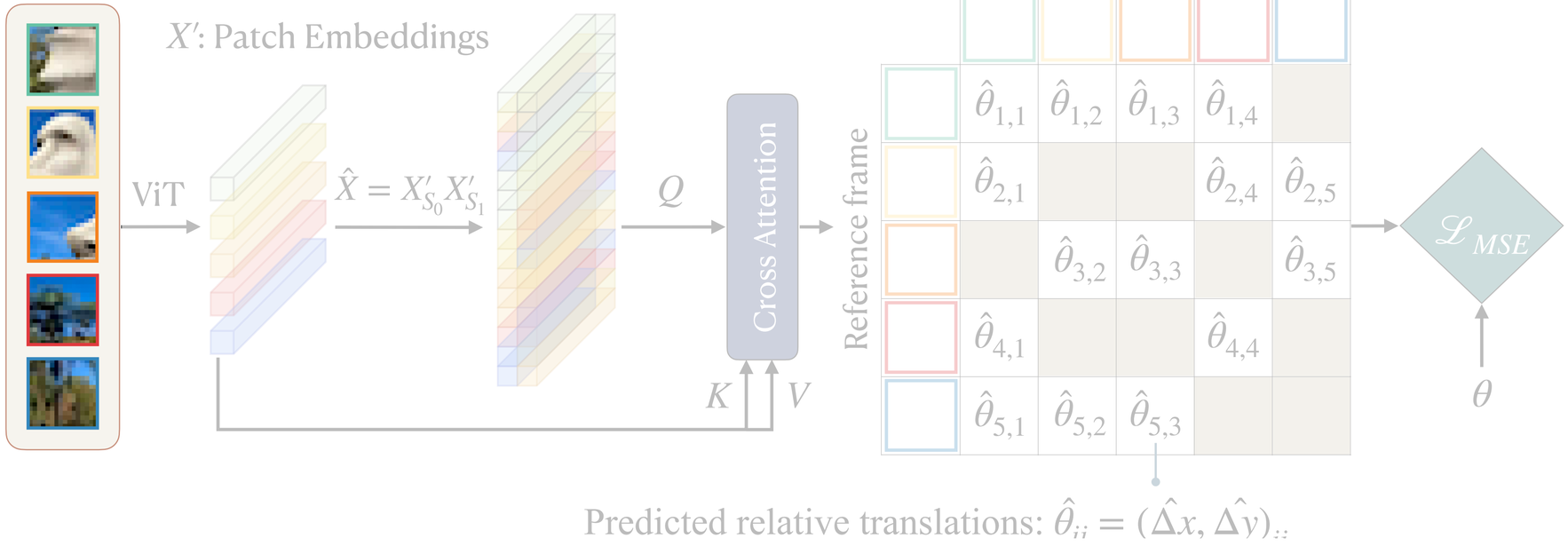
Overview of PART



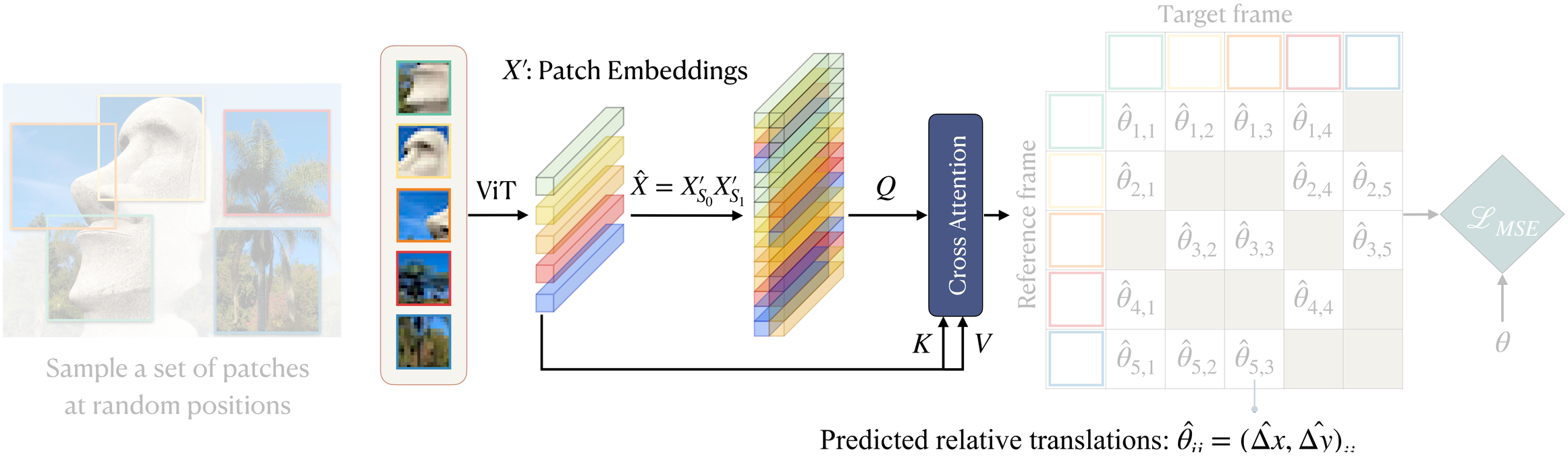
Sample a set of patches
at random positions

PART: Sampling

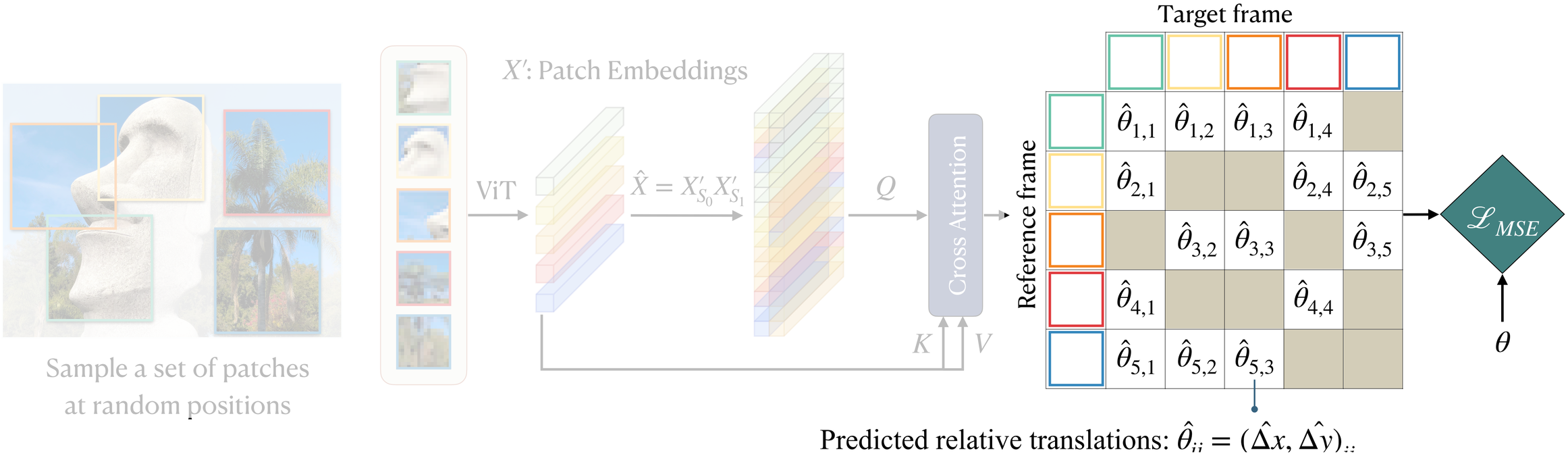
Sample a set of patches
at random positions



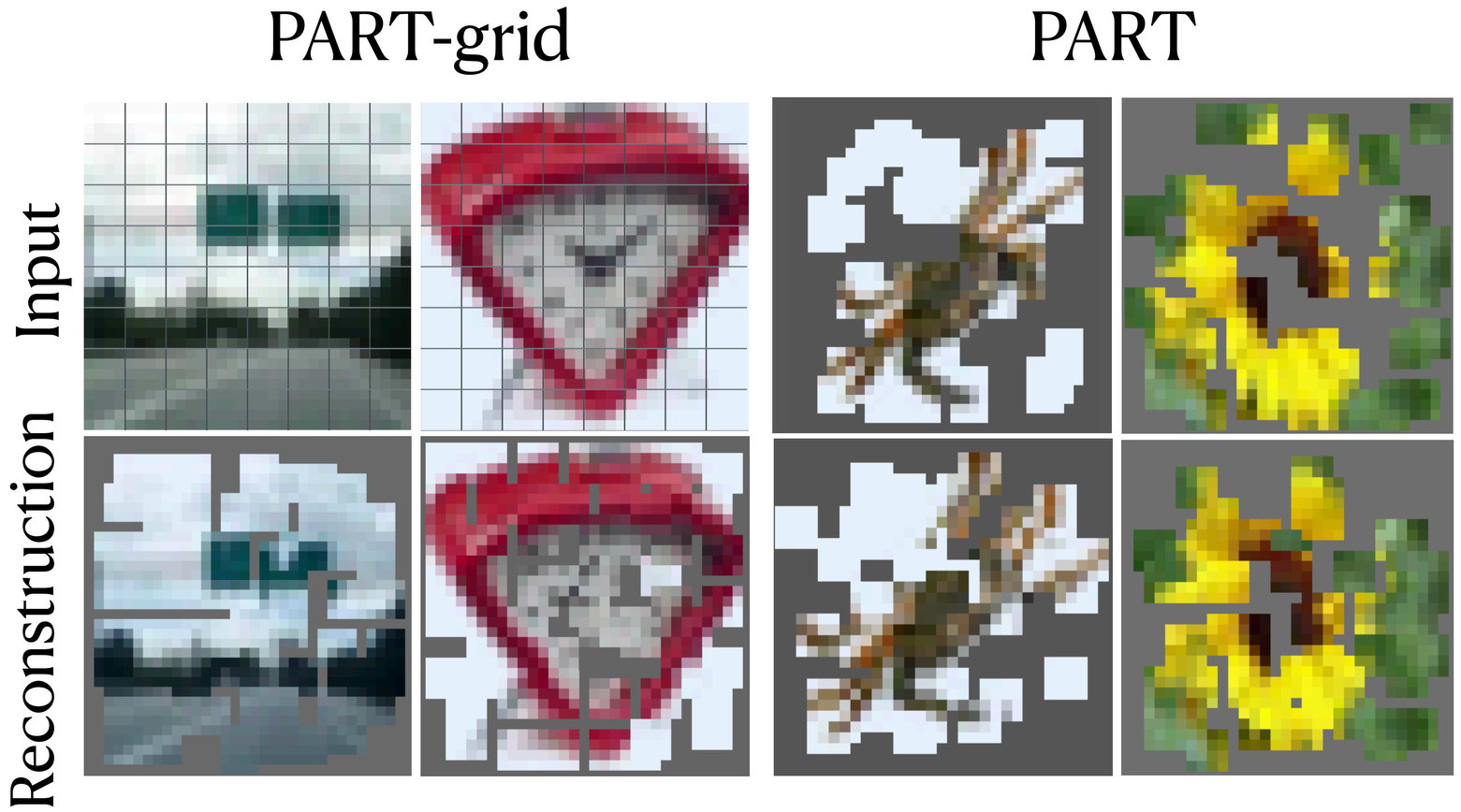
PART: Relative encoder architecture



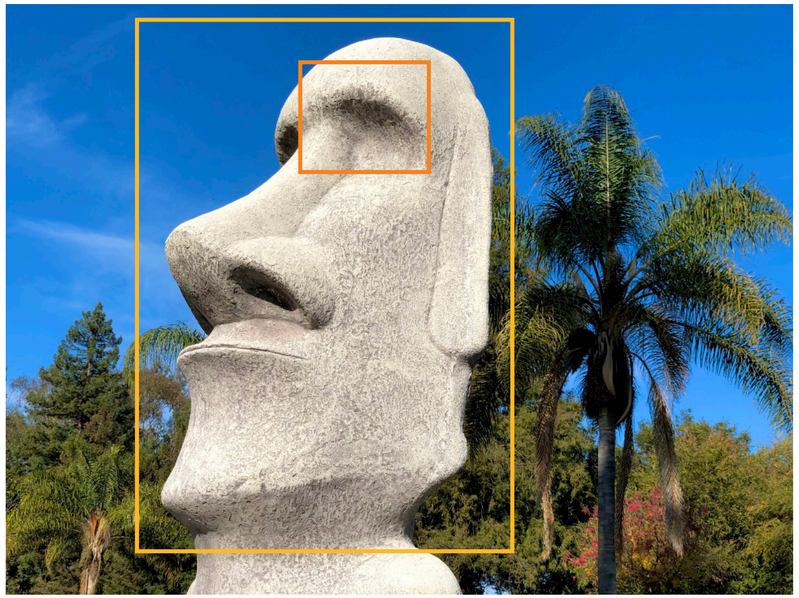
PART: Objective



Capabilities: Off-grid reconstruction



Capabilities: Extension to multiple aspect ratios and scales

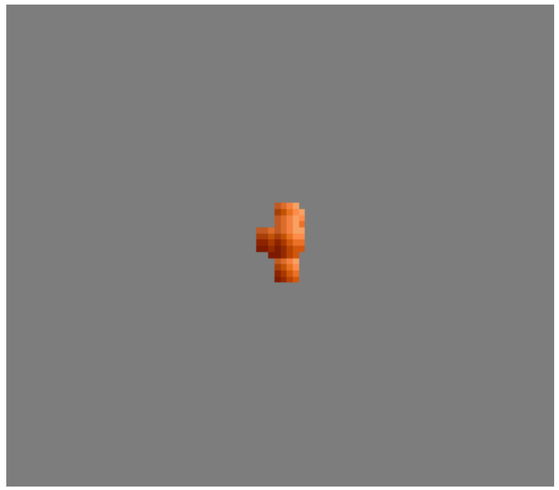


	COCO OD			INet Class.
	AP^b	AP^b_{50}	AP^b_{75}	Accuracy
PART	42.4	62.5	46.8	82.7
PART + aspect ratio + scale	42.0	61.8	46.3	82.6

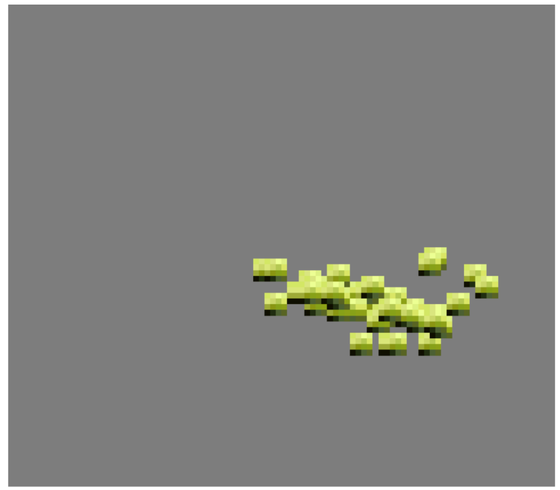
$$\theta_{\text{ref,tgt}} = (\Delta x, \Delta y, \Delta w, \Delta h)_{\text{ref,tgt}} = \Bigg(\underbrace{\frac{x_{\text{tgt}} - x_{\text{ref}}}{w_{\text{ref}}}, \frac{y_{\text{tgt}} - y_{\text{ref}}}{h_{\text{ref}}}}_{\text{relative position}} , \underbrace{\frac{w_{\text{tgt}}}{w_{\text{ref}}}, \frac{h_{\text{tgt}}}{h_{\text{ref}}}}_{\text{relative size}} \Bigg)$$

Capabilities: Patch uncertainty

Original Image

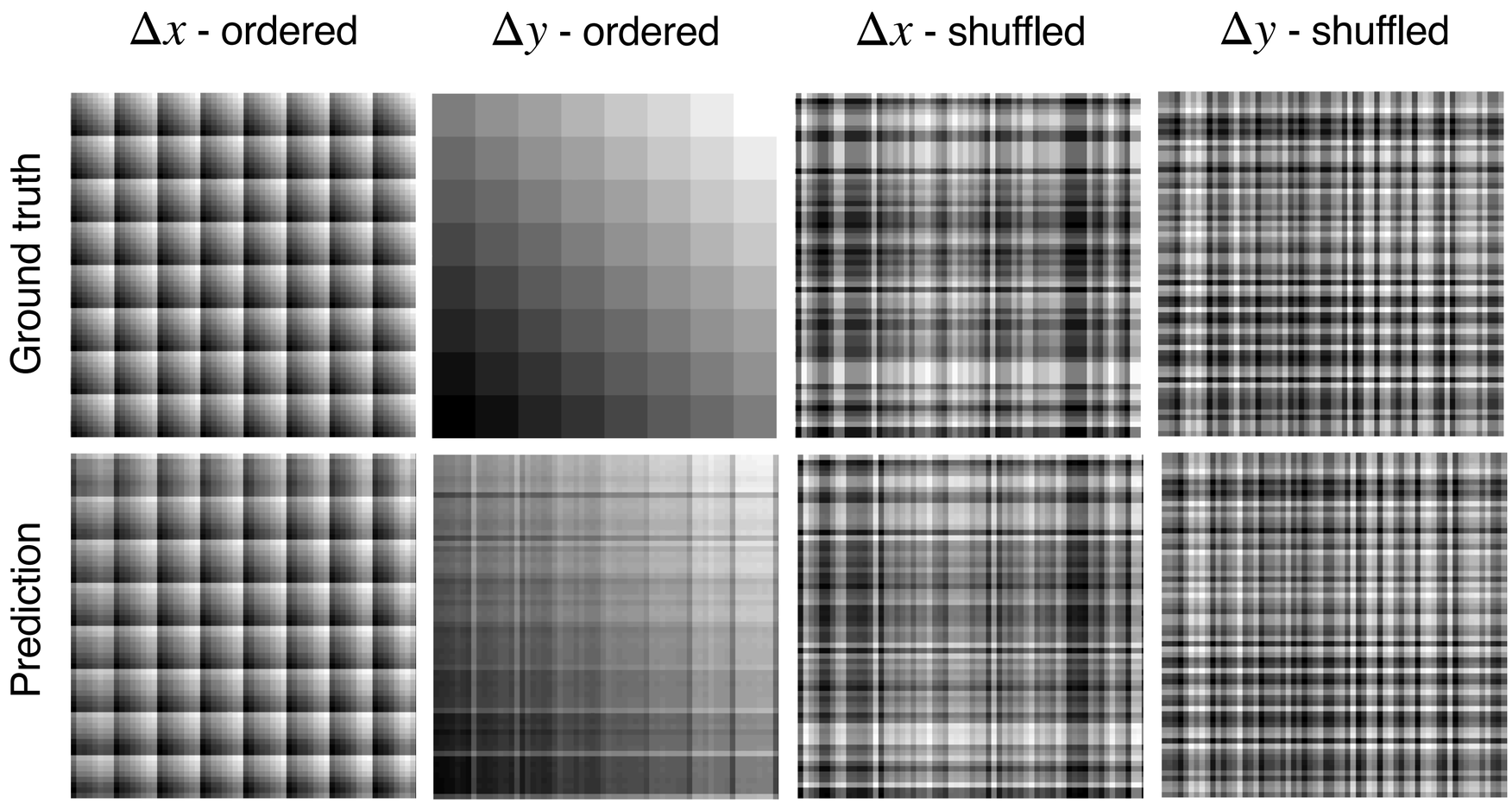


Certain Patch



Uncertain Patch

Capabilities: Symmetry



Comparison to Grid-based: Object detection

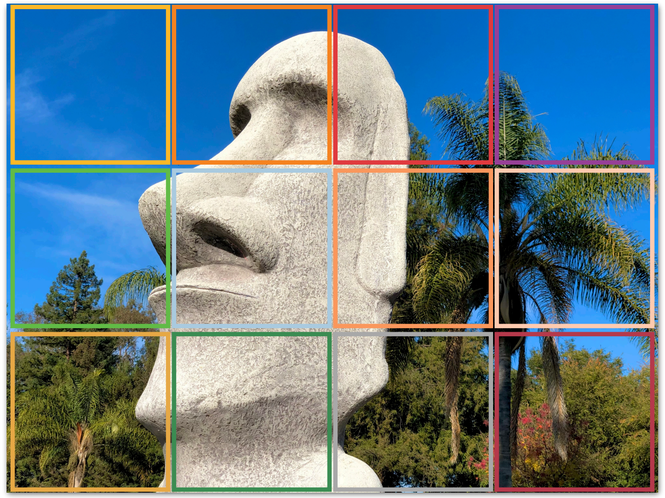
	AP^b	AP^b_{50}	AP^b_{75}
Grid-based			
MAE [20]♯	40.1	60.5	44.1
MP3 [21]†	41.8	61.4	46.0
DropPos [22]	42.1	62.0	46.4
Relative off-grid			
PART	**42.4**	**62.5**	**46.8**

Comparison to Grid-based: Time-series prediction

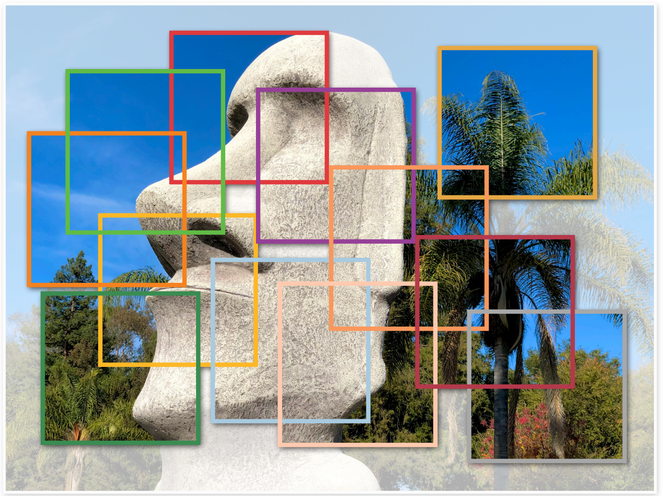
	PT	FT	Cohen's Kappa
Supervised			
Supervised w/ Pos Embed†	0	100	0.531
Grid-based			
MP3 [21]†	1000	100	0.553
DropPos [22]†	1000	100	0.582
MAE [20]†	1000	100	0.595
Relative off-grid			
PART	1000	100	**0.616**

Ablations: Sampling strategies

PART-grid



PART



	COCO	CIFAR-100	IN-1K	Time-series
PART-grid	41.4	82.1	82.43	0.500
PART	**42.4**	**83.0**	**82.7**	**0.616**

Ablations: Relative encoder

	Access to all patches	Shared weights
Fully-connected MLP	✔	✘
Pairwise MLP	✘	✔
Cross-attention	✔	✔

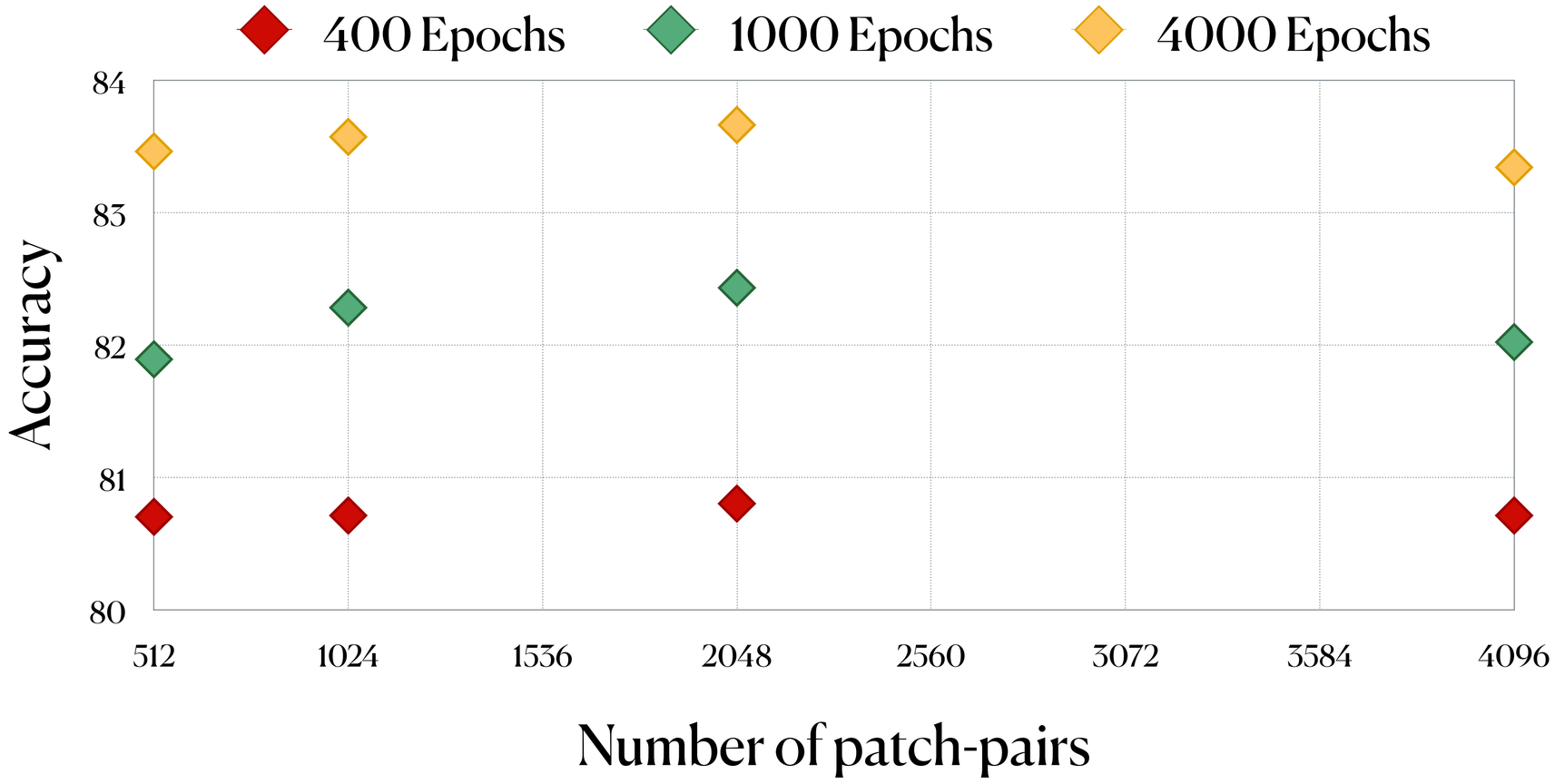
	Error ↓			**Accuracy ↑**
	x	y	Euclidean	
MLP	3.18	2.02	1.68	82.38
Pairwise MLP	2.84	1.76	1.59	82.52
Cross-attention	**1.14**	**0.77**	**0.81**	**83.00**

Does PART come at the cost of image classification?

	Pos Embed	PT	FT	Accuracy
Supervised				
Labelled baseline*	✓	0	300	81.8
Labelled baseline*		0	300	79.1
Contrastive				
MoCo v3 [66]♯	✓	300	150	83.2
DINO [45]♯	✓	300	300	82.8
BEiT [56]♯	✓	800	100	83.2
CIM [25]	✓	300	100	83.1
Grid-based				
MAE [20]*	✓	150	150	82.7
MAE [20]*	✓	1600	100	83.6
MP3 [21]†	✓	400	300	82.6
MP3 [21]		100	300	81.9
DropPos [22]	✓	200	100	83.0
Relative off-grid				
PART		400	300	82.7

	Pos Embed	PT	Accuracy
Supervised			
Labelled baseline♯	✓	0	73.6
Labelled baseline♯		0	64.6
Contrastive			
MoCo v3 [66] ♯	✓	2000	83.3
Grid-based			
MAE [20]♯	✓	2000	84.5
MP3 [21]	✓	2000	84.0
MP3 [21]		2000	82.6
Relative off-grid			
PART		1000	83.0

Ablations: Number of patch pairs



Discussion and Future Work

- Complementary to contrastive learning
- Hierarchical multi-scale learning
- Modeling rotations
- Extension to other tasks
- Universal pretraining across diverse data types and domains

Thank you!

Please reach out to me for discussions and collaborations.

m.ayoughi@uva.nl