ICCV
OCT 19-23, 2025
HONOLULU HAWAII

EQUICAPS: PREDICTOR-FREE POSE-AWARE PRE-TRAINED CAPSULE NETWORKS


Athinoulla Konstantinou
University of Aberdeen

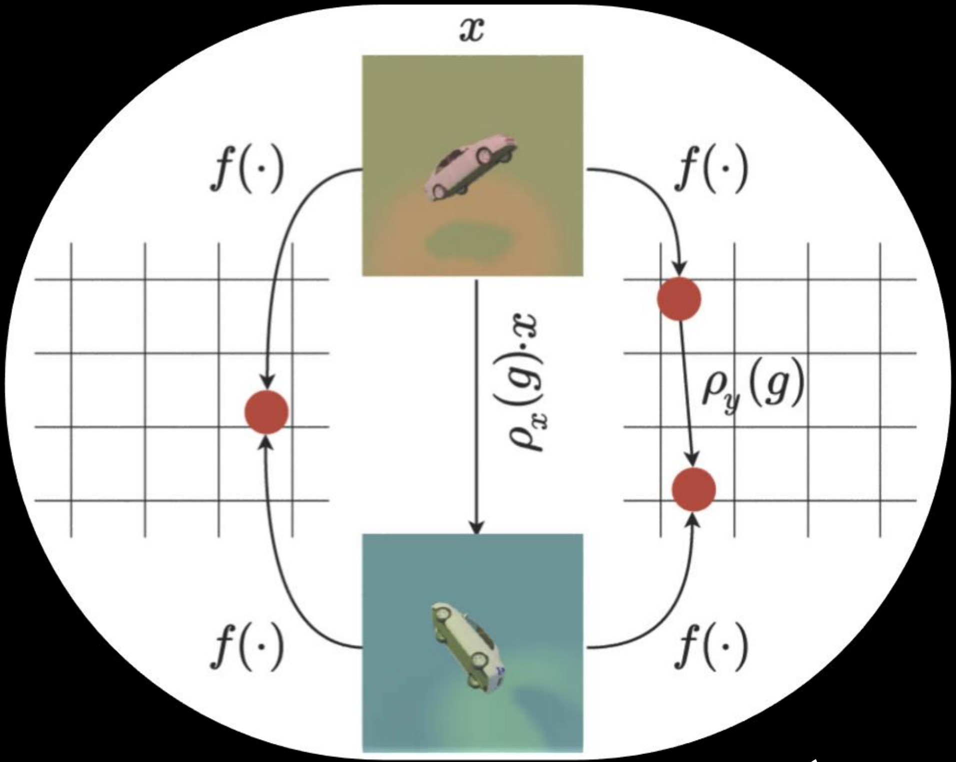

Georgios Leontidis
University of Aberdeen

Mamatha Thota
University of Lincoln


Aiden Durrant
University of Aberdeen

UNIVERSITY OF ABERDEEN
UNIVERSITY OF LINCOLN
UK Research and Innovation
SUSTAIN

INVARIANT VS. EQUIVARIANT SELF-SUPERVISION



Invariance

$$\mathcal{L}_{inv} = \mathcal{L}\big(f(p_x(g) \cdot x), f(x)\big)$$

Equivariance

$$\mathcal{L}_{equi} = \mathcal{L}\Big(f(p_x(g) \cdot x), p_y(g) \cdot f(x)\Big)$$

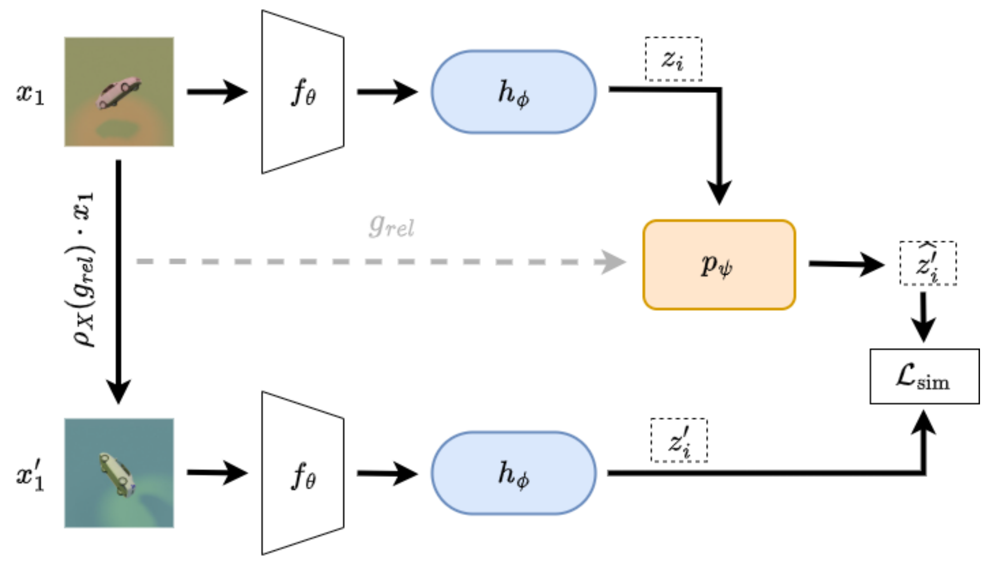
Visual recognition involves not only identifying **what** an object is but also understanding **how** it is presented [1].

[1] Jiayun Wang, Yubei Chen, and Stella Yu. Pose-aware self-supervised learning with viewpoint trajectory regularization. In *ECCV*, 2024.

MOTIVATION

Most equivariant SSL (e.g., SIE, EquiMod) enforce equivariance via objective functions/predictor.

- **Few exploit equivariant architectures in SSL.**
- They use a **predictor** p_ψ s.t. $\widehat{z_i'} = p_\psi(z_i, g_{rel})$,
- produce ad hoc representations that are **hard to interpret and manipulate**,
- rely on architectures (e.g., CNNs) that are **not naturally equivariant**, and
- add **extra complexity** via extra modules.



LEVERAGE CAPSULE NETWORKS' INDUCTIVE BIASES

Most equivariant SSL (e.g., SIE, EquiMod) enforce equivariance via objective functions/predictor.

- Few exploit equivariant architectures in SSL.
- They use a predictor p_ψ s.t. $\hat{z} = p_\psi(z', g_{rel})$,
- rely on architectures (e.g., CNNs) that are not naturally equivariant, and
- produce ad hoc representations that are hard to interpret and manipulate,
- add extra complexity via extra modules.

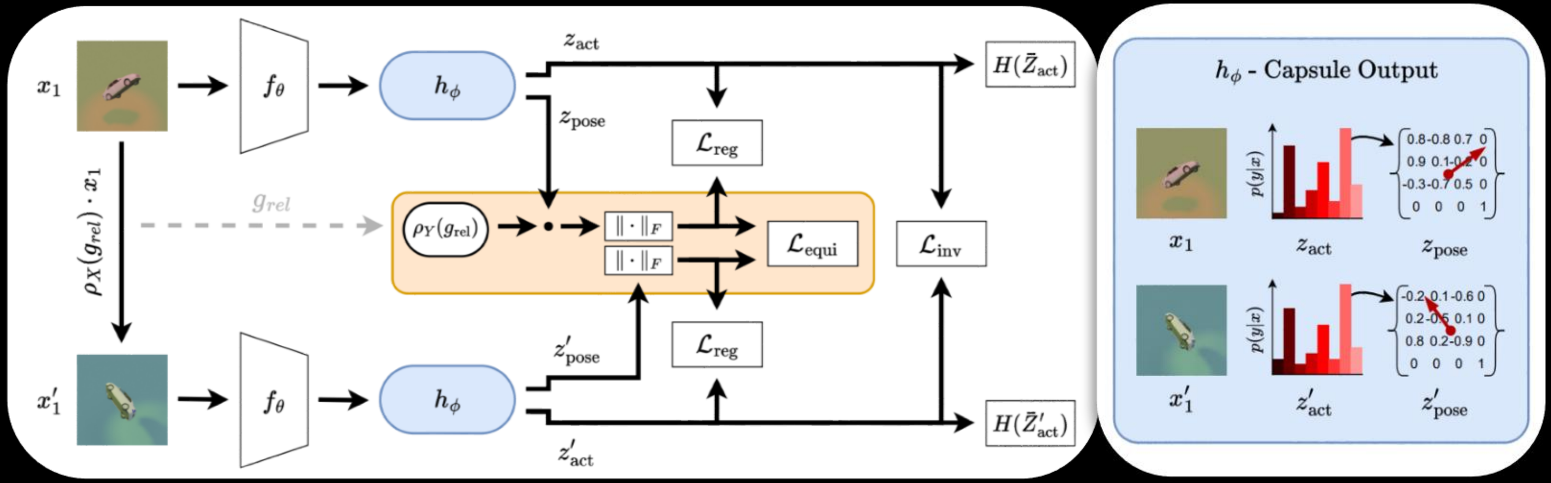
Inductive biases → ✓ ✓ ✓

Using routing based on agreement of **part-whole relationships**, naturally encode both:

- the existence of an entity (**invariance**), and
- its instantiation parameters (**equivariance**).

- **We directly** leverage capsules' equivariant properties,
- gain intuitive **control and interpretability** of the representations (4x4 pose matrices),
- **and keep a streamlined framework.**

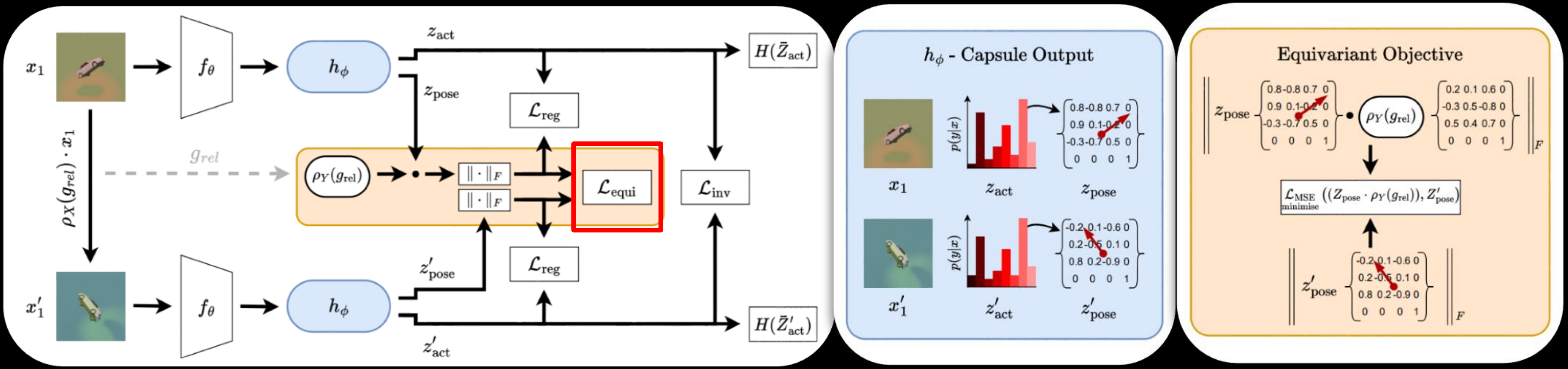
EQUICAPS: PREDICTOR-FREE POSE-AWARE SSL



- To reduce computation, rely on the **non-iterative** self-routing [2] algorithm.
- The activation vectors encode **transformation-invariant** properties
- The pose matrices capture **transformation-equivariant** properties.

[2] Taeyoung Hahn, Myeongjang Pyeon, and Gunhee Kim. Self-routing capsule networks. In *NeurIPS*, 2019.

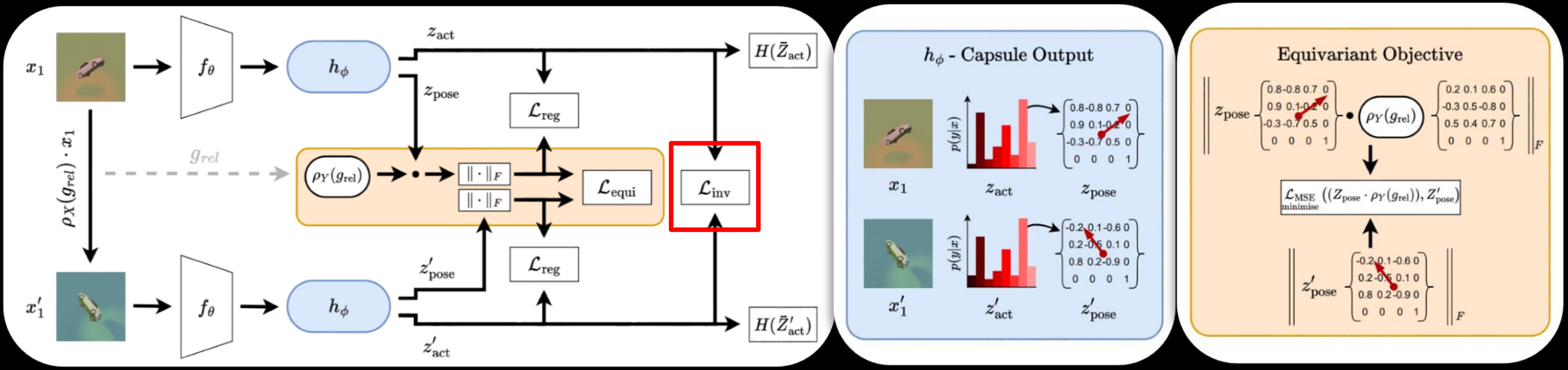
EQUICAPS: PREDICTOR-FREE POSE-AWARE SSL



$$\mathcal{L}_{equi} = \frac{1}{B}\sum_{i=1}^{B}\left\|\frac{Z_{i,pose}\cdot p_y(g_{rel,i})}{\left\|Z_{i,pose}\cdot p_y(g_{rel,i})\right\|_F} - \frac{Z'_{i,pose}}{\left\|Z'_{i,pose}\right\|_F}\right\|_2^2 .$$

This direct manipulation in the latent space removes the need for a predictor.

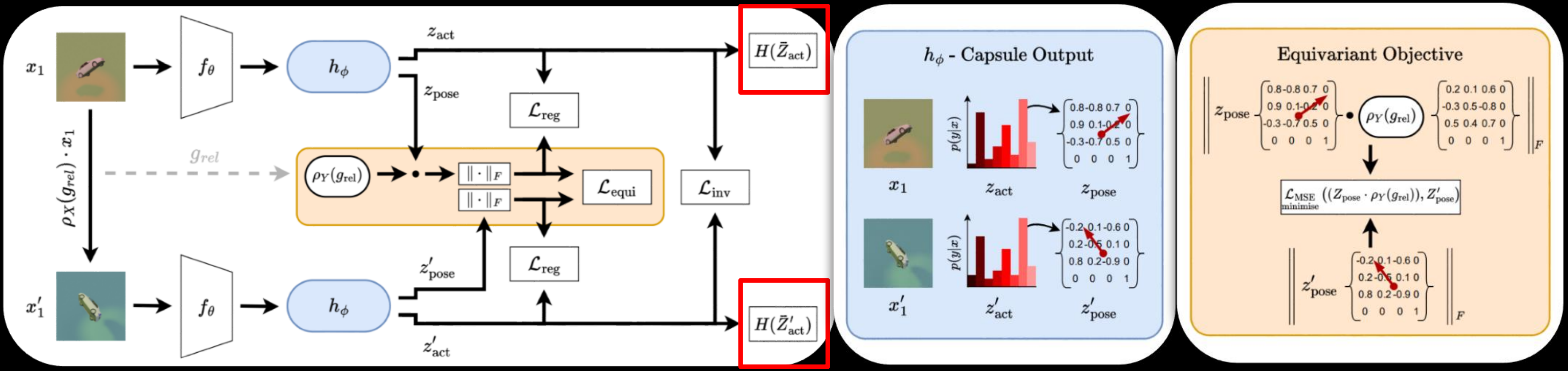
EQUICAPS: PREDICTOR-FREE POSE-AWARE SSL



$$\mathcal{L}_{equi} = \frac{1}{B}\sum_{i=1}^{B}\left\|\frac{Z_{i,pose} \cdot p_y(g_{rel,i})}{\|Z_{i,pose} \cdot p_y(g_{rel,i})\|_F} - \frac{Z'_{i,pose}}{\|Z'_{i,pose}\|_F}\right\|_2^2.$$

$$\mathcal{L}_{inv} = H(Z_{act}, Z'_{act}).$$

EQUICAPS: PREDICTOR-FREE POSE-AWARE SSL

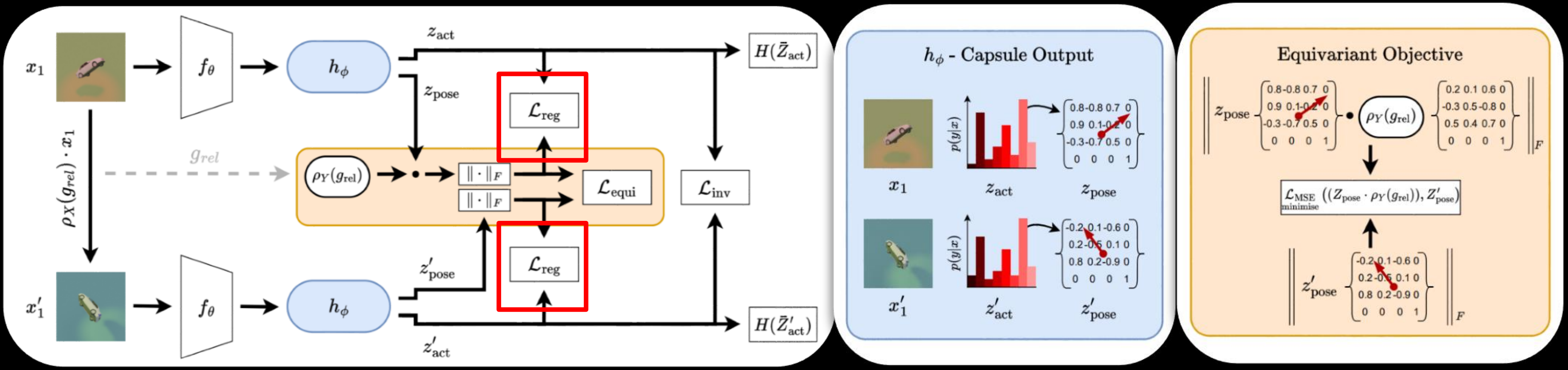


$$\mathcal{L}_{equi} = \frac{1}{B}\sum_{i=1}^{B}\left\|\frac{Z_{i,pose}\cdot p_y(g_{rel,i})}{\|Z_{i,pose}\cdot p_y(g_{rel,i})\|_F} - \frac{Z'_{i,pose}}{\|Z'_{i,pose}\|_F}\right\|_2^2.$$

$$\mathcal{L}_{inv} = H(Z_{act}, Z'_{act}).$$

$$\mathcal{L}_{ME-MAX} = H(\overline{Z}_{act}) + H(\overline{Z}'_{act}).$$

EQUICAPS: PREDICTOR-FREE POSE-AWARE SSL



$$\mathcal{L}_{equi} = \frac{1}{B}\sum_{i=1}^{B}\left\|\frac{Z_{i,pose}\cdot p_y(g_{rel,i})}{\|Z_{i,pose}\cdot p_y(g_{rel,i})\|_F} - \frac{Z'_{i,pose}}{\|Z'_{i,pose}\|_F}\right\|_2^2.$$

$$\mathcal{L}_{inv} = H(Z_{act}, Z'_{act}).$$

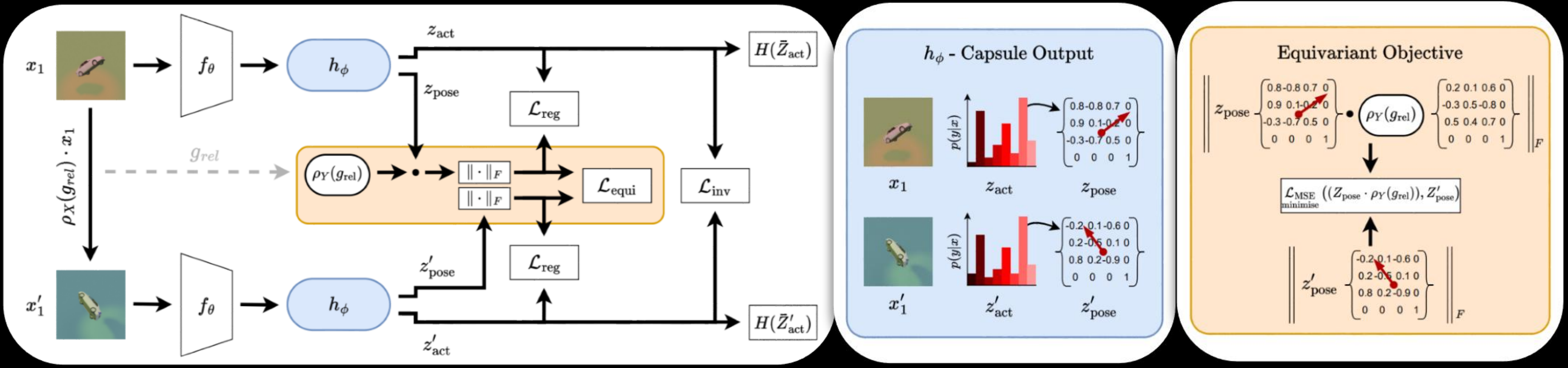
$$\mathcal{L}_{ME-MAX} = H(\bar{Z}_{act}) + H\left(\bar{Z}'_{act}\right).$$

$$\mathcal{L}_{reg}(Z_{cat}) = \lambda_V V(Z_{cat}) + \lambda_C C(Z_{cat}) \text{ where}$$

$$V(Z_{cat}) = \frac{1}{d}\sum_{j=1}^{d} max\left(0, 1 - \sqrt{Var(Z_{cat\cdot,j})}\right),$$

$$C(Z_{cat}) = \frac{1}{d}\sum_{i\neq j} Cov(Z_{cat})_{i,j}^2.$$

EQUICAPS: PREDICTOR-FREE POSE-AWARE SSL



The overall loss is a **combination**:

$$\mathcal{L}_{EquiCaps} = \lambda_{inv} H(Z_{act}, Z'_{act}) + H(\overline{Z}_{act}) + H\left(\overline{Z'}_{act}\right)$$
$$+ \ \lambda_{equi} \frac{1}{B} \sum_{i=1}^{B} \left\| \frac{Z_{i,pose} \cdot p_y(g_{rel,i})}{\|Z_{i,pose} \cdot p_y(g_{rel,i})\|_F} - \frac{Z'_{i,pose}}{\|Z'_{i,pose}\|_F} \right\|_2^2$$
$$+ \mathcal{L}_{reg}(Z_{cat}) + \mathcal{L}_{reg}(Z_{cat}).$$

EquiCaps can theoretically handle **any transformation** which can be expressed as **a matrix** without architectural changes.

3DIEBENCH-T: INVARIANT-EQUIVARIANT BENCHMARK

- Extends 3DIEBench from SO(3) to **SE(3)**, increasing task complexity.
- Comprises:
    - **2,623,600** images
    - **55** classes
    - rendered from **52,472** ShapeNetCoreV2 3D models
    - under **50** (simultaneous SE(3) + colour) transformations per model.

QUANTITATIVE RESULTS

Pre-train for rotation equivariance only

Method	Classification (Top-1)		Rotation (R^2)		Translation (R^2)	Colour (R^2)	
	3DIEBench	3DIEBench-T	3DIEBench	3DIEBench-T	3DIEBench-T	3DIEBench	3DIEBench-T
Supervised Methods							
ResNet-18	86.45	80.13	0.77	0.73	0.67	0.99	0.99
Invariant and Parameter Prediction Methods							
VICReg	84.28	74.71	0.45	0.39	0.22	0.10	0.50
SimCLR	86.73	80.08	0.52	0.44	0.25	0.29	0.50
SimCLR + AugSelf	**87.44**	**80.86**	**0.75**	**0.69**	**0.50**	0.28	0.51
Equivariant Methods							
SEN	86.99	80.20	0.51	0.45	0.26	0.29	0.47
EquiMod	**87.39**	**80.76**	0.50	0.43	0.24	0.29	0.38
SIE	82.94	75.56	0.73	0.45	0.20	0.07	0.46
CapsIE	79.14	75.20	0.74	0.60	0.46	0.01	0.03
EquiCaps	83.24	76.91	**0.78**	**0.73**	**0.60**	0.09	0.05

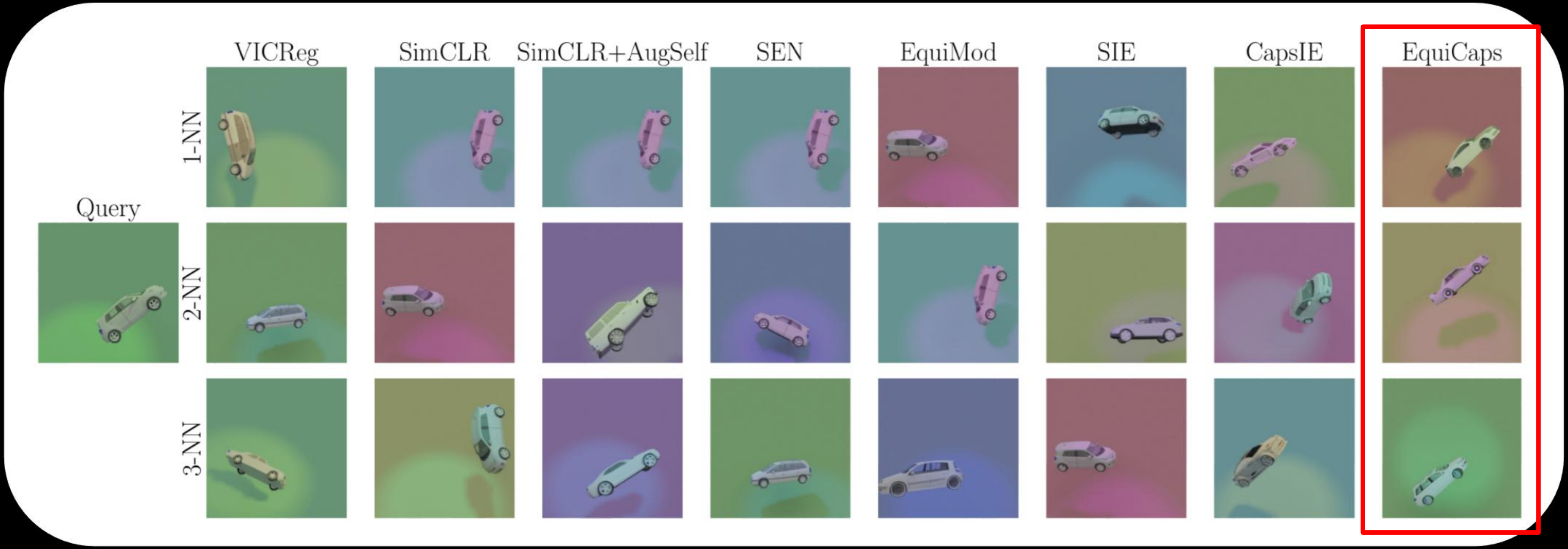
QUANTITATIVE RESULTS

Pre-train for rotation & translation equivariance

Method	Classification (Top-1)	Rotation (R^2)	Translation (R^2)	Colour (R^2)
SimCLR + AugSelf	81.04 $\uparrow$ 0.18	0.69 = 0.00	0.64 $\uparrow$ 0.14	0.51 = 0.00
SEN	80.23 $\uparrow$ 0.03	0.46 $\uparrow$ 0.01	0.28 $\uparrow$ 0.02	0.50 $\uparrow$ 0.03
EquiMod	**80.89** $\uparrow$ 0.13	0.46 $\uparrow$ 0.03	0.37 $\uparrow$ 0.13	0.37 $\downarrow$ 0.01
SIE	75.91 $\uparrow$ 0.35	0.48 $\uparrow$ 0.03	0.22 $\uparrow$ 0.02	0.36 $\downarrow$ 0.10
CapsIE	76.31 $\uparrow$ 1.11	0.62 $\uparrow$ 0.02	0.53 $\uparrow$ 0.07	0.03 = 0.00
EquiCaps	77.88 $\uparrow$ 0.97	**0.71** $\downarrow$ 0.02	**0.61** $\uparrow$ 0.01	0.02 $\downarrow$ 0.03

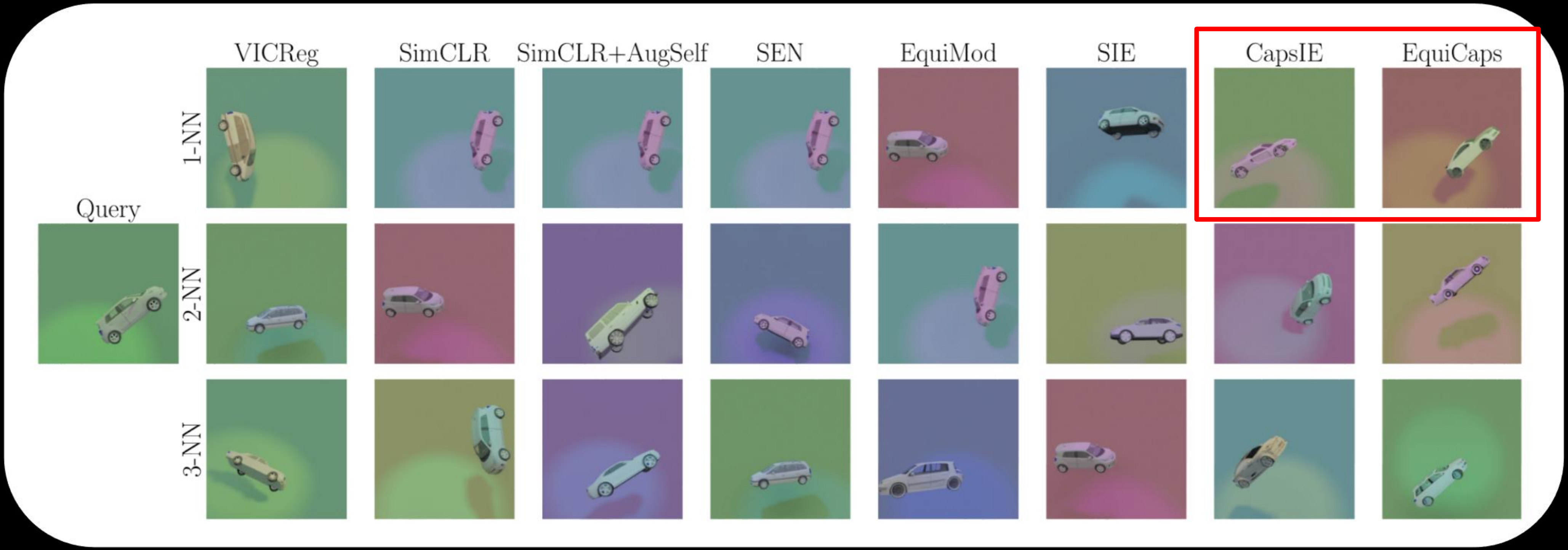
QUALITATIVE RESULTS

k-NN representation retrieval



QUALITATIVE RESULTS

k-NN representation retrieval



MAIN TAKEAWAYS

- **EquiCaps** (predictor-free equivariance)
  - Capsule-based projector
  - Controllable and interpretable latent space
- 3DIEBench-T (**SE(3)** benchmark)
- Extensive experiments
  - **SOTA** on rotation and translation prediction among the equivariant baselines
  - Capsule architectures show **improved generalisation** under combined SE(3) transformations and in transfer learning (including object detection)

THANK YOU

ArXiv

Code

Dataset

Contact: a.konstantinou.24@abdn.ac.uk

This work was supported by UKRI [grant ref: EP/Y03063X/1].