Soft Equivariance Regularization for Invariant Self-Supervised Learning

Joohyung Lee[1]

Changhun Kim[1]

Hyunsu Kim[3]

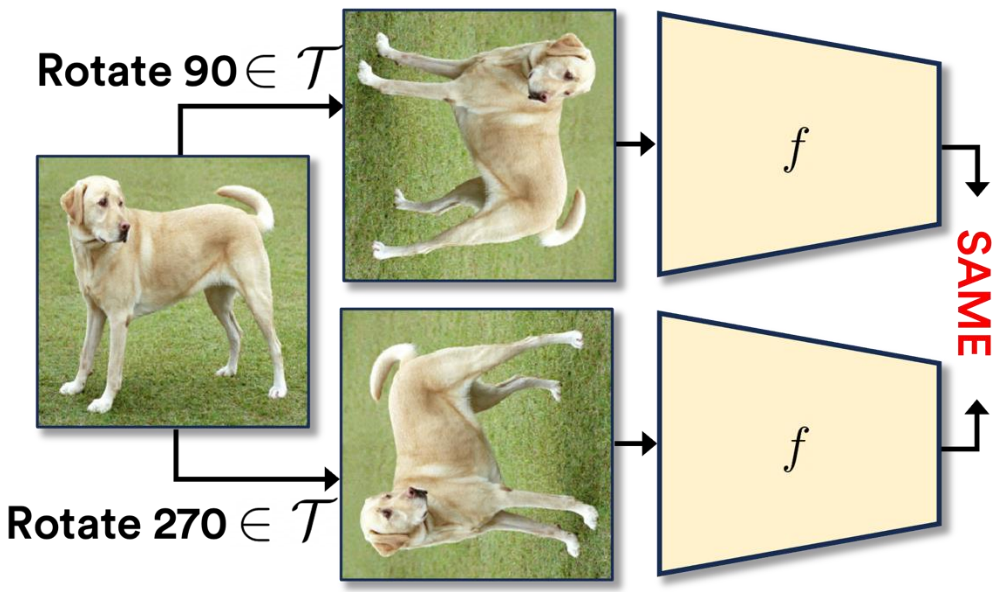
Kwanhyung Lee[1]

Juho Lee[2]

[1] aitrics [2] KAIST [3] VORONOI

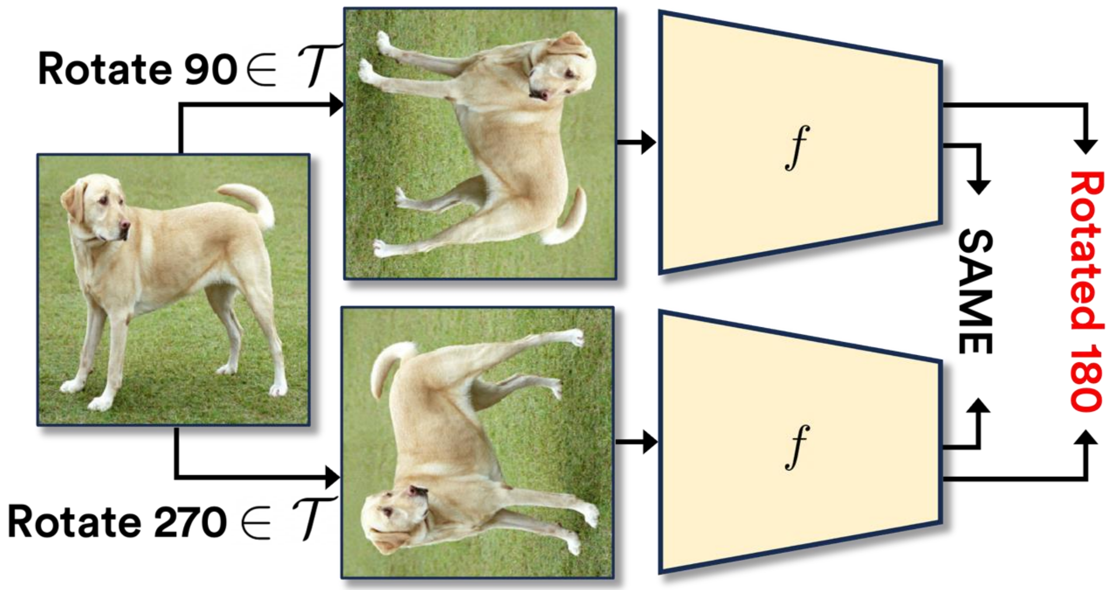
Invariance and Equivariance in Self-Supervised Learning

- <u>Invariance</u> to image transformation is a widely adopted proxy in self-supervised learning (SSL) in vision
  - SimCLR (20'), MoCo (20'), DINO (23'), BarlowTwins (23')



Invariance and Equivariance in Self-Supervised Learning

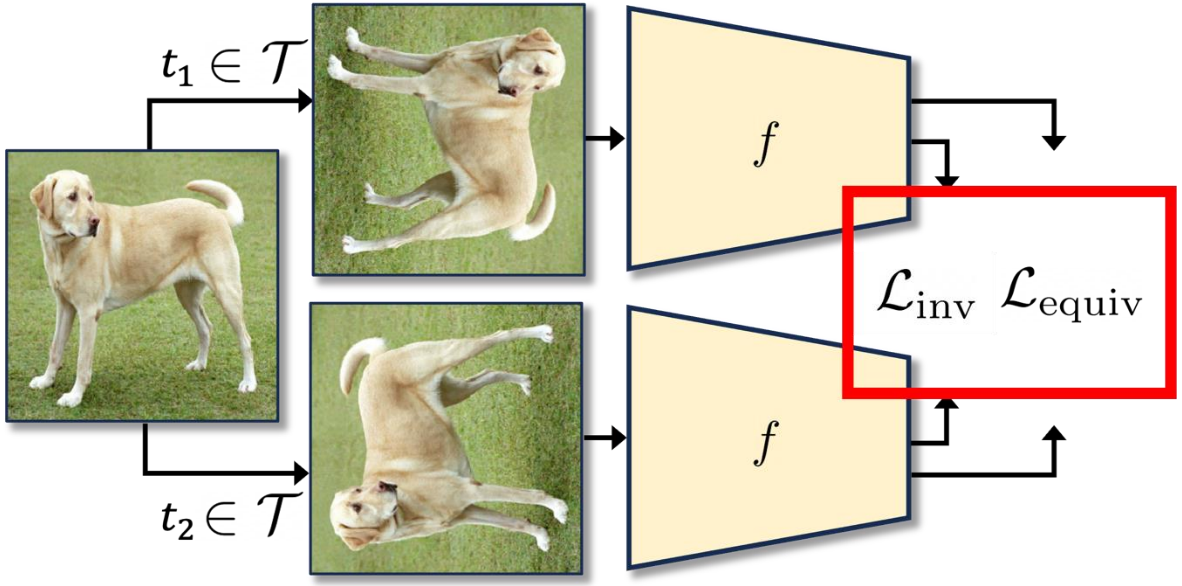
- Enforcing invariance can suppress transformation-dependent structure that is useful in downstream
- A growing body of work thus augments invariance-based SSL with <u>equivariance</u> objective



aitrics

Invariance and Equivariance in Self-Supervised Learning

- Previously proposed invariant-equivariant methods impose both objectives at the final layer.
  - <u>Could this create a conflict?</u>



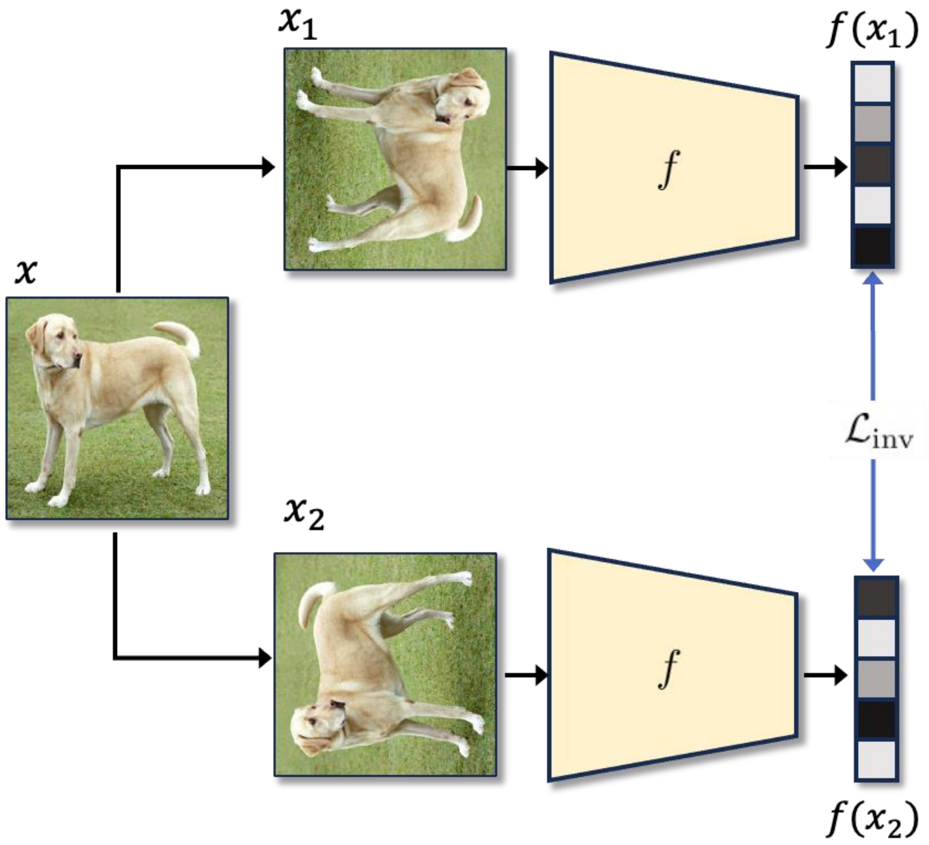
aitrics

Soft Equivariance Regularization (SER)

- SER decouples invariance and equivariance across layer
- SER directly operates on analytically known group action

Soft Equivariance Regularization (SER)

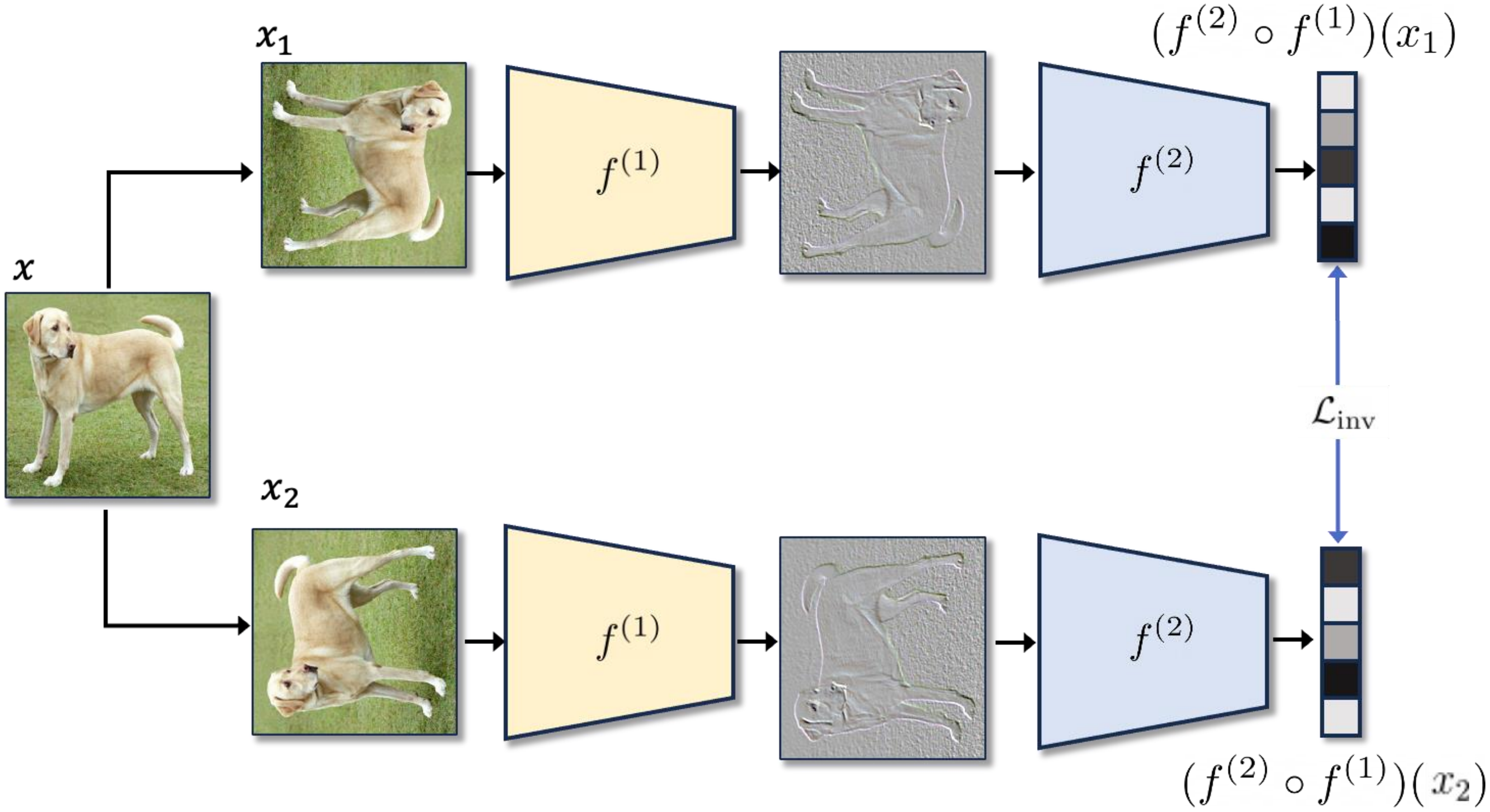
- <u>SER decouples invariance and equivariance across layer</u>



aitrics

Soft Equivariance Regularization (SER)

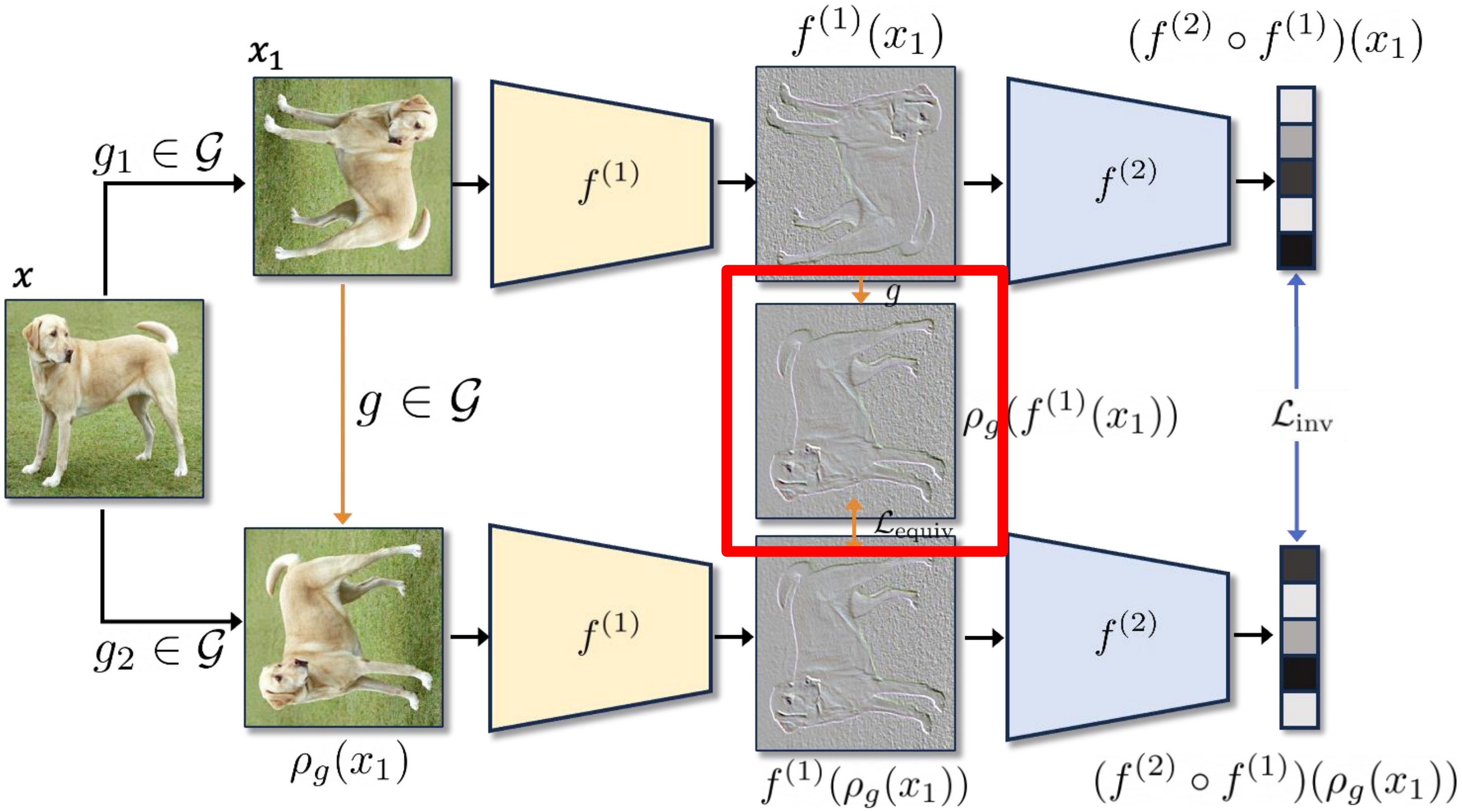
- <u>SER decouples invariance and equivariance across layer</u>



aitrics

Soft Equivariance Regularization (SER)

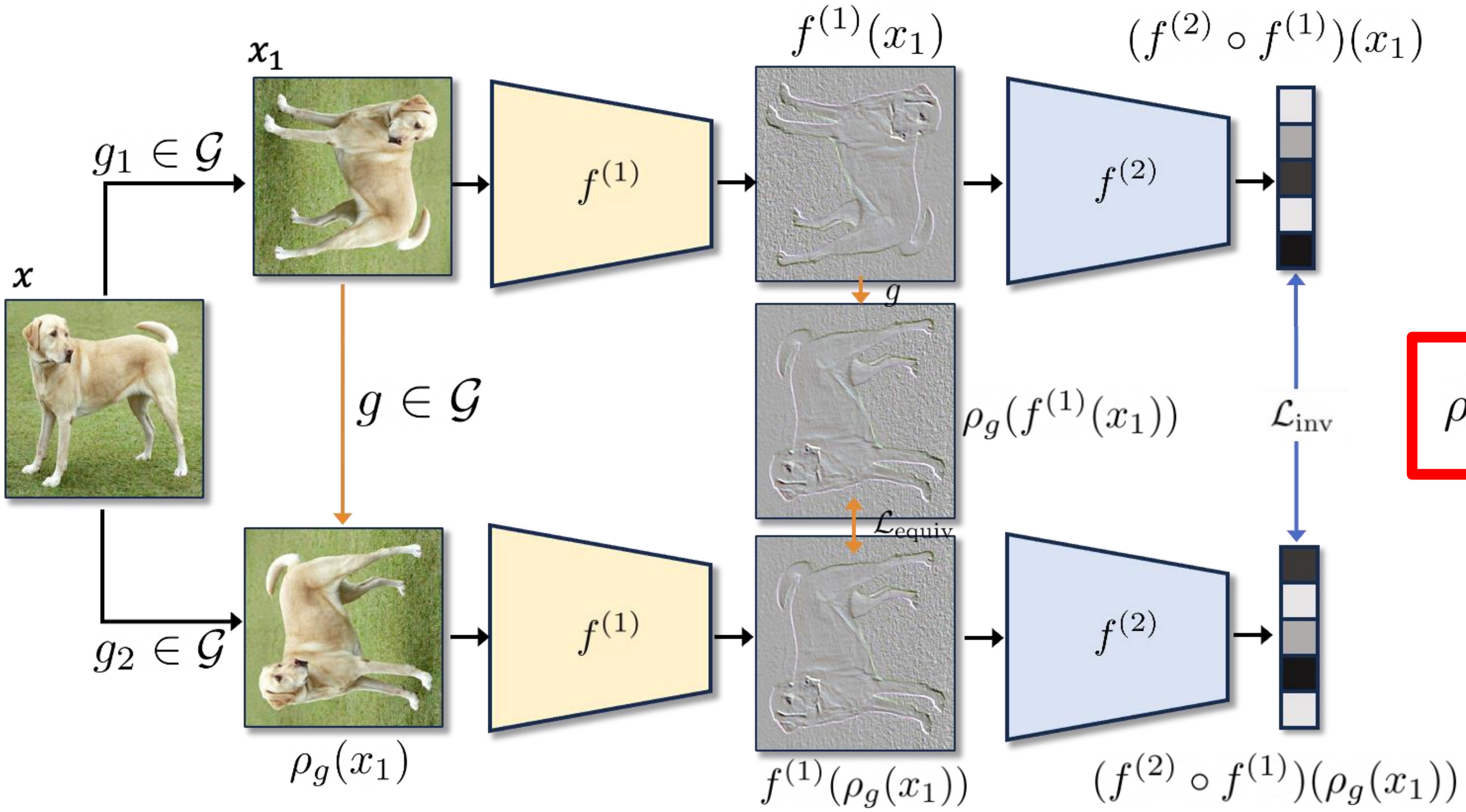
- SER decouples invariance and equivariance across layer



aitrics

Soft Equivariance Regularization (SER)

- SER decouples invariance and equivariance across layer



$$\boxed{\rho_g(f^{(1)}(x_1)) = f^{(1)}(\rho_g(x_1))}$$

aitrics

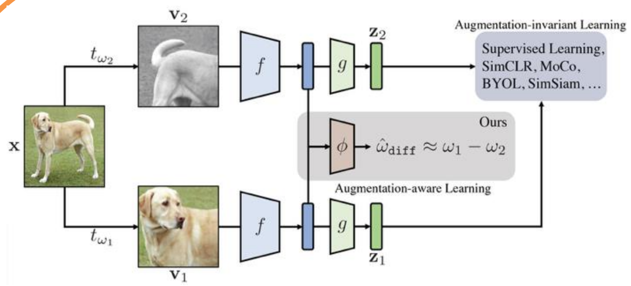
Soft Equivariance Regularization (SER)

- SER decouples invariance and equivariance across layer
- <u>SER directly operates on analytically known group action</u>

Soft Equivariance Regularization (SER)

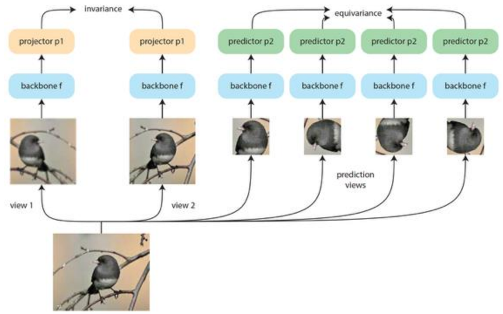
- SER decouples invariance and equivariance across layer
- <u>SER directly operates on analytically known group action</u>

Inv-Equiv SSL: **Implicit** Method



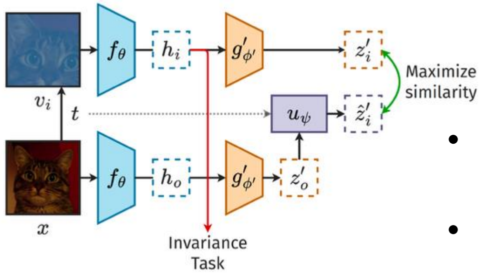
AugSelf (NeurIPS 20')

- Requires auxiliary prediction module
- Does not enforces strict group-action consistency



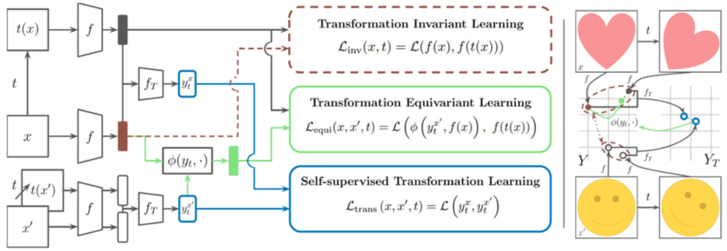
E-SSL (ICLR 21')

Inv-Equiv SSL: **Explicit** Method



EquiMod (ICLR 23')

- Requires auxiliary equiv transformation module
- Does not enforces strict group-action consistency



STL (NeurIPS 23')

aitrics

Soft Equivariance Regularization (SER)

- SER decouples invariance and equivariance across layer
- <u>SER directly operates on analytically known group action</u>
  - No additional transformation module
  - Feature transformations provably respect group properties (closure, identity, inverse, associativity)

Experimental Settings

- Pretraining: ImageNet-1k
- Linear evaluation: ImageNet-1k, 3DIEBench
- Robustness and spatial transfer: ImageNet-C/P/R/V@/Sketch, COCO
- Generalizability across invariant SSL: MoCo-v3, DINO, BarlowTwins
- Non-linear evaluation & finetuning: ImageNet-1k

Questions

- Does SER outperforms other invariance-equivariance methods and invariance baselines in a matching-view setting?
- Is stronger equivariance at the final layer desirable?
- At which layer should we decouple inv and equiv?
- At which layer should the [CLS] token be introduced when decoupling invariance and equivariance?
- Does layer decoupling also benefit other inv-equiv learning methods?

Results

- SER outperforms other inv-equiv methods and inv baselines

View	Algorithm	Param (M)	ImageNet-1k		ImageNet-C		ImageNet-P		ImageNet-Sketch		ImageNet-V2		ImageNet-R		3DIEBench	
			Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
2 view	MoCo-v3	42.9	68.44 ±0.07	88.02 ±0.04	39.47	61.76	63.91	84.67	17.65	31.87	56.54	78.68	18.59	30.08	68.43	91.96
	+ AugSelf	43.7	67.55 ±0.05	87.62 ±0.05	38.93	61.90	63.58	84.69	13.30	25.35	53.74	76.68	17.62	28.66	64.97	90.73
	+ STL	62.2	65.49 ±0.12	85.91 ±0.08	33.33	54.83	61.85	83.30	15.40	28.96	55.43	78.02	17.22	28.49	-	-
	+ SER	43.4	**69.28 ±0.01**	**88.79 ±0.02**	**40.58**	**63.53**	**65.13**	**85.69**	**17.68**	**32.54**	**56.95**	**79.29**	**18.95**	**30.72**	**70.17**	**92.78**
3 view	+ EquiMod	43.3	68.95 ±0.02	88.87 ±0.01	37.45	60.33	64.44	85.66	14.81	28.11	56.31	79.93	16.54	27.32	67.97	91.97
2+4 view	+ E-SSL	43.3	70.6 ±0.04	89.85 ±0.02	42.66	**65.92**	66.58	87.01	19.23	34.77	58.33	**80.93**	19.86	32.36	-	-
	+ SER	43.4	**71.56 ±0.03**	**90.04 ±0.01**	**42.91**	65.48	**68.00**	**87.41**	**19.76**	**34.81**	**59.50**	80.72	**20.27**	**32.54**	**70.91**	**93.15**

Linear Evaluation

Algorithm	ImageNet-1k		ImageNet-Sketch		ImageNet-V2		ImageNet-R	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
MoCo-v3	68.44 ±0.07	88.02 ±0.04	17.65	31.87	56.54	78.68	18.59	30.08
+ SER	**69.28 ±0.01**	**88.79 ±0.02**	**17.68**	**32.54**	**56.95**	**79.29**	**18.95**	**30.72**
DINO	67.37 ±0.02	87.55 ±0.01	17.13	32.09	55.00	77.38	18.28	30.38
+ SER	**67.63 ±0.01**	**87.56 ±0.01**	**18.07**	**34.03**	**55.19**	**77.84**	**18.96**	**31.55**
Barlow Twins	63.34 ±0.03	84.3 ±0.04	10.90	21.17	47.69	70.73	12.30	20.94
+ SER	**64.02 ±0.03**	**84.73 ±0.01**	**12.39**	**24.39**	**50.89**	**74.20**	**13.90**	**23.99**

SER with diverse invariance SSL baselines (2-view)

Metric	MoCo	MoCo + SER	MoCo + STL	MoCo + AugSelf
mAP	0.225	**0.242**	0.221	0.197
mAP@50	0.404	**0.428**	0.400	0.359
mAP@75	0.222	**0.244**	0.218	0.192

Object Detection

Results

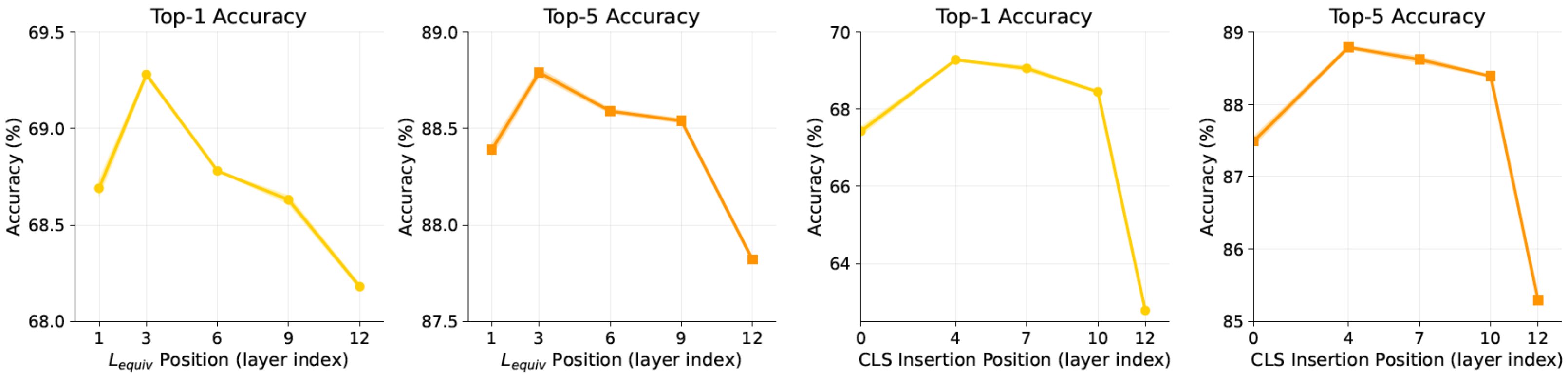
- Stronger equivariance at the final layer can increase the equivariance score but at the cost of its discriminative performance

Algorithm	**Equivariance Loss Layer**	**ImageNet**	**Rotation ↑**	**Hflip ↑**	**Scale ↑**
MoCo-v3	-	68.44	0.804	0.870	0.945
MoCo + SER	Layer 3	**69.28**	0.840	0.881	0.945
MoCo + SER	Layer 9	68.72	0.888	0.886	**0.957**
MoCo + SER	Layer 12	68.18	**0.924**	**0.892**	**0.957**

- Moving equivariance layer toward deeper layers increases the equivariance score of the final representation but reduces ImageNet-1k linear-eval accuracy

Results

- Decoupling inv and equiv shows the highest performance near the middle of the network
- Introducing [CLS] token after equivariance regularization scores the higher performance



- Intermediate “sweet spot” for both the equivariance-loss layer and the [CLS] insertion location

aitrics

Results

- Layer decoupling also benefit other inv-equiv learning methods

Method	Equiv. loss layer	ImageNet-1k Top-1 (%)
EquiMod (Devillers & Lefort, 2023)	12 (final)	68.95 ± 0.02
EquiMod (Devillers & Lefort, 2023)	3 (intermediate)	$\mathbf{69.51 \pm 0.02}$
AugSelf (Lee et al., 2021)	12 (final)	67.55 ± 0.05
AugSelf (Lee et al., 2021)	3 (intermediate)	$\mathbf{68.23 \pm 0.06}$

Decoupling applied

+ 0.56

+ 0.68

aitrics

Check out our paper for more detail

Paper

GitHub

email: chris@aitrics.com

aitrics

KAIST

VORONOI