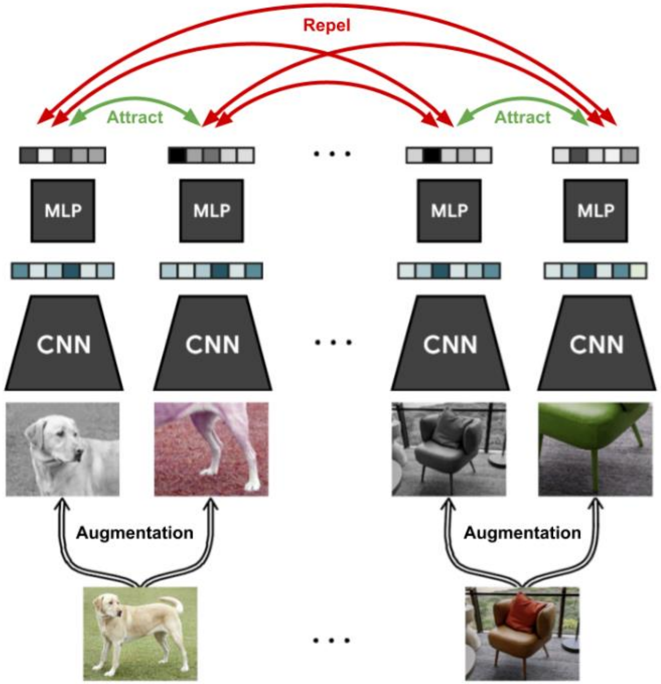
Self-supervised Transformation Learning for Equivariant Representations

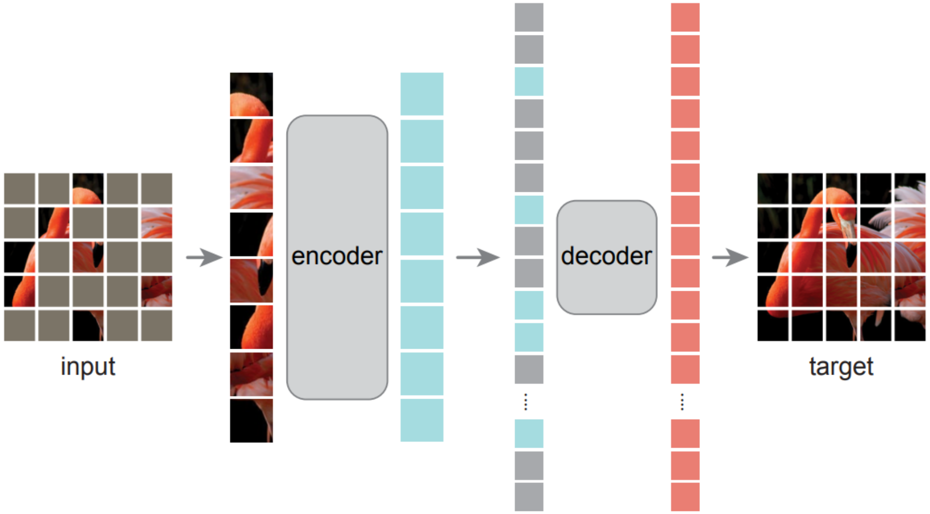
Jaemyung Yu, Jaehyun Choi, Dong-Jae Lee, HyeongGwon Hong, Junmo Kim
Korea Advanced Institute of Science and Technology (KAIST)

jaemyung-u/stl

Self-supervised Learning of Visual Representation



SimCLR (ICML 2020)



MAE (CVPR 2022)

source: Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *International conference on machine learning*. PMLR, 2020.
He, Kaiming, et al. "Masked autoencoders are scalable vision learners." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

Transformation (Augmentation) Invariant Representation

Transformation invariant representation

$$f(x) = f(t(x)) \quad \forall t \in T$$

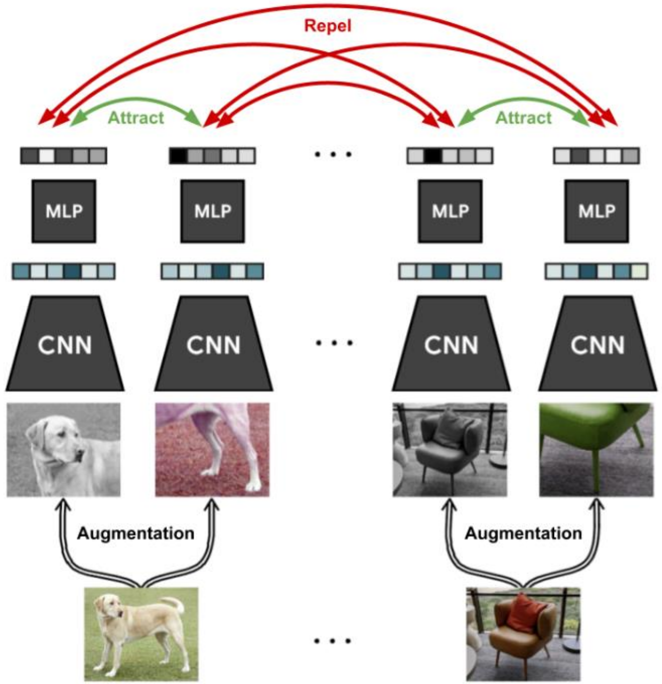
Invariant learning

$$\min_f \mathbb{E}_{x,t}[\mathcal{L}_{\text{inv}}(x,t)]$$

$$\mathcal{L}_{\text{inv}}(x,t) = \mathcal{L}(f(x), f(t(x)))$$

x : image $\quad$ T : group of transformation

f : encoder $\quad$ $\mathcal{L}$: dissimilarity metric (e.g. InfoNCE loss)



source: Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." International conference on machine learning. PMLR, 2020.

Transformation Sensitive Information Matters

Color Information

in Flower Classification

=

≠

Directional Information

in Autonomous Driving

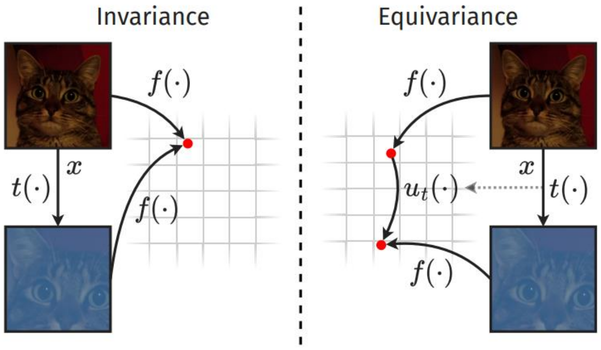
HFlip

Transformation Equivariant Representation

Transformation equivariant representation

$$\exists \phi : T \times Y \to Y \quad \text{s.t.}$$

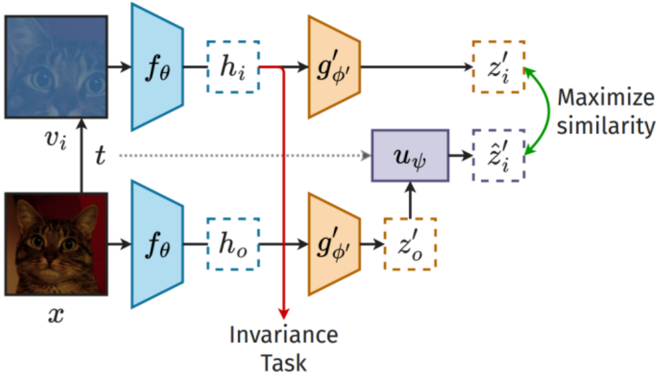
$$f(t(x)) = \phi(t, f(x)) \quad \forall t \in T$$



Equivariant learning (with transformation label)

$$\min_{f,\phi} \mathbb{E}_{x,t}[\mathcal{L}_{\text{equi}}(x,t)]$$

$$\mathcal{L}_{\text{equi}}(x,t) = \mathcal{L}(\phi(t, f(x)),\ f(t(x)))$$



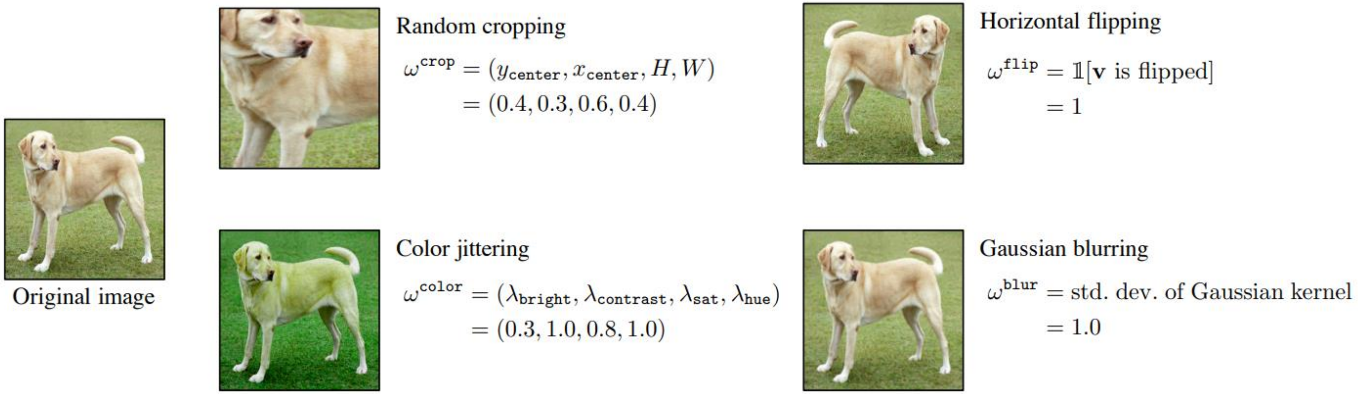
source: Devillers, Alexandre, and Mathieu Lefort. "Equimod: An equivariance module to improve visual instance discrimination." *The Eleventh International Conference on Learning Representations*. 2022.

Limitation of Transformation Label

Imperfect

Transformation Label

hyperparamters of augmentations

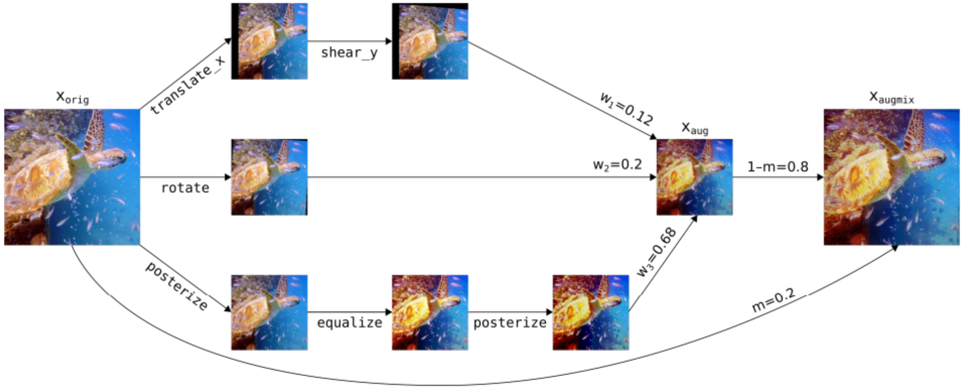


Complex Transformation

with Unknown Structure

AugMix like augmentation,

Complex combination, etc.



AugMix

(ICLR 2020)

Transformation Representation

Equivariant learning **with** transformation label

$$\min_{f,\phi} \mathbb{E}_{x,t}[\mathcal{L}_{\text{equi}}(x,t)] \quad \text{s.t.} \quad \mathcal{L}_{\text{equi}}(x,t) = \mathcal{L}(\phi(t, f(x)),\ f(t(x)))$$

explicit
transformation label

Pairs of representations of original image and transformed image

$$y_t^x = f_T(f(x),\ f(t(x))) \in Y_T \quad \text{for } t \in T \text{ and } x \in X$$

implicit
transformation representation

Equivariant Learning without Transformation Label

Equivariant learning **with** transformation label

$$\min_{f,\phi} \mathbb{E}_{x,t}[\mathcal{L}_{\text{equi}}(x,t)] \quad \text{s.t.} \quad \mathcal{L}_{\text{equi}}(x,t) = \mathcal{L}(\underline{\phi(t, f(x))},\ f(t(x)))$$

$$\boxed{y_t^x = f_T(f(x),\ f(t(x))) \in Y_T \quad \text{for } t \in T \text{ and } x \in X}$$

$$\phi\left(y_t^{x'},\ f(x)\right) = \phi\left(f_T\left(f\left(x'\right), f\left(t(x')\right)\right),\ f(x)\right) \quad \text{for } x \neq x' \in X$$

Equivariant learning **without** transformation representation

$$\min_{f, f_T, \phi} \mathbb{E}_{\underline{x \neq x'}, t}\left[\mathcal{L}_{\text{equi}}(x, x', t)\right]$$

$$\mathcal{L}_{\text{equi}}(x, x', t) = \mathcal{L}\left(\phi\left(y_t^{x'}, f(x)\right),\ f(t(x))\right)$$

prevent trivial solution

$$f(t(x)) = \phi(f_T(f(x), f(t(x))), f(x))$$

Self-supervised Transformation Learning (STL)

Image invariant transformation representation

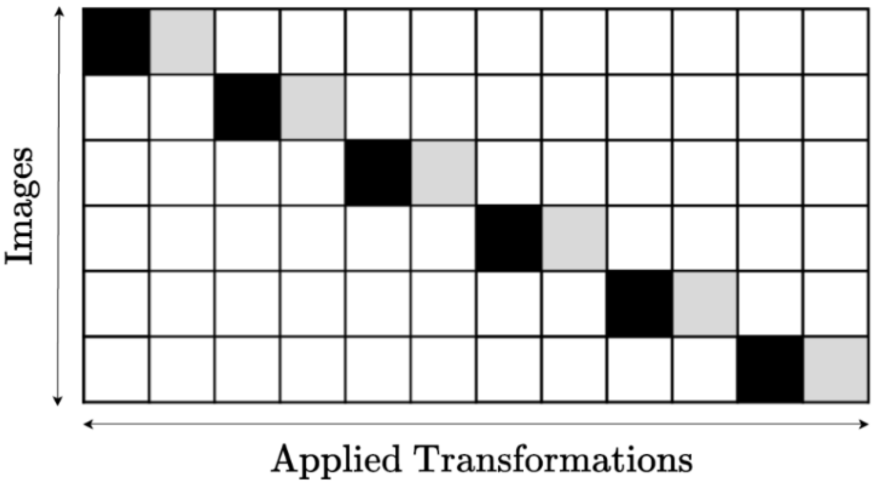
$$y_t^x = y_t^{x'} \quad \forall x \neq x' \in X$$

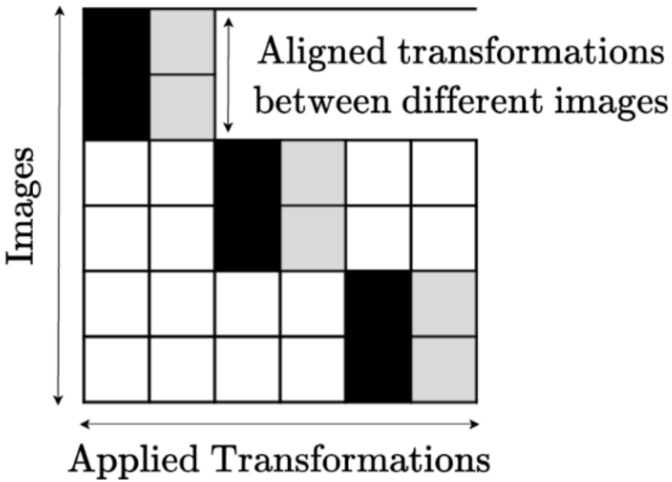
$$y_t^x = f_T(f(x),\ f(t(x))) \in Y_T \quad \text{for } t \in T \text{ and } x \in X$$

Image invariant (transformation representation) learning

$$\min_{f, f_T} \mathbb{E}_{x \neq x', t} \left[\mathcal{L}_{\text{trans}}\left(x, x', t\right)\right] \quad \text{s.t.} \quad \mathcal{L}_{\text{trans}}\left(x, x', t\right) = \mathcal{L}\left(y_t^x, y_t^{x'}\right)$$

Aligned Transformed Batch





Batch size of image = Batch size of transformation

Transformation Equivariant Learning with STL

Dissimilarity metric as $\mathcal{L}_{\text{InfoNCE}}\left(y, y^{+}; g, \tau\right) = -\log \frac{\exp\left(\text{sim}\left(g(y), g(y^{+})\right)/\tau\right)}{\sum_{y' \neq y} \exp\left(\text{sim}\left(g(y), g(y')\right)/\tau\right)}$

$$
\begin{aligned}
\mathcal{L}_{\text{inv}}(x, t) &= \mathcal{L}_{\text{InfoNCE}}\Big(f(x), f(t(x));\ g_{\text{inv}}, \tau_{\text{inv}}\Big), \\
\mathcal{L}_{\text{equi}}\left(x, x', t\right) &= \mathcal{L}_{\text{InfoNCE}}\Big(\phi\big(y_t^{x'}, f(x)\big),\ f(t(x));\ g_{\text{equi}}, \tau_{\text{equi}}\Big), \\
\mathcal{L}_{\text{trans}}(x, x', t) &= \mathcal{L}_{\text{InfoNCE}}\Big(y_t^{x},\ y_t^{x'};\ g_{\text{trans}}, \tau_{\text{trans}}\Big).
\end{aligned}
$$

Overall Objective

$$
\min_{f, f_{\mathcal{T}}, \phi} \mathbb{E}_{x \neq x', t}\Big[\lambda_{\text{inv}}\mathcal{L}_{\text{inv}}(x, t) + \lambda_{\text{equi}}\mathcal{L}_{\text{equi}}(x, x', t) + \lambda_{\text{trans}}\mathcal{L}_{\text{trans}}(x, x', t)\Big]
$$

Overall Framework of STL

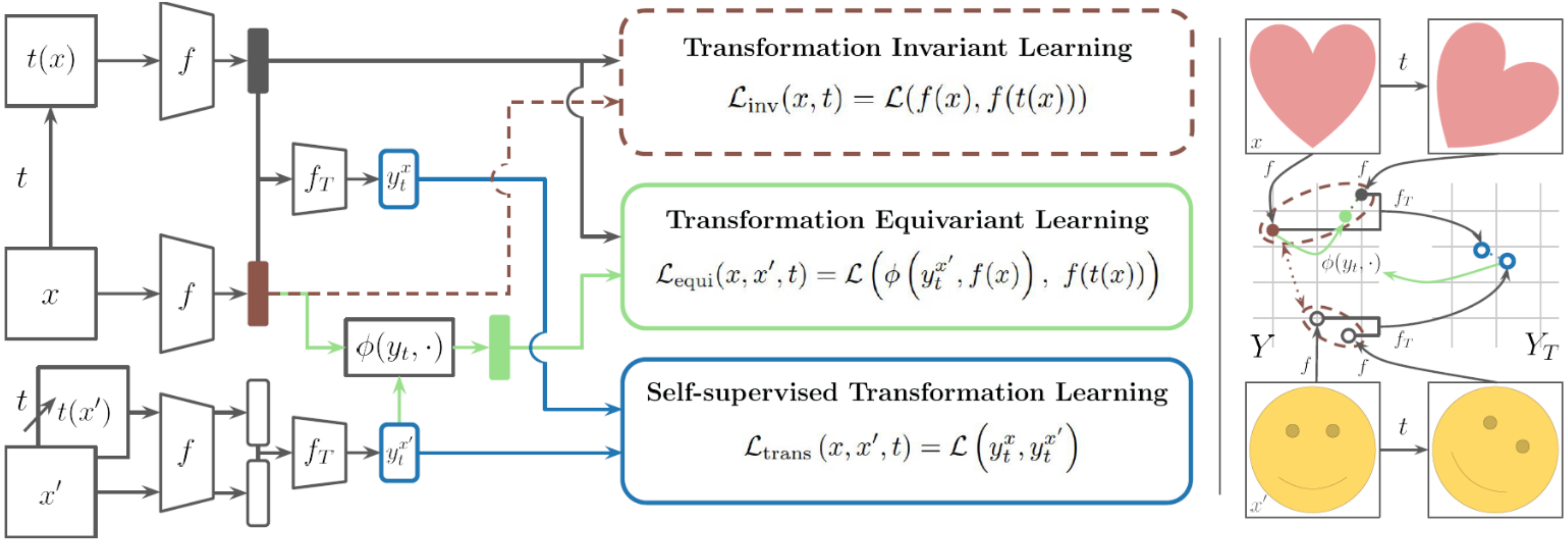


Image Representation Evaluation (Out-domain)

How generalized the learned representation is

Table 2: **Out-domain Classification.** Evaluation of representation generalizability on the out-domain downstream classification tasks. Linear evaluation accuracy (%) is reported for ResNet-50 pretrained on ImageNet100.

Method	CIFAR10	CIFAR100	Food	MIT67	Pets	Flowers	Caltech101	Cars	Aircraft	DTD	SUN397	Mean
Invariant Learning :												
SimCLR	84.24	64.15	59.00	54.78	58.95	91.58	79.32	27.07	36.00	66.01	42.77	60.35
with AugMix	86.90	**67.70**	62.90	57.24	63.75	93.16	83.67	32.37	43.17	67.93	46.15	64.09
Implicit Equivariant Learning :												
E-SSL	85.09	65.74	60.91	56.64	61.00	92.31	80.77	28.84	38.04	66.38	43.49	61.75
AugSelf	85.55	66.09	62.63	57.16	62.61	93.41	82.33	30.71	40.35	68.51	45.24	63.14
Explicit Equivariant Learning :												
SEN	80.68	56.53	52.50	46.79	45.27	79.24	73.42	14.41	27.51	57.45	33.51	51.57
EquiMod	82.89	61.36	56.38	52.84	52.68	87.42	79.17	22.02	34.62	64.10	39.86	57.58
SIE	81.72	58.49	54.04	49.70	47.21	84.37	74.39	16.71	31.68	59.20	35.29	53.89
STL (Ours)	86.55	66.84	64.32	56.64	65.00	94.51	81.83	35.44	45.42	64.68	44.69	64.18
with AugMix (Ours)	**87.19**	**67.70**	**66.12**	**59.70**	**67.10**	**94.87**	**84.61**	**38.48**	**46.14**	**69.57**	**45.75**	**66.11**

Image Representation Evaluation (In-domain)

Whether the learned representation causes trade-offs in the in-domain

Table 3: **In-domain Classification.** Evaluation of representation on in-domain classification task. Linear evaluation accuracy (%) is reported for ResNet-50 pretrained on ImageNet100.

Method	In-domain
Invariant Learning :	
SimCLR	81.20
SimCLR with AugMix	80.54
Implicit Equivariant Learning :	
E-SSL	**82.10**
AugSelf	81.08
Explicit Equivariant Learning :	
SEN	76.32
EquiMod	80.70
SIE	79.40
STL (Ours)	81.10
STL with AugMix (Ours)	81.64

Image Representation Evaluation (Object Detection)

How generalized the learned representation is

Table 4: **Object Detection.** Evaluation of representation generalizability on a downstream object detection task. Average precision is reported for ImageNet100-pretrained ResNet-50 fine-tuned on VOC07+12.

Method	$\mathbf{AP}_{\text{all}}$	$\mathbf{AP}_{50}$	$\mathbf{AP}_{75}$
SimCLR	45.67	72.50	47.83
AugSelf	45.99	72.46	49.23
EquiMod	51.55	78.03	56.17
STL (Ours)	51.95	78.34	56.96
with AugMix (Ours)	**52.70**	**78.81**	**57.76**

Transformation Representation Evaluation (Quantitative)

How the learned equivariant representation reflects the actual transformation

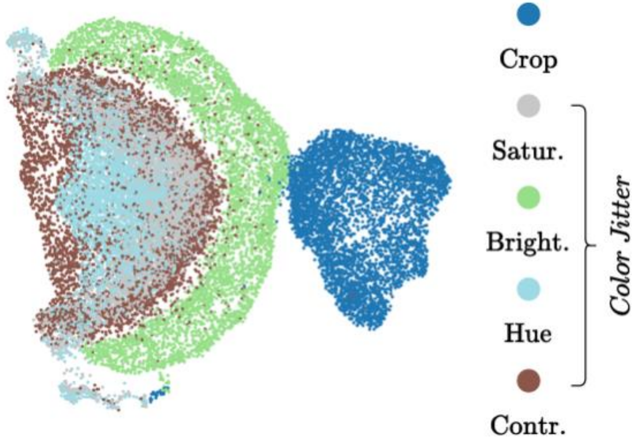
Table 5: **Transformation Prediction.** Evaluation of transformation representation from learned represetation pairs. Regression tasks use MSE loss, and transformation type classification uses accuracy.

	Regression (↓)			**Classification (↑)**
Method	Crop	Color	All	Trans. Type
SimCLR	0.02	0.13	0.08	68.54
AugSelf	**0.01**	0.04	0.03	88.49
EquiMod	**0.01**	0.07	0.04	82.20
STL (Ours)	**0.01**	**0.03**	**0.02**	**93.67**

Transformation Representation Evaluation (Qualitative)

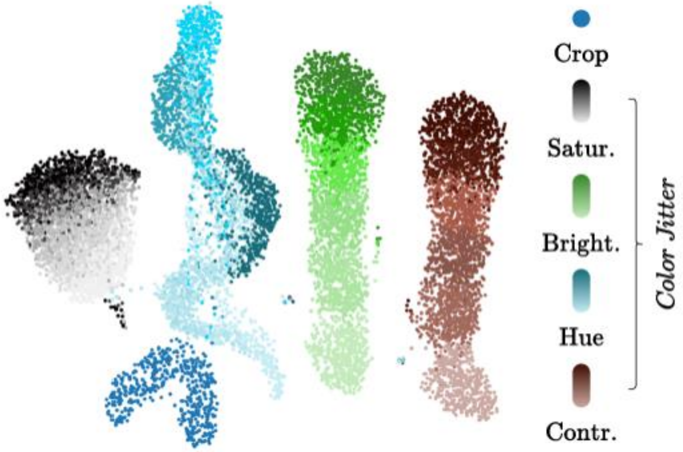
How the learned transformation representation reflects the actual transformation

Inter-relationship of transformations



UMAP Visualization
of transformation representations
by type

Intra-relationship of transformations



UMAP Visualization
of transformation representations
by intensity

Equivariant Transformation Evaluation

How the equivariant transformation reflects the actual trans. in the repr. space

Table 6: **Transformation Equivariance.** Evaluation of the equivariant transformation. Mean Reciprocal Rank (MRR), Hit@k (H@k), and Precision (PRE) metrics on various transformations (crop and color jitter).

	Crop				Color				All			
Method	MRR(↑)	H@1(↑)	H@5(↑)	PRE(↓)	MRR(↑)	H@1(↑)	H@5(↑)	PRE(↓)	MRR(↑)	H@1(↑)	H@5(↑)	PRE(↓)
SEN	0.34	0.15	0.58	0.14	0.18	0.05	0.31	3.69	0.22	0.08	0.37	2.70
EquiMod	**0.37**	0.17	**0.60**	**0.13**	0.16	0.05	0.28	3.72	0.22	0.09	0.36	2.72
SIE	0.33	0.14	0.55	0.33	0.17	0.05	0.28	3.70	0.21	0.08	0.35	2.74
w/o $\mathcal{L}_{\text{trans}}$ (Ours)	0.31	0.18	0.46	0.69	0.27	0.13	0.40	3.37	0.29	0.16	0.43	2.50
STL (Ours)	**0.37**	**0.22**	0.54	0.64	**0.33**	**0.18**	**0.52**	**2.76**	**0.36**	**0.21**	**0.53**	**2.07**

Prediction Retrieval Error (PRE)

The differences b/w the parameters of the equi. trans. and the closest actual trans.

$$\text{PRE} = |\theta_{\text{eq}} - \theta_{\text{real}}|$$

Mean Reciprocal Rank (MRR)

The avg. reciprocal rank of the actual transformed repr. among the closest retrieved reprs.

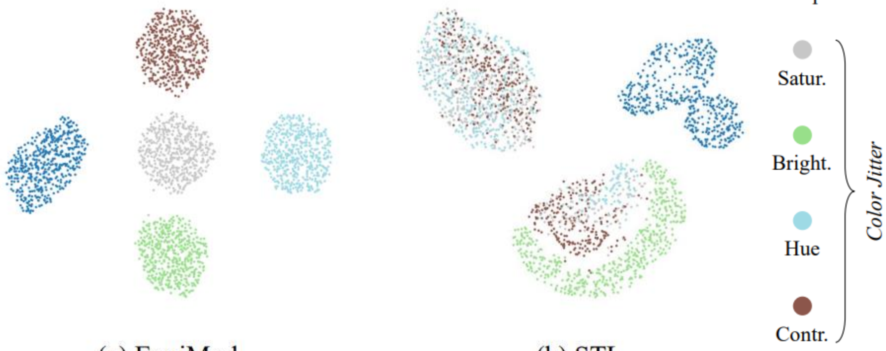
$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\text{rank}_i}$$

Hit Rate at k (H@k)

The proportion of cases where the actual transformed repr ranks within the top k.

$$\text{H@k} = \frac{1}{|Q|}\sum_{i=1}^{|Q|} 1(\text{rank}_i \le k)$$

Equivariant Transformation



(a) EquiMod (b) STL

UMAP Visualization of functional weights

Ablation Study for Modules

Table 7: **Loss Function Ablation Study.** Image classification and transformation prediction results of ResNet-18 pretrained on STL10 with selective inclusion of loss terms for invariant learning ($\mathcal{L}_{\text{inv}}$), equivariant learning ($\mathcal{L}_{\text{equi}}$), and self-supervised transformation learning ($\mathcal{L}_{\text{trans}}$). For image classification, in-domain accuracy (%) and the average accuracy (%) across multiple out-domain datasets are shown. For transformation prediction, MSE is used for regression of crop and color transformations, and accuracy (%) is used for transformation type classification.

	Loss Functions			Image Classification		Transformation Prediction	
Method	$\mathcal{L}_{\text{inv}}$	$\mathcal{L}_{\text{equi}}$	$\mathcal{L}_{\text{trans}}$	In-domain (↑)	Out-domain (↑)	Regression (↓)	Classification (↑)
Only Invariance	✓	-	-	84.74	43.11	0.08	68.54
Only Equivariance	-	✓	-	83.53	**49.99**	**0.02**	93.54
STL w/o $\mathcal{L}_{\text{inv}}$	-	✓	✓	81.86	48.62	**0.02**	93.54
STL w/o $\mathcal{L}_{\text{equi}}$	✓	-	✓	80.99	47.30	**0.02**	**93.92**
STL w/o $\mathcal{L}_{\text{trans}}$	✓	✓	-	**85.11**	48.49	0.08	69.57
STL	✓	✓	✓	84.83	**49.97**	**0.02**	93.67

Ablation Study for Transformations (Augmentation)

Table 8: **Transformation Ablation Study.** Linear evaluation accuracy (%) of ResNet-18 pretrained on STL10 with various transformations used as equivariance targets.

Trans.	Method	CIFAR10	CIFAR100	Food	MIT67	Pets	Flowers	Caltech101	Cars	Aircraft	DTD	SUN397	Mean
crop	AugSelf	82.89	54.92	33.19	**39.70**	44.40	64.96	67.63	15.58	25.38	**41.86**	27.89	45.31
	EquiMod	83.76	55.33	32.01	37.76	41.65	63.00	66.28	14.18	24.96	41.54	26.46	44.27
	STL	**84.94**	**59.12**	**35.15**	39.40	**45.35**	**68.38**	**70.78**	**17.96**	**33.00**	**41.86**	**28.71**	**47.70**
color	AugSelf	**84.33**	57.47	36.57	39.40	**46.80**	71.18	67.91	17.03	**27.12**	43.83	29.37	47.36
	EquiMod	82.22	51.77	31.21	34.18	39.57	61.17	62.07	12.51	21.36	39.52	23.48	41.73
	STL	84.16	**58.71**	**38.49**	**41.34**	45.90	**74.36**	**68.48**	**17.31**	**27.12**	**46.54**	**31.17**	**48.51**
crop + color	AugSelf	84.26	57.78	36.82	40.30	45.46	73.38	68.11	17.22	27.63	45.96	30.38	47.94
	EquiMod	81.35	51.86	33.91	37.76	41.92	66.18	67.38	15.22	25.80	42.50	26.70	44.60
	STL	**85.37**	**61.05**	**39.41**	**41.27**	**46.58**	**76.43**	**71.47**	**19.04**	**30.75**	**46.17**	**32.13**	**49.97**
all	AugSelf	81.76	54.90	36.51	40.90	46.17	71.43	70.14	**18.63**	**30.96**	**45.21**	30.40	47.91
	EquiMod	84.42	56.65	34.23	37.99	42.98	67.16	68.41	15.18	26.91	43.94	26.97	45.89
	STL	**84.96**	**58.91**	**36.71**	**42.09**	**46.25**	**72.41**	**71.01**	17.72	28.44	43.83	**30.99**	**48.48**

Ablation Study for Base Invariant Learning Models

Table 9: **Base Invariant Learning Model Ablation Study.** Linear evaluation accuracy (%) of ResNet-18 pretrained on STL10 with various base models for invariant learning.

Base	Method	CIFAR10	CIFAR100	Food	MIT67	Pets	Flowers	Caltech101	Cars	Aircraft	DTD	SUN397	Mean
BYOL	-	85.55	59.80	37.54	42.61	50.61	73.50	72.46	23.02	31.71	44.95	31.63	50.31
	AugSelf	87.01	64.84	**43.14**	**47.24**	**52.49**	78.88	75.42	25.47	37.02	**48.03**	**34.94**	54.04
	EquiMod	84.64	56.55	32.74	39.18	44.64	66.54	68.37	15.47	24.27	42.71	26.96	45.64
	STL	**86.88**	**65.63**	42.98	46.42	52.33	**79.61**	**76.04**	**28.68**	**39.21**	46.44	34.57	**54.44**
SimSiam	-	83.26	55.69	34.32	40.52	46.52	66.06	69.13	17.15	27.99	41.91	28.97	46.50
	AugSelf	**85.44**	62.20	39.78	43.43	46.77	**77.90**	**71.72**	18.67	**33.30**	45.53	**32.65**	50.67
	EquiMod	81.20	51.23	31.21	37.99	40.53	63.98	64.19	12.22	22.11	40.69	25.76	42.83
	STL	85.20	**62.58**	**40.15**	**44.03**	**48.65**	76.68	71.37	**22.42**	32.37	**45.59**	32.19	**51.02**
Barlow Twins	-	81.67	51.68	27.79	33.13	39.60	57.63	62.17	11.53	19.47	37.13	23.43	40.48
	AugSelf	82.46	51.71	27.83	35.75	39.33	58.24	61.87	11.88	19.77	37.29	23.31	40.86
	EquiMod	81.57	52.15	30.00	36.79	38.70	62.64	63.22	11.80	20.55	40.21	24.92	42.05
	STL	**83.74**	**56.73**	**32.69**	**38.36**	**42.65**	**67.28**	**68.09**	**16.24**	**24.33**	**41.97**	**28.53**	**45.51**

Thank You

https://github.com/jaemyung-u/stl