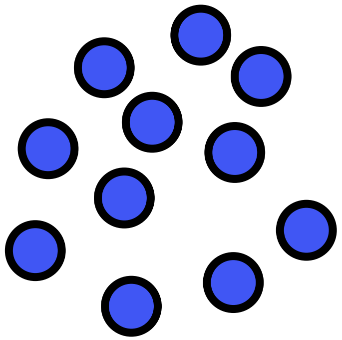
CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Instances

Jihoon Tack*, Sangwoo Mo*, Jongheon Jeong, Jinwoo Shin
Korea Advanced Institute of Science and Technology (KAIST)

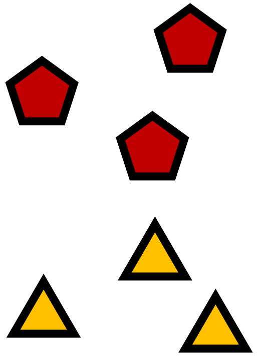
* Equal contribution

Problem: Novelty/Out-of-distribution Detection

- Identifying whether a given sample belongs to the data distribution

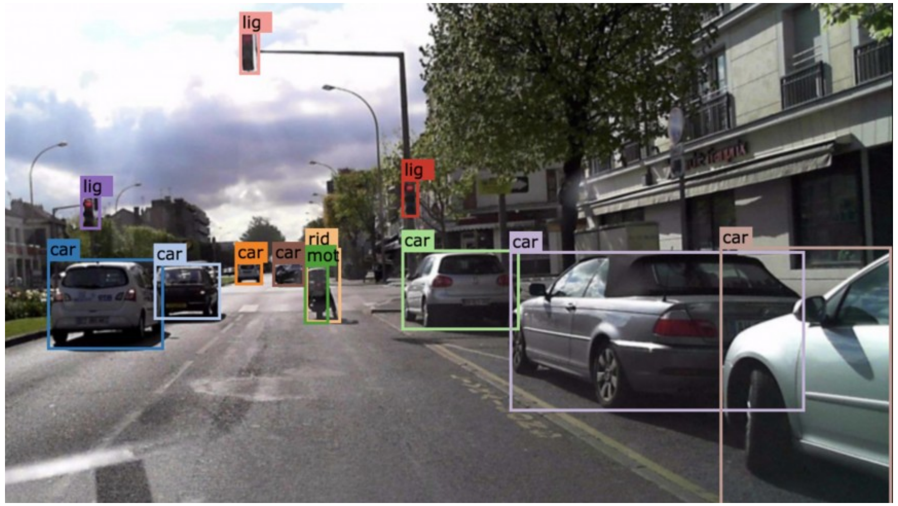


Training samples



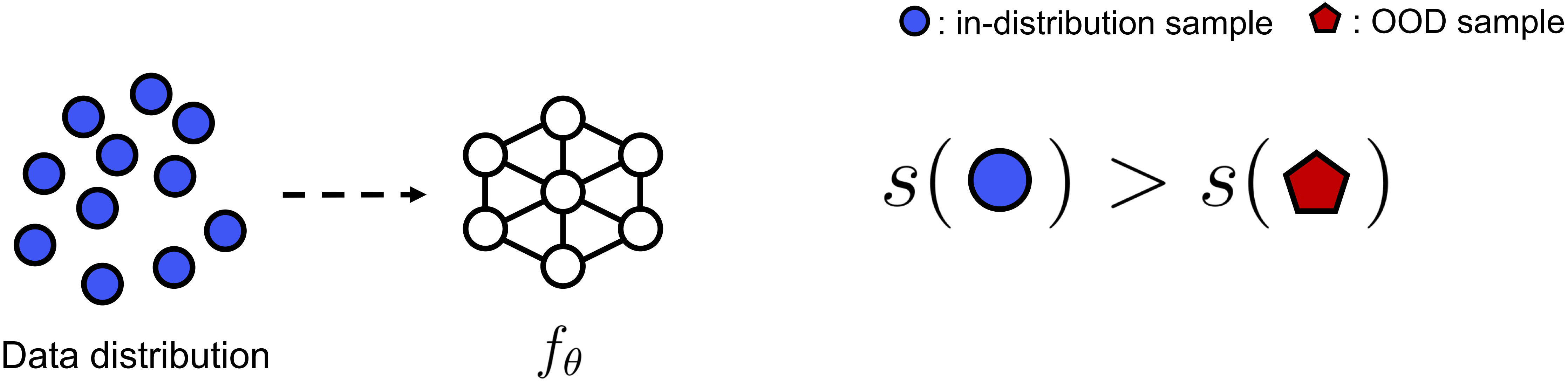
Can be dangerous
for misprediction!

Out-of-distribution samples



Problem: Novelty/Out-of-distribution Detection

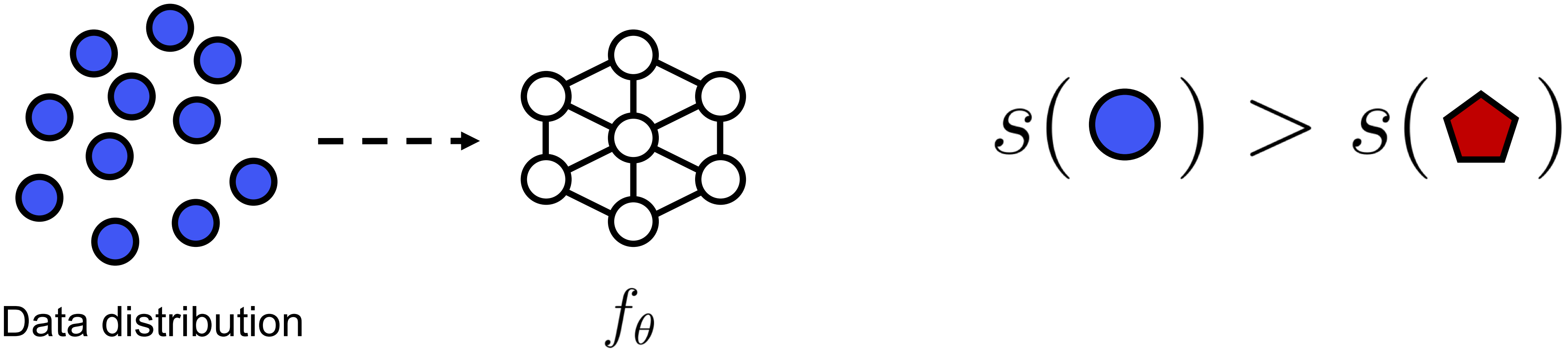
- Identifying whether a given sample belongs to the data distribution



- **Learn a representation f_θ** from the data distribution
- **Define a detection score $s(.)$** utilizing the representation f_θ

Related work: Novelty/Out-of-distribution Detection

- Related works can be categorized as
  - (a) density-based (b) reconstruction-based (c) one-class classifier (d) self-supervised method



- Directly modeling the distribution $p(x)$ is an ideal solution, but known to be hard
- Recently, **self-supervised learning** method shows dramatic performance
  - E.g., Learn to classify applied transformation (data augmentation)

Related work: Novelty/Out-of-distribution Detection

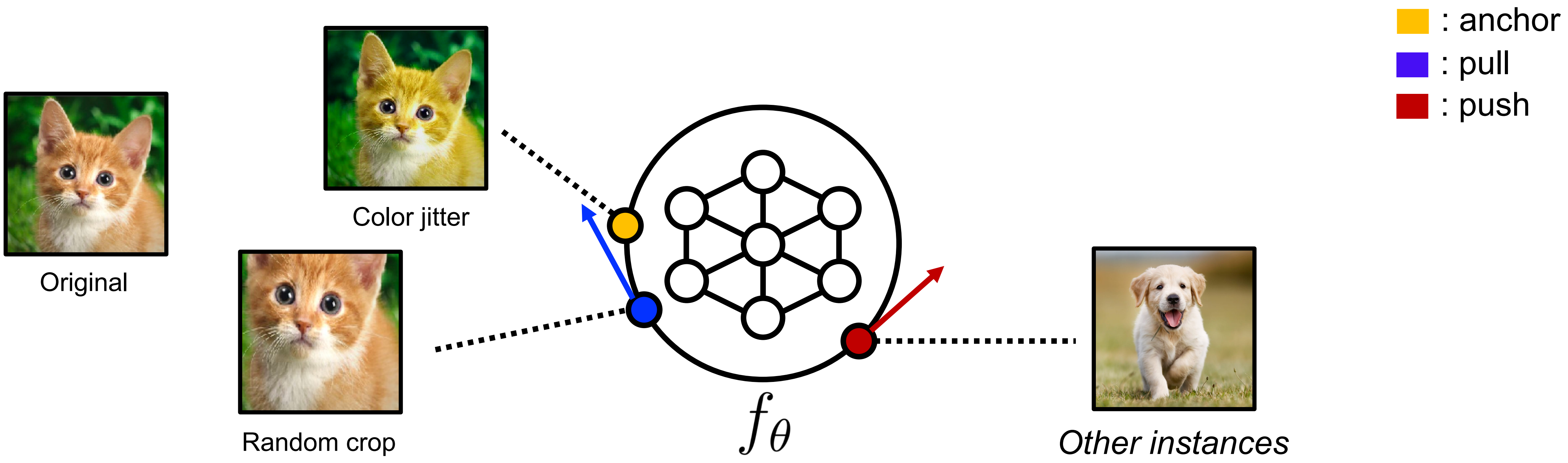
- Related works can be categorized as
  - (a) density-based (b) reconstruction-based (c) one-class classifier (d) self-supervised method

Can we use more advanced self-supervised learning framework?

- Directly modeling the distribution $p(x)$ is an ideal solution, but known to be hard
- Recently, **self-supervised learning** method shows dramatic performance
  - E.g., Learn to classify applied transformation (data augmentation)

Self-Supervised Learning: Contrastive Learning

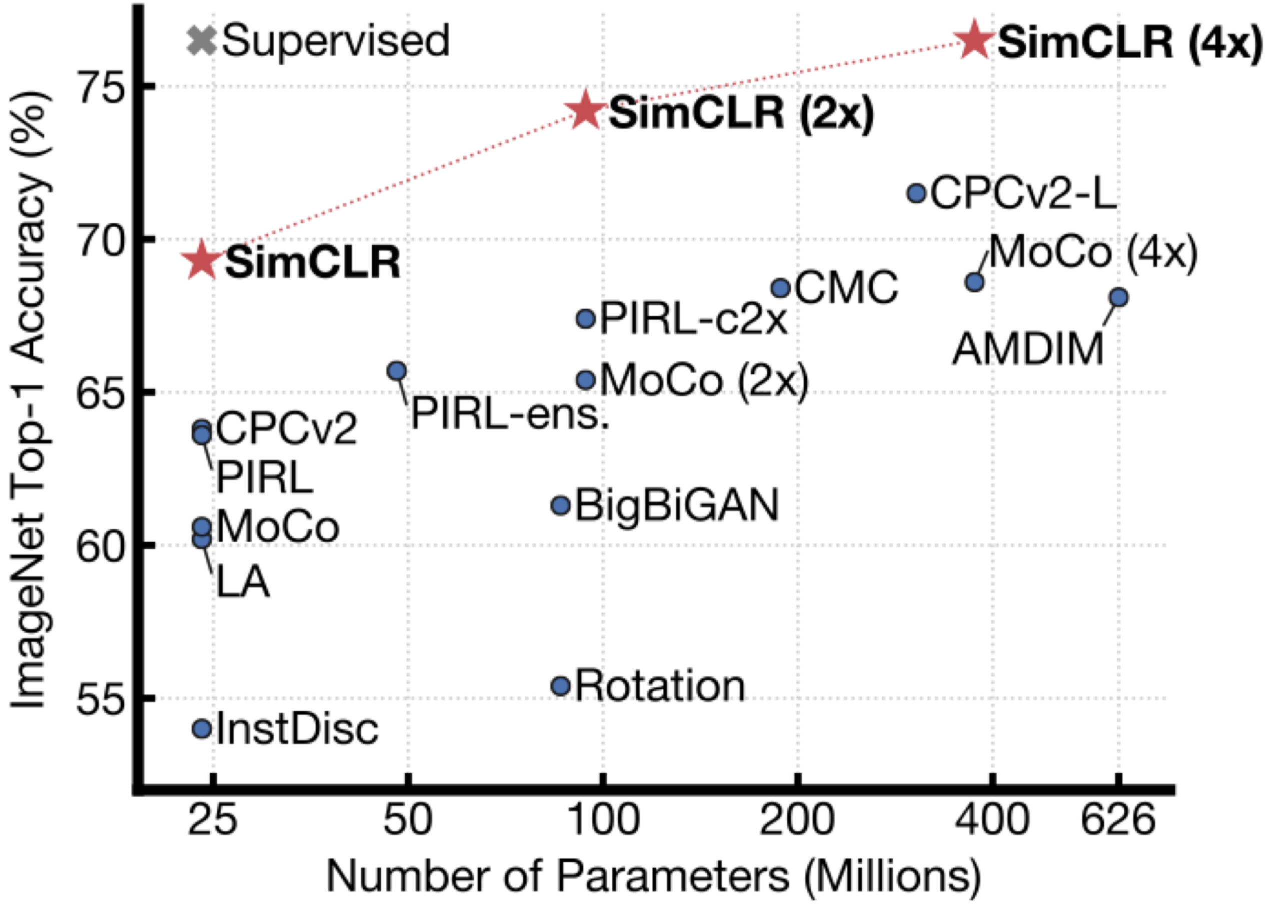
- Learn the representation that **encodes the similarity between data points**
- We use **simple contrastive learning (SimCLR)** [1]:
  - **pull** the <u>same samples</u> of *different augmentations*
  - **push** the <u>different samples</u>



[1] Chen et al. A simple framework for contrastive learning of visual representations. ICML 2020.

Self-Supervised Learning: Contrastive Learning

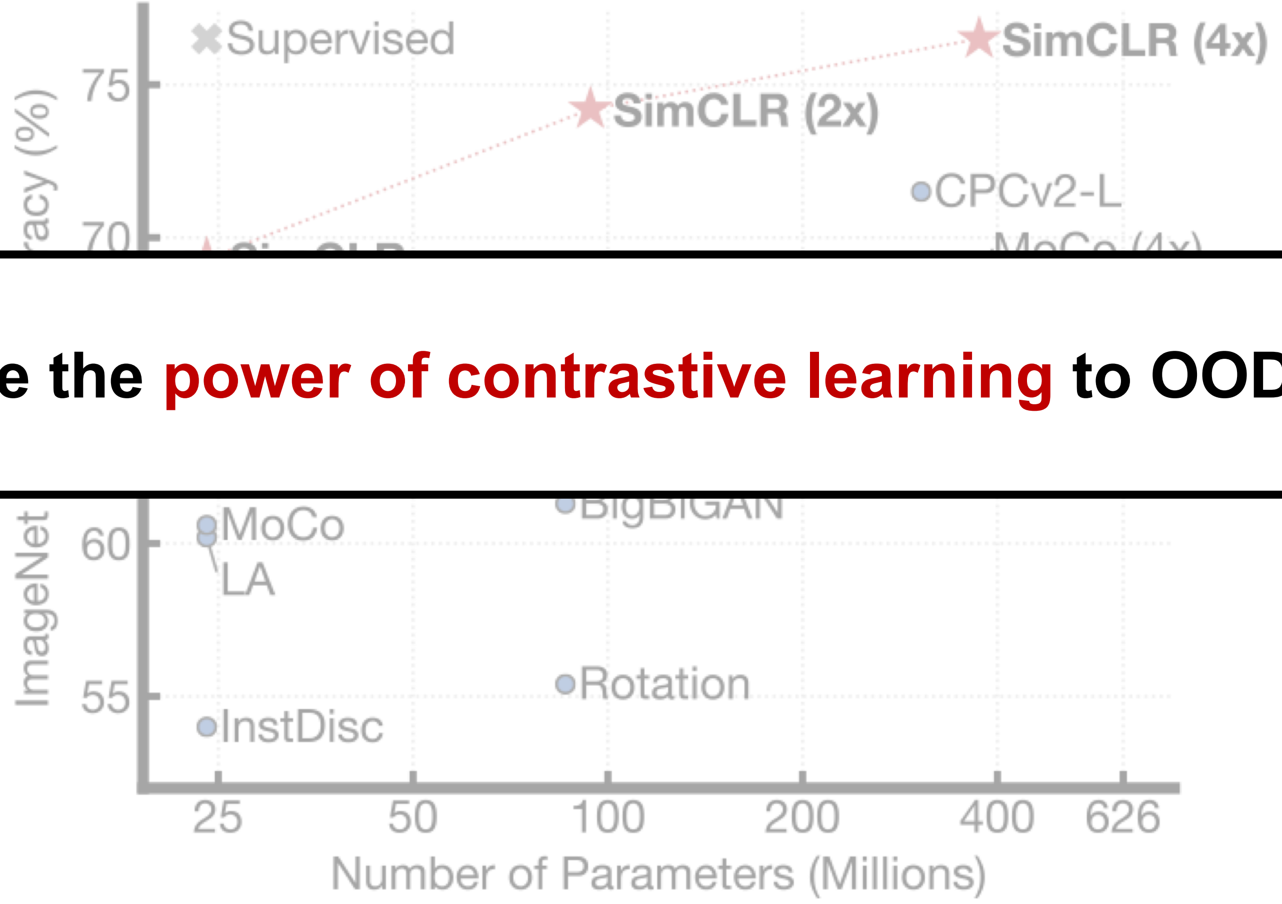
- We can learn discriminative representation **without any label**



[1] Chen et al. A simple framework for contrastive learning of visual representations. ICML 2020.

Self-Supervised Learning: Contrastive Learning

- We can learn discriminative representation without any label

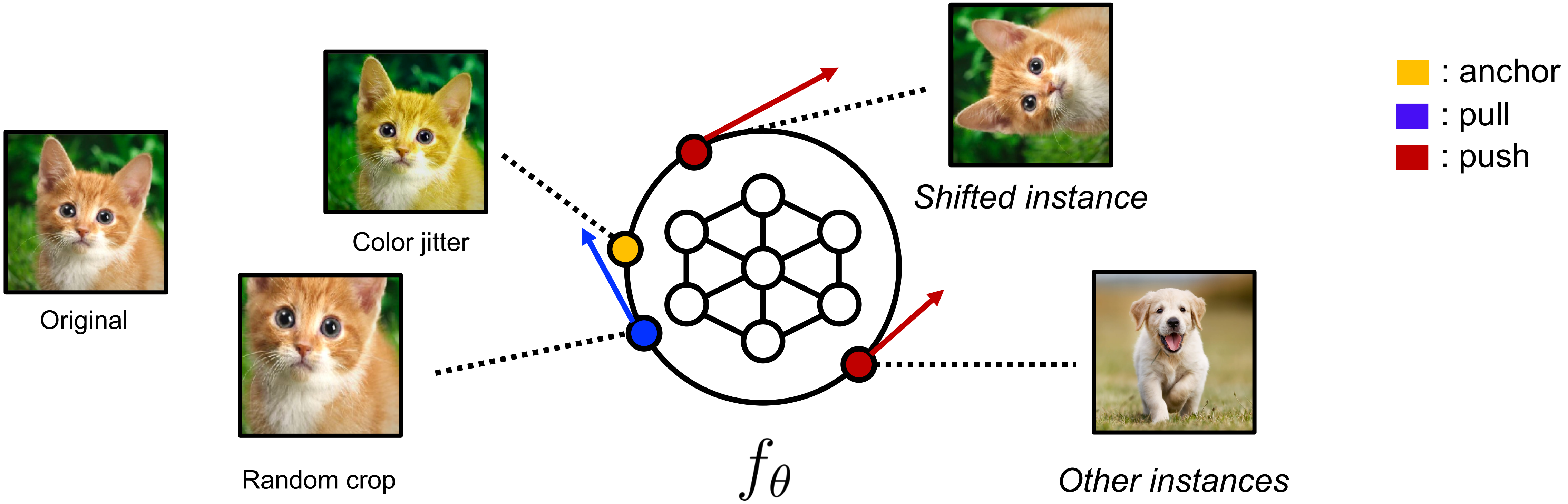


Lets utilize the power of contrastive learning to OOD detection

[1] Chen et al. A simple framework for contrastive learning of visual representations. ICML 2020.

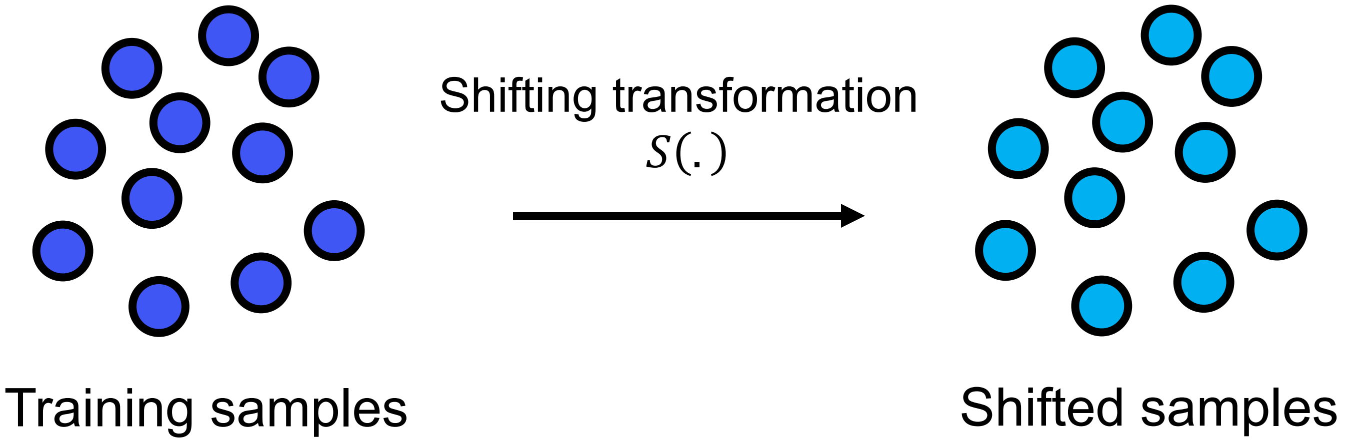
Summary: Contrasting Shifted Instances

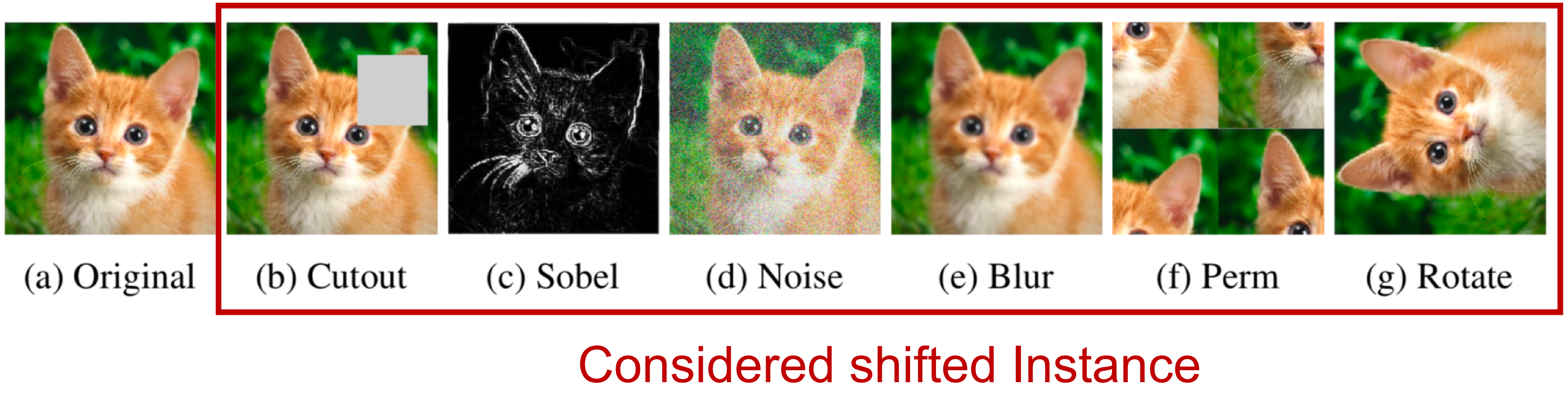
- We utilize the power of **contrastive learning** for OOD detection
- We further improve OOD detection by using **shifted instances**



Contrasting Shifted Instances (CSI): Representation

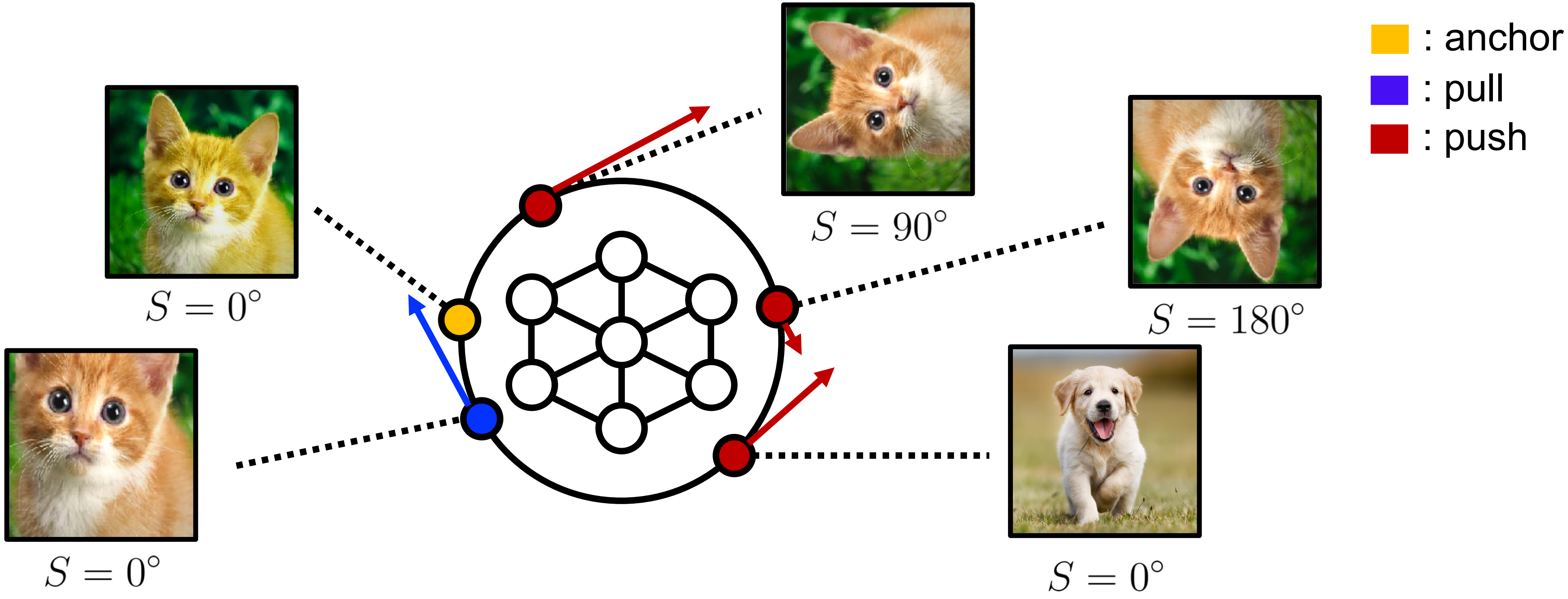
- We train the representation via **contrastive learning with shifted instances:**





Contrasting Shifted Instances (CSI): Representation

- We train the representation via **contrastive learning with shifted instances:**
  - We found **contrastively learned representation** [1] is already effective at OOD detection
  - CSI further improves by **pushing the shifted samples** in addition to the different samples
  - Additionally **classify the shifting transformation**



[1] Chen et al. A simple framework for contrastive learning of visual representations. ICML 2020.

Contrasting Shifted Instances (CSI): Detection Score

- Detection score for **contrastively learned representation:**
  - The ***cosine similarity*** to the nearest training sample
  - The ***norm*** of the representation

$$s_{\texttt{con}}(x;\{x_m\}) := \max_{m}\ \mathrm{sim}(z(x_m), z(x))\cdot \parallel z(x) \parallel .$$

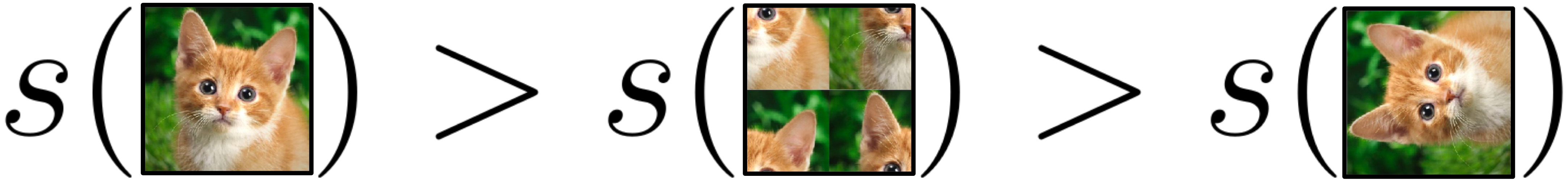
score: cosine similarity * norm

- Further improving the detection score by **utilizing the shifting transformation:**
  - $s_{\mathrm{con-SI}}(x,\{x_m\})$: ensemble the score $s_{\mathrm{con}}(x;\{x_m\})$ **over all shifting transformation**
  - $s_{\mathrm{cls-SI}}(x)$: confidence of the **shifting transformation classifier**

$$s_{\texttt{CSI}}(x;\{x_m\}) := s_{\texttt{con-SI}}(x;\{x_m\}) + s_{\texttt{cls-SI}}(x)$$

Contrasting Shifted Instances (CSI): OOD-ness

- **OOD-ness**: How to choose the shifting transformation?
  - The transformation that generates the most **OOD-like yet semantically meaningful samples**
  - We choose the transformation with the high OOD-ness (AUROC on vanilla SimCLR)



Original sample

Low OOD-ness

High OOD-ness: Most OOD-like transformation

Contrasting Shifted Instances (CSI): Extension

- We also **extend CSI** for training <u>confidence-calibrated classifier</u> [2]:
  - Accurate on predicting label y when input x is in-distribution
  - *Confidence* $s_{\text{sup}}(x) := \max_y p(y|x)$ of the classifier is well-calibrated

: in-distribution *correct* sample : in-distribution *in-correct* sample : OOD sample

- We adapt the **idea of CSI** to the <u>supervised contrastive learning</u> (SupCLR) [3]:
  - SupCLR contrasts samples in *class-wise*, instead of in instance-wise
  - Similar to CSI, sup-CSI consider **shifted instance as a different class's sample**

[2] Lee et al. Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples. ICLR 2018.
[3] Khosla et al. Supervised contrastive learning. NeurIPS 2020.

Experiments: Unlabeled One-class OOD

- **CSI** achieves the state-of-the-art performance in **all tested scenarios**:
  - For unlabeled one-class OOD detection, outperforms prior methods in **every classes**

(a) One-class CIFAR-10

Method	Network	Plane	Car	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck	Mean
OC-SVM* [64]	-	65.6	40.9	65.3	50.1	75.2	51.2	71.8	51.2	67.9	48.5	58.8
DeepSVDD* [60]	LeNet	61.7	65.9	50.8	59.1	60.9	65.7	67.7	67.3	75.9	73.1	64.8
AnoGAN* [63]	DCGAN	67.1	54.7	52.9	54.5	65.1	60.3	58.5	62.5	75.8	66.5	61.8
OCGAN* [55]	OCGAN	75.7	53.1	64.0	62.0	72.3	62.0	72.3	57.5	82.0	55.4	65.7
Geom* [17]	WRN-16-8	74.7	95.7	78.1	72.4	87.8	87.8	83.4	95.5	93.3	91.3	86.0
Rot* [27]	WRN-16-4	71.9	94.5	78.4	70.0	77.2	86.6	81.6	93.7	90.7	88.8	83.3
Rot+Trans* [27]	WRN-16-4	77.5	96.9	87.3	80.9	92.7	90.2	90.9	96.5	95.2	93.3	90.1
GOAD* [2]	WRN-10-4	77.2	96.7	83.3	77.7	87.8	87.8	90.0	96.1	93.8	92.0	88.2
Rot [27]	ResNet-18	78.3±0.2	94.3±0.3	86.2±0.4	80.8±0.6	89.4±0.5	89.0±0.4	88.9±0.4	95.1±0.2	92.3±0.3	89.7±0.3	88.4
Rot+Trans [27]	ResNet-18	80.4±0.3	96.4±0.2	85.9±0.3	81.1±0.5	91.3±0.3	89.6±0.3	89.9±0.3	95.9±0.1	95.0±0.1	92.6±0.2	89.8
GOAD [2]	ResNet-18	75.5±0.3	94.1±0.3	81.8±0.5	72.0±0.3	83.7±0.9	84.4±0.3	82.9±0.8	93.9±0.3	92.9±0.3	89.5±0.2	85.1
CSI (ours)	ResNet-18	**89.9**±0.1	**99.1**±0.0	**93.1**±0.2	**86.4**±0.2	**93.9**±0.1	**93.2**±0.2	**95.1**±0.1	**98.7**±0.0	**97.9**±0.0	**95.5**±0.1	**94.3**

(b) One-class CIFAR-100 (super-class)

Method	Network	AUROC
OC-SVM* [64]	-	63.1
Geom* [17]	WRN-16-8	78.7
Rot [27]	ResNet-18	77.7
Rot+Trans [27]	ResNet-18	79.8
GOAD [2]	ResNet-18	74.5
CSI (ours)	ResNet-18	**89.6**

(c) One-class ImageNet-30

Method	Network	AUROC
Rot* [27]	ResNet-18	65.3
Rot+Trans* [27]	ResNet-18	77.9
Rot+Attn* [27]	ResNet-18	81.6
Rot+Trans+Attn* [27]	ResNet-18	84.8
Rot+Trans+Attn+Resize* [27]	ResNet-18	85.7
CSI (ours)	ResNet-18	**91.6**

Experiments: Unlabeled Multi-class OOD

- **CSI** achieves the state-of-the-art performance in **all tested scenarios**:
  - For unlabeled multi-class OOD detection, outperforms prior methods in **every OOD datasets**

(a) Unlabeled CIFAR-10

Method	Network	CIFAR10 → SVHN	LSUN	ImageNet	LSUN (FIX)	ImageNet (FIX)	CIFAR-100	Interp.
Likelihood*	PixelCNN++	8.3	-	64.2	-	-	52.6	52.6
Likelihood*	Glow	8.3	-	66.3	-	-	58.2	58.2
Likelihood*	EBM	63.0	-	-	-	-	-	70.0
Likelihood Ratio* [55]	PixelCNN++	91.2	-	-	-	-	-	-
Input Complexity* [61]	PixelCNN++	92.9	-	58.9	-	-	53.5	-
Input Complexity* [61]	Glow	95.0	-	71.6	-	-	73.6	-
Rot [25]	ResNet-18	97.6±0.2	89.2±0.7	90.5±0.3	77.7±0.3	83.2±0.1	79.0±0.1	64.0±0.3
Rot+Trans [25]	ResNet-18	97.8±0.2	92.8±0.9	94.2±0.7	81.6±0.4	86.7±0.1	82.3±0.2	68.1±0.8
GOAD [2]	ResNet-18	96.3±0.2	89.3±1.5	91.8±1.2	78.8±0.3	83.3±0.1	77.2±0.3	59.4±1.1
CSI (ours)	ResNet-18	**99.8**±0.0	**97.5**±0.3	**97.6**±0.3	**90.3**±0.3	**93.3**±0.1	**89.2**±0.1	**79.3**±0.2

(b) Unlabeled ImageNet-30

Method	Network	ImageNet-30 → CUB-200	Dogs	Pets	Flowers	Food-101	Places-365	Caltech-256	DTD
Rot [25]	ResNet-18	76.5±0.7	77.2±0.5	70.0±0.5	87.2±0.2	72.7±1.5	52.6±1.4	70.9±0.1	89.9±0.5
Rot+Trans [25]	ResNet-18	74.5±0.5	77.8±1.1	70.0±0.8	86.3±0.3	71.6±1.4	53.1±1.7	70.0±0.2	89.4±0.6
GOAD [2]	ResNet-18	71.5±1.4	74.3±1.6	65.5±1.3	82.8±1.4	68.7±0.7	51.0±1.1	67.4±0.8	87.5±0.8
CSI (ours)	ResNet-18	**90.5**±0.1	**97.1**±0.1	**85.2**±0.2	**94.7**±0.4	**89.2**±0.3	**78.3**±0.3	**87.1**±0.1	**96.9**±0.1

Experiments: Labeled Multi-class OOD

- **CSI** achieves the state-of-the-art performance in **all tested scenarios**:
  - For labeled multi-class OOD detection, outperforms prior methods in **every OOD datasets**

(a) Labeled CIFAR-10

			CIFAR10 →						
Train method	Test acc.	ECE	SVHN	LSUN	ImageNet	LSUN (FIX)	ImageNet (FIX)	CIFAR100	Interp.
Cross Entropy	93.0±0.2	6.44±0.2	88.6±0.9	90.7±0.5	88.3±0.6	87.5±0.3	87.4±0.3	85.8±0.3	75.4±0.7
SupCLR [30]	93.8±0.1	5.56±0.1	97.3±0.1	92.8±0.5	91.4±1.2	91.6±1.5	90.5±0.5	88.6±0.2	75.7±0.1
CSI (ours)	94.8±0.1	4.40±0.1	96.5±0.2	96.3±0.5	96.2±0.4	92.1±0.5	92.4±0.0	90.5±0.1	78.5±0.2
CSI-ens (ours)	**96.1**±0.1	**3.50**±0.1	**97.9**±0.1	**97.7**±0.4	**97.6**±0.3	**93.5**±0.4	**94.0**±0.1	**92.2**±0.1	**80.1**±0.3

(b) Labeled ImageNet-30

			ImageNet-30 →						
Train method	Test acc.	ECE	CUB-200	Dogs	Pets	Flowers	Food-101	Places-365	Caltech-256	DTD
Cross Entropy	94.3	5.08	88.0	96.7	95.0	89.7	79.8	90.5	90.6	90.1
SupCLR [30]	96.9	3.12	86.3	95.6	94.2	92.2	81.2	89.7	90.2	92.1
CSI (ours)	97.0	2.61	93.4	97.7	96.9	96.0	87.0	92.5	91.9	93.7
CSI-ens (ours)	**97.8**	**2.19**	**94.6**	**98.3**	**97.4**	**96.2**	**88.9**	**94.0**	**93.2**	**97.4**

Experiments: Ablation study

- We verified the effectiveness of **shifting transformation selection scheme**
  - **Higher OOD-ness valued transformation** leads to higher detection performance

(a) Original (b) Cutout (c) Sobel (d) Noise (e) Blur (f) Perm (g) Rotate

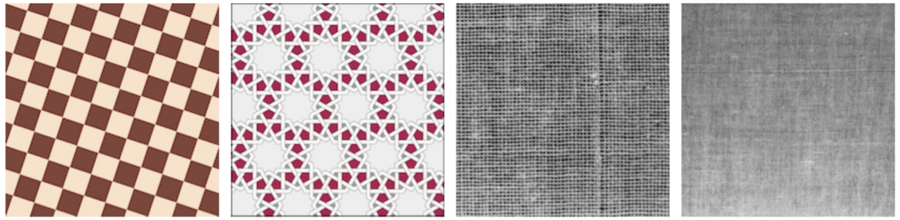
	Cutout	Sobel	Noise	Blur	Perm	Rotate
OOD-ness	79.5	69.2	74.4	76.0	83.8	85.2

Higher OOD-ness → Higher performance

Base		Cutout	Sobel	Noise	Blur	Perm	Rotate
87.9	+Align	84.3	85.0	85.5	88.0	73.1	76.5
	+Shift	88.5	88.3	89.3	89.2	90.7	94.3

Experiments: Ablation study

- We verified the effectiveness of **shifting transformation selection scheme**
  - **Higher OOD-ness valued transformation** leads to higher detection performance
  - Our method works on rotation-invariant datasets *i.e., rotation is not shifting transformation*



(a) OOD-ness		(b) AUROC		
Rot.	Noise	Base	CSI(R)	CSI(N)
50.6	**75.7**	70.3	65.9	**80.1**

- We also verified the effectiveness of **each component**

(a) Training objective

	SimCLR	Con.	Cls.	AUROC
$\mathcal{L}_{\texttt{SimCLR}}$ (2)	✓	-	-	87.9
$\mathcal{L}_{\texttt{con-SI}}$ (3)	✓	✓	-	91.6
$\mathcal{L}_{\texttt{cls-SI}}$ (4)	-	-	✓	88.6
$\mathcal{L}_{\texttt{CSI}}$ (5)	✓	✓	✓	**94.3**

(b) Detection score

	Con.	Cls.	Ensem.	AUROC
$s_{\texttt{con}}$ (6)	✓	-	-	91.3
$s_{\texttt{con-SI}}$ (7)	✓	-	✓	93.3
$s_{\texttt{cls-SI}}$ (8)	-	✓	✓	93.8
$s_{\texttt{CSI}}$ (9)	✓	✓	✓	**94.3**

Conclusion

- We propose **Contrasting Shifted Instances (CSI)** for OOD detection
  - We extend the power of contrastive learning for OOD detection
  - We further improve the OOD detection by utilizing shifting transformations

- CSI shows **outstanding performance** under various OOD detection scenarios

- We believe CSI would guide various future directions in **OOD detection** & **self-supervised learning** as an important baseline.

Thank you for your attention ☺

Paper: arxiv.org/abs/2007.08176

Code: https://github.com/alinlab/CSI