CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Instances

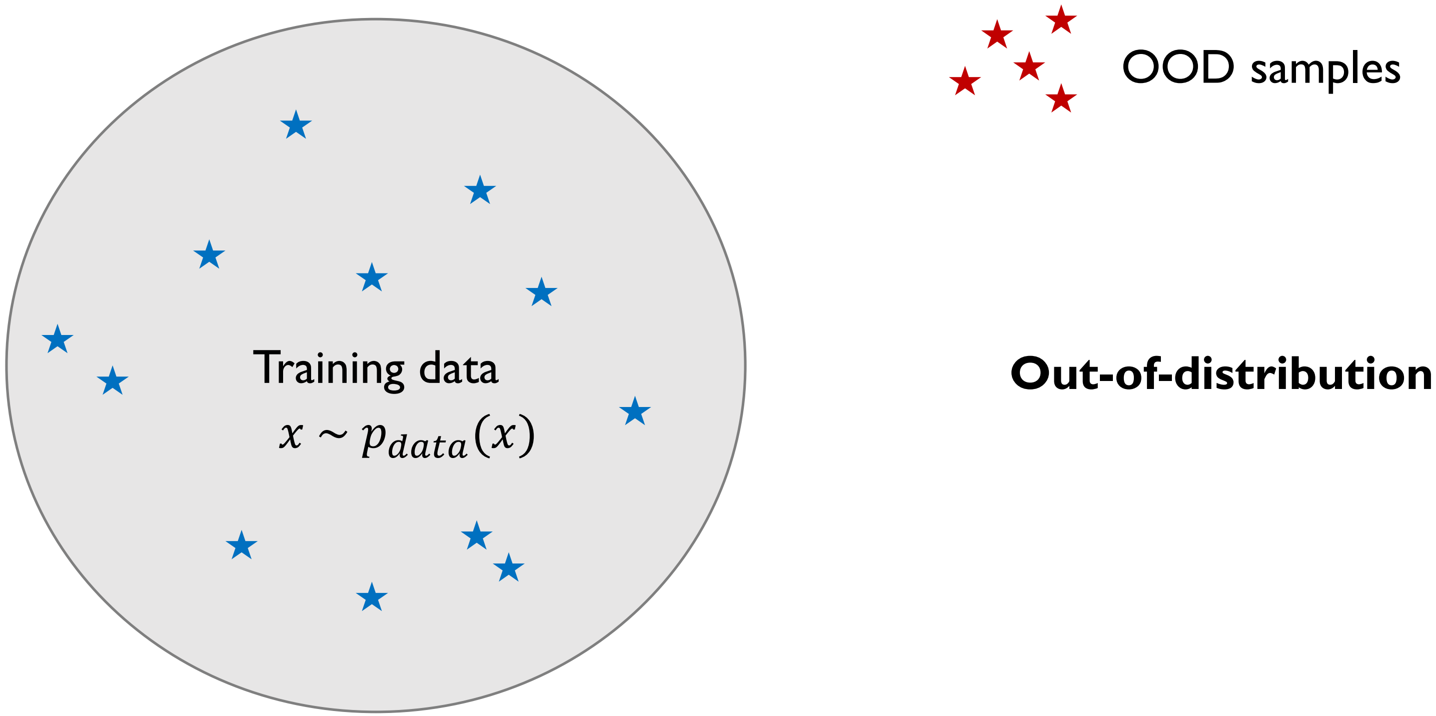
Jihoon Tack, Sangwoo Mo, Jongheon Jeong , Jinwoo Shin
2020 NIPS

Presenter: jeonghoon park

Out-Of-Distribution Detection

- Dataset $\{x_m\}_{m=1}^{M}$

In-distribution



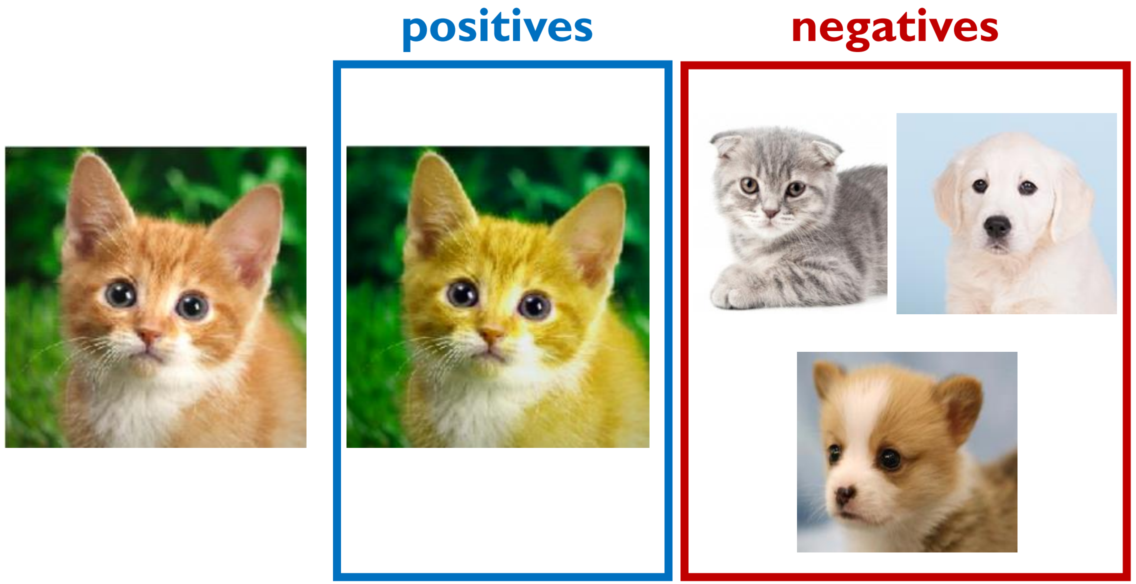
Contrastive learning

- Primitive form of the contrastive loss

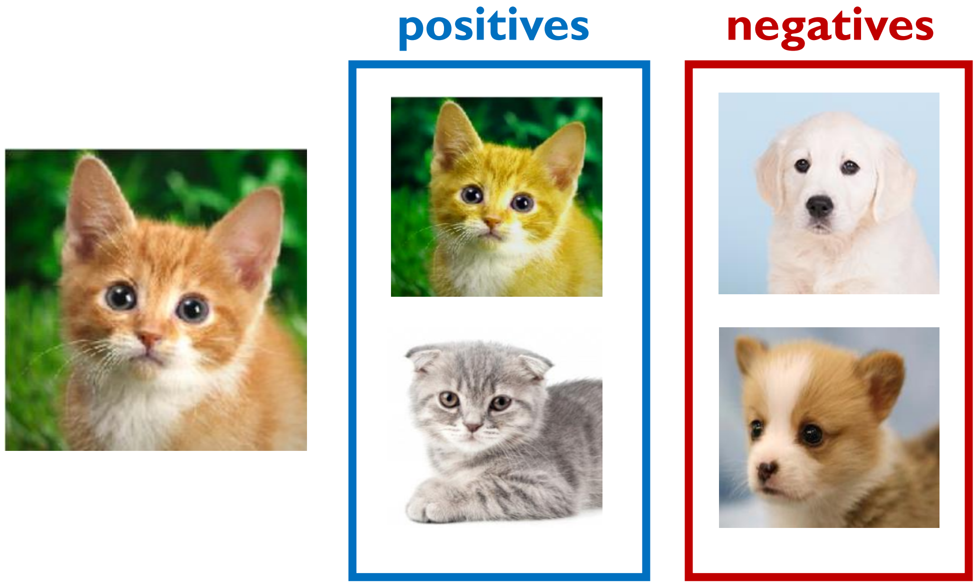
$$\mathcal{L}_{\text{con}}(x, \{x_+\}, \{x_-\}) := -\frac{1}{|\{x_+\}|} \log \frac{\sum_{x' \in \{x_+\}} \exp(\mathrm{sim}(z(x), z(x'))/\tau)}{\sum_{x' \in \{x_+\} \cup \{x_-\}} \exp(\mathrm{sim}(z(x), z(x'))/\tau)},$$

where x: query, $\{x_+\}$: set of positive samples, $\{x_-\}$: set of negative samples, $\mathrm{sim}(z, z') = z \cdot z' / \|z\| \|z'\|$

- SimCLR (w/o label)

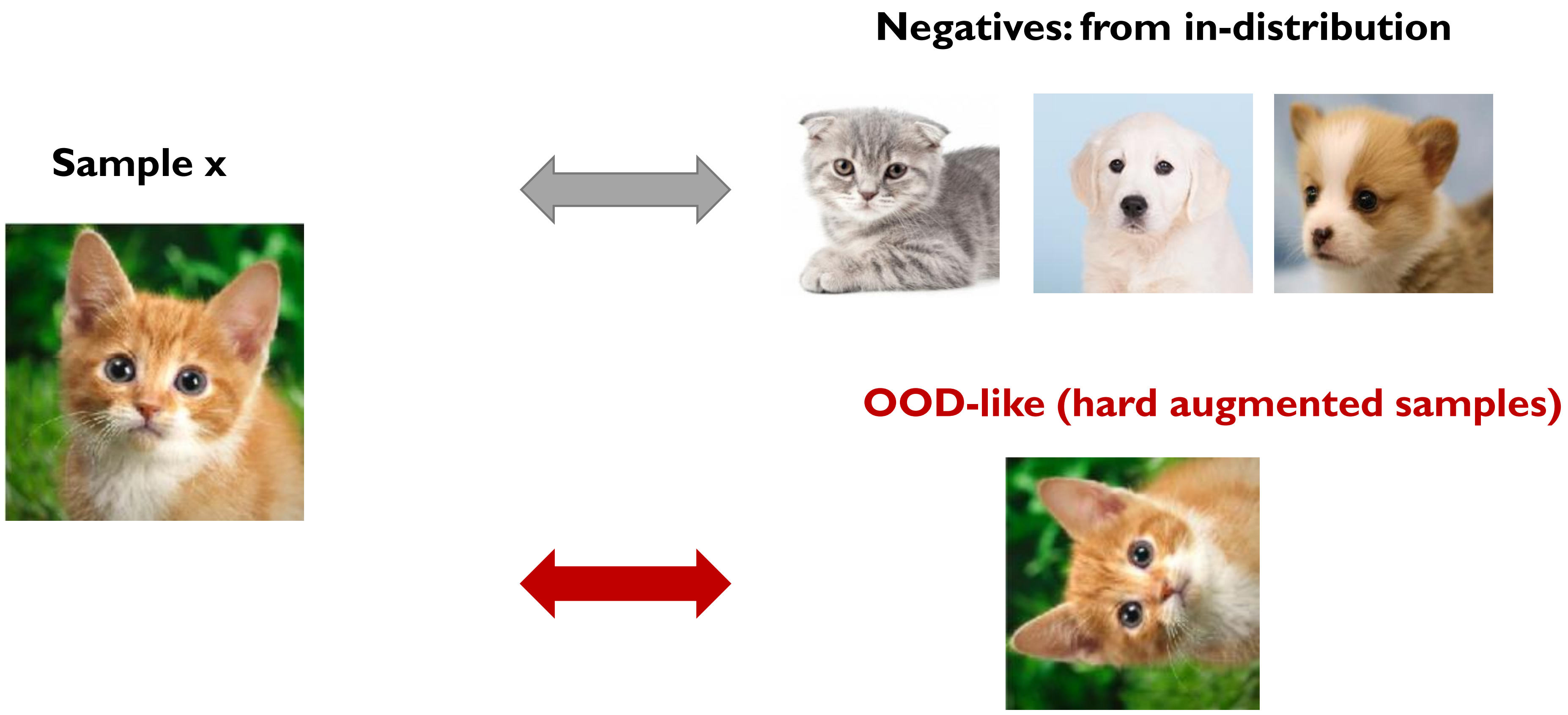


- SupCLR (w/ label)



Key Idea

- Discriminating within in-distribution => discriminating between in- and out-distribution



SimCLR

- SimCLR: for a given batch $\mathcal{B} := \{x_i\}_{i=1}^{B}$

$$\mathcal{L}_{\texttt{SimCLR}}(\mathcal{B};\mathcal{T}) := \frac{1}{2B}\sum_{i=1}^{B}\mathcal{L}_{\texttt{con}}(\tilde{x}_i^{(1)}, \tilde{x}_i^{(2)}, \tilde{\mathcal{B}}_{-i}) + \mathcal{L}_{\texttt{con}}(\tilde{x}_i^{(2)}, \tilde{x}_i^{(1)}, \tilde{\mathcal{B}}_{-i}),$$

$$\tilde{\mathcal{B}} := \{\tilde{x}_i^{(1)}\}_{i=1}^{B} \cup \{\tilde{x}_i^{(2)}\}_{i=1}^{B} \text{ and } \tilde{\mathcal{B}}_{-i} := \{\tilde{x}_j^{(1)}\}_{j\neq i} \cup \{\tilde{x}_j^{(2)}\}_{j\neq i}$$

where $\tilde{x}_i^{(1)} = T_1(x_i)$ and $\tilde{x}_i^{(2)} = T_2(x_i)$, $T_1, T_2 \sim \mathcal{T}$

	x_1	x_2	x_3
T_1	$\tilde{x}_1^{(1)}$	$\tilde{x}_2^{(1)}$	$\tilde{x}_3^{(1)}$
T_2	$\tilde{x}_1^{(2)}$	$\tilde{x}_2^{(2)}$	$\tilde{x}_3^{(2)}$

positives **negatives**

Contrastive learning for distribution-shifting transformations

(1) Contrasting shifted instances loss

- ✓ Key finding: some augmentations can be useful for OOD detection by considering them as negatives
- ✓ Family of augmentations S: distribution-shifting transformations, $S = \{ S_0 = I, S_1 \ldots, S_{k-1} \}$
- ✓ Distributionally-shifted samples are considered as an OOD

$$\mathcal{L}_{\texttt{con-SI}} := \mathcal{L}_{\texttt{SimCLR}}\left(\bigcup_{S \in \mathcal{S}} \mathcal{B}_S; \mathcal{T}\right), \quad \text{where } \mathcal{B}_S := \{S(x_i)\}_{i=1}^{B}.$$

	x_1	x_2	x_3	$S_1(x_1)$	$S_1(x_2)$	$S_1(x_3)$	$S_2(x_1)$	$S_2(x_2)$	$S_2(x_3)$
T_1	$\tilde{x}_1^{(1)}$	$\tilde{x}_2^{(1)}$	$\tilde{x}_3^{(1)}$	$\widetilde{S_1(x_1)}^{(1)}$	$\widetilde{S_1(x_2)}^{(1)}$	$\widetilde{S_1(x_3)}^{(1)}$	$\widetilde{S_2(x_1)}^{(1)}$	$\widetilde{S_2(x_2)}^{(1)}$	$\widetilde{S_2(x_3)}^{(1)}$
T_2	$\tilde{x}_1^{(2)}$	$\tilde{x}_2^{(2)}$	$\tilde{x}_3^{(2)}$	$\widetilde{S_1(x_1)}^{(2)}$	$\widetilde{S_1(x_2)}^{(2)}$	$\widetilde{S_1(x_3)}^{(2)}$	$\widetilde{S_2(x_1)}^{(2)}$	$\widetilde{S_2(x_2)}^{(2)}$	$\widetilde{S_2(x_3)}^{(2)}$

positives **negatives**

Contrastive learning for distribution-shifting transformations

(2) Classifying shifted instances loss

✓ Predicts shifting transformation $y^S \in \mathcal{S}$ for a given input x

✓ Add an linear layer to $f(\theta)$ for a classifier $p_{\mathrm{cls-SI}}(y^S|x)$

✓ $\tilde{B}_S$: batch augmented from B_S via SimCLR

$$\mathcal{L}_{\texttt{cls-SI}} := \frac{1}{2B}\frac{1}{K}\sum_{S\in\mathcal{S}}\sum_{\tilde{x}_S\in\tilde{\mathcal{B}}_S} -\log p_{\texttt{cls-SI}}(y^{\mathcal{S}} = S \mid \tilde{x}_S).$$

(3) Overall loss

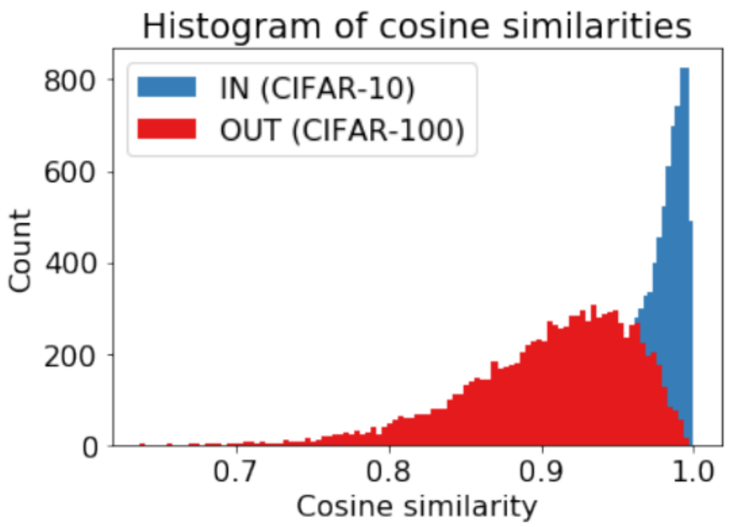
$$\mathcal{L}_{\texttt{CSI}} = \mathcal{L}_{\texttt{con-SI}} + \lambda \cdot \mathcal{L}_{\texttt{cls-SI}}$$

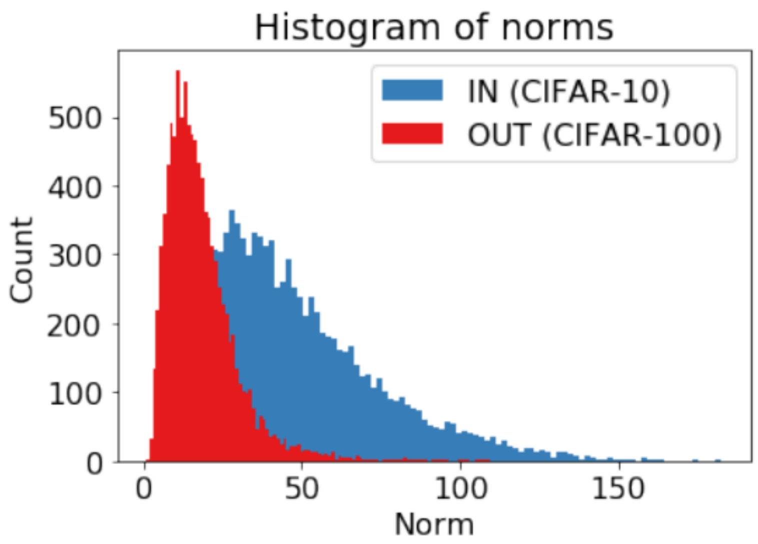
Score functions for detecting out-of-distribution

Detection score

✓ Cosine similarity to the nearest training sample
✓ Norm of the representation

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





Score functions for detecting out-of-distribution

Utilizing shifting transformations

$$s_{\texttt{con-SI}}(x;\{x_m\}) := \sum_{S\in\mathcal{S}} \lambda_S^{\texttt{con}}\, s_{\texttt{con}}(S(x);\{S(x_m)\}),$$

$$\lambda_S^{\texttt{con}} := M/\textstyle\sum_m s_{\texttt{con}}(S(x_m);\{S(x_m)\}) = M/\sum_m \|z(S(x_m))\|$$

$$s_{\texttt{cls-SI}}(x) := \sum_{S\in\mathcal{S}} \lambda_S^{\texttt{cls}}\, W_S f_\theta(S(x)),$$

$$\lambda_S^{\texttt{cls}} := M/\textstyle\sum_m [W_S f_\theta(S(x_m))]$$

W_S is the weight vector in the linear layer of $p_{\text{cls-SI}}(y^S|x)$

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

Ensembling over random augmentations

✓ Ensembling over random augmentations T

$$s_{\texttt{CSI-ens}}(x) := \mathbb{E}_{T\sim\mathcal{T}}[s_{\texttt{CSI}}(T(x))]$$

Experiments

Experimental Setting

✓ ResNet-18
✓ Data augmentations T: Inception crop, horizontal flip, color jitter, and grayscale
✓ Distribution-shifting transformations S: random rotation 0°, 90°, 180°, 270°
Shifting transformation : the most OOD-like yet semantically meaningful samples.
Cutout/Sobel filtering/Gaussian noise/Gaussian blur/Rotation: reported to be ineffective in SimCLR
Such transformations shift the in-distribution => considering as positive samples can be harmful.
OOD-ness: the AUROC between in-distribution vs. transformed samples under vanilla SimCLR (one-class CIFAR-10)

(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

Most shift the distribution!

Experiments

Experimental Setting

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

✓ Such transformations shift the in-distribution => considering as positive samples can be harmful.
✓ Using hard augmented samples as negative samples improves OOD detection 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

Results

Unlabeled one-class datasets

- ✓ In-distribution: one of the classes
- ✓ OOD: remaining classes

(a) One-class CIFAR-10

Method	Network	Plane	Car	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck	Mean
OC-SVM* [59]	-	65.6	40.9	65.3	50.1	75.2	51.2	71.8	51.2	67.9	48.5	58.8
DeepSVDD* [56]	LeNet	61.7	65.9	50.8	59.1	60.9	65.7	67.7	67.3	75.9	73.1	64.8
AnoGAN* [58]	DCGAN	67.1	54.7	52.9	54.5	65.1	60.3	58.5	62.5	75.8	66.5	61.8
OCGAN* [52]	OCGAN	75.7	53.1	64.0	62.0	72.3	62.0	72.3	57.5	82.0	55.4	65.7
Geom* [15]	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* [25]	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* [25]	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 [25]	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 [25]	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* [59]	-	63.1
Geom* [15]	WRN-16-8	78.7
Rot [25]	ResNet-18	77.7
Rot+Trans [25]	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* [25]	ResNet-18	65.3
Rot+Trans* [25]	ResNet-18	77.9
Rot+Attn* [25]	ResNet-18	81.6
Rot+Trans+Attn* [25]	ResNet-18	84.8
Rot+Trans+Attn+Resize* [25]	ResNet-18	85.7
CSI (ours)	ResNet-18	**91.6**

Results

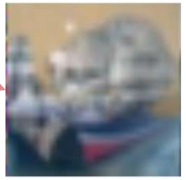
Unlabeled multi-class datasets

✓ In-distribution: multi-class dataset w/o labels

✓ OOD: external 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

Results

Unlabeled multi-class datasets

- ✓ In-distribution: multi-class dataset w/o labels
- ✓ OOD: external datasets

Figure 5: Current benchmark datasets: resized LSUN (left two) and ImageNet (right two).

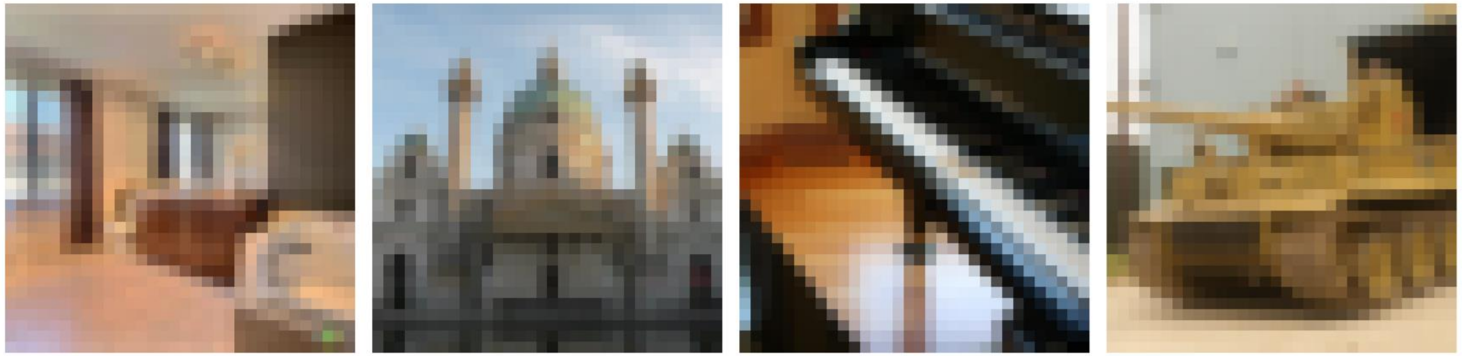


Figure 6: Proposed datasets: LSUN (FIX) (left two) and ImageNet (FIX) (right two).

Ablation study

(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

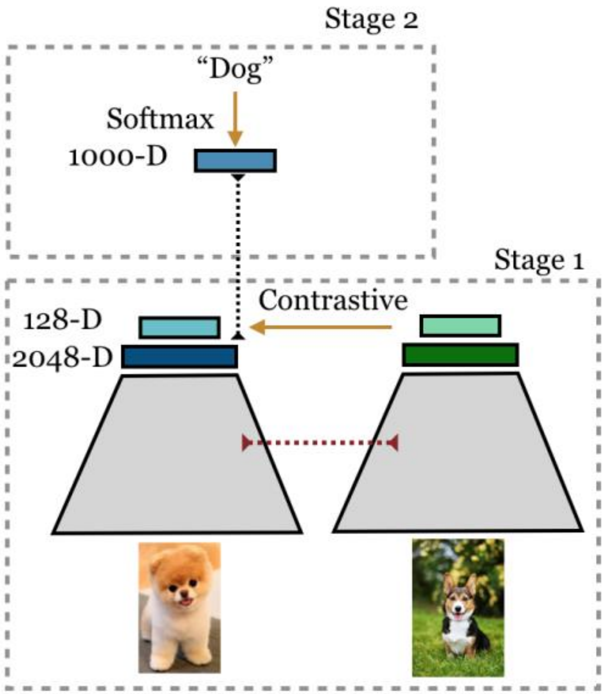
	Con.	Cls.	Ensem.	AUROC
$s_{\texttt{con}}$ (6)	✓	-	-	91.3
$s_{\texttt{con-SI}}$ (7)	✓	-	✓	93.3
$s_{\texttt{cls-SI}}$ (8)	-	✓	✓	93.8
$s_{\texttt{CSI}}$ (9)	✓	✓	✓	**94.3**

SupCLR

- ✓ Supervised Contrastive Learning
- ✓ For training confidence-calibrated classifiers
- ✓ For a given batch $\mathcal{C} = \{(x_i, y_i)\}_{i=1}^{B}$, $\tilde{\mathcal{C}} = \tilde{\mathcal{C}}_y \cup \tilde{\mathcal{C}}_{-y}$

$$\mathcal{L}_{\mathtt{SupCLR}}(\mathcal{C}; \mathcal{T}) := \frac{1}{2B} \sum_{j=1}^{2B} \mathcal{L}_{\mathtt{con}}(\tilde{x}_j, \tilde{\mathcal{C}}_{y_j} \setminus \{\tilde{x}_j\}, \tilde{\mathcal{C}}_{-y_j}).$$

- ✓ Add a linear layer on $f_\theta(x)$ to classify $p_{\text{SupCLR}}(y|x)$

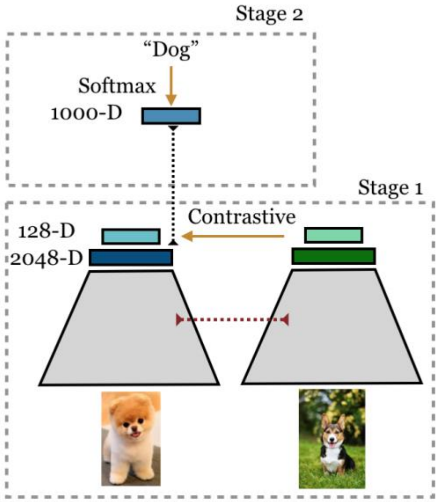


Class	y_1		y_2	
samples	x_1	x_2	x_3	x_4
T_1	$\tilde{x}_1^{(1)}$	$\tilde{x}_2^{(1)}$	$\tilde{x}_3^{(1)}$	$\tilde{x}_4^{(1)}$
T_2	$\tilde{x}_1^{(2)}$	$\tilde{x}_2^{(2)}$	$\tilde{x}_3^{(2)}$	$\tilde{x}_4^{(2)}$

positives **negatives**

Supervised extension of CSI

- ✓ For training confidence-calibrated classifiers
- ✓ Family of augmentations S: distribution-shifting transformations, $S = \{ S_0 = I, S_1 \ \ldots, S_{k-1}\}$
- ✓ Distributionally-shifted samples are considered as an OOD



$$\mathcal{L}_{\texttt{sup-CSI}} := \mathcal{L}_{\texttt{SupCLR}}\left(\bigcup_{S\in\mathcal{S}} \mathcal{C}_S; \mathcal{T}\right), \quad \text{where } \mathcal{C}_S := \{(S(x_i),(y_i,S))\}_{i=1}^{B}.$$

- ✓ Now, an added linear layer on $f_\theta(x)$ classifies the shifted instances with joint labels (y_i, S)

Class	y_1		y_2		y_1		y_2	
samples	x_1	x_2	x_3	x_4	$S_1(x_1)$	$S_1(x_2)$	$S_1(x_3)$	$S_1(x_4)$
T_1	$\tilde{x}_1^{(1)}$	$\tilde{x}_2^{(1)}$	$\tilde{x}_3^{(1)}$	$\tilde{x}_4^{(1)}$	$\widetilde{S_1(x_1)}^{(1)}$	$\widetilde{S_1(x_2)}^{(1)}$	$\widetilde{S_1(x_3)}^{(1)}$	$\widetilde{S_1(x_4)}^{(1)}$
T_2	$\tilde{x}_1^{(2)}$	$\tilde{x}_2^{(2)}$	$\tilde{x}_3^{(2)}$	$\tilde{x}_4^{(2)}$	$\widetilde{S_1(x_1)}^{(2)}$	$\widetilde{S_1(x_2)}^{(2)}$	$\widetilde{S_1(x_3)}^{(2)}$	$\widetilde{S_1(x_4)}^{(2)}$

positives **negatives**

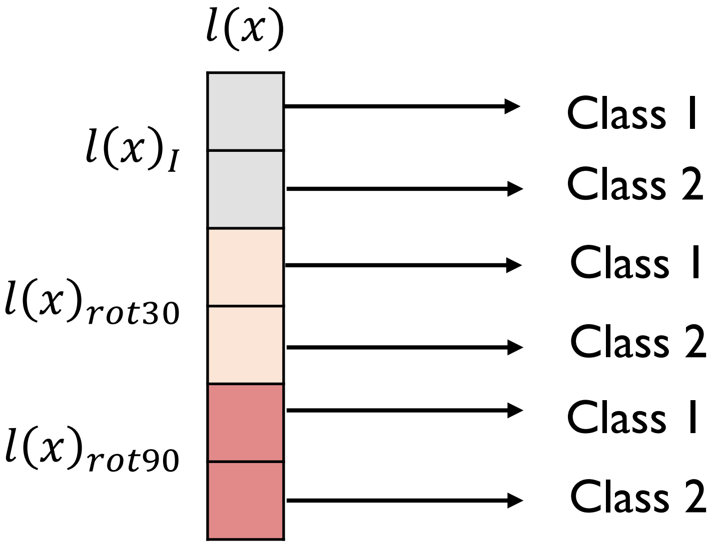
Supervised extension of CSI: confidence score

- Additionally train two linear classifiers: $p_{\text{CSI}}(y|x), p_{\text{CSI}-\text{joint}}(y, y^{\mathcal{S}}|x)$
- "CSI": Confidence computed by p_{CSI}

$$s_{\text{sup}}(x) = \max_y p_{\text{CSI}}(y|x)$$

- "CSI-ens": Confidence computed by $p_{\text{CSI}-\text{joint}}$

$$s_{\text{sup}}(x) = \max_y p_{\text{CSI}-\text{ens}}(y|x)$$



$$p_{\texttt{CSI-ens}}(y|x) := \sigma\left(\frac{1}{K}\sum_k \underline{l(S_k(x))_k}\right)$$

Logit values correspond to

$p_{\texttt{CSI-joint}}(y, y^{\mathcal{S}} = S_k|x)$

$l(x)_k \in \mathbb{R}^C$

Results

Labeled multi-class datasets

- ✓ In-distribution: multi-class dataset w/ labels
- ✓ OOD: external datasets
- ✓ Calibrated well?

(a) Labeled CIFAR-10

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

(b) Labeled ImageNet-30

			ImageNet-30 $\rightarrow$							
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**

Summary

- ✓ Utilize contrastive learning for OOD detection by discriminating between in- and out-distribution
- ✓ Verify the effectiveness under various environments (unlabeled one-class, unlabeled multi-class, labeled multi-class)
- ✓ Larger improvement in harder OOD samples (verify with fixed version of the dataset)