模式分析与机器智能
工业和信息化部重点实验室
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

ParNeC 模式识别与神经计算研究组
PAttern Recognition and NEural Computing

# Mining In-distribution Attributes in Outliers for Out-of-distribution Detection

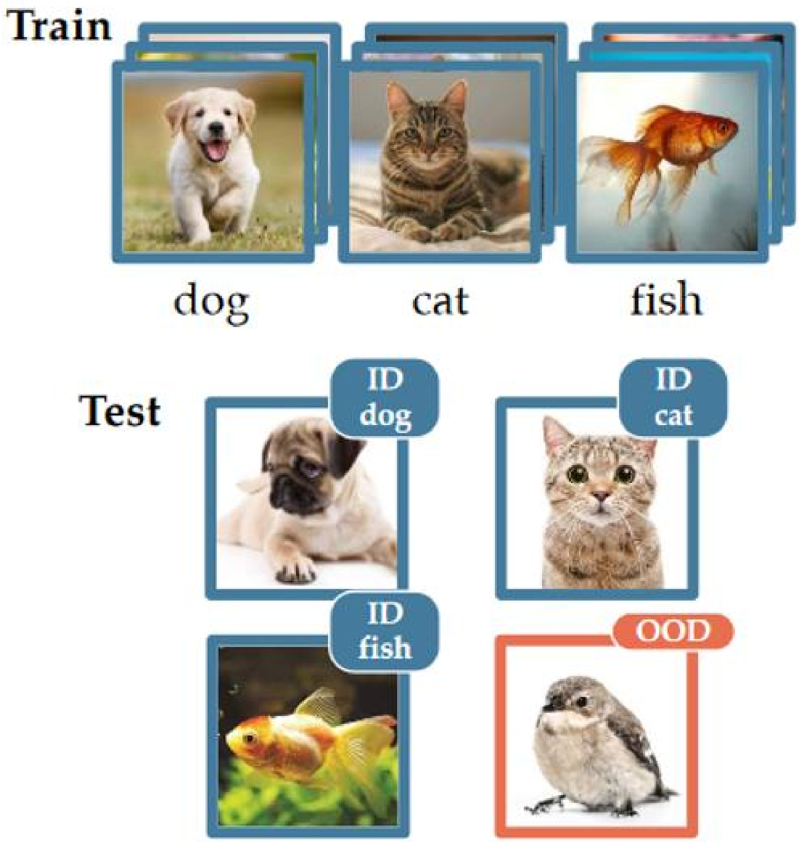
Yutian Lei, Luping Ji*, Pei Liu

AAAI 2025

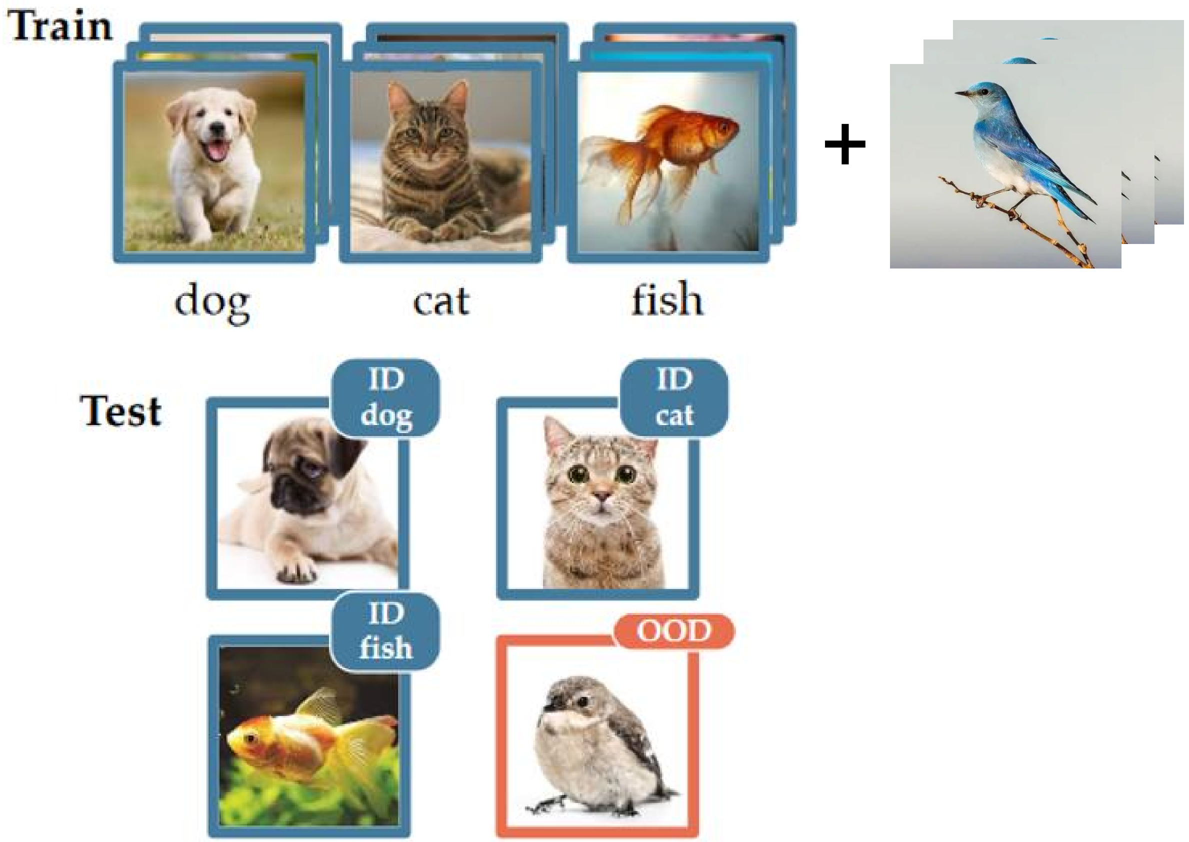
# Background



## OOD Detection




## OOD Detection with outlier exposure (OE)

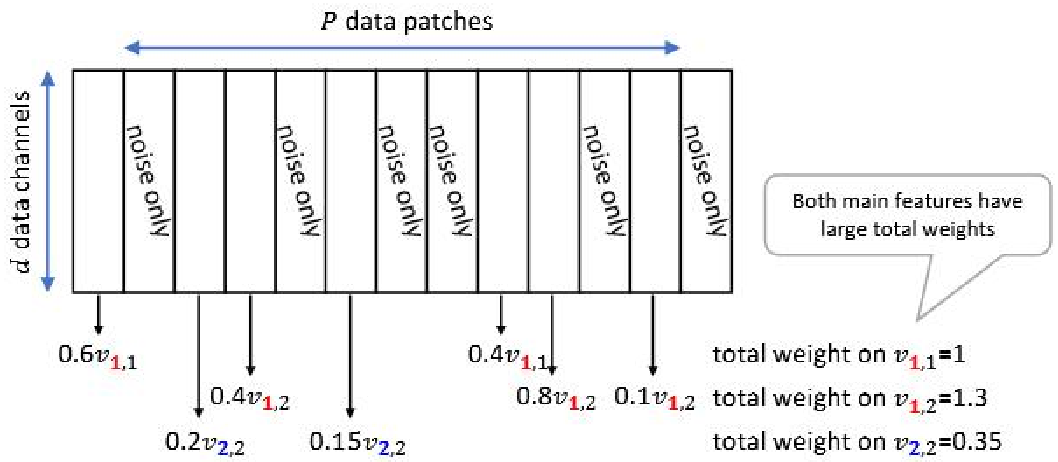


# Multi-view Data Model

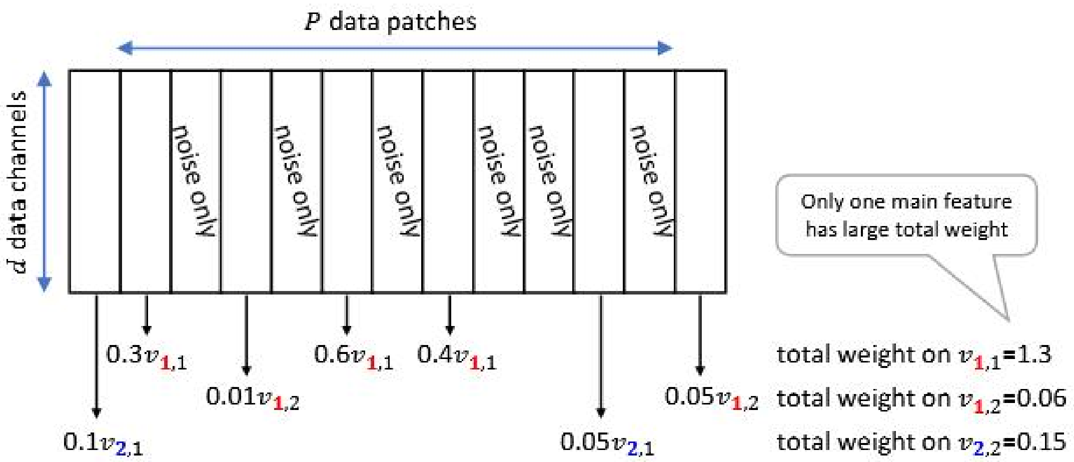
a data point comprises main features, minor features, and noise








(a) example of a *multi-view* data with label 1




(b) example of a *single-view* data with label 1

# Multi-view Data Model



## Example

- When the label is class 1, then:

$$\begin{cases} \text{both } v_1, v_2 \text{ appears with weight 1, one of } v_3, v_4 \text{ appears with weight 0.1} & \text{w.p. } 80\%; \\ \text{only } v_1 \text{ appears with weight 1, one of } v_3, v_4 \text{ appears with weight 0.1} & \text{w.p. } 10\%; \\ \text{only } v_2 \text{ appears with weight 1, one of } v_3, v_4 \text{ appears with weight 0.1} & \text{w.p. } 10\%. \end{cases}$$

- When the label is class 2, then

$$\begin{cases} \text{both } v_3, v_4 \text{ appears with weight 1, one of } v_1, v_2 \text{ appears with weight 0.1} & \text{w.p. } 80\%; \\ \text{only } v_3 \text{ appears with weight 1, one of } v_1, v_2 \text{ appears with weight 0.1} & \text{w.p. } 10\%; \\ \text{only } v_4 \text{ appears with weight 1, one of } v_1, v_2 \text{ appears with weight 0.1} & \text{w.p. } 10\%. \end{cases}$$

# Multi-view Data Model



## Definition

**Definition 1** (data distributions $D_m^{in}$ and $D_s^{in}$). *Given* $D \in \{D_m^{in}, D_s^{in}\}$, *we define* $(X^{in}, y) \sim D$ *as follows. First, choose the label* $y \in [k]$ *uniformly at random. Then,* $X^{in}$ *is generated as follows:*

**1.** *Denote* $\mathcal{V}(X^{in}) = \{v_{y,1}, v_{y,2}\} \cup \mathcal{V}'$ *as the set of feature vectors used in this data vector* $X$, *where* $\{v_{y,1}, v_{y,2}\}$ *are* ***main ID features*** *and* $\mathcal{V}'$ *is a set of* ***minor ID features*** *uniformly sampled from* $\{v_{j,1}, v_{j,2}\}_{j\in[k]\setminus\{y\}}$.

**2.** *For each* $v \in \mathcal{V}(X)$, *pick many disjoint patches in* $[P]$ *and denote them as* $\mathcal{P}_v(X^{in}) \subset [P]$. *We denote* $\mathcal{P}(X^{in}) = \bigcup_{v\in\mathcal{V}(X^{in})} \mathcal{P}_v(X^{in})$.

**3.** *If* $D = D_s^{in}$ *is the single-view distribution, pick a value* $\hat{\ell} = \ell(X^{in}) \in [2]$ *uniformly at random.*

**4.** *For each* $v \in \mathcal{V}(X^{in})$ *and* $p \in \mathcal{P}_v(X^{in})$, $x_p = z_p v +$ *"noise"* $\in \mathbb{R}^d$. *These random coefficients* $z_p \geq 0$ *satisfy:*

· *In the case of multi-view distribution* $D = D_m^{in}$:

*1)* $\sum_{p\in\mathcal{P}_v(X^{in})} z_p \in [1, O(1)]$ *when* $v \in \{v_{y,1}, v_{y,2}\}$;

*2)* $\sum_{p\in\mathcal{P}_v(X^{in})} z_p \in [\Omega(1), 0.4]$ *when* $v \in V(X^{in}) \setminus \{v_{y,1}, v_{y,2}\}$.

· *In the case of single-view distribution* $D = D_s^{in}$:

*1)* $\sum_{p\in\mathcal{P}_v(X^{in})} z_p \in [1, O(1)]$ *when picked* $v = v_{y,\hat{\ell}}$;

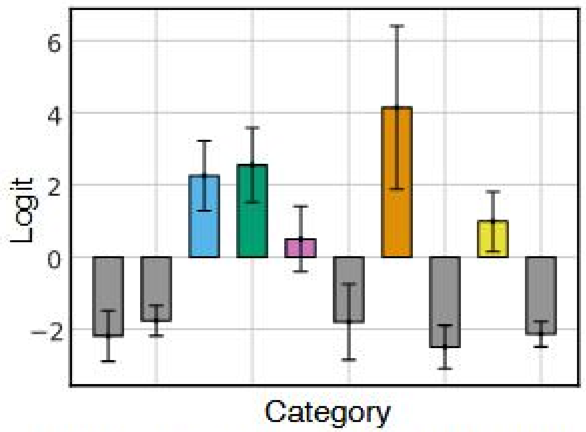
*2)* $\sum_{p\in\mathcal{P}_v(X^{in})} z_p$ *is much smaller than that of* $v_{y,\hat{\ell}}$ *and can be ignored when* $v \in V(X^{in}) \setminus \{v_{y,\hat{\ell}}\}$.

**5.** *For each* $p \in P \setminus P(X^{in})$, $x_p$ *consists only of "noise".*

# Motivation

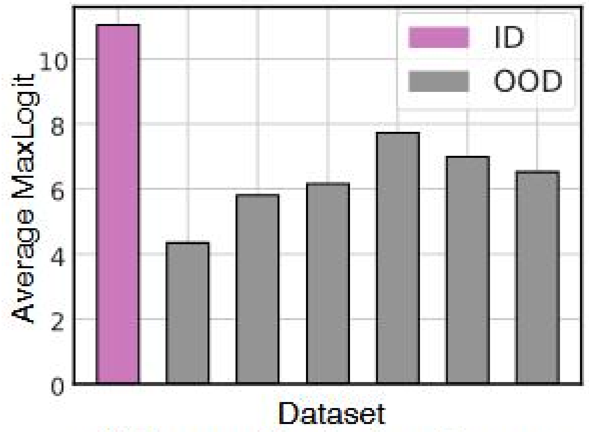


outliers could contain ID attributes




(a) Produced Logits for an OOD Data

outliers mainly consist of minor ID features and noise




(b) Average MaxLogit on Datasets

# Extended Multi-view Data Model

## Definition

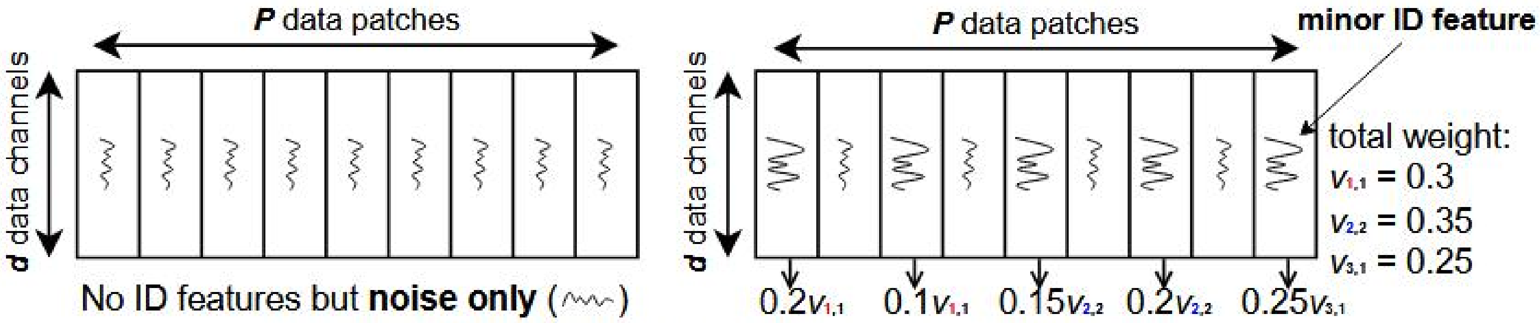
**Definition 3** (Out-of-distribution $D^{out}$). *We define $X^{out} \sim D^{out}$ as follows. $X^{out}$ is generated by:*

*1. Denote $\mathcal{V}(X^{out})$ as the set of **minor ID feature** vectors used in this data vector $X^{out}$, which are uniformly sampled from $\{v_{j,1}, v_{j,2}\}_{j\in[k]}$.*

*2. For each $v \in \mathcal{V}(X^{out})$, pick many disjoint patches in $[P]$ and denote it as $P_v(X^{out}) \subset [P]$. We denote $P(X^{out}) = \bigcup_{v\in\mathcal{V}(X^{out})} \mathcal{P}_v(X)$.*

*3. For each $v \in \mathcal{V}(X^{out})$ and $p \in \mathcal{P}_v(X^{out})$, we set $x_p = z_p v +$ "noise" $\in \mathbb{R}^d$. These random coefficients $z_p \geq 0$ satisfy that: $\sum_{p\in\mathcal{P}_v(X^{out})} z_p \in [\Omega(1), 0.4]$.*

*4. For each $p \in [P] \setminus \mathcal{P}(X^{out})$, $x_p$ consists only of "noise".*




(a) Underlying **OOD data structure** in OE (left) and our MVOL (right).

# Interpretability of MaxLogit



**MaxLogit based OOD detector.** For a sample $(X^{in}, y) \sim D^{in}$ or $X^{out} \sim D^{out}$ and a neural network $F$. Feeding $X \in \{X^{in}, X^{out}\}$ into $F$, we get logit outputs $F(X) = (F_1(X), \ldots, F_k(X)) \in \mathbb{R}^k$. Then, the MaxLogit scoring function is given as follows.

$$\mathrm{MaxLogit}(X; F) = \max(F_1(X), \ldots, F_k(X)).$$

Then MaxLogit can be used in the following OOD detector:

$$G(X; \tau, F) = \begin{cases} 0 & \text{if } \mathrm{MaxLogit}(X; F) \leq \tau, \\ 1 & \text{if } \mathrm{MaxLogit}(X; F) > \tau, \end{cases} \tag{1}$$

$$I(X) = \mathrm{argmax}_{j \in [k]} \sum_{p \in P_{v_{j,1}}(X) \bigcup P_{v_{j,2}}(X)} z_p; \tag{2}$$

$$z(X) = \max_{j \in [k]} \sum_{p \in P_{v_{j,1}}(X) \bigcup P_{v_{j,2}}(X)} z_p \tag{3}$$

$I(X)$ is the category with the largest sum of coefficients on associated features. $z(X)$ is this sum value.

**Proposition 1.** *For every $X^{out} \sim D^{out}$, every $(X_s^{in}, y_s) \sim D_s^{in}$, and every $(X_m^{in}, y_m) \sim D_m^{in}$, we have:*

$$z(X^{out}) < z(X_s^{in}) \quad \textit{and} \quad z(X^{out}) < z(X_m^{in})$$

$F_{I(X^{out})}(X^{out}) < F_{I(X_s^{in})}(X_s^{in})$ and $F_{I(X^{out})}(X^{out}) < F_{I(X_m^{in})}(X_m^{in})$ with Proposition 1, which corresponds to relation among MaxLogit scores.

# Multi-view-based Learning Objective



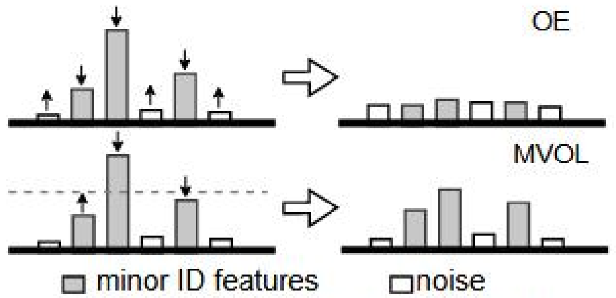
confidence loss in OE

$$\mathcal{L}_{\text{OE}} = \frac{1}{N}\sum_{j=1}^{N} -\log P_\theta(\hat{y}=y|X_j^{in}) + \frac{\beta}{M}\sum_{j=1}^{M}\sum_{i=1}^{k} -\frac{1}{k}\log P_\theta(\hat{y}=i|X_j^{out})$$

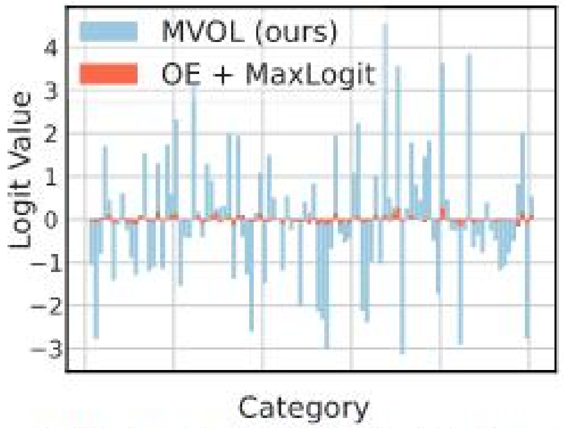
$$\mathcal{L}_{\text{MVOL}}^{(t)} = \frac{1}{N}\sum_{j=1}^{N} -\log P_\theta(\hat{y}=y|X_j^{in}) \tag{5}$$

$$+ \frac{\beta}{M}\sum_{j=1}^{M}\sum_{i=1}^{k} -p_{j,i}^{(t)}\log P_\theta(\hat{y}=i|X_j^{out}),$$

$$\text{where} \quad p_{j,i}^{(t)} = \min(\text{logit}_i(F^{(t)}, X_j^{out}), \epsilon), \tag{6}$$




(b) Optimization objective on **logits** of OE and MVOL for outliers.




(a) Visualization of Optimization Effects

# Experiment



| Setting | Category | Method | CIFAR-10 FPR95 ↓ | CIFAR-10 AUROC ↑ | CIFAR-10 ID-Acc ↑ | CIFAR-100 FPR95 ↓ | CIFAR-100 AUROC ↑ | CIFAR-100 ID-Acc ↑ |
|---|---|---|---|---|---|---|---|---|
| Single Model Setting | Post Hoc | MSP | 56.29 ± 1.62 | 89.59 ± 0.63 | 94.27 ± 0.14 | 80.75 ± 0.81 | 74.70 ± 1.01 | 74.69 ± 0.21 |
| | | Energy | 41.20 ± 5.28 | 89.70 ± 1.93 | 94.27 ± 0.14 | 72.58 ± 1.78 | 79.01 ± 1.12 | 74.69 ± 0.21 |
| | | MaxLogit | 41.68 ± 4.99 | 89.69 ± 1.88 | 94.27 ± 0.14 | 73.21 ± 1.69 | 78.88 ± 1.11 | 74.69 ± 0.21 |
| | | ODIN | 41.75 ± 3.86 | 87.38 ± 2.41 | 94.27 ± 0.14 | 68.13 ± 1.83 | 79.36 ± 0.91 | 74.69 ± 0.21 |
| | | Mahalanobis | 23.96 ± 1.26 | 92.81 ± 0.32 | 94.27 ± 0.14 | 46.40 ± 3.73 | 87.44 ± 1.12 | 74.69 ± 0.21 |
| | | KNN | 30.89 ± 2.76 | 94.53 ± 0.44 | 94.27 ± 0.14 | 82.02 ± 2.58 | 75.84 ± 1.35 | 74.69 ± 0.21 |
| | | ASH | 40.03 ± 5.18 | 90.01 ± 1.70 | 94.26 ± 0.11 | 63.31 ± 1.91 | 79.35 ± 1.09 | 74.23 ± 0.31 |
| | Outlier Synthesis | VOS | 34.67 ± 5.01 | 91.54 ± 1.92 | **94.75** ± 0.17 | 70.17 ± 2.52 | 81.73 ± 1.78 | **75.94** ± 0.20 |
| | | ATOL | 12.86 ± 0.59 | 97.34 ± 0.07 | 93.89 ± 0.17 | 64.67 ± 1.73 | 80.17 ± 1.34 | 72.70 ± 0.17 |
| | Outlier Exposure | Energy w/Aux | 4.70 ± 0.50 | 97.77 ± 0.06 | 90.74 ± 0.24 | 52.43 ± 3.51 | 88.40 ± 1.16 | 62.13 ± 0.27 |
| | | OE | 4.25 ± 0.15 | 98.56 ± 0.07 | 94.47 ± 0.13 | 46.51 ± 3.65 | 89.78 ± 0.98 | 74.02 ± 0.04 |
| | | OE + MaxLogit | 4.12 ± 0.20 | 98.58 ± 0.07 | 94.47 ± 0.13 | 46.20 ± 3.53 | 90.59 ± 0.87 | 74.02 ± 0.04 |
| | | **MVOL** (ours) | **3.30** ± 0.19 | **98.70** ± 0.05 | 94.68 ± 0.09 | **42.96** ± 0.86 | **90.69** ± 0.26 | 74.29 ± 0.33 |
| nsemble Distillation Model Setting | Post Hoc | MSP | 55.82 ± 2.46 | 89.52 ± 0.53 | 94.36 ± 0.11 | 80.36 ± 0.72 | 74.72 ± 0.79 | 76.99 ± 0.15 |
| | | Energy | 38.23 ± 1.72 | 89.97 ± 0.57 | 94.36 ± 0.11 | 71.97 ± 1.02 | 79.66 ± 0.57 | 76.99 ± 0.15 |
| | | MaxLogit | 38.64 ± 2.00 | 89.94 ± 0.57 | 94.36 ± 0.11 | 72.71 ± 1.08 | 79.46 ± 0.60 | 76.99 ± 0.15 |
| | | ODIN | 40.46 ± 2.20 | 86.94 ± 1.07 | 94.36 ± 0.11 | 67.71 ± 0.61 | 78.71 ± 0.78 | 76.99 ± 0.15 |
| | | Mahalanobis | 23.06 ± 0.64 | 92.94 ± 0.19 | 94.36 ± 0.11 | 42.36 ± 1.93 | 88.39 ± 0.49 | 76.99 ± 0.15 |
| | | KNN | 41.98 ± 2.47 | 92.70 ± 0.55 | 94.36 ± 0.11 | 84.67 ± 2.72 | 73.67 ± 1.78 | 76.99 ± 0.15 |
| | | ASH | 36.50 ± 1.43 | 91.55 ± 0.55 | 94.23 ± 0.09 | 59.19 ± 1.75 | 80.20 ± 0.49 | 76.41 ± 0.26 |
| | Outlier Synthesis | VOS | 30.58 ± 4.34 | 92.23 ± 1.07 | **95.02** ± 0.11 | 72.01 ± 1.81 | 79.86 ± 1.90 | **77.18** ± 0.28 |
| | | ATOL | 28.14 ± 1.79 | 93.60 ± 0.43 | 94.19 ± 0.07 | 74.07 ± 0.98 | 77.82 ± 0.66 | 74.79 ± 0.09 |
| | Outlier Exposure | Energy w/Aux | 4.10 ± 0.24 | 98.07 ± 0.04 | 91.48 ± 0.19 | 52.81 ± 3.52 | 89.14 ± 0.81 | 68.27 ± 0.48 |
| | | OE | 3.95 ± 0.23 | 98.56 ± 0.07 | 94.67 ± 0.23 | 47.04 ± 0.73 | 89.35 ± 0.21 | 75.01 ± 0.13 |
| | | OE + MaxLogit | 3.61 ± 0.24 | **98.62** ± 0.06 | 94.67 ± 0.23 | 46.92 ± 0.75 | **90.79** ± 0.26 | 75.01 ± 0.13 |
| | | **MVOL** (ours) | **3.34** ± 0.20 | 98.61 ± 0.06 | 94.68 ± 0.20 | **36.62** ± 1.36 | 90.37 ± 0.43 | 76.27 ± 0.33 |

# Experiment



| Noise | Method | Single Model Setting FPR95 ↓ | AUROC ↑ | ID-Acc ↑ | Ensemble Distillation Model Setting FPR95 ↓ | AUROC ↑ | ID-Acc ↑ |
|---|---|---|---|---|---|---|---|
| - | MaxLogit | 47.39 ± 2.93 | 89.81 ± 0.55 | 91.38 ± 0.24 | 40.59 ± 2.50 | 90.21 ± 0.71 | 91.94 ± 0.09 |
| $\alpha = 0$ | WOODS | 21.97 ± 1.83 | 96.02 ± 0.32 | 91.20 ± 0.22 | 51.50 ± 3.99 | 84.85 ± 0.99 | 89.79 ± 0.19 |
| | OE + MaxLogit | 18.04 ± 1.34 | **96.57** ± 0.15 | **92.24** ± 0.08 | 18.86 ± 2.10 | 96.12 ± 0.31 | **92.32** ± 0.15 |
| | **MVOL** (ours) | **17.34** ± 2.86 | 96.21 ± 0.30 | 91.71 ± 0.08 | **12.96** ± 0.95 | **96.45** ± 0.14 | 91.96 ± 0.13 |
| $\alpha = 0.05$ | WOODS | 22.04 ± 2.36 | 96.02 ± 0.34 | 91.27 ± 0.16 | 51.49 ± 3.97 | 84.86 ± 0.99 | 89.79 ± 0.19 |
| | OE + MaxLogit | 22.11 ± 1.26 | 95.64 ± 0.26 | 91.29 ± 0.20 | 23.11 ± 3.90 | 95.67 ± 0.48 | 91.51 ± 0.18 |
| | **MVOL** (ours) | **19.55** ± 1.04 | **96.22** ± 0.16 | **91.75** ± 0.23 | **12.87** ± 1.11 | **96.49** ± 0.11 | **91.74** ± 0.30 |
| $\alpha = 0.1$ | WOODS | 22.38 ± 2.30 | 95.98 ± 0.35 | 91.27 ± 0.21 | 51.59 ± 4.03 | 84.85 ± 0.98 | 89.81 ± 0.19 |
| | OE + MaxLogit | 25.49 ± 1.60 | 94.97 ± 0.30 | 90.92 ± 0.31 | 26.90 ± 3.04 | 95.24 ± 0.41 | 91.01 ± 0.22 |
| | **MVOL** (ours) | **18.05** ± 1.58 | **96.16** ± 0.20 | **91.55** ± 0.31 | **13.96** ± 0.80 | **96.07** ± 0.15 | **91.64** ± 0.28 |
| $\alpha = 0.3$ | WOODS | 22.03 ± 2.45 | 95.99 ± 0.40 | 91.24 ± 0.19 | 51.58 ± 4.03 | 84.86 ± 0.99 | 89.80 ± 0.21 |
| | OE + MaxLogit | 37.67 ± 3.85 | 92.64 ± 0.44 | 89.04 ± 0.30 | 51.20 ± 6.17 | 91.83 ± 0.57 | 88.60 ± 0.13 |
| | **MVOL** (ours) | **20.71** ± 1.86 | **95.94** ± 0.32 | **91.24** ± 0.16 | **13.79** ± 0.93 | **96.43** ± 0.16 | **91.76** ± 0.07 |
| $\alpha = 0.5$ | WOODS | **22.31** ± 3.05 | **95.93** ± 0.50 | **91.28** ± 0.12 | 51.49 ± 3.96 | 84.86 ± 0.99 | 89.80 ± 0.20 |
| | OE + MaxLogit | 45.14 ± 2.94 | 90.22 ± 0.63 | 88.04 ± 0.18 | 54.23 ± 3.77 | 89.93 ± 0.21 | 86.80 ± 0.21 |
| | **MVOL** (ours) | 25.53 ± 1.23 | 95.23 ± 0.18 | 90.79 ± 0.21 | **14.85** ± 1.86 | **96.44** ± 0.32 | **91.81** ± 0.19 |

# Thanks