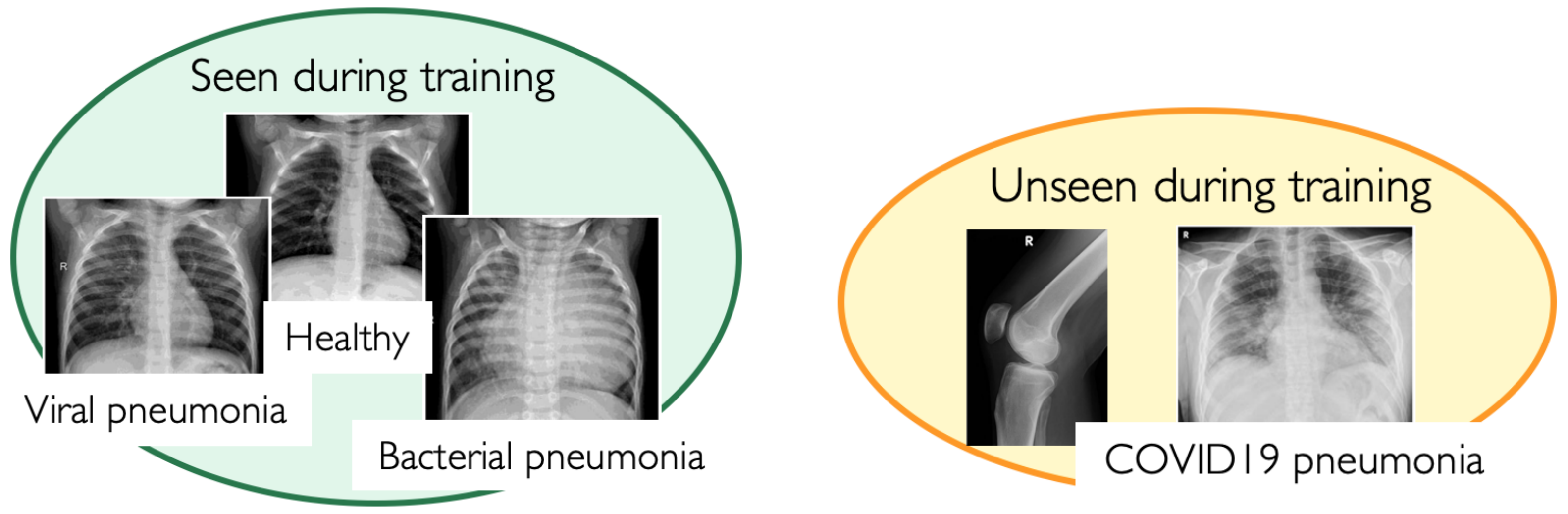
Novel disease detection using ensembles with regularized disagreement

Alexandru Țifrea, Eric Stavarache, Fanny Yang
Department of Computer Science, ETH Zurich

Novel classes as OOD data

Problem: Classifier predictions are incorrect on novel classes.
→ Flag data from unseen classes as out-of-distribution (OOD).



→ Novel classes are often similar to in-distribution (ID) classes
⇒ difficult to distinguish ID and OOD data.

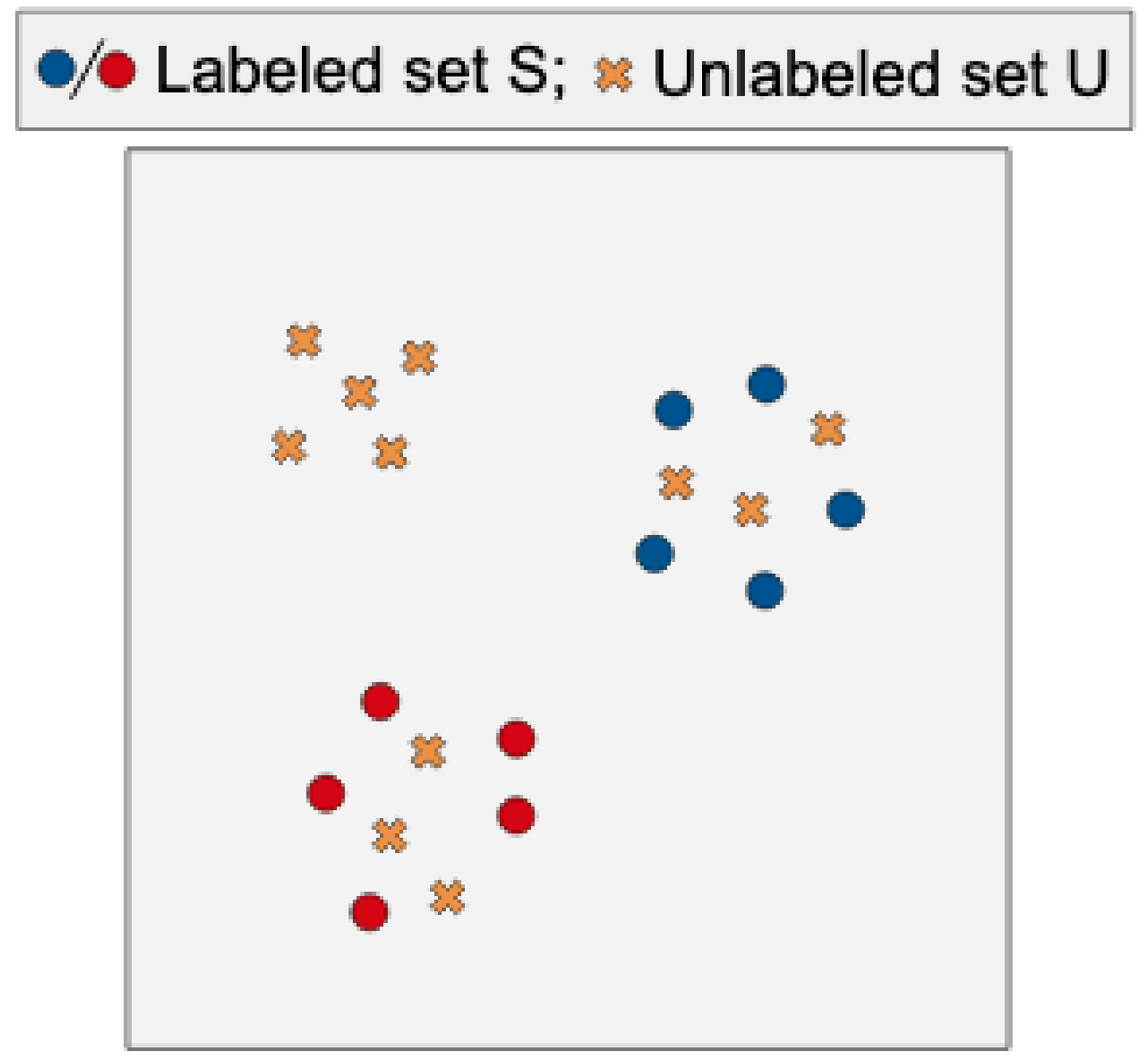
Existing OOD detection methods (assuming different access to OOD data) **perform poorly on novel-class detection**.

Our setting

Available data:

- Labeled set with ID samples.
  → e.g. the training set for the prediction task.
- Unlabeled set with unknown mixture of ID and OOD data.
  → e.g. hospital collects all X-rays performed during the day.

Unknown OOD setting:



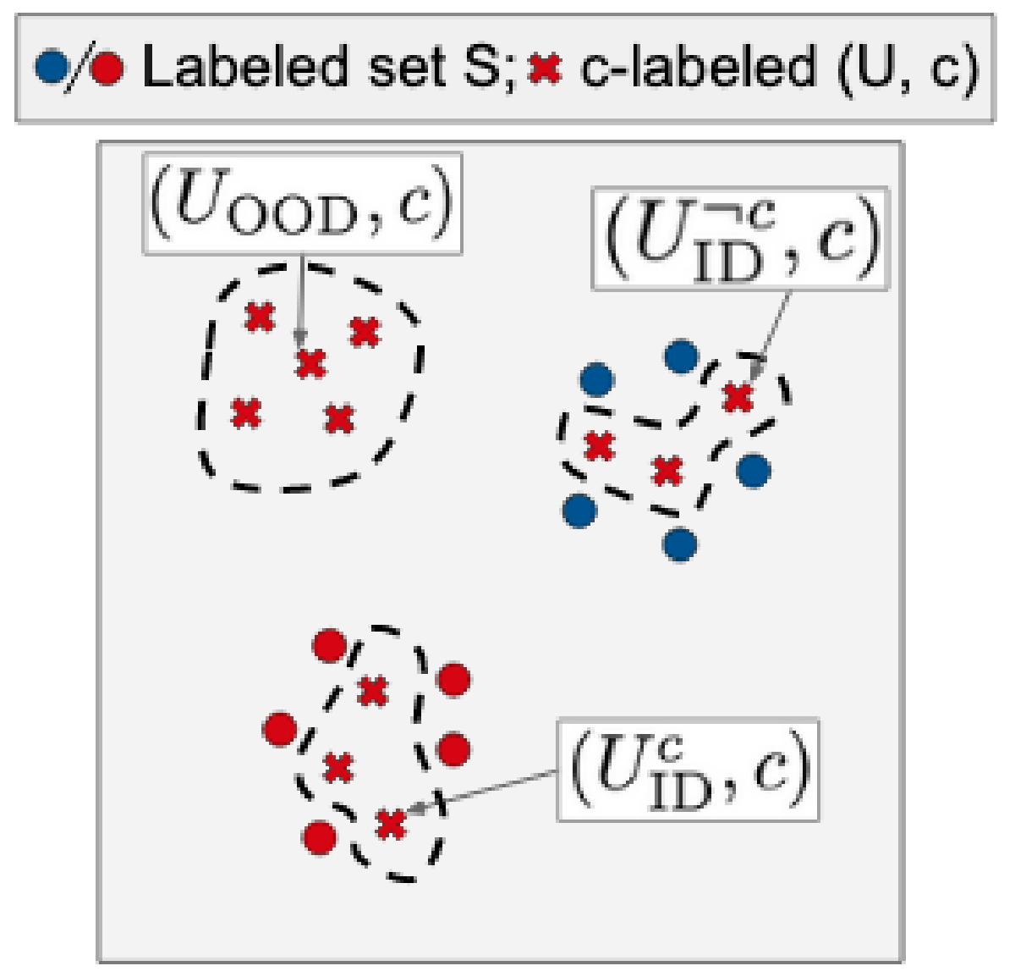
Unknown OOD setting

Previous methods that employ the Unknown OOD setting (e.g. nnPU, MCD) **fail to leverage unlabeled data effectively**.

Our approach

Idea: Train an **Ensemble w/ Regularized Disagreement**.

```
Algorithm 1: Fine-tuning the ERD ensemble
Input : Train set $S$, Validation set $V$, Unlabeled set $U$,
        Weights $W$ pretrained on $S$, Ensemble size $K$
Result: ERD ensemble $\{f_{y_i}\}_{i=1}^K$
Sample $K$ different labels $\{y_1, ..., y_K\}$ from $\mathcal{Y}$
for $c \leftarrow \{y_1, ..., y_K\}$ do // fine-tune $K$ models
    $f_c \leftarrow$ Initialize$(W)$
    $(U, c) \leftarrow \{(x, c) : x \in U\}$
    $f_c \leftarrow$ FinetuneWithEarlyStopping $(f_c, S \cup (U, c); V$
return $\{f_{y_i}\}_{i=1}^K$
```


●/● Labeled set S; ✖ c-labeled (U, c)
(U_{OOD}, c)
$(U_{\text{ID}}^{\neg c}, c)$
(U_{ID}^{c}, c)

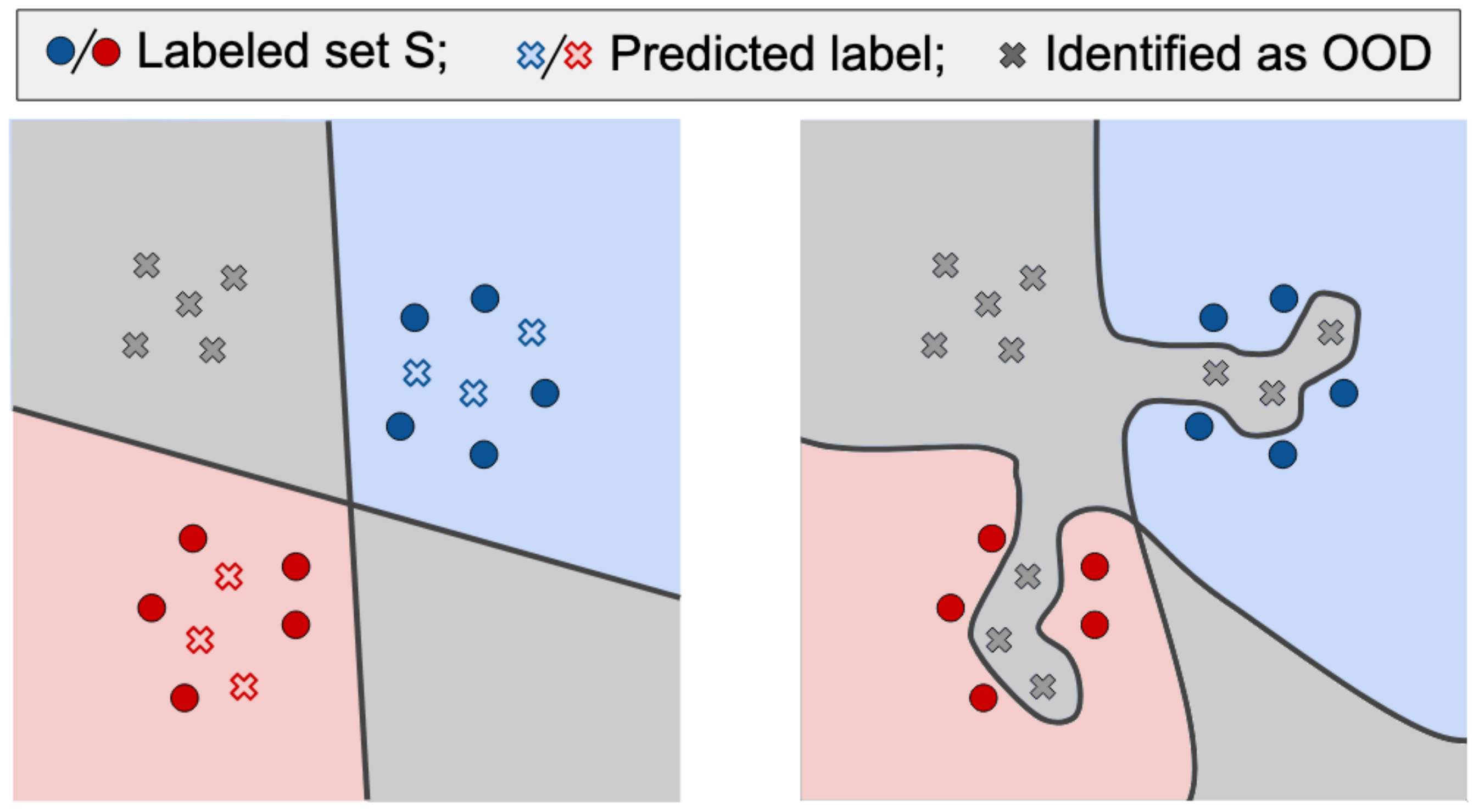
Subsets of the c-labeled U

At test time:

- For a test sample x, use outputs $f_1(x), ..., f_k(x)$ to compute the **average pairwise disagreement score** (details later).
  → Flag as OOD samples with score larger than threshold τ.

Key ingredients

1) Regularization: Prevent complex models from interpolating on $S \cup (U, c)$.
→ We **early stop** at epoch with highest ID validation accuracy.



ERD - good func. complexity

ERD - large func. complexity

2) Average pairwise disagreement score:

$$(\text{Avg} \circ \rho)(f_1(x), ..., f_K(x)) := \frac{2}{K(K-1)} \sum_{i \neq j} \rho\left(f_i(x), f_j(x)\right)$$

→ e.g. ρ = total variation distance

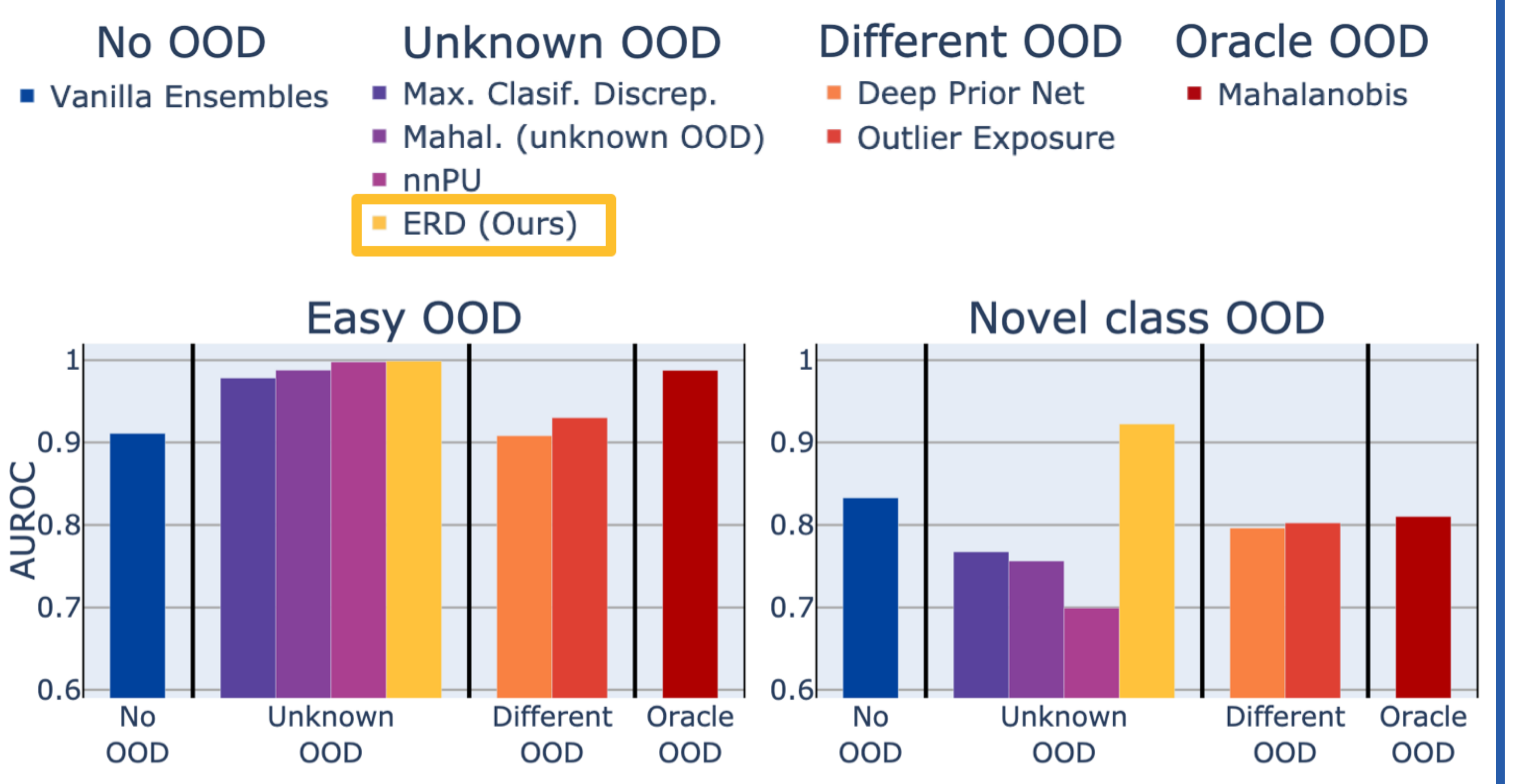
- Unlike prior OOD metrics (e.g. entropy of average predictor), our score exploits ensemble diversity.

Experiments

Evaluation metric: Area under the ROC curve (AUROC).
→ higher is better.
→ TP = correctly identified OOD; FP = ID flagged as OOD.

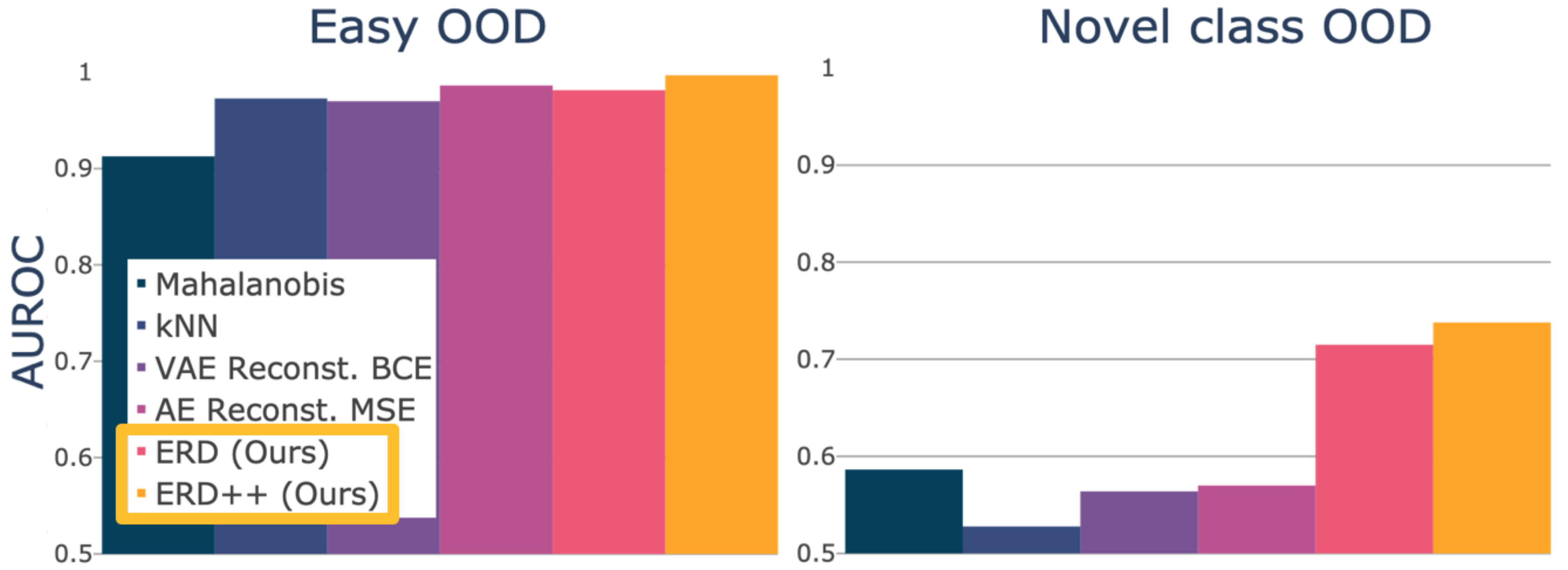
1) Natural images

Easy OOD: SVHN vs CIFAR10, CIFAR10 vs SVHN etc
Novel class OOD: CIFAR100[0-49] vs CIFAR100[50-99] etc



2) Medical images

Easy OOD: Chest X-ray vs Knee X-ray etc
Novel class OOD: Novel disease as OOD



→ **Image modalities:** Frontal and lateral chest X-rays and retinal images.

→ **ERD++:** our approach trained from random initializations instead of pretrained weights.