LoCoOp:
Few-Shot Out-of-Distribution Detection via Prompt Learning

Atsuyuki Miyai, Qing Yu, Go Irie, Kiyoharu Aizawa (NeurIPS 2023)

POSTECH EffL Lab
Presented by Jiyun Bae
2024. 03. 14.

Contents

1. Motivation
2. Methodology
3. Experiment
4. Analysis

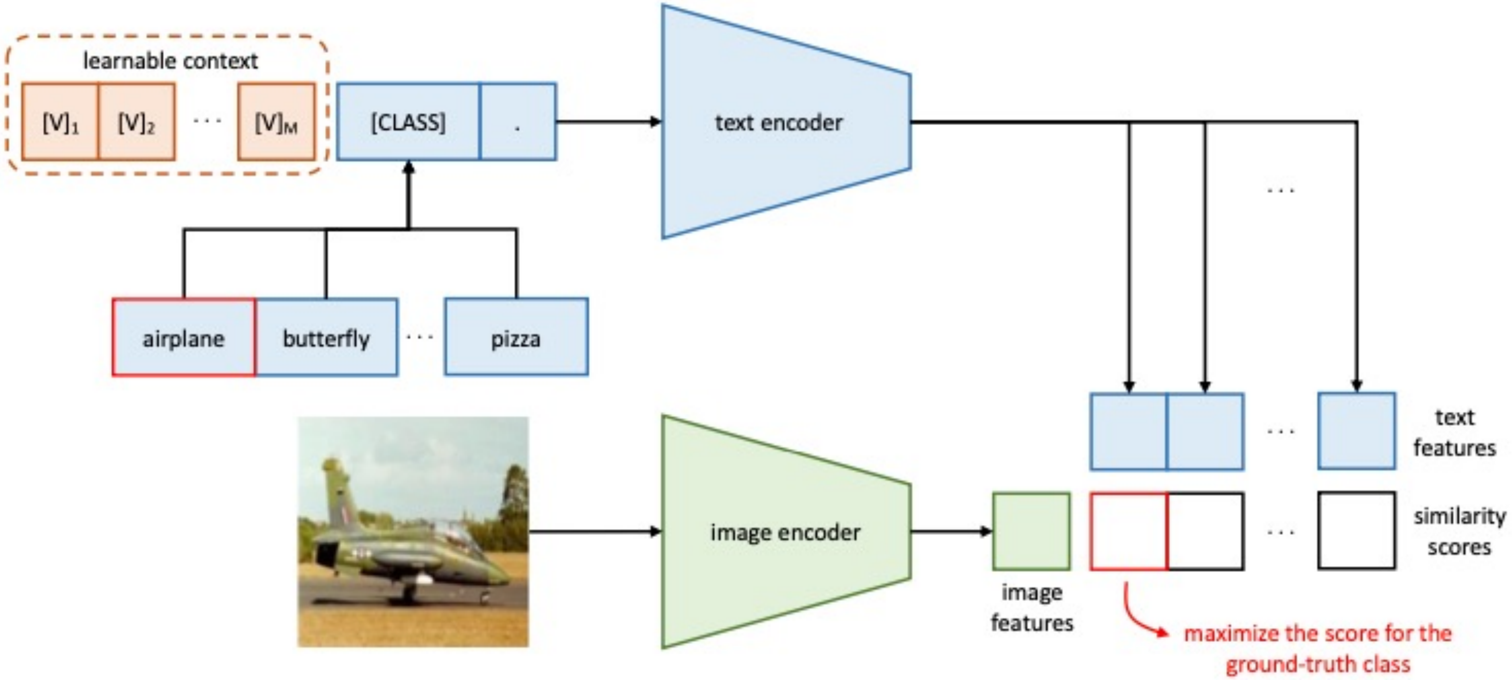
Motivation

- Detecting out-of-distribution (OOD) samples is crucial in real-world scenarios, where new classes of samples can naturally arise and require caution
- Previous studies on CLIP-based OOD detection
  - Zero-shot methods: encountering a domain gap with ID downstream data
  - Fully supervised methods: may destroying the rich representations of CLIP, also requiring enormous training costs

→ Few-shot OOD detection method: utilizing a few ID training images for OOD detection

Motivation

- CoOp has limitations in OOD detection due to the potential presence of ID-irrelevant information in text embeddings



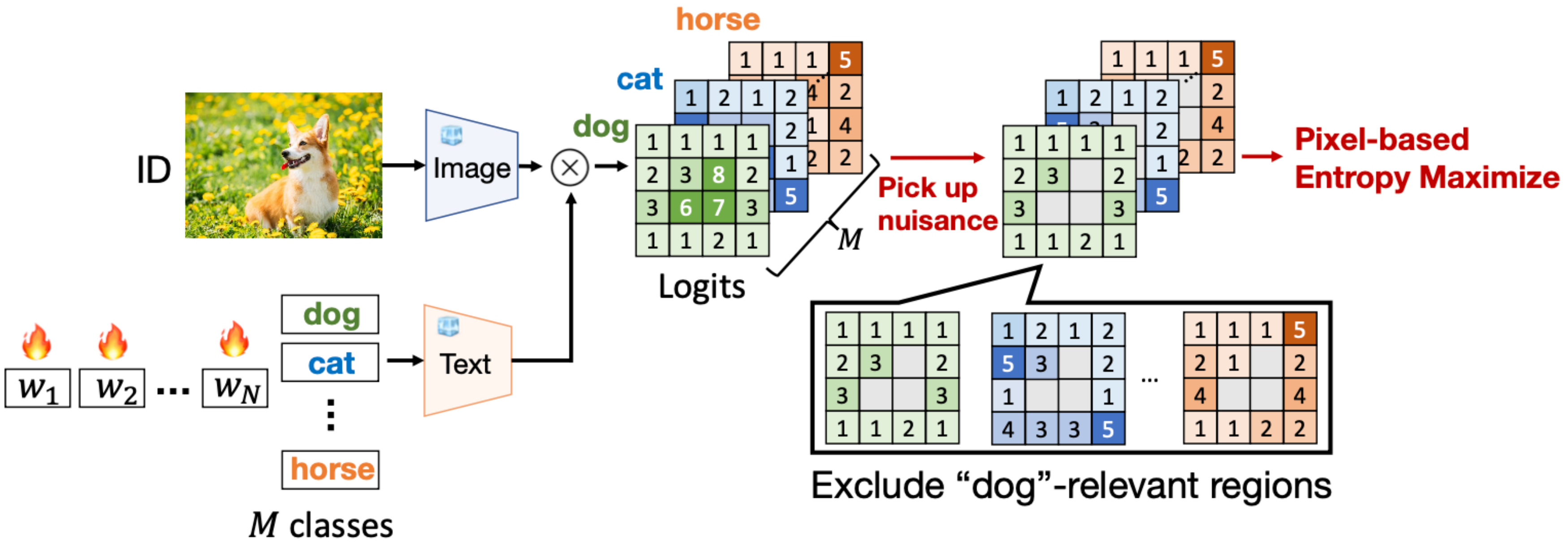
→ The text embeddings, which may contain ID-irrelevant information, result in incorrectly high confidence scores for OOD images

Methodology

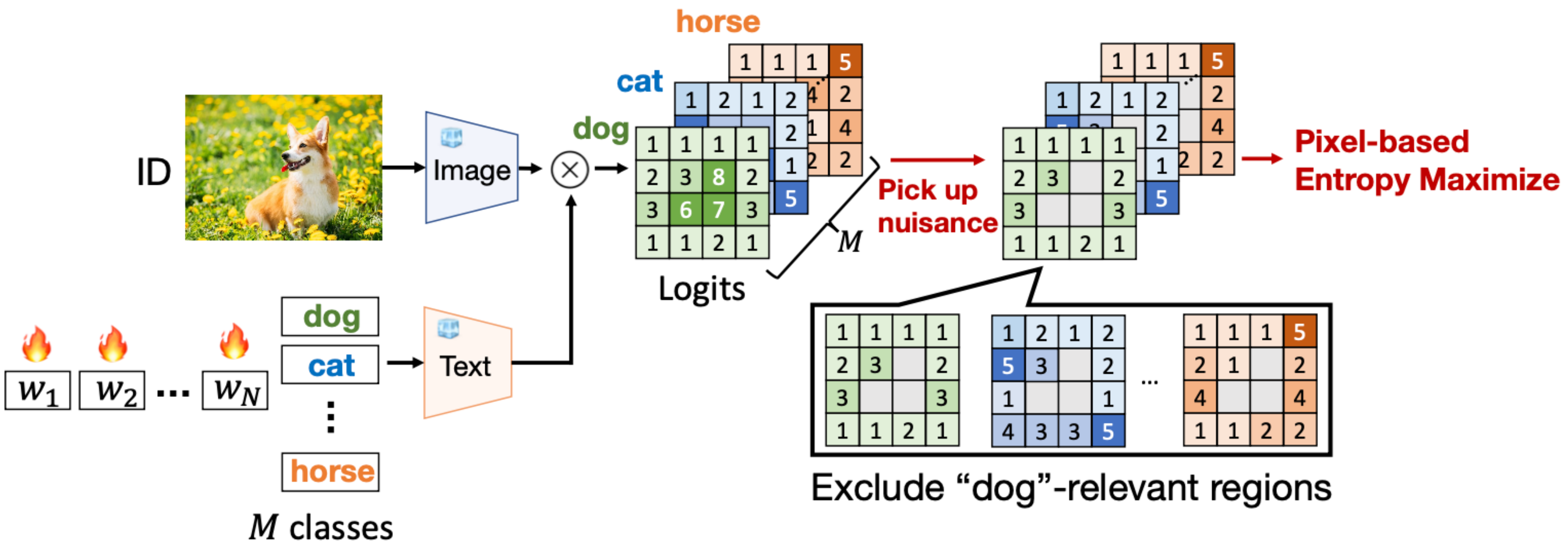
- **LoCoOp** (Local regularized Context Optimization)
  - Treating ID-irrelevant nuisances as OOD
  - Learning to push ID-irrelevant nuisances away from the ID class text embeddings

→ Removing unnecessary information from the text embeddings of ID classes

(= preventing the model from producing high ID confidence scores for the OOD features)



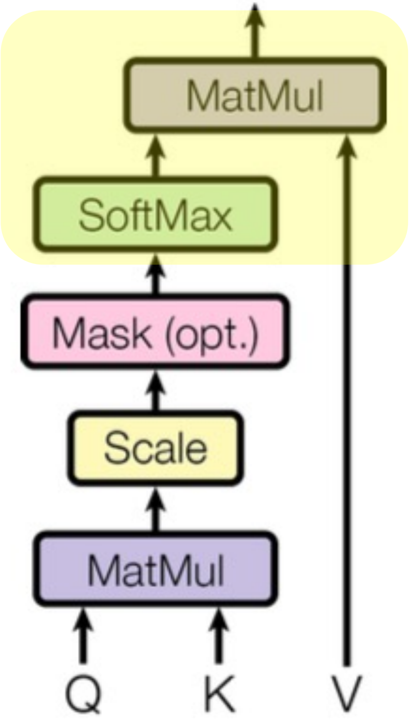
Methodology



- **Three Questions will arise** 🤔
  - How to obtain local features from CLIP?
  - How to extract object-irrelevant regions from ID images?
  - How to use OOD regularization loss during training?

Methodology:

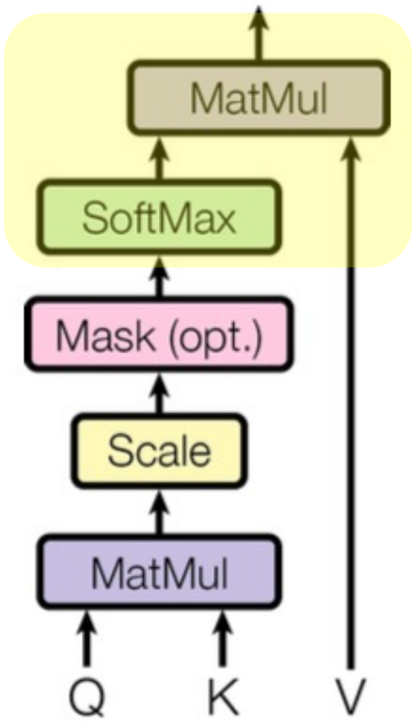
How to obtain local features from CLIP



$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Methodology:

How to obtain local features from CLIP



$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Global average pooling to x

Global feature: $$\begin{aligned}\mathrm{AttnPool}(\mathbf{x}) &= \mathrm{Proj}_{v\to t}\left(\sum_i \mathrm{softmax}\left(\frac{q(\bar{\mathbf{x}})k\left(\mathbf{x}_i\right)^T}{D}\right)\cdot v\left(\mathbf{x}_i\right)\right) \\ &= \sum_i \mathrm{softmax}\left(\frac{q(\bar{\mathbf{x}})k\left(\mathbf{x}_i\right)^T}{D}\right)\cdot \mathrm{Proj}_{v\to t}(v\left(\mathbf{x}_i\right)) \\ &= \mathrm{Pool}(\mathrm{Proj}_{v\to t}(v\left(\mathbf{x}_i\right))),\end{aligned}$$

Visual feature of each region i

Local feature

Methodology:
How to extract object-irrelevant regions from ID images

- The classification prediction probabilities for each region i

$$p_i(y = m \mid \boldsymbol{x}^{\text{in}}) = \frac{\exp\left(\text{sim}\left(\boldsymbol{f}_i^{\text{in}}, \boldsymbol{g}_m\right)/\tau\right)}{\sum_{m'=1}^{M} \exp\left(\text{sim}\left(\boldsymbol{f}_i^{\text{in}}, \boldsymbol{g}_{m'}\right)/\tau\right)}.$$

$$\downarrow$$

- Identifying ID-irrelevant regions j

$$J = \{i \in I : \text{rank}(p_i(y = y^{\text{in}} | \boldsymbol{x}^{\text{in}})) > K\}$$

Methodology:
How to extract object-irrelevant regions from ID images

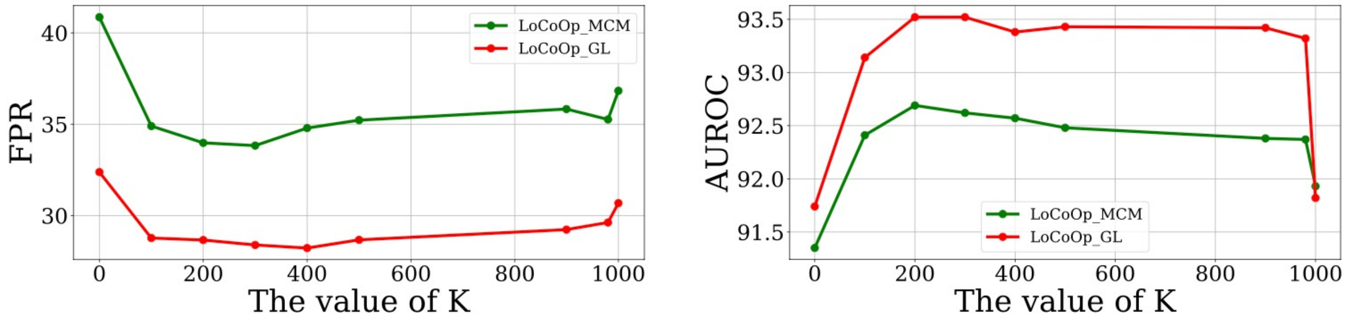
- The classification prediction probabilities for each region i

$$p_i(y=m \mid \boldsymbol{x}^{\text{in}}) = \frac{\exp\left(\operatorname{sim}\left(\boldsymbol{f}_i^{\text{in}}, \boldsymbol{g}_m\right)/\tau\right)}{\sum_{m'=1}^{M}\exp\left(\operatorname{sim}\left(\boldsymbol{f}_i^{\text{in}}, \boldsymbol{g}_{m'}\right)/\tau\right)}.$$

Top- K prediction;
if a region i is ID-irrelevant, y^{in} should not appear among the top-K prediction

- Identifying ID-irrelevant regions j

$$J = \{i \in I : \operatorname{rank}(p_i(y = y^{\text{in}} | \boldsymbol{x}^{\text{in}})) > K\}$$



(a) FPR95 (b) AUROC

Methodology:
How to use OOD regularization loss during training

- Entropy maximization
  - Making the entropy of $p_j(y|x^{in})$ larger and enables the ODD image features f_j^{in} to be dissimilar to any ID text embedding

- The loss function for this regularization: $\mathcal{L}_{\text{ood}} = -H(p_j)$

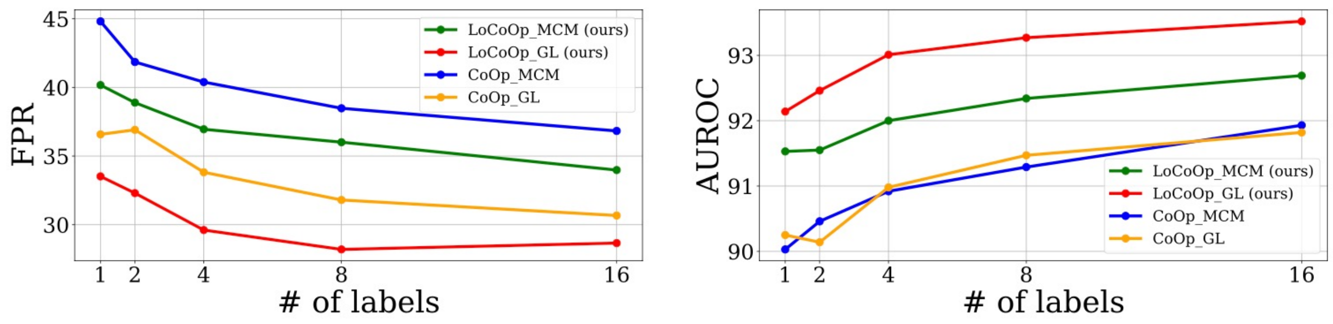
- Final objective: $\mathcal{L} = \mathcal{L}_{\text{coop}} + \lambda\mathcal{L}_{\text{ood}}$

Experiment:
setup

- Baselines
  - Baseline prompt learning methods: CoOp
  - Zero-shot detection methods: MCM, GL-MCM
  - Fully-supervised detection methods: NPOS, ODIN, ViM, KNN
- Evaluation Metrics
  - FPR95
  - AUROC
- Few-shot training
- Datasets
  - ID data: ImageNet-1K dataset
  - OOD datasets: iNaturalist, SUN, Places, TEXTURE

Experiment:
Main Result

- Test-time OOD detection



(a) FPR95 (b) AUROC

Figure 2: **Few-shot OOD detection results** with different numbers of ID labeled samples. We report average FPR95 and AUROC scores on four OOD datasets in Table 1. The lower value is better for FPR95, and the larger value is better for AUROC. We find that in all settings, our proposed LoCoOp with GL-MCM (red one) outperforms CoOp by a large margin.

- MCM score: utilizing the softmax score of global image features and text features
- GL-MCM score: utilizing the softmax score of both global and local image features and text features

Experiment:
Main Result

Table 1: **Comparison results on ImageNet OOD benchmarks.** We use ImageNet-1K as ID. We use CLIP-B/16 as a backbone. Bold values represent the highest performance. † is cited from [46]. * is our reproduction. We find that LoCoOp with GL-MCM (LoCoOp$_{GL}$) is the most effective method.

Method	iNaturalist		SUN		Places		Texture		**Average**	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Zero-shot										
MCM [30]*	30.94	94.61	37.67	92.56	44.76	89.76	57.91	86.10	42.82	90.76
GL-MCM [33]*	15.18	96.71	30.42	93.09	38.85	89.90	57.93	83.63	35.47	90.83
Fine-tuned										
ODIN [28]†	30.22	94.65	54.04	87.17	55.06	85.54	51.67	87.85	47.75	88.80
ViM [50]†	32.19	93.16	54.01	87.19	60.67	83.75	53.94	87.18	50.20	87.82
KNN [45]†	29.17	94.52	35.62	92.67	39.61	91.02	64.35	85.67	42.19	90.97
NPOS$_{MCM}$ [46]†	16.58	96.19	43.77	90.44	45.27	89.44	46.12	88.80	37.93	91.22
NPOS$_{MCM}$ [46]*	19.59	95.68	48.26	89.70	49.82	88.77	51.12	87.58	42.20	90.43
NPOS$_{GL}$*	18.70	95.36	38.99	90.33	41.86	89.36	47.89	86.44	36.86	90.37
Prompt learning				*one-shot (one label per class)*						
CoOp$_{MCM}$	43.38	91.26	38.53	91.95	46.68	89.09	50.64	87.83	44.81	90.03
CoOp$_{GL}$	21.30	95.27	31.66	92.16	40.44	89.31	52.93	84.25	36.58	90.25
LoCoOp$_{MCM}$ (ours)	38.49	92.49	33.27	93.67	39.23	91.07	49.25	89.13	40.17	91.53
LoCoOp$_{GL}$ (ours)	24.61	94.89	25.62	94.59	34.00	**92.12**	49.86	87.49	33.52	92.14
				16-shot (16 labels per class)						
CoOp$_{MCM}$	28.00	94.43	36.95	92.29	43.03	89.74	**39.33**	91.24	36.83	91.93
CoOp$_{GL}$	**14.60**	96.62	28.48	92.65	36.49	89.98	43.13	88.03	30.67	91.82
LoCoOp$_{MCM}$ (ours)	23.06	95.45	32.70	93.35	39.92	90.64	40.23	**91.32**	33.98	92.69
LoCoOp$_{GL}$ (ours)	16.05	**96.86**	**23.44**	**95.07**	**32.87**	91.98	42.28	90.19	**28.66**	**93.52**

Experiment:
Visualization of extracted OOD regions

- The performance of OOD extraction is key to LoCoOp method
- Rank-based approach can accurately identify OOD regions

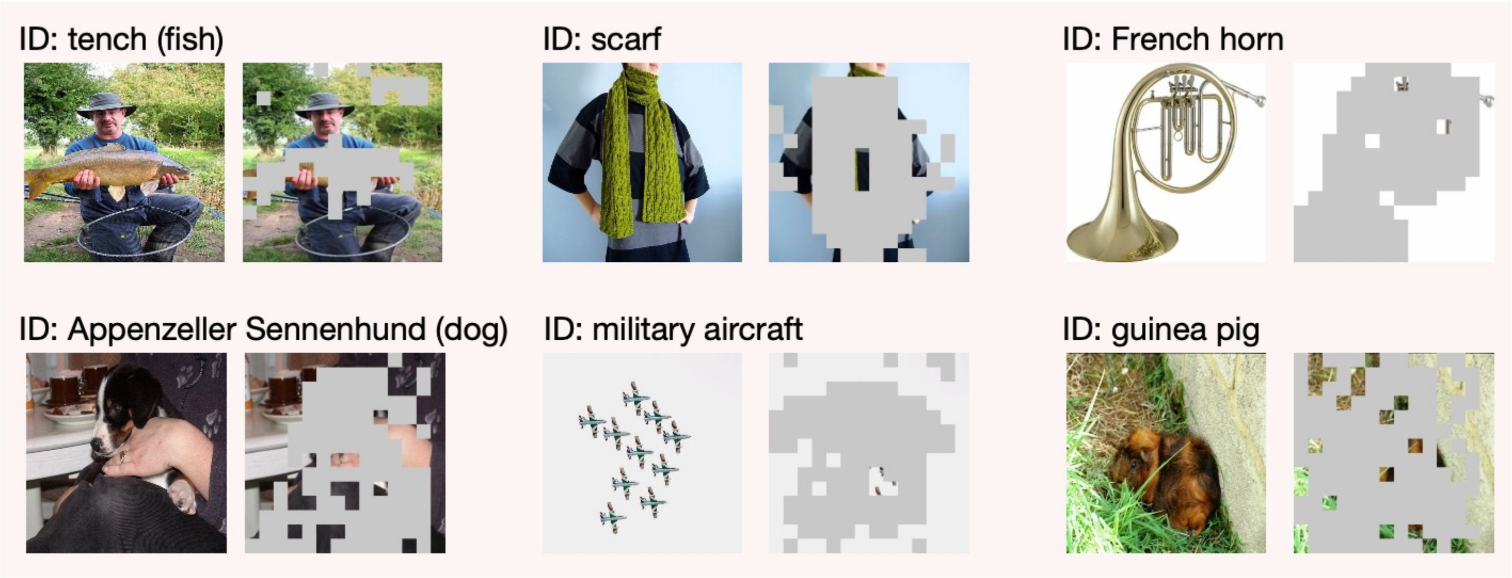
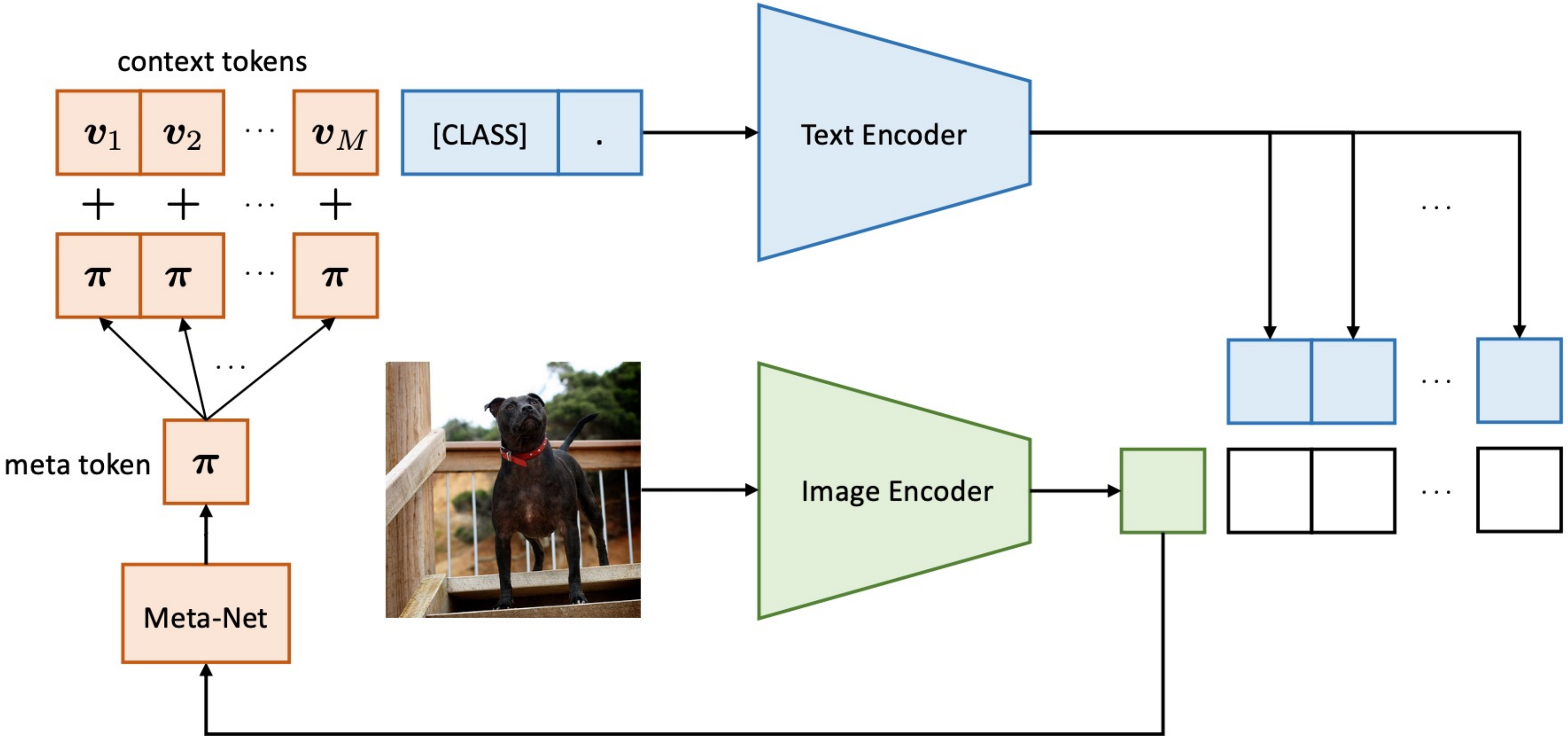


Figure 4: **Visualization of extracted OOD regions.** We find that our approach can correctly extract ID-irrelevant regions.

Analysis:
Comparison with CoCoOp



- Generating an input-conditional text token for each image

Analysis

- Comparison with CoCoOp

Table 2: Comparison results with CoCoOp [61]. We report average FPR and AUROC scores on four OOD datasets in Table 1.

Method	Infer time↓	Average FPR95↓	Average AUROC↑
$\text{CoCoOp}_{\text{MCM}}$ [61]	149 ms	35.53	91.99
$\text{LoCoOp}_{\text{MCM}}$	**2.59 ms**	33.98	92.69
$\text{LoCoOp}_{\text{GL}}$	5.97 ms	**28.66**	**93.52**

- ID accuracy of LoCoOp and CoOp

Table 3: Comparison in ID accuracy on ImageNet-1K validation data.

Method	Top-1 Accuracy
CoOp	**72.1**
LoCoOp	71.7