Detecting Out-of-Distribution Objects through Class-Conditioned Inpainting Supplementary Material

A. Data processing

Following [3, 4], we use PascalVOC [5] with 20 categories[1] and BDD100k [18] with 10 categories[2]. For OOD data, we sample from MS-COCO [8] and OpenImages [6], ensuring no overlap by removing all images containing ID categories. Additionally, to improve alignment with CLIP's visual-textual representations, we standardize certain British labels (e.g., "Aeroplane", "Couch") to their US counterparts ("Airplane", "Sofa").

As our method is zero-shot and requires no training, all experiments are performed on subsets drawn from the test sets of VOC, BDD100k, COCO, and OpenImages. To ensure fairness, thoroughness, and a balanced ID–OOD distribution, we sample 200 images from the BDD100k test set and 400 images each from PascalVOC, MS-COCO, and OpenImages. The smaller sample size for BDD100k accounts for its higher object density per image.

B. Implementation Details for Baselines

For discriminative zero-shot baselines, we adapt MCM [12], CLIPN [16], TAG [10], OLE [2], and GL-MCM [13]—originally designed for image-level OOD detection—to the object level by cropping each detection and treating it as an individual image. Each baseline is then applied directly on top of the object detector's predictions using its default setting, configuration, or conditioning prompt, without any specific modification. ODIN [7] and Energy Score [9], developed for non-zero-shot classifier-based settings, are adapted similarly by using CLIP for zero-shot classification of detection crops. For generative baselines, we synthesize objects using Stable Diffusion 2, extract features via SimCLRv2 [1], and apply standard OOD detectors such as Mahalanobis [17] and KNN [15] for fair comparison. For SIREN [3] and VOS [4], we follow their original protocols using features from their trained detectors.

C. Additional Ablation Studies

C.1. Impact of Masking Ratios

Extending the ablation study in Tab. 5, we study the impact of the masking ratio on OOD detection by varying the inpainting mask size, covering from 25% to 100% of the predicted bounding box. Quantitative and qualitative results in Tab. 1 and Fig. 1 suggest that despite being robust across masking ratios, RONIN achieves optimal performance when 80–90% of the object is masked for inpainting. As shown in Fig. 1, low masking ratios preserve strong original objects' appearances in both ID and OOD cases, leading to high similarity scores. In contrast, too large a masking ratio leads to excessive deviation and ambiguity in both cases, leading to poor similarity scores. Covering 75–90% strikes a balance, as the inpainting outcomes not only retain similar inpainting for ID samples but also produce vivid dissimilar inpaintings for OOD ones, enabling RONIN to distinguish ID and OOD objects effectively.

C.2. Robust across Generative Models

Extending the ablation study in Tab. 6 to the many-step regime, we further evaluate RONIN by replacing the default Stable Diffusion 2 with Kandinsky 2.1 [14] and Diffusion Model 1, both using 20 denoising steps. We also include results from the few-step regime study, with the one-step diffusion model InstaFlow [11] and 5-step Stable Diffusion 2 for comparison. As shown in Tab. 2, RONIN remains consistently robust across different diffusion models and configurations. Notably, Stable Diffusion 2 provides the best trade-off between performance and runtime, even in the few-step generative mode. These

[1] **PascalVOC ID labels**: Person, Car, Bicycle, Boat, Bus, Motorbike, Train, Aeroplane, Chair, Bottle, Dining Table, Potted Plant, TV Monitor, Couch, Bird, Cat, Cow, Dog, Horse, Sheep. [2] **BDD100k ID labels**: Pedestrian, Rider, Car, Truck, Bus, Train, Motorcycle, Bicycle, Traffic Light, Traffic Sign.

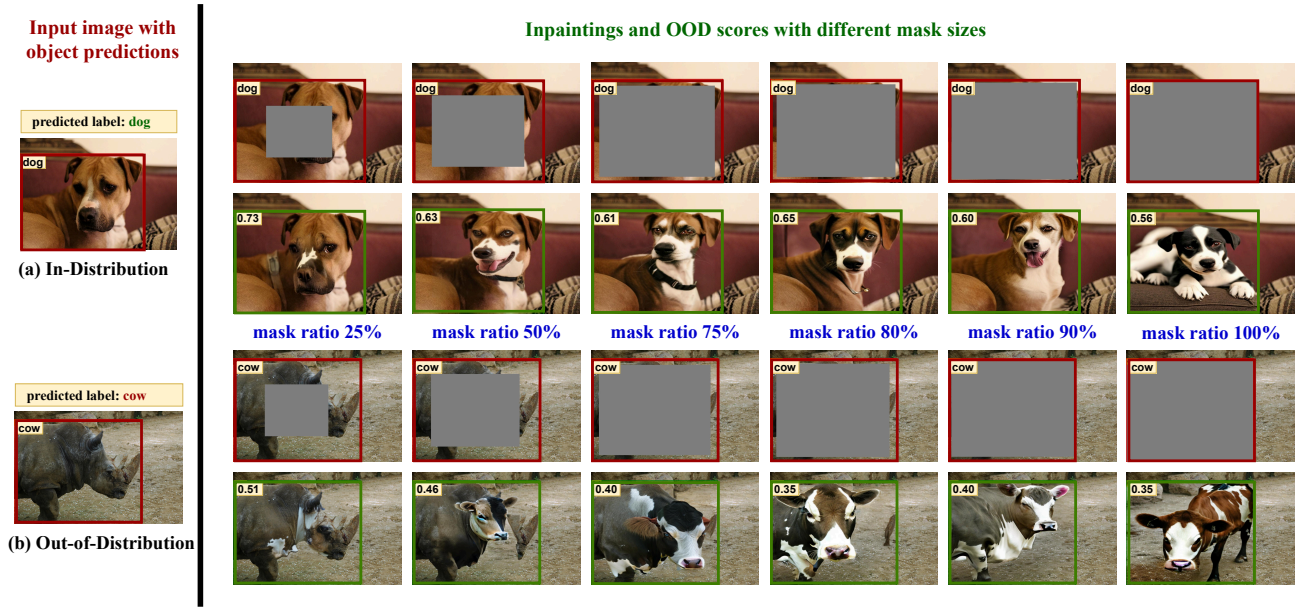

Figure 1. **RONIN inpainting performance across masking ratios**. Small masks preserve object structure, yielding high similarity for both ID and OOD. Large masks cause excessive distortion, lowering similarity in both cases. A balanced 80–90% ratio retains ID integrity while amplifying OOD deviation, enabling clear ID-OOD separation.

Table 1. **Different OOD detection performances when varying mask size for inapinting**. RONIN maintains consistent and robust performance across different masking ratios, with the best performance attained when 80%-90% of each object is masked.

Covered masking ratio m	MS-COCO		OpenImages	
	FPR@95 (↓)	AUROC (↑)	FPR@95 (↓)	AUROC (↑)
25%	42.47	88.53	36.74	90.40
50%	34.23	89.39	23.91	92.45
75%	30.10	90.43	21.74	92.41
80%	25.80	91.32	17.74	93.84
90%	25.57	91.53	19.57	92.46
100%	29.68	91.61	20.29	92.61

Table 2. **RONIN performances across different diffusion models.** The consistently robust performances across inpainting models, including one-step inpainting, demonstrate RONIN effectiveness without any dependence on specific models or inpainting quality.

		Denoising Process (steps)	COCO / OpenImages		Avg runtime/image (seconds)
			FPR@95 (↓)	AUROC (↑)	
Many-step Regime	Kandinsky 2.1 [3]	20	28.99 / 23.96	89.92 / 91.62	1.12
	Stable Diffusion 1.5 [4]	20	29.90 / 23.70	91.36 / 92.36	1.20
	Stable Diffusion 2 (***default choice***)	20	25.84 / 18.91	92.28 / 93.30	0.83
	Stable Diffusion 3 [5]	20	21.09 / 18.89	94.01 / 94.43	2.86
Few-step Regime	Stable Diffusion 2	5	27.94 / 20.00	92.35 / 92.32	0.19
	InstaFlow	1	31.34 / 25.74	88.03 / 90.35	0.03

findings suggest that by subtly leveraging semantic misalignment between original and synthesized objects for zero-shot OOD detection, RONIN performs stably and reliably without relying on any specific diffusion model or high-quality inpainted outputs. Even when models produce diverse inpainting results, our framework remains both robust and effective.

References

[1] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020. 1

[2] Choubo Ding and Guansong Pang. Zero-shot out-of-distribution detection with outlier label exposure. In *2024 International Joint Conference on Neural Networks*, 2024. 1

[3] Xuefeng Du, Gabriel Gozum, Yifei Ming, and Yixuan Li. SIREN: Shaping Representations for Detecting Out-of-distribution Objects. *Advances in Neural Information Processing Systems*, 35:20434–20449, 2022. 1

[4] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. VOS: Learning What You Don't Know by Virtual Outlier Synthesis. *Proceedings of the International Conference on Learning Representations*, 2022. 1

[5] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International journal of computer vision*, 88:303–338, 2010. 1

[6] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 1

[7] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *International Conference on Learning Representations*, 2018. 1

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[9] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020. 1

[10] Xixi Liu and Christopher Zach. Tag: Text prompt augmentation for zero-shot out-of-distribution detection. In *European Conference on Computer Vision*, pages 364–380, 2024. 1

[11] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *International Conference on Learning Representations*, 2024. 1

[12] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into Out-of-distribution Detection with Vision-language Representations. *Advances in Neural Information Processing Systems*, 35:35087–35102, 2022. 1

[13] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Gl-mcm: Global and local maximum concept matching for zero-shot out-of-distribution detection. *International Journal of Computer Vision*, pages 1–11, 2025. 1

[14] Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. *arXiv preprint arXiv:2310.03502*, 2023. 1

[15] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-Distribution Detection with Deep Nearest Neighbors. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 20827–20840. PMLR, 17–23 Jul 2022. 1

[16] Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. CLIPN for zero-shot OOD detection: Teaching CLIP to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1802–1812, 2023. 1

[17] Zhisheng Xiao, Qing Yan, and Yali Amit. Do we really need to learn representations from in-domain data for outlier detection? *arXiv preprint arXiv:2105.09270*, 2021. 1

[18] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 1