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

Quang-Huy Nguyen[1*] Jin Peng Zhou[2*] Zhenzhen Liu[2*]
Khanh-Huyen Bui[3] Kilian Q. Weinberger[2] Wei-Lun Chao[1] Dung D. Le[4]
nguyen.2959@osu.edu, jz563@cornell.edu, zl535@cornell.edu
huyenbk2@fpt.com, kilian@cornell.edu, chao.209@osu.edu, dung.ld@vinuni.edu.vn
[1]The Ohio State University, Columbus, Ohio, USA [2]Cornell University, Ithaca, New York, USA
[3]FPT Software AI Center, Hanoi, Vietnam [4]VinUniversity, Hanoi, Vietnam
[*]Equal contribution

Abstract

Recent object detectors have achieved impressive accuracy in identifying objects seen during training. However, real-world deployment often introduces novel and unexpected objects, referred to as out-of-distribution (OOD) objects, posing significant challenges to model trustworthiness. Modern object detectors are typically overconfident, making it unreliable to use their predictions alone *for OOD detection. To address this, we propose leveraging an* auxiliary *model as a complementary solution. Specifically, we utilize an off-the-shelf text-to-image generative model, such as Stable Diffusion, which is trained with objective functions distinct from those of discriminative object detectors. We hypothesize that this fundamental difference enables the detection of OOD objects by measuring inconsistencies between the models. Concretely, for a given detected object bounding box and its predicted in-distribution class label, we perform class-conditioned inpainting on the image with the object removed. If the object is OOD, the inpainted image is likely to deviate significantly from the original, making the reconstruction error a robust indicator of OOD status. Extensive experiments demonstrate that our approach consistently surpasses existing zero-shot and non-zero-shot OOD detection methods, establishing a robust framework for enhancing object detection systems in dynamic environments. Our implementation is available at* https://github.com/quanghuy0497/RONIN.

1. Introduction

Object detection systems have achieved remarkable progress in recent years, becoming integral to a wide range of applications in both online and offline settings. These include, but are not limited to, environmental monitoring [2], industrial manufacturing [1], and healthcare [39]. As these systems increasingly influence decision-making processes, ensuring their reliability and robustness has become a key challenge.

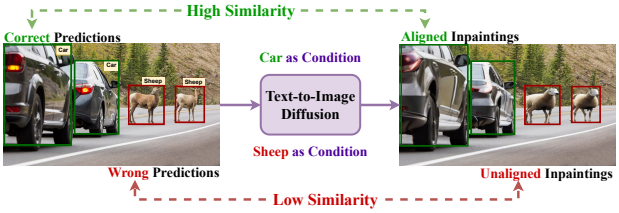

Figure 1. **Intuition behind RONIN for OOD detection**. Object detectors can overconfidently make wrong predictions on unseen objects based on a pre-defined set of ID labels; in this case, predicting two OOD *deers* as "*sheep*". RONIN leverages the label predictions to condition the resynthesizing process of an off-the-shelf text-to-image diffusion model, producing similar inpaintings for correct predictions and dissimilar inpaintings for incorrect ones. Under a similarity measurement, RONIN can identify the wrong prediction of unseen objects, therefore ODD detection.

One common failure mode arises when object detectors encounter object categories not seen during training, referred to as out-of-distribution (OOD) cases. Such occurrences are inevitable due to the dynamic and ever-changing nature of the real world. Unfortunately, modern object detectors often exhibit overconfidence, frequently misclassifying OOD objects as belonging to in-distribution (ID) categories, thereby resulting in false positives [37]. This overconfidence underscores a fundamental limitation: the internal responses of object detectors cannot reliably identify these errors. Consequently, alternative strategies are needed to enhance the trustworthiness of these systems.

A popular strategy is to "rehearse" these error cases during detector training. For example, [36, 37] introduced auxiliary losses to regularize model confidence, while [7, 9] proposed methods to learn more distinguishable detector features between ID and OOD categories. While promising, these strategies require retraining and are not widely imple-

mented in pre-trained, publicly accessible object detectors. Given that many stakeholders rely on off-the-shelf detectors for their applications, there is a strong demand for effective *post-hoc* solutions to address OOD detection challenges.

In this paper, we propose leveraging an additional pre-trained model to provide auxiliary information for effective OOD detection. Specifically, we utilize off-the-shelf text-conditioned generative models [42, 43], which are trained to synthesize realistic images from text prompts, such as category names. We hypothesize that the distinct training objectives of object detectors and generative models lead to inconsistent responses for OOD cases. These inconsistencies thus serve as a robust post-hoc indicator for OOD detection, offering a practical and accessible solution to enhance the reliability of object detection systems without requiring retraining.

More specifically, given a detected object bounding box and its predicted ID class label by the object detector, we propose performing class-conditioned inpainting for OOD detection. We mask the image portion within the box and use the ID class label as the prompt to resynthesize it. If the masked-out object is ID, the reconstructed image portion should be similar to the original, both visually and semantically. In contrast, for an OOD object, the inpainted portion prompted by the misclassified ID label will significantly deviate from the original. This discrepancy provides a valuable cue for distinguishing ID and OOD objects, requiring neither retraining nor access to ID or OOD training examples. We name our approach Ze**R**o-shot C**ON**ditional **IN**painting (**RONIN**), as illustrated in Fig. 1.

We evaluate RONIN on real-world benchmark datasets, including PASCAL VOC [11], BDD-100K [54], MS-COCO [27], and OpenImages [20], covering indoor, outdoor, and in-the-wild scenarios for comprehensive evaluation. The empirical results demonstrate RONIN's superior OOD detection performance compared to existing methods. Notably, RONIN shows strong potential in near-OOD scenarios by incorporating potential nearest non-ID concepts as exclusions, which leads to larger reconstruction errors when the original object belongs to one of these concepts and thus address near-OOD detection. Extensive ablation studies, analyses, and visualizations further validate the effectiveness of RONIN.

In summary, RONIN offers three notable advantages, positioning itself as a *plug-and-play* solution for OOD detection:

- **Retraining-free and zero-shot:** RONIN requires no access to training data for either ID or OOD categories, making it *highly versatile* and *data-efficient*.
- **Ease of implementation and compatibility:** RONIN is *straightforward to implement* and *compatible with most existing object detectors*, including open-vocabulary object detectors [29], which still rely on a pre-defined set of ID concepts to limit their scope.
- **Leveraging off-the-shelf generative models:** By utilizing pre-trained generative models, RONIN continuously benefits from advancements in generative modeling, *ensuring improved effectiveness and efficiency over time*.

Remark. While straightforward to implement, RONIN is far from a trivial contribution. It effectively leverages the inconsistency between discriminative and generative models as a post-hoc signal for identifying OOD objects. Unlike [33], RONIN extends applicability to object detection without requiring diffusion models to be trained on ID data. With the growing availability of powerful pre-trained models, using them as auxiliary tools to enhance OOD detection in domain-specific pipelines is both practical and scalable.

A potential limitation of RONIN lies in runtime, as diffusion models are generally slower than object detectors. However, this can be mitigated by reducing denoising steps—since our focus is on detecting OOD objects, not generating high-fidelity images, a slight quality trade-off is acceptable. We investigate this in Sec. 5, showing that fewer denoising steps significantly reduce runtime with minimal performance drop.

Importantly, object detectors are not always used in real-time. Many applications, such as wildlife monitoring with camera traps, involve offline post-processing where accuracy and generalizability matter more than speed. Our experiments across indoor and outdoor scenes validate RONIN 's strong performance in such settings, highlighting its readiness for practical deployment.

As generative modeling continues to evolve—with efficiency gains from methods like one-step diffusion—RONIN stands to benefit further. By prioritizing flexibility, generalizability, and practicality in both design and empirical results, RONIN offers a meaningful contribution to zero-shot out-of-distribution object detection.

2. Related Work

Out-of-distribution Detection. The goal of the out-of-distribution (OOD) detection task is to determine whether a given sample belongs to a certain distribution. It has been widely studied at the *image level*, where a whole image is treated as a sample. Common approaches include leveraging classifier-specific information such as confidence scores [5, 15, 17, 21, 26, 30, 50] or learned features [4, 22, 45, 47, 48, 53], or directly modeling an image distribution using generative models [14, 24, 33, 41, 44, 46, 52, 56]. Recent works [10, 34, 49] have also explored using CLIP [38] to identify OOD examples in a zero-shot manner, bypassing the need to learn from in-domain data explicitly.

OOD detection can also be extended to the *object-level*, where objects within an image are treated as individual samples. Most existing research in this setting focuses on

training-time interventions. For instance, [7, 9] improve detector features to make them more separable between in-distribution (ID) and OOD data; [51] uses adversarial examples to train an MLP for classifying ID and OOD instances. In contrast, our approach RONIN explores the object-level setting through post-hoc interventions instead.

Text-to-Image Generative Models. Recent advances in generative modeling have made large-scale text-to-image models widely available [40, 42, 43]. These models exhibit a deep understanding of language and are highly effective at generating or editing high-quality images from diverse prompts, offering a promising approach to data synthesis across various tasks. Furthermore, recent works have shown that these models can also enhance discriminative tasks. For instance, [19, 23] demonstrate that Stable Diffusion [42] can function as an effective zero-shot classifier, achieving accuracy comparable to or surpassing CLIP and various trained discriminative classifiers. There has also been recent progress in applying text-conditioned diffusion models for image-level OOD detection [8, 12, 13]. Unlike these works, RONIN focuses on exploring the use of such models for object-level tasks.

3. Problem Formulation

We address the task of object-level out-of-distribution (OOD) detection, which is motivated by the fact that object detectors can recognize unseen objects as seen ones. Specifically, given an object detector trained to detect a predefined set of categories (*e.g.* different kinds of vehicles), we aim to identify the error cases when it wrongly recognizes a novel object (*e.g.* a wild animal) as one of the predefined categories and detects it overconfidently. We refer to the pre-defined set of categories as the in-distribution (ID) classes, following [9]; the novel categories as the OOD classes.

Formally, given an image $\boldsymbol{x}$ and an object detector $f(.;\ \theta)$ trained for the ID classes $\mathcal{Y}^{in}$, $f(\boldsymbol{x};\ \theta)$ outputs a list of bounding boxes $\boldsymbol{b} = \{b_1,\ b_2, \ldots,\ b_n\}$ and their associated ID class labels $\hat{\boldsymbol{y}} = \{\hat{y}_1,\ \hat{y}_2, \ldots,\ \hat{y}_n\}$, where $\hat{y}_i \in \mathcal{Y}^{in}$. Object-level OOD detection is then formulated as a binary classification problem, classifying whether the object within b_i truly belongs to ID classes. Typically, one would develop a scoring function $\boldsymbol{g}$ such that given a bounding box and its predicted label $(b,\ \hat{y})$, $\boldsymbol{g}$ gives a higher score $\boldsymbol{g}(b,\ \hat{y})$ if b outlines an ID object and a lower score if it outlines an OOD object.

Existing works [7, 9] proposed a specific training process for the object detector so that its responses (*e.g.* the feature vector of each bounding box) would better distinguish between ID and OOD classes. However, doing so requires re-training the object detector, assuming access to the original training data and making it infeasible to off-the-shelf object detectors.

Aim. To address this limitation, we aim to design an OOD detection mechanism that is ***post-hoc***, without the need to modify (*e.g.* fine-tune) the pre-trained object detector, and ***zero-shot*** [34], without the need to access the original training data or any ID-class data. This sharply contrasts several prior post-hoc methods that need ID-class data [41, 52].

Approach. The emergence of vision-language foundation models [25, 28, 38] trained on abundant images and free-form text pairs has gradually removed the boundary between the closed-set and open-set settings [23]. For example, while not perfect, CLIP [38] can match an image with unbounded concepts. Such a zero-shot capability has been leveraged in prior work to detect OOD samples given a set of ID concepts [10, 34, 49]. In this work, we leverage another kind of foundation model, text-to-image generative models [42], capable of generating images given free-form text. While not designed for object detection nor specifically optimized for the ID data, we surmise their built-in, generic capability would facilitate object-level OOD detection.

4. RONIN: Zero-Shot OOD Conditional Inpainting

Our proposed framework, RONIN, builds on the assumption that generative models and object detectors behave differently when encountering unseen OOD objects due to their distinct training objectives. Object detectors are trained to separate foreground from background and classify objects into pre-defined ID categories, learning $P(y|x)$. This results in soft decision boundaries that can overconfidently assign unseen objects to seen classes. In contrast, generative models aim to capture the full data distribution $P(x, y)$ by modeling nuanced category characteristics, making them sensitive to deviations from training data and thus effective at flagging outliers. Leveraging this, we use generative models as auxiliary signals to improve zero-shot OOD detection in conjunction with object detection outputs.

RONIN is illustrated in Fig. 2. Given an object with its predicted bounding box and ID label, RONIN uses off-the-shelf class-conditioned inpainting models to regenerate the object, aligning it with the in-distribution domain. The regenerated image is then compared with the original using a vision-language triplet similarity metric. High similarity implies visual and semantic consistency with the ID category (in-distribution), while low similarity suggests a mismatch (out-of-distribution). We detail each component of this approach below.

4.1. Class-conditional Inpainting

Our key idea behind RONIN is to harness diffusion models for inpainting, refine predicted objects, and bring them closer to the ID domain. Specifically, given a detected bounding box $\boldsymbol{b}$ and a predicted ID class $\hat{y} \in \mathcal{Y}^{in}$ by the object detector, we generate a masked region $\boldsymbol{m}$ within $\boldsymbol{b}$ and use $\hat{y}$ as a condition for reconstruction. Since $\boldsymbol{m}$ is always contained

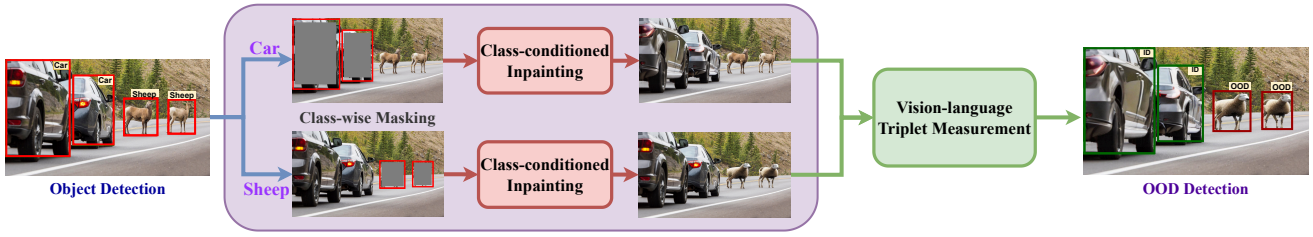

Figure 2. **Overall framework of RONIN**, including (i) *Class-conditioned Inpainting* and (ii) *Vision-language Triplet Measurement*. Given an image with bounding boxes and predicted labels, RONIN masks and reconstructs objects via inpainting, then evaluates alignment using triplet similarity for zero-shot OOD detection.

within $\boldsymbol{b}$, the inpainting model receives sufficient context from both the background and the detected object as hints for reconstructing the original object. This ensures that the inpainted outcomes closely align with the original ID objects, not only in semantic meaning but also in visual appearance, while still being distinct in OOD cases.

More specifically, let the semantic background $\boldsymbol{s}$ be the entire image region excluding the masked region $\boldsymbol{m}$, RONIN takes input of $\boldsymbol{m}$ (also known as the inpainting mask), the predicted label $\boldsymbol{c} = \hat{y}$, and the semantic background $\boldsymbol{s}$ for inpainting. We choose to use pre-trained text-conditioned diffusion models such as DDPM [16]. The diffusion models starts with a random Gaussian noise over $\boldsymbol{m}$ and perform the reverse diffusion process with T steps, guided by $\boldsymbol{c}$, to obtain the outcome image $\boldsymbol{x}_{inp} = \boldsymbol{x}_0$, as in Eq. (1), where μ_θ and Σ_θ are priorly learned by pre-training a neural network with parameter θ.

$$p_\theta(\tilde{\boldsymbol{x}}_{t-1}|\boldsymbol{x}_t,\ \boldsymbol{c}) = \mathcal{N}\big(\tilde{\boldsymbol{x}}_{t-1};\ \mu_\theta(\boldsymbol{x}_t,\ t,\ \boldsymbol{c}), \Sigma_\theta(\boldsymbol{x}_t,\ t,\ \boldsymbol{c})\big) \tag{1}$$

$$\boldsymbol{x}_{t-1} = \tilde{\boldsymbol{x}}_{t-1} \odot \boldsymbol{m} + \boldsymbol{s} \tag{2}$$

While the inpainted objects are synthesized based on $\hat{y}$ over the foreground region $\boldsymbol{m}$, the information retained from $\boldsymbol{s}$ guides the DDPM to preserve some visual features of the original object, as described in Eq. (2). As a result, the inpainted outcome object retains the same visual and semantic characteristics as the original object, despite being presented differently. This enables RONIN the generation of objects closely aligned with the originals in ID cases and significantly deviating in OOD cases, as shown in Fig. 2. We further analyze how varying the inpainted mask $\boldsymbol{m}$ based on $\boldsymbol{b}$ impacts OOD detection in Sec. 6.

Class-wise Masking and Inpainting. Intuitively, a natural approach to inpainting is treating each object independently, or "*object-wise*" inpainting. However, this is highly inefficient when many objects share the same predicted label and denoising condition, making the process redundant. To address this, we propose "*class-wise*" inpainting, where objects are grouped by the predicted label $\hat{y}$ for a single inpainting process, ensuring the computational cost at most scales with $\mathcal{Y}^{in}$ the worst case. This significantly reduces computational cost while maintaining performance, as the shared denoising condition ensures effective synthesis. We further evaluate the two strategies in Sec. 6.

4.2. Vision-language Triplet Measurement

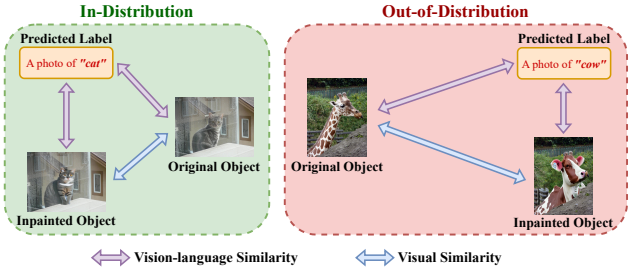

Figure 3. **Triplet similarity relationships** between (i) the original object, (ii) the inpainted outcome, and (iii) the predicted label. ID samples show strong alignments across all three, whereas OOD samples exhibit weak alignments, aiding effective OOD detection.

Similarity Aligments. After inpainting, we obtain the original object $\mathbf{x_{ori}}$, its predicted label $\hat{y} \in \mathcal{Y}^{in}$, and the corresponding inpainted object $\mathbf{x_{inp}}$ which is synthesized based on $\hat{y}$. Intuitively, for in-distribution (ID) objects, $\mathbf{x_{inp}}$ tends to closely resemble $\mathbf{x_{ori}}$, while in contrast, out-of-distribution (OOD) objects exhibit lower similarity, indicating divergence from the original given a wrong label.

However, due to the stochastic nature of diffusion inpainting and subtle distinctions in semantic meanings between OOD samples and predicted ID labels, the similarity between $\mathbf{x_{ori}}$ and $\mathbf{x_{inp}}$ can vary greatly. To mitigate this, we introduce a triplet of similarity scores for a more robust ID vs. OOD distinction:

$$\begin{aligned} \textit{similarity}(ori,\ \hat{y}) &= \mathbf{cosine}\left(f_v(\mathbf{x_{ori}}), f_t(\hat{y})\right) \\ \textit{similarity}(inp,\ \hat{y}) &= \mathbf{cosine}\left(f_v(\mathbf{x_{inp}}), f_t(\hat{y})\right) \\ \textit{similarity}(ori,\ inp) &= \mathbf{cosine}\left(f_i(\mathbf{x_{ori}}), f_i(\mathbf{x_{inp}})\right) \end{aligned} \tag{3}$$

Here, **cosine** means the cosine similarity between two vectors and f_v and f_t represent visual and textual embeddings from large vision-language contrastive models (e.g., CLIP [38]), while f_i is derived from a visual contrastive model (e.g., SimCLRv2 [3]). Cosine similarity ef-

fectively measures the alignment between these representations. Specifically, $similarity(ori,\ inp)$ emphasizes ***visual alignment*** for ID retention while $similarity(ori,\ \hat{y})$ supports ***semantic alignment*** for OOD recognition. In contrast, $similarity(inp,\ \hat{y})$ helps ***normalize*** variations across different labels. Fig. 3 illustrates the relationships between these components for both ID and OOD cases.

Triplet Measurement. Based on the relationship between the three similarities, we then formulate a triplet OOD score:

$$S_{\text{triplet}} = \frac{similarity(ori,\ \hat{y})^{\boldsymbol{\alpha}} \times similarity(ori,\ inp)^{\boldsymbol{\beta}}}{similarity(inp,\ \hat{y})} \quad (4)$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are hyperparameters controlling the balance between visual-language and visual similarities. A higher S_{triplet} suggests a greater likelihood of the object being ID, while lower values indicate OOD cases. In practice, a thresholding method can be applied for effective OOD detection depending on different tasks.

4.3. Refined Prompt for Near-OOD

Identifying near-OOD cases poses a significant challenge for conventional OOD detection methods, as unseen objects often share closely related semantic meanings with certain in-distribution (ID) categories, making separation difficult. However, off-the-shelf generative and CLIP models inherently possess the ability to distinguish such cases because of their extensive pre-training on large, diverse datasets. To validate this, we proposed a simple approach that, given a predefined set of ID labels, we use knowledge graphs or taxonomies from large language models to retrieve closely related non-ID concepts and explicitly instruct the generative model not to inpaint these concepts as part of the conditioning. For example, if the ID label is "*horse*", we can retrieve highly related but OOD classes such as—*donkey, zebra, mule, pony, and camel*—and explicitly condition the inpainting model to generate a horse that remains distinct from these related categories. We believe this "*refined*" prompting technique not only supports RONIN to generate outcomes that align better with the ID label but also improve its ability to do near-OOD detection.

5. Experiments

5.1. Experimental Setups

Datasets. We use Pascal-VOC [11] and Berkerly DeepDrive (BDD-100k) [54] as ID data, and MS-COCO [27] and OpenImages [20] as OOD data. The combinations of these data covered indoor, outdoor, and in-the-wild objects and contexts, allowing us to evaluate RONIN comprehensively. We follow the splits from [7, 9], where *overlapping categories in the ID and OOD datasets are removed*. We select a subset of 200 images from the test set of BDD-100k and 400 each from the test sets of Pascal-VOC, MS-COCO, and OpenImages. Data processing details can be found in Appendix A.

Evaluation Metrics. We evaluate the OOD detection performance following the two standard OOD metrics: the area under the Receiver Operating Characteristic curve (**AUROC**) and the false positive rate of OOD objects at 95% true positive rate of in-distribution objects (**FPR@95**).

Baselines. We consider three types of baselines: (1) *Discriminative zero-shot approaches*: **MCM** [34], CLIP-based **ODIN** [26], CLIP-based **Energy Score** [30], **CLIPN** [49], **TAG** [31], **OLE** [6] and **GL-MCM** [35]. (2) *Alternative generative zero-shot approaches*: We first generate synthetic data and then apply standard OOD detection methods such as **Mahalanobis** [53] and **KNN** [47]. (3) *Training-based approaches*: for a complete comparison, we also include two representative training-based object-level methods **VOS** [9] and **SIREN** [7], resulting in a total of **eleven** baselines. Implementation details can be found in Appendix B.

Implementation Details. For RONIN, we perform class-wise inpainting with Stable Diffusion 2 Inpainting [42] with 20 steps. Each object is masked with a center mask covering 0.9 of the height and width of its original bounding box. Object-label similarities are determined by the distance between OpenCLIP [18] features, and object-object similarities by the SimCLRv2 [3] features.

For the main experiments in Sec. 5.2, we use detected objects obtained from Deformable-DETR [55], as implemented by [7], to enable direct comparisons with training-based baselines. However, it is important to note that RONIN is compatible with various types of detectors. In Sec. 5.3, we demonstrate RONIN 's performance using detections from the pre-trained open-vocabulary detector GroundingDINO [29], showcasing its flexibility and compatibility. The Triplet Similarity (Eq. (4)) is computed with $\boldsymbol{\alpha} = 2$ and $\boldsymbol{\beta} = 1$. All experiments were conducted on a single NVIDIA GeForce RTX 2080Ti 11GB.

5.2. Quantitative and Qualitative performance

Tab. 1 shows the performance of RONIN and baselines based on DeformableDETR detections across four settings. RONIN achieves the highest results in three scenarios, even with just 5 denoising steps, outperforming training-based methods like VOS and SIREN without retraining on in-domain data. Increasing the number of denoising steps further improves performance, demonstrating that RONIN consistently surpasses all baselines across most settings.

Fig. 4 presents side-by-side visualizations of RONIN with Pascal-VOC as the ID setting. Overall, ID objects are well reconstructed while OOD objects appear distinctly different, enabling effective separation via the triplet score. However, we identify two common types of failures: *inpainting failures*, where small, occluded, or poorly lit objects pre-

Table 1. **Object-level OOD detection on four settings**. We highlight the **best** and second best performance. Even with 5 denoising steps, RONIN is able to outperform all the baselines in three out of four settings. Denoising with more steps further improve the performance.

In-Distribution	Method	MS-COCO		OpenImages	
		FPR@95 (↓)	AUROC (↑)	FPR@95 (↓)	AUROC (↑)
PASCAL-VOC	ODIN [26]	41.65	88.22	55.87	86.46
	Energy Score [30]	29.48	90.26	24.57	91.24
	Mahalanobis [53]	63.30	83.26	45.22	87.11
	KNN [47]	58.56	85.52	45.00	84.62
	VOS [9]	48.15	88.75	54.63	83.65
	SIREN [7]	64.70	78.68	66.69	75.12
	MCM [34]	62.47	83.15	71.52	81.45
	CLIPN [49]	43.09	85.45	41.74	89.31
	TAG [31]	61.03	78.35	48.48	87.69
	OLE [6]	54.23	86.13	49.13	88.07
	GL-MCM [35]	78.56	78.64	69.35	82.89
	RONIN (5 steps)	27.94	92.35	20.00	92.32
	RONIN (10 steps)	29.07	91.10	20.20	92.80
	RONIN (15 steps)	25.57	91.63	19.87	92.84
	RONIN (20 steps)	**25.36**	**92.31**	**18.91**	**93.10**
BDD-100k	ODIN [26]	96.51	55.18	95.56	57.87
	Energy Score [30]	71.75	74.49	53.33	73.09
	Mahalanobis [53]	31.75	90.74	**23.33**	**93.81**
	KNN [47]	35.87	86.58	31.11	92.75
	VOS [9]	65.45	78.34	59.23	80.42
	SIREN [7]	42.86	89.37	37.97	91.78
	MCM [34]	95.56	55.82	92.22	57.05
	CLIPN [49]	28.49	92.14	44.76	85.78
	TAG [31]	46.43	90.85	55.30	76.67
	OLE [6]	57.14	78.14	40.00	85.36
	GL-MCM [35]	96.83	50.82	94.44	58.88
	RONIN (5 steps)	27.94	92.59	26.67	91.23
	RONIN (10 steps)	26.67	92.73	25.56	91.33
	RONIN (15 steps)	26.03	92.83	24.44	91.65
	RONIN (20 steps)	**26.03**	**92.90**	**23.33**	91.90

vent meaningful reconstruction, and *similarity-score failures*, where generated objects remain visually or semantically too close to the original, reducing the effectiveness of the triplet score for OOD detection.

5.3. Near-OOD Detection

To demonstrate the potential of RONIN for near-OOD detection, we constructed a small-scale dataset featuring visually and semantically similar ID and OOD objects. Specifically, we curated 411 samples by pairing ID objects from PascalVOC with closely related OOD counterparts from MS-COCO and OpenImages, forming categories such as "*horse vs. zebra*", "*dog vs. fox*", and "*cat vs. raccoon/tiger*". Using the open-vocabulary detector GroundingDINO [29], we obtained 245 ID and 265 OOD predictions. For the "refined" prompt, we simply leveraged ChatGPT (GPT-4) to generate five nearest non-ID concepts for each ID label as exclusions, with the prompt structure and results are presented in Fig. 5.

Table 2. **Near-OOD detection performance**. With more information, the "refined" prompt performs better on challenging cases.

	FPR@95 (↓)	AUROC (↑)
RONIN (*Simple* prompting)	7.79	96.86
RONIN (*Refined* prompting)	5.66	98.03

Tab. 2 presents the near-OOD performance under two prompting strategies. Further shown in Fig. 5, the "refined"

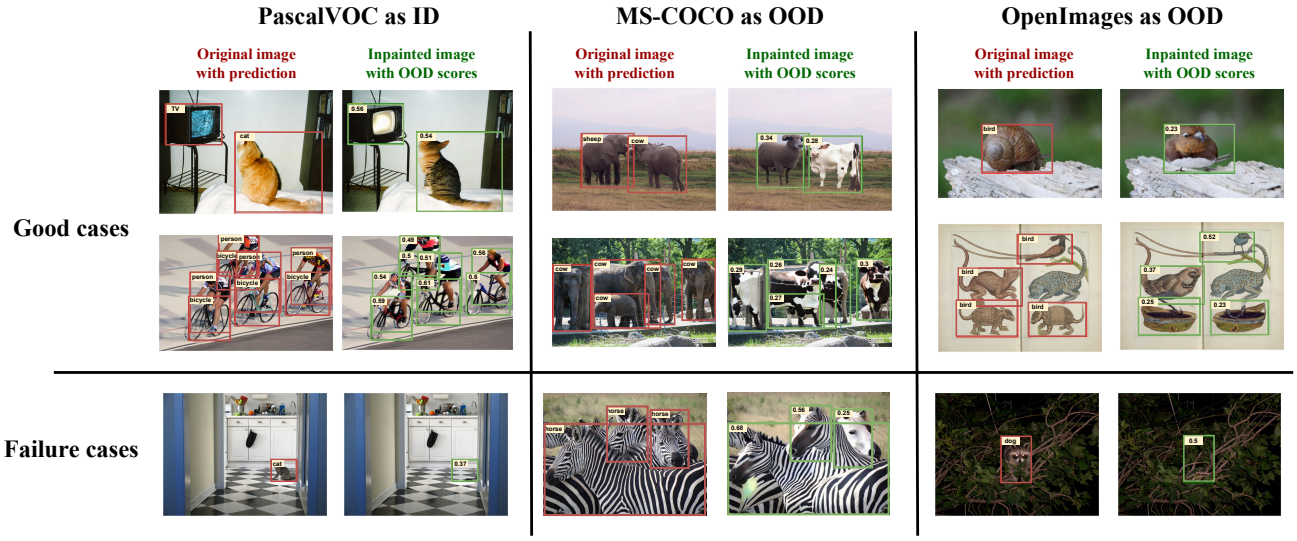

Figure 4. **Side-by-side quantitative visualization.** Good cases show synthesized objects consistent with predicted labels and clear OOD score separation; bad cases show inpainting failures or OOD objects too resembling the originals, leading to ineffective OOD score.

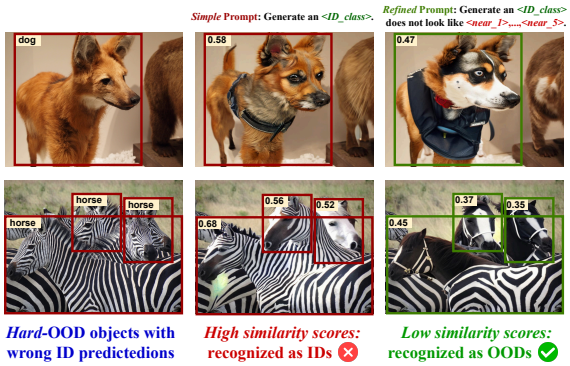

Figure 5. **RONIN performance on near-OOD with *refined* prompting**. With distinct inpainting by providing more context, RONIN is able to yield lower scores for near-OOD detection.

prompt enables the inpainting model to generate outputs more closely aligned with the predicted labels, thereby improving near-OOD detection. These results support our hypothesis that pre-trained generative and CLIP models already possess the capability for fine-grained OOD detection, given sufficient context.

6. Ablation Studies

Performance-Speed Trade off. Fig. 6 analyzes the trade-off between RONIN 's performance and runtime by varying the number of denoising steps and the image resolutions used during the inpainting process. For relative performance comparison, we include CLIPN [49] as the best-performing baseline, though we do not compare absolute runtimes between the two methods. Reducing the number of denoising steps and downscaling image sizes significantly accelerates RONIN, with only minor performance degradation. These results highlight the flexibility of our approach and its potential suitability for low-resource compute environments, with reasonable performance drops.

Table 3. **RONIN with object-wise and class-wise inpainting.** Class-wise inpainting retains performance with faster runtime.

	FPR@95 (↓)		AUROC (↑)	
	Object-wise	Class-wise	Object-wise	Class-wise
VOC - COCO	26.60	25.36	91.48	92.31
VOC - OpenImages	20.00	18.91	92.68	93.10
BDD - COCO	28.89	26.03	93.74	92.90
BDD - OpenImages	26.67	23.33	92.42	91.90

Class-wise vs. Object-wise Inpainting. Tab. 3 compares the "*object-wise*" inpainting strategy, which processes object individually, with our default "*class-wise*" approach, which inpaints objects categorically. By leveraging label-wise grouping, class-wise inpainting achieves similar performances while offering more efficient computation cost.

Triplet Similarity Measurement. Tab. 4 presents ablation studies on the triplet similarity score (Eq. (4)). The upper half (rows 1–4) assesses the effect of removing individual similarity components, confirming that all are essential for optimal performance. The lower half (rows 5–7) explores the weighting between semantic alignment $similarity(ori,\ \hat{y})$ (weighted by $\boldsymbol{\alpha}$) and visual alignment $similarity(ori,\ inp)$ (weighted by $\boldsymbol{\beta}$). As shown in Fig. 3, emphasizing semantic alignment ($\boldsymbol{\alpha} > \boldsymbol{\beta}$) yields the best results, guiding our default choice ($\boldsymbol{\alpha} = 2$, $\boldsymbol{\beta} = 1$). However, the sensitivities of these weights suggest potential adaptability to more challenging settings.

Impact of Masking Ratios. Tab. 5 examines the impact of mask size on OOD performance. While smaller masks

Table 4. **Ablation on triplet similarity**. All three similarities are essential for effective score (underlined). Emphasizing visual-language semantic alignment (α) improves performance (**bolded**).

	COCO / OpenImages	
	FPR@95 (↓)	AUROC (↑)
without $similarity(inp, \hat{y})$	44.54 / 37.39	89.77 / 91.42
without $similarity(ori, \hat{y})$	65.15 / 68.91	69.72 / 78.53
without $similarity(ori, inp)$	30.93 / 36.96	86.14 / 90.10
triplet similarity, $\boldsymbol{\alpha} = \boldsymbol{\beta} = 1$	35.46 / 24.78	87.75 / 91.43
triplet similarity, $\boldsymbol{\alpha} = 1; \boldsymbol{\beta} = 2$	43.93 / 33.91	85.40 / 89.50
triplet similarity, $\boldsymbol{\alpha} = 2; \boldsymbol{\beta} = 1$	**25.36 / 18.91**	**92.31 / 93.10**
triplet similarity, $\boldsymbol{\alpha} = 2; \boldsymbol{\beta} = 2$	35.46 / 25.43	90.32 / 92.32

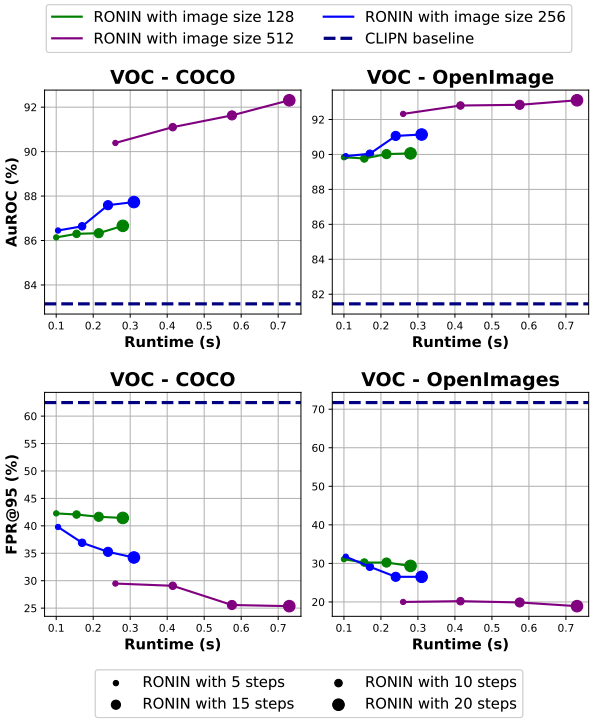

Figure 6. **Performance-speed tradeoff of RONIN**. Reducing image size and denoising steps speeds up inference with minimal performance loss. At full settings, RONIN achieves top accuracy with competitive runtime. CLIPN is for performance comparison.

Table 5. **Trade-off performance with different mask sizes**. Sufficient mask size (80% bounding box) retains cues for distinguishing object synthesis toward effective OOD performances.

Covered masking ratio m	MS-COCO/OpenImages	
	FPR@95 (↓)	AUROC (↑)
75%	30.10 / 21.74	90.43 / 92.41
80%	25.80 / 17.74	91.32 / 93.84
100%	29.68 / 20.29	91.61 / 92.61

lead to minimal inpainting changes for effective OOD object synthesis, larger masks do not retain sufficient information for accurate ID object reconstruction, leading to suboptimal performance. More analysis is provided in Appendix C.1.

Table 6. **RONIN performances with few-step generative models.** The on-par performances demonstrate RONIN effectiveness with minimal reliance on specific models or synthesis quality.

	COCO / OpenImages	
	FPR@95 (↓)	AUROC (↑)
Stable Diffusion 2 (5 steps)	27.94 / 20.00	92.35 / 92.32
InstaFlow (1 steps)	31.34 / 25.74	88.03 / 90.35

Robust across Generative Models. Tab. 6 evaluates RONIN across various generative models, focusing on the few-step regime. RONIN shows strong performance even with the one-step generative model InstaFlow [32], suggesting minimal reliance on specific models or inpainting quality for effective OOD detection. Additional ablations in the many-step regime are provided in Appendix C.2.

Table 7. **Recall and falsy reject ratio over ID data**, showing RONIN minimal impact on ID objects.

ID set	ID object	Recall	ID Rejected
Pascal-VOC	1007	91.57%	8.43%
BDD	2632	96.67%	4.33%

End-to-end impact on ID object. Tab. 7 reports Recall and ID rejection ratios after the detection stage, showing that RONIN maintains stable OOD detection without degrading ID accuracy. While BDD mainly contains small, similar vehicle categories, Pascal VOC spans far more diverse objects and animals, making it a more challenging dataset.

7. Conclusion

Overall, RONIN delivers consistently strong performance across diverse settings, from indoor scenes to in-the-wild scenarios, demonstrating clear flexibility and generality. By exploiting subtle misalignments between generative and discriminative outputs, RONIN achieves robust OOD performance without reliance on specific diffusion models or inpainting results, and it remains adaptable to a wide range of object detectors as a training-free, plug-and-play solution. Although current diffusion models limit real-time deployment, many practical applications operate in offline or post-processing pipelines where accuracy and generalizability matter most—areas where RONIN excels. With strong generality and high adaptability across experiments, RONIN offers a meaningful step forward for zero-shot out-of-distribution object detection.

References

[1] Hafiz Mughees Ahmad and Afshin Rahimi. Deep learning methods for object detection in smart manufacturing: A survey. *Journal of Manufacturing Systems*, 64:181–196, 2022. 1

[2] Sara Beery, Dan Morris, and Siyu Yang. Efficient pipeline for camera trap image review. *arXiv preprint arXiv:1907.06772*, 2019. 1

[3] 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. 4, 5

[4] Taylor Denouden, Rick Salay, Krzysztof Czarnecki, Vahdat Abdelzad, Buu Phan, and Sachin Vernekar. Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance. *arXiv preprint arXiv:1812.02765*, 2018. 2

[5] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018. 2

[6] Choubo Ding and Guansong Pang. Zero-shot out-of-distribution detection with outlier label exposure. In *2024 International Joint Conference on Neural Networks*, 2024. 5, 6

[7] 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, 3, 5, 6

[8] Xuefeng Du, Yiyou Sun, Xiaojin Zhu, and Yixuan Li. Dream the Impossible: Outlier Imagination with Diffusion Models. In *Advances in Neural Information Processing Systems*, 2023. 3

[9] 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, 3, 5, 6

[10] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot Out-of-distribution Detection based on the pre-trained model CLIP. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 6568–6576, 2022. 2, 3

[11] 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. 2, 5

[12] Kun Fang, Qinghua Tao, Zuopeng Yang, Xiaolin Huang, and Jie Yang. Ddos: Diffusion distribution similarity for out-of-distribution detection. *arXiv preprint arXiv:2409.10094*, 2024. 3

[13] Ruiyuan Gao, Chenchen Zhao, Lanqing Hong, and Qiang Xu. Diffguard: Semantic mismatch-guided out-of-distribution detection using pre-trained diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1579–1589, 2023. 3

[14] Mark S. Graham, Walter H.L. Pinaya, Petru-Daniel Tudosiu, Parashkev Nachev, Sebastien Ourselin, and Jorge Cardoso. Denoising Diffusion Models for Out-of-Distribution Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2948–2957, June 2023. 2

[15] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 2

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 4

[17] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data, 2020. 2

[18] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip. https://doi.org/10.5281/zenodo.5143773, 2021. 5

[19] Priyank Jaini, Kevin Clark, and Robert Geirhos. Intriguing properties of generative classifiers. In *International Conference on Learning Representations*, 2024. 3

[20] 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. 2, 5

[21] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017. 2

[22] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018. 2

[23] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2206–2217, 2023. 3

[24] Jingyao Li, Pengguang Chen, Zexin He, Shaozuo Yu, Shu Liu, and Jiaya Jia. Rethinking out-of-distribution (ood) detection: Masked image modeling is all you need. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11578–11589, 2023. 2

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 3

[26] 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. 2, 5, 6

[27] 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. 2, 5

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee.

Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3

[29] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2, 5, 6

[30] 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. 2, 5, 6

[31] 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. 5, 6

[32] 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. 8

[33] Zhenzhen Liu, Jin Peng Zhou, Yufan Wang, and Kilian Q Weinberger. Unsupervised out-of-distribution detection with diffusion inpainting. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22528–22538. PMLR, 23–29 Jul 2023. 2

[34] 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. 2, 3, 5, 6

[35] 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. 5, 6

[36] Muhammad Akhtar Munir, Muhammad Haris Khan, M Sarfraz, and Mohsen Ali. Towards improving calibration in object detection under domain shift. *Advances in Neural Information Processing Systems*, 35:38706–38718, 2022. 1

[37] Bimsara Pathiraja, Malitha Gunawardhana, and Muhammad Haris Khan. Multiclass confidence and localization calibration for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19734–19743, 2023. 1

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 4

[39] Mohammed Gamal Ragab, Said Jadid Abdulkadir, Amgad Muneer, Alawi Alqushaibi, Ebrahim Hamid Hasan Sumiea, Rizwan Qureshi, Safwan Mahmood Al-Selwi, and Hitham Seddiq Alhassan Alhussian. A comprehensive systematic review of yolo for medical object detection (2018 to 2023). *IEEE Access*, 12:57815–57836, 2024. 1

[40] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 3

[41] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *Advances in neural information processing systems*, 32, 2019. 2, 3

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2, 3, 5

[43] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2, 3

[44] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017. 2

[45] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. *arXiv preprint arXiv:2103.12051*, 2021. 2

[46] Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. *arXiv preprint arXiv:1909.11480*, 2019. 2

[47] 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. 2, 5, 6

[48] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems*, 33:11839–11852, 2020. 2

[49] 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. 2, 3, 5, 6, 7

[50] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International Conference on Machine Learning*, pages 23631–23644. PMLR, 2022. 2

[51] Samuel Wilson, Tobias Fischer, Feras Dayoub, Dimity Miller, and Niko Sünderhauf. SAFE: Sensitivity-aware Features for Out-of-distribution Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23565–23576, 2023. 3

[52] Zhisheng Xiao, Qing Yan, and Yali Amit. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. *Advances in neural information processing systems*, 33:20685–20696, 2020. 2, 3

[53] 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. 2, 5, 6

[54] 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. 2, 5

[55] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 5

[56] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations*, 2018. 2