Dream-Box: Object-wise Outlier Generation for Out-of-Distribution Detection

Brian K.S. Isaac-Medina[1], Toby P. Breckon[1,2]
Department of {Computer Science[1], Engineering[2]}, Durham University, Durham, UK

Abstract

*Deep neural networks have demonstrated great generalization capabilities for tasks whose training and test sets are drawn from the same distribution. Nevertheless, out-of-distribution (OOD) detection remains a challenging task that has received significant attention in recent years. Specifically, OOD detection refers to the detection of instances that do not belong to the training distribution, while still having good performance on the in-distribution task (*e.g.*, classification or object detection). Recent work has focused on generating synthetic outliers and using them to train an outlier detector, generally achieving improved OOD detection than traditional OOD methods. In this regard, outliers can be generated either in feature or pixel space. Feature space driven methods have shown strong performance on both the classification and object detection tasks, at the expense that the visualization of training outliers remains unknown, making further analysis on OOD failure modes challenging. On the other hand, pixel space outlier generation techniques enabled by diffusion models have been used for image classification using, providing improved OOD detection performance and outlier visualization, although their adaption to the object detection task is as yet unexplored. We therefore introduce Dream-Box, a method that provides a link to object-wise outlier generation in the pixel space for OOD detection. Specifically, we use diffusion models to generate object-wise outliers that are used to train an object detector for an in-distribution task and OOD detection. Our method achieves comparable performance to previous traditional methods while being the first technique to provide concrete visualization of generated OOD objects.*

1. Introduction

Out-of-distribution (OOD) detection has emerged as a critical challenge in the deployment of deep neural networks, particularly in tasks such as classification [4, 11, 21, 28] and object detection [22, 27, 30] in addition to having broader impact potential for practical outlier detection across a range of application tasks [1, 8–10, 14]. While these models exhibit remarkable generalization capabilities within their training distribution, their ability to identify and handle data that deviates from this distribution remains a significant limitation. OOD detection aims to address this issue by distinguishing between in-distribution and out-of-distribution instances, ensuring robust performance in real-world scenarios where the input data may not conform to the training set [29].

Traditional approaches for OOD detection include using simple softmax probabilities [12] or the Mahalanobis distance [16] to leverage the statistical properties of feature distributions, identifying OOD instances by measuring the distance of a given sample from the inlier distribution in the feature space. Another approach consists of using Gram matrices [24] to capture the correlations between feature maps, providing a robust representation to discriminate between in-distribution and OOD data. A recent approach that has shown great performance for OOD detection is to use the energy score [20], *i.e.*, the *log-sum-exp* operation over the class logits, framing OOD detection as a density estimation problem by assigning lower energy scores to in-distribution samples and higher scores to outliers. Whilst these methods have shown promising results, they often lack the ability to leverage training outliers (*e.g.*, synthetically generated outlier instances). This limitation has spurred the development of outlier generation approaches, which not only improve detection performance but in some cases offer the added benefit of interpretable outlier visualization [5, 6].

In terms of outlier synthesis, other works have focused on creating synthetic outliers for training an in-distribution/OOD binary classifier. Among these approaches, feature space outlier generation methods, such as Virtual Outlier Synthesis (VOS) [5] and Feature Flow Synthesis (FFS) [15], have demonstrated strong performance in both classification and object detection tasks. These techniques generate synthetic outliers in the feature space, enabling the training of outlier classifiers that enhance OOD detection capabilities. However, a notable limitation of these methods is the lack of interpretability and visualization of the generated outliers, which hinders a deeper understanding of their failure modes and limits further analysis towards improved performance. On the other hand, pixel space outlier generation techniques have shown promise in image classification tasks. For instance, Dream-OOD [6]

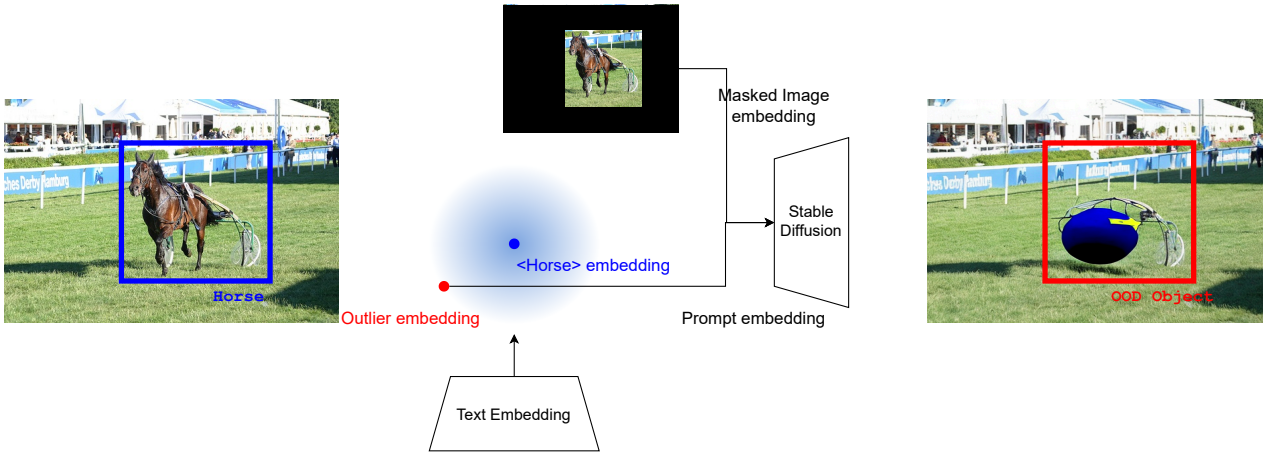

Figure 1. Dream-Box enables object-wise OOD detection by generating objects using embeddings far from the class-name text embeddings.

leverages diffusion models to generate synthetic outlier images directly in pixel space. This technique provides meaningful visualizations of the generated outliers, offering insights into their characteristics, while improving OOD detection. Despite these advancements, the application of pixel space outlier generation to object-wise OOD detection remains largely unexplored.

This work aims to close this gap by introducing a novel approach leveraging diffusion models for object-wise pixel space outlier generation in the context of object detection. Our method generates object-wise outliers (Fig. 1), which are then used to train an object detector capable of performing both in-distribution tasks and OOD detection. Our method, dubbed Dream-Box, achieves OOD detection performance comparable to state-of-the-art traditional methods while providing interpretable visualizations of OOD objects. This work builds on the foundations laid by Dream-OOD [6], which explores the use of generative models for OOD detection, but further extends its applicability to object detection tasks. Our contributions not only advance the state-of-the-art in OOD detection but also open new avenues for research in outlier visualization and analysis, encouraging for more robust and interpretable deep learning systems. Reference software code is available at:

https://github.com/KostadinovShalon/dream-box

2. Related Work

Several works investigate OOD detection using different approaches. For instance, some works rely on estimating the probability density of the training data and flagging low-density regions as OOD. For instance Lee *et al*. [16] proposed using the Mahalanobis distance in the feature space of deep neural networks, achieving strong performance in OOD detection tasks. Nonetheless, Mahalanobis distance may not generalize well for complex distributions, such as for object detection. Other methods, such as ODIN [17] and Gram matrices [24], modify pre-trained models to improve OOD detection without retraining. While computationally efficient, these methods often rely on heuristics and may not generalize across diverse datasets.

A recent approach that has shown promising results uses the free energy score for OOD detection [20]. Energy-based models treat OOD detection as a density estimation problem by assigning lower energy scores to in-distribution samples and higher scores to outliers, leveraging the logits of a pre-trained classifier to detect OOD samples, demonstrating high performance on benchmark datasets. Other works have used the energy method while using synthetically generated outliers. For instance, Virtual Outlier Synthesis (VOS) [5] generates outliers in the feature space by learning Gaussian distributions over the feature representation of different classes and sampling from low-probablity regions. Similarly, Feature Flow Synthesis (FFS) [15] learns a reversible transformation from the feature space to a class-agnostic normalized space, where outliers are generated, demonstrating improvement over VOS.

Additionally, Isaac-Medina *et al*. [14] extended the VOS and FFS frameworks introducing OLN-SSOS for class-agnostic open-world OOD detection in object detection. In this context, these methods are the first methods to show a significant performance for the object detection task, but often lack interpretability and visualization capabilities.

On the other hand, Dream-OOD [6] uses diffusion models to generate full-image synthetic outliers in the pixel space, offering improved OOD detection performance and interpretable visualizations, enhancing the robustness of OOD detectors. Nonetheless, its application to object detection via object-wise outlier generation remains an area of investigation and hence forms the focus of the OOD study presented here in this paper.

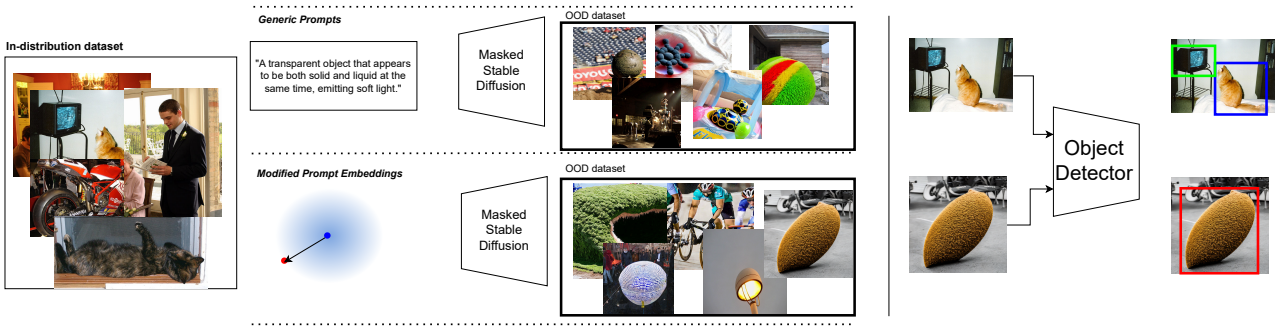

Figure 2. Dream-Box overview. We generate outlier objects using two prompt strategies and leveraging in-distribution objects. Subsequently, we train an object detector with a classification output head for in-distribution/OOD.

3. Dream-Box

In general, object-wise OOD refers to detecting objects in an image and labeling them as either an in-distribution class or an OOD instance, in a semi-supevised mannner, without having any ground-truth (true) OOD samples available during training. Therefore, we introduce Dream-Box, a framework for object bounding-box based OOD that leverages diffusion models for object-wise outlier generation in the pixel space. We develop our strategy for outlier synthesis in Sec. 3.1, while we outline our OOD detection technique in Sec. 3.2. Finally, the implementation details are discussed in Sec. 3.3.

3.1. Outlier Generation Strategy

Motivated by Dream-OOD [6], Dream-Box uses Stable Diffusion [23] to generate object-wise outliers in the pixel space, enabling object-aware OOD detection. An overview of Dream-Box is presented in Fig. 2.

Our method consists in augmenting the training dataset by replacing in-distibution objects with synthetic outliers that are labeled as OOD. Specifically, we consider an image $\mathbf{x}$ drawn from the in-distribution $\mathbf{x} \sim \mathcal{D}_{in}$ and consider its ground truth bounding boxes $\mathcal{B} = \{b_k\}_{k=1}^{M}$, where M is the number of ground truth bounding boxes in the image, and each bounding box consists of a class label c_k and four position values, *i.e.*, the top-left coordinates (x_k, y_k) and the width and height (w_k, h_k). An image $\tilde{\mathbf{x}}$ with OOD objects is generated by inpainting an object for each $b_k \in \mathcal{B}$ using a masked generator model f that generates an object given a mask m_k and a prompt $\rho(c_k)$ in function of the class label, such that $\tilde{\mathbf{x}}(b_k) = f(\mathbf{x}, \rho(c_k), m_k)$. In this sense, we take the mask as the region described by the bounding box. If we consider all objects within the input image, then the generated OOD image consists of generating one object at a time in a sequential manner:

$$\tilde{\mathbf{x}}_i = f(\tilde{\mathbf{x}}_{i-1}, \rho(c_i), m_i)\,. \tag{1}$$

Dream-Box samples N images from $\mathcal{D}_{in}$ and uses the process from Eq. (1) to generate an OOD dataset $\mathcal{D}_{ood}$ which will enable object-wise OOD detection (Sec. 3.2).

We consider two prompting strategies to use with Eq. (1), namely *generic textual prompts* and *distance-based modified prompts*, described as follows:

Generic textual prompts: Since the goal is to perform OOD detection while still having a good performance in the in-distribution task (object detection), we propose a simple sampling mechanism that create objects similar to the in-distribution dataset but with features that make them OOD. Therefore, we use generic prompts taking the class name and describing unrealistic and/or impossible characteristics. We choose 20 different prompts, given in Tab. 1, where {} indicates the class name. The motivation of this strategy is that we can learn to recognize objects that do not look like the *normal* as in the in-distribution dataset but can be still detected as objects since they might look like the in-distribution data.

Distance-based modified prompts: Following Dream-OOD, this strategy aims to perturb the text embedding of the class name in a region relatively far from the embedding. Therefore, given a class name c_k, we first obtain its text embedding $\zeta(c) = \text{CLIP}(c)$ and perturb it using random noise, such that the prompt embedding becomes:

$$\rho(c) = \zeta(c) + \sigma\epsilon\,, \tag{2}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, I)$ and σ is the standard deviation. The aim of this strategy is to create near-anomalies that, while similar to the in-distribution, serve as cues as what means to be a normal object instance. Compared with the generic textual prompts strategy, this method does not require an explicit description of what an OOD object should look like, then avoiding to introduce any bias in the object detector.

3.2. Out-of-distribution Detection

The Dream-Box framework enables an augmented dataset $\mathcal{D} = \mathcal{D}_{in} \cup \mathcal{D}_{ood}$ with images containing in-distribution and

No.	Prompt
1	A {} that defies the laws of physics, floating in mid-air with strange edges.
2	A mechanical {} with organic, plant-like growths intertwining through its structure.
3	A transparent {} that appears to be both solid and liquid at the same time, emitting soft light.
4	A {} that changes its shape continuously, with shifting low contrast colors and textures.
5	A futuristic {} that blends digital and physical elements, glowing with an otherworldly light.
6	A {} with an impossible texture, smooth like liquid but solid like metal, floating in space.
7	A {} that merges two unrelated materials, seamlessly integrating them in an abstract form.
8	A floating {} with intricate geometric patterns constantly changing on its surface.
9	A {} made of plastic that constantly reconfigures itself into different shapes.
10	A {} that appears to be in multiple states at once, existing in two places simultaneously.
11	A strange {} that casts light in ugly but usual colors, transforming its appearance as it moves.
12	A {} that is both solid and ethereal, with strange veins of energy running through it.
13	A {} suspended in time, frozen in mid-motion, with particles of light trailing behind it.
14	A mysterious floating {} with a strange core, surrounded by shifting shadows.
15	A {} made of multiple contrasting materials that somehow coexist harmoniously.
16	A complex {} with multiple layers, each one having a different texture and ugly color that shifts over time.
17	A {} that seems to have multiple dimensions, existing in more than one space at once.
18	A {} with a constantly rotating surface, covered in strange markings and symbols.
19	A {} that looks like it's part of the natural world, but is made entirely of artificial materials.
20	A smooth {} that seems to be melting and reforming simultaneously, surrounded by mist.

Table 1. List of Prompts for the Generic Prompts strategy. {} indicates the name of a class.

OOD objects. Therefore, any standard object detector can be used for the in-distribution object recognition. With regards to the OOD object detection task, a naive approach of simply learning OOD objects as belonging to an additional class, we aim to detect unseen objects as predict them as OOD. Therefore, a supervised approach as such might not generalize to unseen instances since they might look quite different form the generated OOD instances. In this sense, energy-based methods have shown to have a good performance for OOD in object detection [5, 14, 15]. Specifically, the energy score of an object feature representation $\mathbf{v}$ is given by:

$$E(\mathbf{v}) = -\log \sum_{i=1}^{K} \exp(g_k(\mathbf{v}))\,, \tag{3}$$

where K is the number of classes and $g_k(\mathbf{x})$ is the logit of the k-th class. Then, this energy is passed to a small multi-layer perceptron $\phi(E)$ that is trained for in-distribution/OOD classification using binary cross entropy, such that we add the extra loss term to the object detector:

$$\begin{aligned}\mathcal{L}_{ood} = \mathbb{E}_{\mathbf{v}_{in}\sim\mathcal{D}_{in}}\left[-\log\frac{e^{\phi(E(\mathbf{v}_{in}))}}{1+e^{\phi(E(\mathbf{v}_{in}))}}\right] + \\ \mathbb{E}_{\mathbf{v}_{ood}\sim\mathcal{D}_{ood}}\left[-\log\frac{1}{1+e^{\phi(E(\mathbf{v}_{ood}))}}\right]\,.\end{aligned} \tag{4}$$

The outline of this approach is shown in Fig. 3.

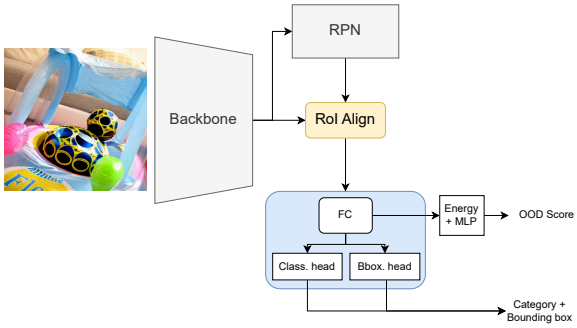

Figure 3. Object detector with modified OOD output (head).

For this work, we use Faster R-CNN [22] with a ResNet50 [11] backbone. Following previous works in energy-based OOD detection [5, 6, 14, 15, 20], the feature representation of an object $\mathbf{v}$ comes from the penultimate layer of the object classification head. Additionally, since the generated outliers are still intended to be detected as objects, the Region Proposal (sub-)Network within the Faster RCNN architectire is trained with these instances as objects (therefore, increasing their objectness score). Nevertheless, to avoid a negative impact on the in-distribution task, the classification and bounding box heads are not trained for OOD instances.

Method	FPR95 (%)	AUROC (%)	mAP (ID) (%)
MSP [12]	70.99	83.45	48.7
ODIN [17]	59.82	82.20	48.7
Mahalanobis [16]	96.46	59.25	48.7
Energy score [20]	56.89	83.69	48.7
Gram matrices [24]	62.75	79.88	48.7
Generalized ODIN [13]	59.57	83.12	48.1
VOS [5]	47.77	89.00	51.5
FFS [15]	44.15	89.71	51.8
Generic Prompts (*Ours*)	59.37	80.43	48.2
Distance-based modified embeddings (*Ours*)	65.03	79.27	49.5

Table 2. Performance comparison of Dream-Box to current OOD techniques for object detection.

3.3. Evaluation and Implementation details

Diffusion model. We use Stable Diffusion fine-tuned for image inpainting. Specifically, the Stable Diffusion v2 model [23] is fine-tuned for 200k epochs using the LaMa strategy for masked generation [25]. Regarding the *distance-based modified prompt* strategy, choosing a standard deviation value in Eq. (2) would require us to know the how far the OOD objects are, which is not known *a priori*. Therefore, we try different standard deviation values, such that $\sigma \in \{0.01, 0.1, 1.0, 2.5, 5.0\}$. We sample $N = 5,000$ images with repetition for each experiment. Finally, since we observed that Stable Diffusion v2 might erase objects when the mask area is too small, we only synthesize new objects whose bounding boxes have an area $A > 2,000$ pixels.

Object Detection. We use MMDetection [2] for training Faster RCNN [22] with a ResNet50 [11] backbone pre-trained on the ImageNet [3]. All of our models are trained using Stochastic Gradient Descent for 18 epochs (following Du *et al*. [5]), with a batch size of 16, weight decay of 1×10^{-5} and initial learning rate of 0.02 that is decreased by a factor of 0.1 after 12 and 16 epochs. The OOD classifier is trained using Focal Loss [19] with a loss weight of 10.0. Each model is trained using a single NVIDIA A100 GPU.

Datasets. We use the PASCAL VOC 2007/12 dataset [7], comprising 20 object classes, as the in-distribution dataset. The OOD dataset consists of the testing partition of the MS-COCO dataset [18], removing all images that contain any of the 20 in-distribution classes, with the final OOD testing dataset consisting of 930 images.

Evaluation Metrics We report the area under the receiver operating characteristic (AUROC) curve for OOD detection, and the false positive rate at 95% confidence (FPR95) of in-distribution detection, *i.e*., the rate of OOD instances labeled as in-distribution given an OOD score threshold such that 95% of in-distribution instances are correctly classified. Additionally, we report average precision (AP) in the in-distribution object detection task to evaluate the effect of Dream-Box in the main task.

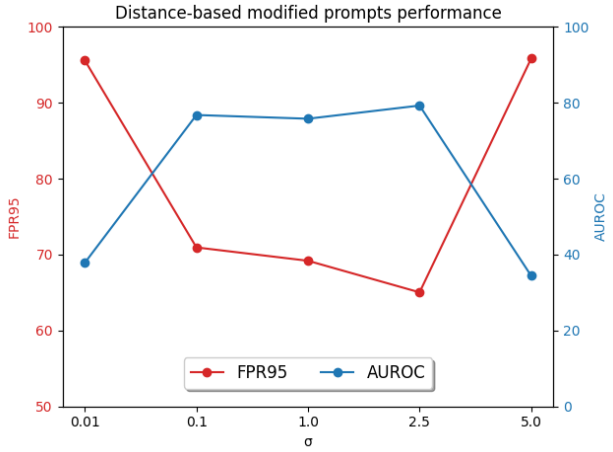

Figure 4. Performance of the distance-based modified prompting strategy with different σ values.

4. Results

Tab. 2 shows the results of our method compared to other standard OOD approaches in object detection. The results shown for the distance-based modified prompt embeddings, corresponding to $\sigma = 2.5$, achieved the best performance. Although our method has a lower performance compared with the state-of-the-art methods (VOS and FFS), it still achieves comparable performance with respect to other traditional approaches, indicating that the generated outliers are providing cues into what are in-distribution samples or not. A key principle in VOS and FFS is that the outliers are generated from the feature representation, therefore learning a compact representation of the feature space and classifying other instances as OOD. This indicates that the feature representation of the objects in Faster RCNN has strong signals of whether an object is part of the distribution or

Figure 5. Exemplar generated outlier objects of the generic prompt and distance-based modified prompting strategies.

not. However, our method provides a sense of explainability by enabling visualization of outliers in the pixel space. Fig. 5 shows the generated outliers for both strategies. The improved performance of the *generic prompts* is directly related to the outliers being closer to the actual object distribution than the *distance-based modified prompt embeddings* approach. For instance, the dog in the third row completely disappears, indicating that the σ value used to generate the embeddings is far from the class embedding. Therefore, the generated images from the *generic prompts* approach indicate that showing images near the in-distribution data but with OOD features helps in OOD detection for object detection.

Ablation studies corresponding to different values of σ in the distance-based modified prompt strategy are presented at Fig. 4, where it is observed that the best performance is achieved at $\sigma = 2.5$. The ablations also show that using too small or too large values is detrimental to Dream-Box performance. For small σ values, the generated outliers might look quite similar to the original in-distribution data, which

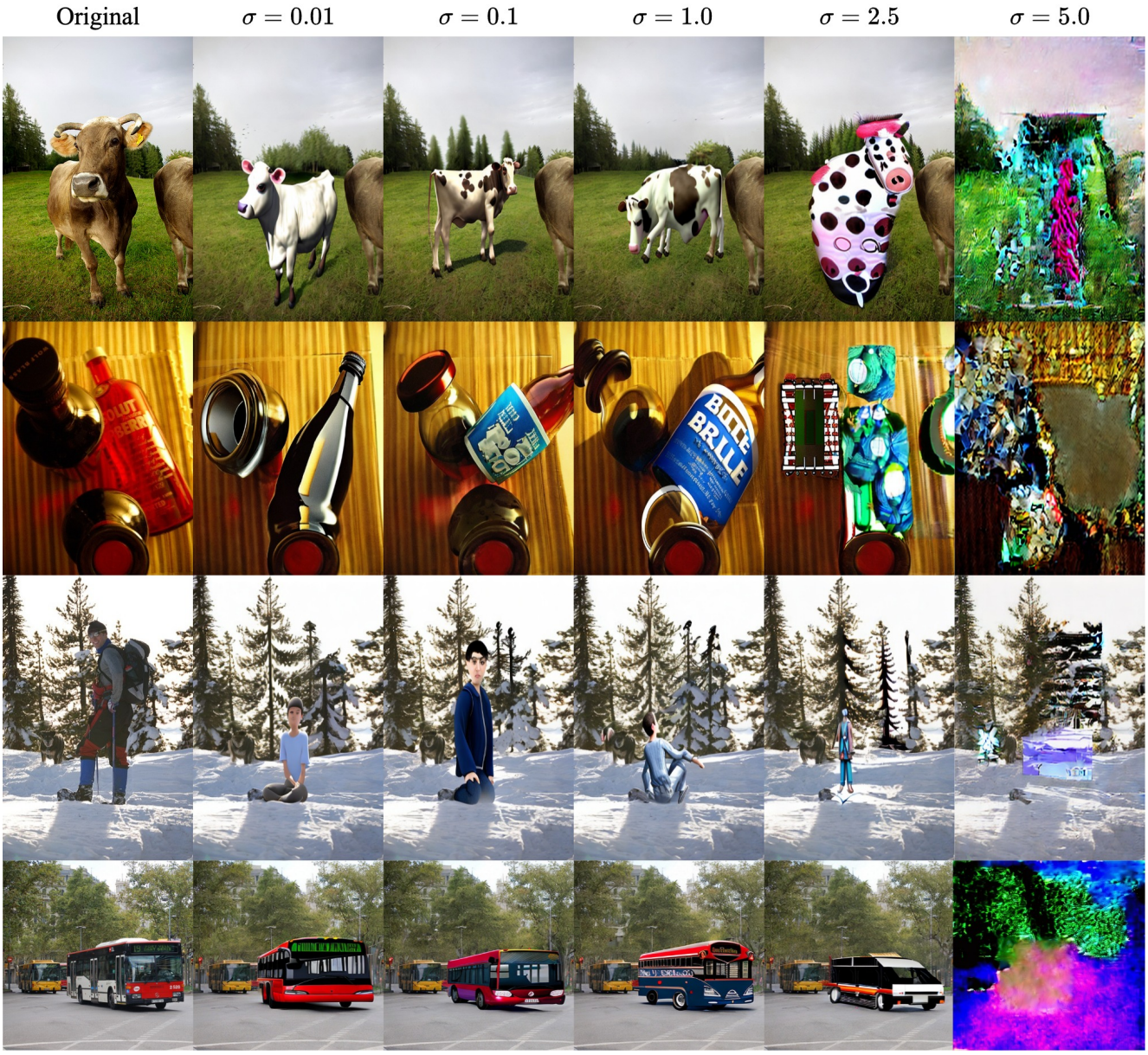

Figure 6. Outlier objecr generation comparison by varying σ in the distance-based modified prompting strategy.

is opposite to the goal of generating synthetic outliers. On the other hand, too large values of σ create very dissimilar objects that might be too far from the object distribution (*i.e.*, they do not look like objects anymore).

This is further supported by Fig. 6, where we show different outliers by varying σ values. For instance, most of the objects with $\sigma \leq 1.0$ are similar to what in-distribution objects look like, whereas for $\sigma = 5.0$ the images break and no longer show identifiable objects. On the other hand, while objects generated with $\sigma = 2.5$ can still be identified as an object with some resemblance to the original class, they present anomalous features that make them ideal candidates for OOD training. Nonetheless, it is also observed that objects might look still like the original class in some instances (such as the bus in the last row), or completely unrecognizable (*e.g.*, the bottle in the second row). This indicates that identifying the proper distance to the class-embedding centre for outlier synthesis might be critical for proper class-aware OOD object synthesis.

Recent works [6, 26] have shown other strategies for OOD sampling without relying to explicit distances from the class name embedding, but their application to object-

wise ODD are as yet un-investigated. Therefore, the study of improved outlier sampling strategies for object-wise OOD remains an area for future work.

5. Conclusion

This work introduces Dream-Box, a novel framework for outlier object generation in the pixel space for object-wise out-of-distribution (OOD) detection. By leveraging diffusion models, our method synthesizes pixel-space outliers, addressing the lack of interpretability which is the key limitation of alternative feature-space approaches. Unlike prior methods such as VOS [5] and FFS [15], which rely on abstract feature representations, Dream-Box provides a more intuitive and explainable means of generating synthetic outliers, thereby enhancing the explainability of OOD detection models.

Experimental evaluations demonstrate that Dream-Box achieves competitive OOD detection performance with traditional OOD approaches, despite not surpassing state-of-the-art feature-space methods. Notably, the generic prompt strategy yields improved OOD classification results compared to distance-based modified prompt embeddings, suggesting that generating outliers closer to the decision boundary contributes positively to detection accuracy. The ability to visualize synthetic outliers offers additional insights into the failure modes of object detection models, an aspect previously unexplored in pixel-space OOD generation for this task. Further analysis shows that while the distance-based modified prompt strategy for outlier generation underperforms the generic prompt strategy, it provides a tunable parameter for controlling the objects anomalous appearance. Additionally, the use of improved sampling methods in the class embedding space may improve such a strategy, although its application to object detection remains unexplored.

These findings underscore the potential of pixel-space outlier generation for interpretable OOD detection in object detection. Future research directions include refining the generation process to better balance in-distribution performance with OOD separability, extending Dream-Box to additional vision tasks, and exploring more adaptive prompt engineering techniques to improve the quality of synthesized outliers, including the exploration of the text embedding space for OOD prompt embeddings, akin Dream-OOD [6].

References

[1] Jack W. Barker, Neelanjan Bhowmik, and Toby P. Breckon. Semi-supervised surface anomaly detection of composite wind turbine blades from drone imagery. In *Proc. Int. Conf. on Computer Vision Theory and Applications*, pages 868–876. IEEE, 2022. 1

[2] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 5

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[5] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *arXiv preprint arXiv:2202.01197*, 2022. 1, 2, 4, 5, 8

[6] Xuefeng Du, Yiyou Sun, Jerry Zhu, and Yixuan Li. Dream the impossible: Outlier imagination with diffusion models. *Advances in Neural Information Processing Systems*, 36:60878–60901, 2023. 1, 2, 3, 4, 7, 8

[7] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2):303–338, 2010. 5

[8] Yona F.A. Gaus, Neelanjan Bhowmik, Brian K.S. Isaac-Medina, Amir Atapour-Abarghouei, Hubert P.H Shum, and Toby P. Breckon. Region-based appearance and flow characteristics for anomaly detection in infrared surveillance imagery. In *Proc. Conf. Computer Vision and Pattern Recognition Workshops*, pages 2995–3005. IEEE/CVF, 2023. 1

[9] Yona F.A. Gaus, Brian K.S. Isaac-Medina, Neelanjan Bhowmik, Yee T. Lam, and Toby P. Breckon. Semi-supervised object-wise anomaly detection for firearm and firearm component detection in x-ray security imagery. In *Proc. Computer Vision Pattern Recognition Workshops*. IEEE/CVF, 2025.

[10] Simon G.E. Gökstorp and Toby P. Breckon. Temporal and non-temporal contextual saliency analysis for generalized wide-area search within unmanned aerial vehicle (uav) video. *The Visual Computer*, 38: 2033–2040, 2021. 1

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 4, 5

[12] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 1, 5

[13] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10951–10960, 2020. 5

[14] Brian KS Isaac-Medina, Yona Falinie A Gaus, Neelanjan Bhowmik, and Toby P Breckon. Towards open-world object-based anomaly detection via self-supervised outlier synthesis. In *European Conference on Computer Vision*, pages 196–214. Springer, 2024. 1, 2, 4

[15] Nishant Kumar, Siniša Šegvić, Abouzar Eslami, and Stefan Gumhold. Normalizing flow based feature synthesis for outlier-aware object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5156–5165, 2023. 1, 2, 4, 5, 8

[16] 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. 1, 2, 5

[17] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017. 2, 5

[18] 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. 5

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 5

[20] 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, 2, 4, 5

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39 (6):1137–1149, 2016. 1, 4, 5

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3, 5

[24] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In *International Conference on Machine Learning*, pages 8491–8501. PMLR, 2020. 1, 2, 5

[25] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 5

[26] Leitian Tao, Xuefeng Du, Jerry Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *The Eleventh International Conference on Learning Representations*, 2023. 7

[27] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7464–7475, 2023. 1

[28] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, pages 23965–23998. PMLR, 2022. 1

[29] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132 (12):5635–5662, 2024. 1

[30] 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. 1