World Feature Simulation: Supplementary Material

Aming Wu, Cheng Deng*
School of Electronic Engineering, Xidian University, Xi'an, China
amwu@xidian.edu.cn, chdeng@mail.xidian.edu.cn

To accelerate the safe application of detectors in real scenarios, the models should be promoted to learn robust discriminative features. This paper explores a new challenging problem, i.e., Open-Domain Unknown Object Detection (ODU-OD), and proposes World Model-based method (i.e., World Feature Simulation (WFS)) for representation learning. Particularly, WFS aims to emulate human beings to leverage perception and memory to imagine unknown characteristics. In Supplementary Material, we will provide more experimental details, experimental analysis, and detection results.

1. More Experimental Details

Datasets. For unsupervised OOD-OD, we adopt PASCAL VOC [2] and Berkeley DeepDrive (BDD-100k) [10] as the ID data for training. Meanwhile, MS-COCO [6] and Open-Images [5] are taken as the OOD datasets to evaluate the trained model. And the OOD datasets are manually examined to guarantee they do not contain ID categories. Meanwhile, for ODU-OD, the training data is kept unchanged. And we only employ the work Clipstyler [8] to render the testing data of unsupervised OOD-OD into three different styles, i.e., 'Acrylic', 'Purple-Brush', and 'Sketch'. Besides, for OSOD, we follow the work [4] and utilize 20 VOC classes and 60 non-VOC classes in COCO to evaluate our method under different open-set conditions. Finally, for Single-DGOD, we follow the settings of the work [9]. And the object detector is trained on the daytime-sunny weather and is tested on the night-sunny, daytime-rainy, night-rainy, and daytime-foggy weather.

Metrics. For ODU-OD and OOD-OD, we report: (1) the false positive rate (FPR95) of OOD objects when the true positive rate of ID objects is at 95%; (2) the area under the receiver operating characteristic curve (AUROC). For OSOD, we use Wilderness Impact (WI) [1] to measure the degree of unknown objects misclassified to known classes. And we also utilize Absolute Open-Set Error (AOSE) [7] to count the number of misclassified unknown objects. For Single-DGOD [9], mean average precision (mAP) is utilized as the metric.

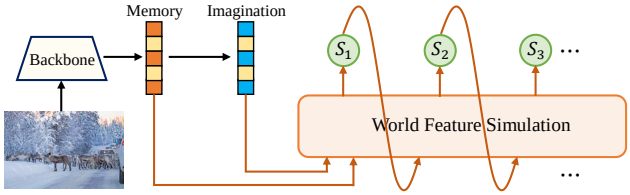

Figure 1. In this paper, World Feature Simulation takes the captured memory and imagination as the input and iteratively synthesizes unknown virtual features.

Implementation Details of OSOD and Single-DGOD. To further demonstrate that our method could simultaneously enhance robustness and discrimination, we verify our method on OSOD and Single-DGOD. Here, our method is directly plugged into three baseline methods and does not calculate the uncertainty loss. The training details are the same as the baselines.

Specifically, in order to sufficiently exploit the synthesized virtual OOD features, we train a binarized classifier, i.e., the output of the known category is 1, and the output of the virtual OOD features is 0. Meanwhile, we still employ a memory recorder to enhance current representation. By these operations and minimizing the cross-entropy loss, the discrimination ability of the object classifier could be strengthened effectively.

2. Further Discussion of Unknown Features

Beyond mere perception, humans possess the uncanny ability to predict the outcomes of their actions, envision potential futures, and abilities that underpin interaction with the world. To this end, world models have emerged as a critical solution that aims to bridge the cognitive divide between human and machine intelligence [3].

As shown in Fig. 1, we consider unknown objects as the future and design a World Feature Simulation for iteratively synthesizing virtual features. Particularly, in Fig. 6 of the submitted paper, we show three visualization examples within unseen domains. We can observe that compared with the calculated memory-enhanced features, the synthesized unknown features mainly pay attention to the background regions. We consider this is reasonable. Compared

*Corresponding author.

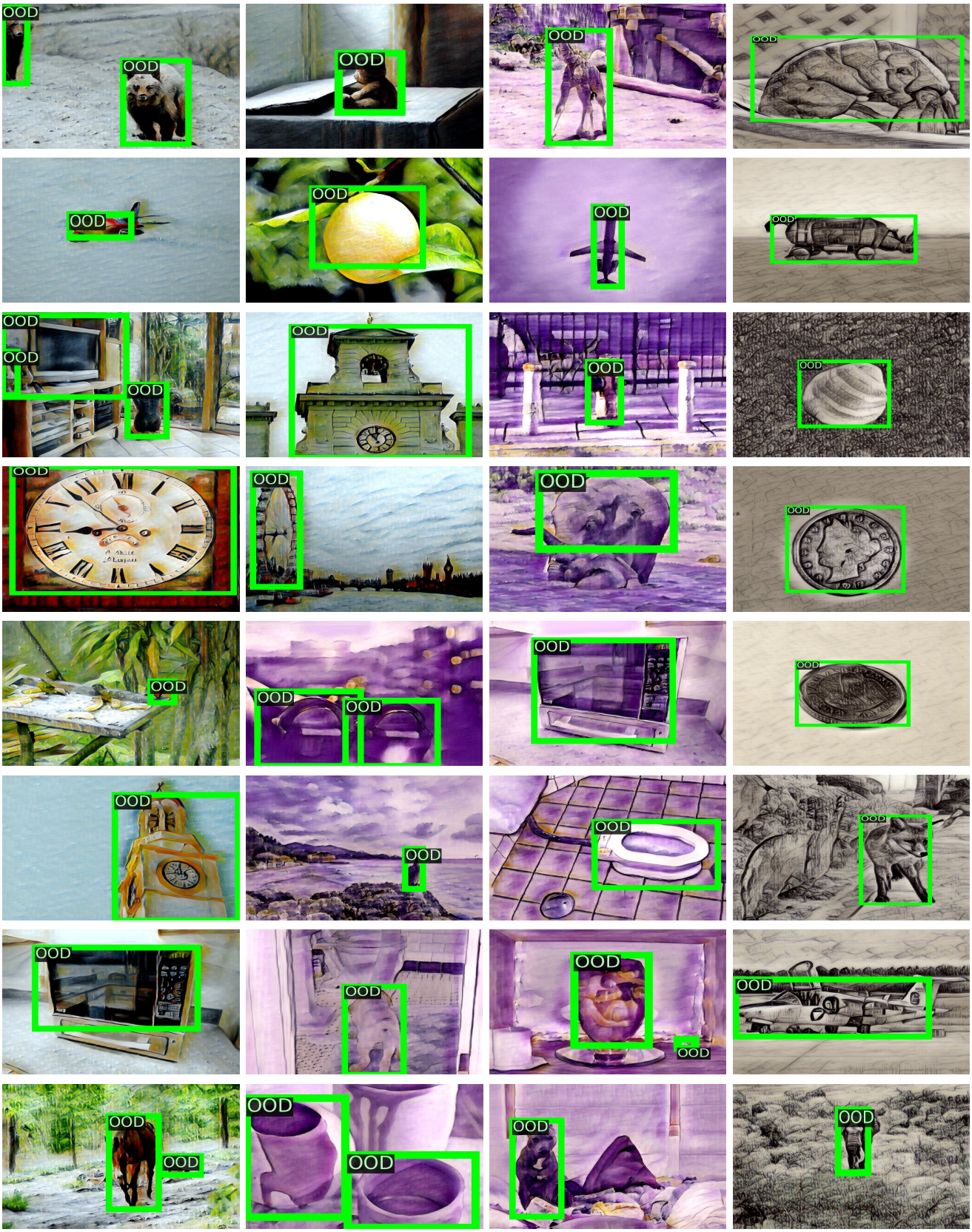

Figure 2. More ODU-OD detection examples based on our method. We can observe that our method accurately distinguishes OOD objects within unseen domains, which shows the effectiveness of our method.

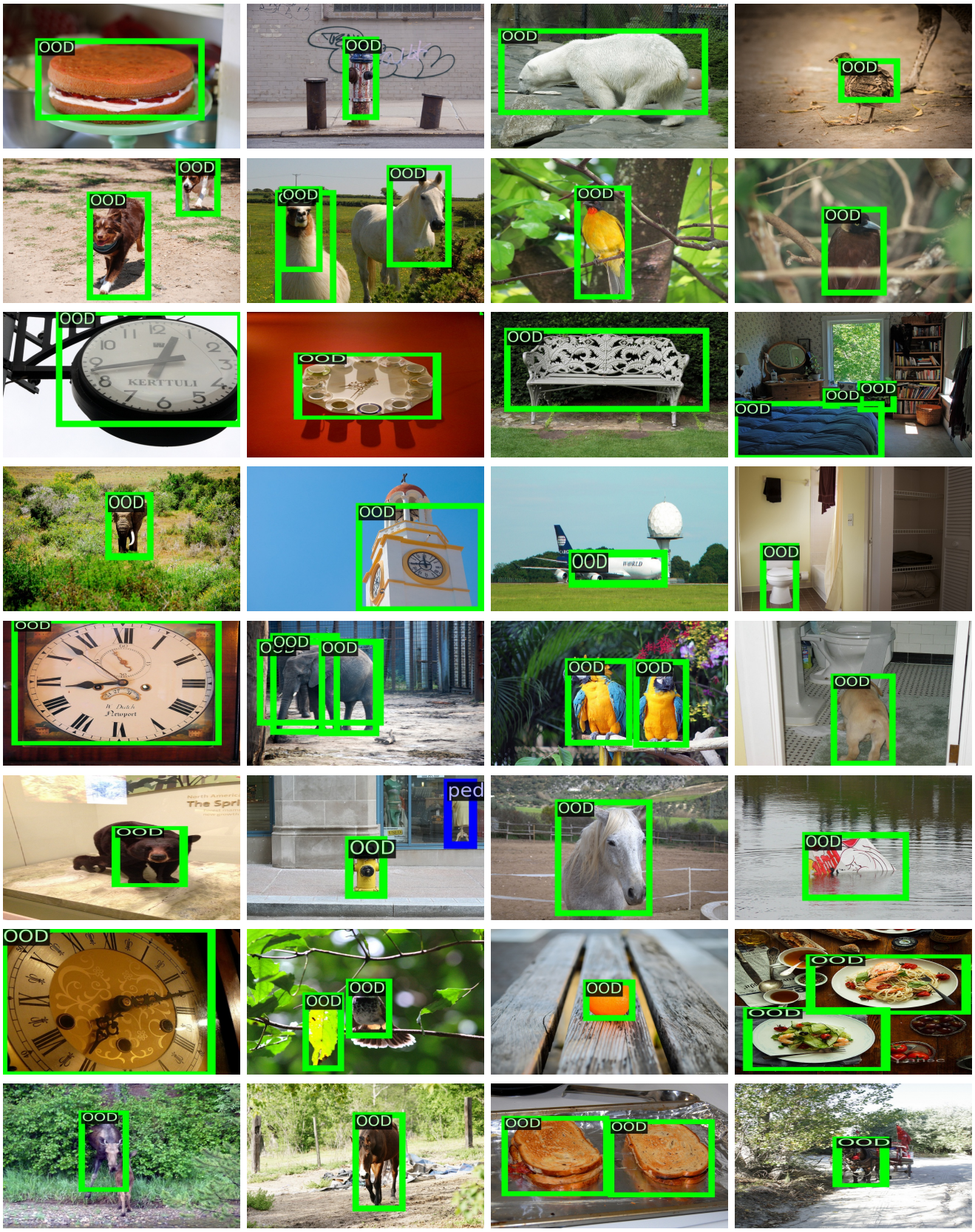

Figure 3. More OOD-OD detection examples based on our method. We can observe that our method accurately distinguishes OOD objects within the nature domains, which shows the effectiveness of our method.

Figure 4. Detection results based on PASCAL VOC. We can see that our method accurately localizes and recognizes objects in these images, e.g., the cow, dog, bird, and person, which shows that our method is effective for in-distribution data.

with the foreground regions, the background regions contain plentiful visual information. Obviously, leveraging the background visual information is instrumental in synthesizing virtual unknown features, improving the ability of detecting unknown objects within unseen domains.

Algorithm 1 World Feature Simulation for ODU-OD

Input: ID data $\{X, Y\}$, randomly initialized detector with parameter φ, randomly initialized modulator with parameter γ and β, randomly initialized unknown-feature generator, weight α for the loss $\mathcal{L}_{contrast}$, weight λ for the uncertainty loss $\mathcal{L}_{\text{unty}}$.
Output: Object detector with parameter φ^*, and OOD detector $\mathcal{C}$.
while *train* **do**
  Sample images from the ID dataset $\{X, Y\}$.
  Perform the multi-level perception using Eq. (1) to obtain F_v.
  Build the memory recorder using Eq. (2) to obtain M.
  Reason the imagination bank using Eq. (3) to obtain $\mathcal{I}$.
  Calculate the variational encoding operation using Eq. (5) to obtain Z and G.
  for t = 1, ..., N **do**
    $z_t = \sigma(W_{zz} * Z_t + W_{oz} * G_t + W_{sz} * S_{t-1} + b_z)$,
    $g_t = \sigma(W_{zg} * Z_t + W_{og} * G_t + W_{sg} * S_{t-1} + b_g)$,
    $\hat{S}_t = \text{Tanh}(W_s * Z_t + W_h * (g_t \odot S_{t-1} + (1 - g_t) \odot G_t))$,
    $S_t = (1 - z_t) \odot S_{t-1} + z_t \odot \hat{S}_t$,
  end
  Calculate the overall training objective $\mathcal{L}$ using Eq. (4), (7), (8), and (9).
  Update the model parameters based on Eq. (9).
end
while *eval* **do**
  Calculate the OOD uncertainty score using the left part of Eq. (10).
  Perform thresholding comparison using the right part of Eq. (10).
end

Besides, since there is no OOD information available, our method could effectively select certain characteristics to synthesize expected OOD features, and it is not a simple spatial division. In Fig. 5, taking the first image in the second row as an example, as it is hard to leverage foot regions to recognize ID objects, the imaged unknown content mainly focuses on these regions.

3. More Ablation Analysis

In Eq. (9), we utilize two hyperparameters, i.e., α and λ, to adjust the contrastive loss and uncertainty loss. Since the uncertainty loss $\mathcal{L}_{\text{unty}}$ is directly related to the current task, the value of λ should be set larger than α. Here, we make an ablation analysis of the two hyper-parameters.

Analysis of α. The hyper-parameter α is used to balance the contrastive losses consisting of $\mathcal{L}_{im}$, $\mathcal{L}_{unk}$, and $\mathcal{L}_{\text{dis}}$. In the experiments, we observe that when α is set to 0.01, 0.001, and 0.0001, the corresponding performance of FPR95 is 81.03%, 80.16%, and 80.83%.

Analysis of λ. In this paper, the hyper-parameter λ in Eq. (9) is to constrain the uncertainty loss $\mathcal{L}_{\text{unty}}$. In the experiments, we observe that when λ is set to 0.5, 0.1, and 0.01, the corresponding performance of FPR95 is 81.64%, 80.16%, and 80.77%.

4. More Detection Results

Algorithm 1 shows the training and evaluation details of our method. By performing multi-level perception and building the memory recorder, the unknown-feature generator could recurrently generate virtual features. With the help of specific constraints, the virtual features are promoted to contain expected characteristics. Finally, in Fig. 2, 3, and 4, we show more detection examples. We can see that our method could accurately localize and recognize OOD objects within unseen and seen domains, demonstrating the superiorities of our WFS method.

References

[1] Akshay Dhamija, Manuel Gunther, Jonathan Ventura, and Terrance Boult. The overlooked elephant of object detection: Open set. In *WACV*, pages 1021–1030, 2020.

[2] 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(2):303–338, 2010.

[3] Yanchen Guan, Haicheng Liao, Zhenning Li, Guohui Zhang, and Chengzhong Xu. World models for autonomous driving: An initial survey. *arXiv preprint arXiv:2403.02622*, 2024.

[4] Jiaming Han, Yuqiang Ren, Jian Ding, Xingjia Pan, Ke Yan, and Gui-Song Xia. Expanding low-density latent regions for open-set object detection. In *CVPR*, pages 9591–9600, 2022.

[5] 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. *International journal of computer vision*, 128(7):1956–1981, 2020.

[6] 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 *ECCV*, pages 740–755. Springer, 2014.

[7] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. Dropout sampling for robust object detection in open-set conditions. In *ICRA*, pages 3243–3249, 2018.

[8] Vidit Vidit, Martin Engilberge, and Mathieu Salzmann. Clip the gap: A single domain generalization approach for object detection. In *CVPR*, pages 3219–3229, 2023.

[9] Aming Wu and Cheng Deng. Single-domain generalized object detection in urban scene via cyclic-disentangled self-distillation. In *CVPR*, pages 847–856, 2022.

[10] 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 *CVPR*, pages 2636–2645, 2020.