DiffGuard: Semantic Mismatch-Guided Out-of-Distribution Detection using Pre-trained Diffusion Models

Ruiyuan Gao[†] Chenchen Zhao[†] Lanqing Hong[‡] Qiang Xu[†]

[†]The Chinese University of Hong Kong [‡]Huawei Noah's Ark Lab

ICCV23 PARIS

Abstract

Given a classifier, the inherent property of **semantic Out-of-Distribution (OOD)** samples is that their contents differ from all legal classes in terms of semantics, namely **semantic mismatch**. As diffusion models are much easier to train and amenable to various conditions compared to cGANs, in this work, we propose to directly use **pre-trained diffusion models** for semantic mismatch-guided OOD detection, named DiffGuard. Specifically, given an OOD input image and the predicted label from the classifier, we try to enlarge the **semantic difference** between the reconstructed OOD image under these conditions and the original input image. We also present several test-time techniques to **further strengthen such differences**. Experimental results show that DiffGuard is effective on both Cifar-10 and hard cases of the large-scale ImageNet, and it can be easily combined with existing OOD detection techniques to achieve state-of-the-art OOD detection results.

Motivations

Task: Given a classifier trained with semantic labels $\mathcal{Y}$, **semantic OOD detection** is to differentiating samples without any semantics in $\mathcal{Y}$.

Semantic mismatch: the contents of semantic OOD samples differ from all legal classes in terms of semantics.

- **semantic mismatch** is the inherent property of semantic OOD samples and is promising for OOD detection.
- **conflict conditions**: conditional GAN can construct semantic mismatch, but is hard to train.
- **Diffusion models** can combine different conditions easily.

Preliminaries

❓ How can diffusion models introduce two conditions?

For label condition **Conditional Diffusion Models** can synthesize images according to semantic conditions with two strategies:

- **classifier guidance** (with a separately trained **noisy classifier** $\log p_\phi$)

$$\hat{\epsilon}(\boldsymbol{x}_t) := \epsilon(\boldsymbol{x}_t) + s\sqrt{1-\alpha_t} \cdot \nabla_{\boldsymbol{x}_t} \log p_\phi(\boldsymbol{y}|\boldsymbol{x}_t), \tag{1}$$

- **classifier-free guidance** (by training conditional diffusion models $\bar{\epsilon}$)

$$\tilde{\epsilon}(\boldsymbol{x}_t, \boldsymbol{y}) := \bar{\epsilon}(\boldsymbol{x}_t, \emptyset) + \omega[\bar{\epsilon}(\boldsymbol{x}_t, \boldsymbol{y}) - \bar{\epsilon}(\boldsymbol{x}_t, \emptyset)], \tag{2}$$

For image condition **The Inversion Problem of Diffusion Models.** Input image as a condition for synthesis can be done by solving the inversion problem. Such a latent can be used to **reconstruct the input** through the denoising process.

$$\boldsymbol{x}_{t+1} = \sqrt{\alpha_{t+1}}\left(\frac{\boldsymbol{x}_t - \sqrt{1-\alpha_t}\epsilon(\boldsymbol{x}_t)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t+1}}\epsilon(\boldsymbol{x}_t), \text{ where } t \in [0, ..., T-1]. \tag{3}$$

Method

As diffusion models are much easier to train and amenable to various conditions compared to cGANs, in this work, we propose to directly use pre-trained diffusion models for semantic mismatch-guided OOD detection.

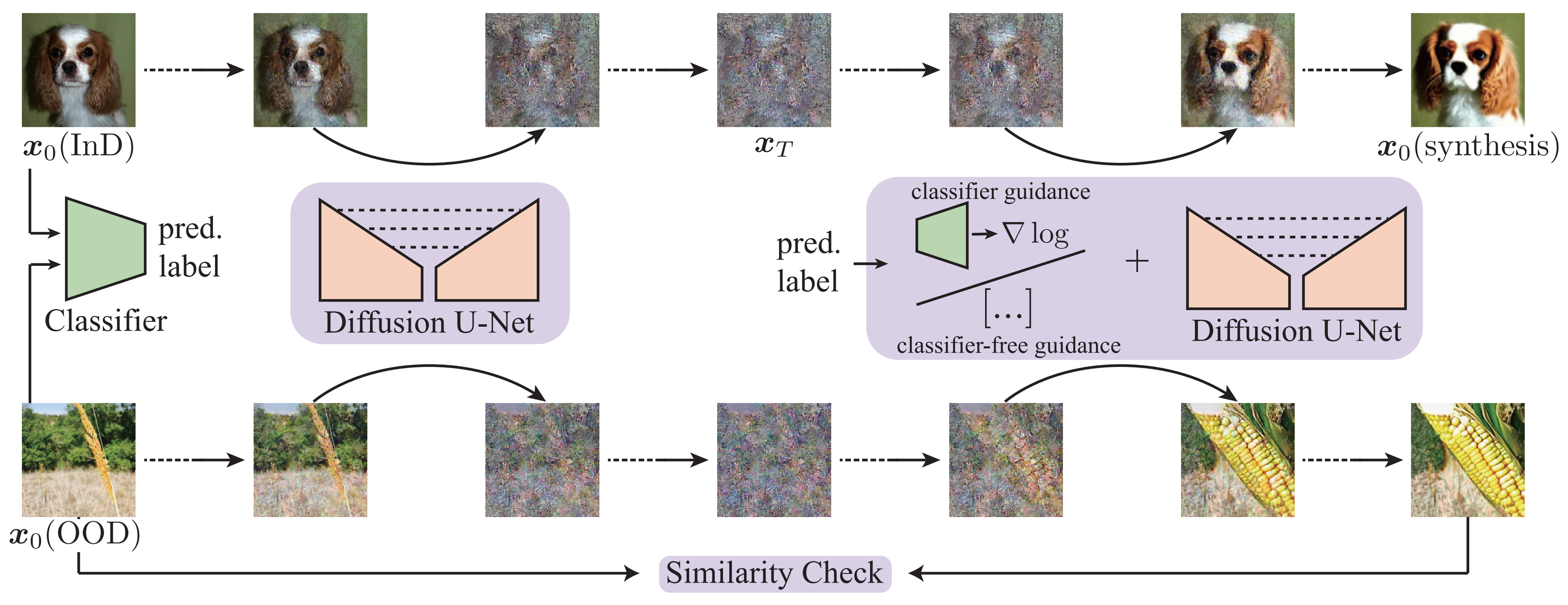


Figure 1. Overview of DiffGuard for OOD Detection. We first use DDIM inversion to get the latent embedding ($\boldsymbol{x}_T$) of the input ($\boldsymbol{x}_0$ left). Then, we apply conditional image synthesis towards the label predicted by the **classifier-under-protection**. Finally, we differentiate OODs based on the similarity between the input and the synthesis.

Classifier guidance

❓ How to use the clean classifier (*i.e.*, **classifier-under-protection**, $\log p_{\phi_n}$) for guidance?

Enhance Label **Tech #1: Clean Grad**. We first change the noisy $\boldsymbol{x}_t$ to a clearer estimation $\hat{\boldsymbol{x}}_0 = \frac{\boldsymbol{x}_t - \sqrt{1-\alpha_t}\epsilon_\theta^{(t)}(\boldsymbol{x}_t)}{\sqrt{\alpha_t}}$ and use the gradient given be the normal classifier for guidance:

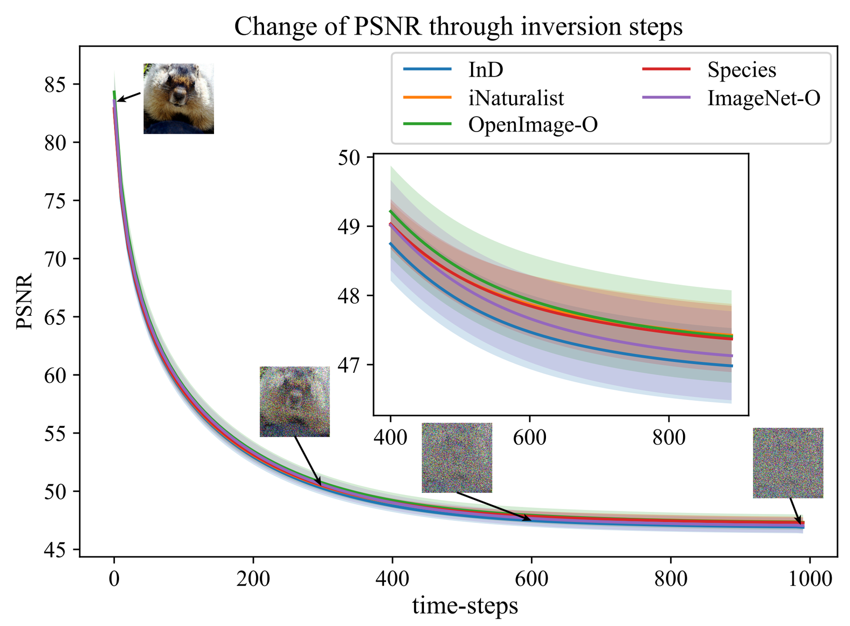
$$\nabla_{\boldsymbol{x}_t} \log p_\phi(y|\boldsymbol{x}_t) := \nabla_{\boldsymbol{x}_t} \log p_{\phi_n}(y|\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t))$$

Weaken Image **Tech #2: Adaptive early-Stop**. We early stop the diffusion process by measuring current noisy level, shown in Fig. 2a.

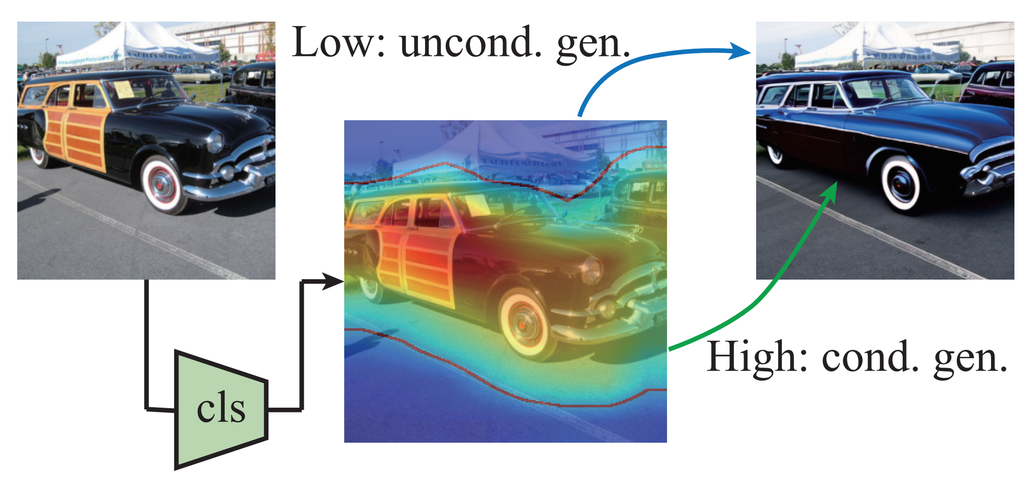
Classifier-free guidance

❗ Consider the information from **classifier-under-protection**.

Enhance Image **Tech #3: Distinct Semantic Guidance** (DSG). We use GradCAM to balance the fidelity and controllability of generation, shown in Fig. 2b.



(a) The effect of adaptive early-stop.



(b) We use GradCAM for classifier-free guidance.

Figure 2. Different techniques in DiffGuard.

Results

Method	Species		iNaturalist		OpenImage-O		ImageNet-O		Average Over 4 OODs	
	AUROC ↑	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑	FPR@95 ↓
EBO	72.04	82.33	90.61	53.83	89.15	57.10	41.91	100.00	73.43	73.31
KNN	76.38	76.19	85.12	68.41	86.45	57.56	75.37	84.65	80.83	71.70
ViM	70.68	83.94	88.40	67.85	89.63	57.56	70.88	85.30	79.90	73.66
MLS	72.89	80.87	91.15	50.80	89.26	57.11	40.85	100.00	73.54	72.20
Ours(GDM)	73.19±0.18	83.68±0.22	85.81±0.16	71.23±0.54	82.32±0.30	74.80±0.38	65.23±0.19	87.74±0.20	76.64±0.13	79.36±0.12
Ours(LDM)	65.87	91.70	75.64	79.06	73.92	81.19	68.57	84.35	71.00	84.08
Ours(GDM)+KNN	**77.81**+1.43	71.04-5.15	90.19+5.07	48.79-19.62	87.80+1.35	52.80-4.76	**75.68**+0.31	**80.85**-3.80	**82.87**+2.04	63.37-8.33
Ours(GDM)+ViM	74.48+3.80	72.26-11.68	92.50+4.10	39.09-28.76	**91.11**+1.48	45.02-12.54	72.42+1.54	82.30-3.00	82.63+2.73	59.67-14.00
Ours(LDM)+ViM	71.08+0.40	82.20-1.74	89.39+0.99	61.01-6.84	89.65+0.02	55.83-1.73	74.85+3.97	81.95-3.35	81.24+1.35	70.25-3.41
Ours(GDM)+MLS	75.95+3.06	**70.31**-10.56	**93.03**+1.88	**30.74**-20.06	90.74+1.48	**40.61**-16.50	65.72+24.87	87.05-12.95	81.36+7.82	**57.18**-15.02
Ours(LDM)+MLS	73.69+0.80	75.91-4.96	91.55+0.40	43.56-7.24	89.61+0.35	50.61-6.50	69.33+28.48	84.00-16.00	81.05+7.51	63.52-8.68

Table 1. The OOD detection performance with ImageNet as InD. GDM uses classifier guidance, while LDM uses classifier-free guidance. All the values are in percentages. ↑/↓ indicates that a higher/lower value is better. The best results are in **bold**. We highlight the comparisons with colors when combining DiffGuard with other baselines. For AUROC with Ours(GDM), we present the average and standard deviation over four runs. There is no randomness in LDM.

Figure 3. Visualization for InD and OOD cases with their syntheses according to the predicted labels. Images are from the ImageNet benchmark. We use LDM in this figure, i.e., classifier-free guided diffusion. We can identify a clear similarity difference between InDs and OODs by comparing the inputs with their syntheses.

Resources

Code ArXiv About Author Our Group