Category-Extensible Out-of-Distribution Detection via Hierarchical Context Descriptions

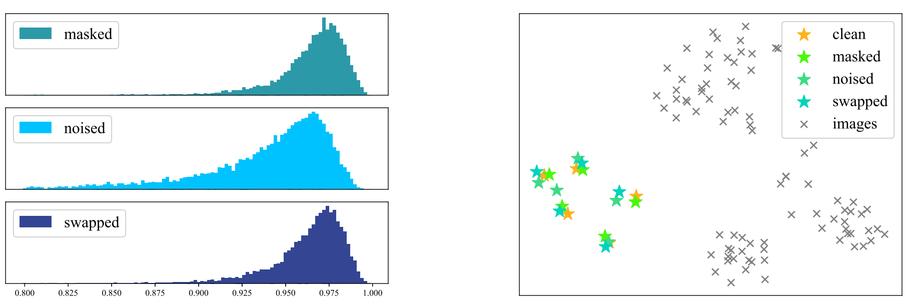
Supplementary Materials

A Implementation Details

A.1 Perturbation Guidance

In the manuscript, we present a perturbation-guided approach to synthesize spurious samples to train the hierarchical contexts. Recall that for arbitrary k-th ID category, we use the spurious context to explicitly describe a corresponding spurious category, and a critical consideration is how to synthesize training samples spurious to that k-th ID category. Recently, generating adversarial data samples have been widely studied, including GAN networks [19, 14], diffusion models [22, 23], image attacks [17, 38], and feature-space sampling [7, 26]. For simplicity, we just take the tractable feature-space sampling as NPOS [26] to generate spurious candidates. In practice, we calculate the k-NN distance for each ID sample in the specific category, and generate spurious candidates by sampling from a multivariate Gaussian distribution around those samples with largest distances (basically away from the clustering center). Then, we leverage the perturbed descriptions of perceptual context to guide the further filtering for high-quality spurious syntheses.

Given the perceptual context $\mathbf{v}_k^p = [v_{k,1}^p; v_{k,2}^p; \cdots; v_{k,m}^p])$ of k-th ID category, we randomly apply a perturbation u onto one arbitrary $v_{k,j}^p$ to produce a perturbed description $\mathring{\mathbf{w}}_k^p$ through text-encoder. Intuitively, there are two ways to perturb a context $\mathbf{v}_k^p$: erasing or replacing the specific visual character $v_{k,j}^p$. Consequently, we design three types of perturbation: (1) masking with a placeholder $u = [\texttt{MASK}]$, (2) noise from a Gaussian distribution $u = \sigma$, and (3) swapping with another category $u = v_{k',j'}^p$. And the perturbed text-feature is produced by: $\mathring{\mathbf{w}}_k^p = \mathcal{T}([v_{k,1}^p; \cdots; u; \cdots; v_{k,m}^p; \texttt{CLS}_k])$. We also conduct empirical experiments to verify the effectiveness of those perturbations.



(a) similarity distribution. (b) t-SNE visualization

Figure A1: **Statistics of perturbations**, including (a) similarities between original and perturbed text-features, and (b) distribution of original text-feature, perturbed text-features, and image-features.

As shown in Fig. A1, all of the perturbed text-features $\mathring{\mathbf{w}}_k^p$ slightly deviate from the original $\mathbf{w}_k^p$ while keep the affinity (*e.g.*, shares a 97% similarity against the original one.) Specifically, the noised $\mathring{\mathbf{w}}_k^p$ leads to a greater deviation, since the noised visual character $v_{k,j}^p := u$ is more unpredictable than the masked or swapped ones.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

In addition, now that every perturbation can directly produce the description (*i.e.*, text-feature) of an unknown spurious category, one may try to take the perturbed description as a substitute for the learned spurious contexts to execute OOD detection. That is, use the perturbed $\mathring{\mathbf{w}}_k^p$ to replace the learned $\mathbf{w}_k^s$ in Eq.(5) in the manuscript. And the results are shown in Tab. A1, where ImageNet-100 [18] is the ID dataset. Given a baseline model [37] learned with perceptual contexts only, simply using the perturbed descriptions (denoted as +*Perturb-Desc.*) brings slight improvements (*e.g.*, 0.2% decrease on FPR95). The insignificant advantage is not due to the limited capacity of only one perturbed description for each ID category. Because ensembling [18] several perturbed descriptions for an ID category at once (denoted as +*Perturb-Ensem.*) dose not bring remarkable improvements. In contrast, our proposed CATEX can significantly enhance the OOD detection performance, which demonstrates it is still necessary to explicitly learn the spurious contexts for each ID category.

Table A1: Comparison with directly using perturbed descriptions for OOD detection.

Method	FPR95↓	AUROC↑
baseline [37]	13.07	97.42
+Perturb-Desc.	12.84	97.43
+Perturb-Ensem.	12.87	97.45
CATEX (Ours)	**10.31**	**97.82**

A.2 Cross-ID-Domain Generalization

As indicated in the manuscript, the precise category boundary learned by our method shows robust OOD performance when the ID data is shifted. In fact, the shifted ID classification can be further boosted by our proposed integrated inference strategy (Eq.(5) in the manuscript), as shown in Tab. A2. It implies the regularization item γ successfully modulates the relative similarities between input images and learned perceptual descriptions for each category, leading to more precise boundaries.

Table A2: Additionally improved ID accuracy on shifted datasets.

Method	Target Datasets		
	ImageNet-A	ImageNet-R	ImageNet-Sketch
CATEX	50.87	76.67	48.59
+IntegInfer.	**50.98**	**76.72**	**48.65**

However, our method only takes the secondary place on ImageNet-Sketch [29] on both ID classification (inferior to NPOS [26]) and OOD detection (inferior to MCM [18]). It is mainly because of the huge domain gap between vanilla ImageNet-1K [4] and shifted ImageNet-Sketch. As shown in Fig. A2, compared to the shifted ImageNet-A [10] and ImageNet-R [13], images from ImageNet-Sketch only preserve objects' shape and main texture, while the color information is totally vanished. We leave the generalization to heavily-shifted ID datasets as future work.

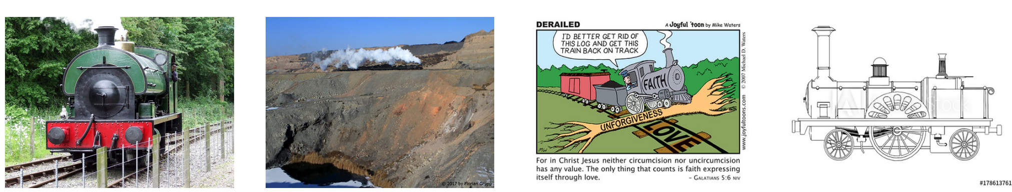


Figure A2: Left to right: examples from ImageNet, ImageNet-A, ImageNet-R, and ImageNet-Sketch.

A.3 Cross-ID-Task Generalization

To verify the efficacy of our proposed framework, we conduct a category-extended experiment in Sec.4.1 and Tab.3. Here more implementation details are provided for reproducibility.

Given two models independently trained on the separated ImageNet-100 (I) and ImageNet-100 (II), how to test them on the union ImageNet-200 (I ∪ II) with our CATEX is simple. In the vision-language prompt-tuning framework, the image-encoder $\mathcal{I}$ and text-encoder $\mathcal{T}$ are frozen, and we only learn the perceptual and spurious contexts (*i.e.*, $\mathbf{v}^p$ and $\mathbf{v}^s$). And the l_2-normalized text-feature can be pre-extracted with the 100 category names in each subset, taking the perceptual descriptions for example, which are denoted as $\{\mathbf{w}^p_{\mathtt{I},k} = \mathcal{T}(\mathbf{v}^p_{\mathtt{I},k}; \mathtt{CLS}_{\mathtt{I},k})\}_{k=1}^{100}$ and $\{\mathbf{w}^p_{\mathtt{II},k} = \mathcal{T}(\mathbf{v}^p_{\mathtt{II},k}; \mathtt{CLS}_{\mathtt{II},k})\}_{k=1}^{100}$. During inference, one may concatenate the 200 text-features together as $\{\mathbf{w}^p_k\}_{k=1}^{200}$. Given an input image I, the l_2-normalized image-feature is extracted by $\mathbf{x} = \mathcal{I}(I)$, and the perceptual image-text similarities are computed as $\mathbf{s}^p = [\langle \mathbf{w}^p_1, \mathbf{x}\rangle, \langle \mathbf{w}^p_2, \mathbf{x}\rangle, \cdots, \langle \mathbf{w}^p_{200}, \mathbf{x}\rangle] \triangleq [s^p_1, s^p_2, \cdots, s^p_{200}]$. Similarly, the spurious similarities become $\mathbf{s}^s = [s^s_1, s^s_2, \cdots, s^s_{200}]$. Then we can leverage the measurement defined in Eq.(5) for both ID classification and OOD detection.

As for the competitors, (*e.g.*, VOS [7] and NPOS [26]), two image-encoders are trained separately (denoted as $\mathcal{I}_{\mathtt{I}}$ and $\mathcal{I}_{\mathtt{II}}$). And for each input image I, there are two corresponding image-features: $\mathbf{x}_{\mathtt{I}} = \mathcal{I}_{\mathtt{I}}(I)$ and $\mathbf{x}_{\mathtt{II}} = \mathcal{I}_{\mathtt{II}}(I)$. Consequently, there also two sets of image-text similarity vector: $\mathbf{s}_{\mathtt{I}} = [\langle \mathbf{w}_1, \mathbf{x}_{\mathtt{I}}\rangle, \langle \mathbf{w}_2, \mathbf{x}_{\mathtt{I}}\rangle, \cdots, \langle \mathbf{w}_{200}, \mathbf{x}_{\mathtt{I}}\rangle] = \{\langle \mathbf{w}_k, \mathbf{x}_{\mathtt{I}}\rangle\}_{k=1}^{200}$ and $\mathbf{s}_{\mathtt{II}} = \{\langle \mathbf{w}_k, \mathbf{x}_{\mathtt{II}}\rangle\}_{k=1}^{200}$ (the superscript ^{p} is hidden for simplicity). For compatibility, we choose the one for ID classification and OOD detection according to its highest image-text similarity. $\mathbf{s} = \begin{cases} \mathbf{s}_{\mathtt{I}} & max(\mathbf{s}_{\mathtt{I}}) > max(\mathbf{s}_{\mathtt{II}}) \\ \mathbf{s}_{\mathtt{II}} & \mathtt{otherwise} \end{cases}$. Now, the performance of our method and other rivals are evaluated under the same measurements.

Note that since we only take one image encoder throughout, the inference time is fixed (because the text-features can be pre-extracted). In contrast, applying other methods brings multiple time cost (*e.g.*, twice slower than ours in this case). When the training subsets extend intensely (*e.g.*, from ImageNet-1K to ImageNet-21K in our manuscript), our method still keeps a fast speed (*e.g.*, 100FPS on V100) during inference, which can even enable real-time applications in practice.

A.4 Error Bars

To verify the robustness, we repeat the training of our method and the rivals on ImageNet-100 [18] with CLIP-B/16 for 3 times, and the results are shown in Tab. A3. Our CATEX consistently outperforms the rivals on OOD detection by a significant margin.

Table A3: Error Bars on ImageNet-100 after 3 runs

Method	ACC↑	FPR95↓	AUROC↑
MSP [8]	**94.77** (±0.05)	41.90 (±0.61)	93.38 (±0.05)
Energy [16]	**94.77** (±0.05)	31.89 (±0.50)	94.53 (±0.18)
VOS [7]	94.75 (±0.07)	24.48 (±0.71)	96.04 (±0.36)
NPOS [26]	94.34 (±0.12)	17.32 (±0.87)	96.46 (±0.13)
CATEX	94.11 (±0.03)	**10.97** (±0.79)	**97.75** (±0.07)

A.5 Software and Hardware

We use Python 3.7.13 and PyTorch 1.8.1, and 2 NVIDIA V100-32G GPUs.

B Experiments on CIFAR Benchmarks

To further verify the robustness of our method, we conduct additional experiments on CIFAR-10 and CIFAR-100 datasets, and evaluate the OOD detection performance on SCOOD [34] benchmark. We train our CATEX for 20 epochs, and the other settings are the same as Sec.4 in the manuscript. The results are shown in Tab. A4 and Tab. A5, where "Surr." means the extra TinyImages80M [27] is adopted for surrogate OOD training set. Accordingly, our CATEX consistently ourperforms the competitors as well, and even surpasses those who adopts the extra OOD training data. It implies the pre-trained knowledge for large-scale CLIP [20] model leveraged by our method is capable enough to detect the OOD samples in the open-world. The efficacy of our CATEX is further demonstrated.

Table A4: Performance on CIFAR-10.

Method	Surr.	ID-ACC ↑	FPR95↓	AUROC↑
MCM [18]	✗	90.79	23.14	94.68
ODIN [15]	✗	95.36	52.00	82.00
Energy [16]	✗	95.36	50.03	83.83
OE [12]	✓	94.90	50.53	88.93
UDG [34]	✓	94.71	36.22	93.78
CATEX	✗	**95.57**	**21.17**	**95.33**

Table A5: Performance on CIFAR-100.

Method	Surr.	ID-ACC ↑	FPR95↓	AUROC↑
MCM [18]	✗	66.91	71.93	79.39
ODIN [15]	✗	81.84	81.89	77.98
Energy [16]	✗	81.84	83.66	79.31
OE [12]	✓	81.31	80.06	78.46
UDG [34]	✓	80.89	75.45	79.63
CATEX	✗	**81.99**	**67.95**	**84.04**

C Combination with Post-hoc Enhancements

Recently, post-hoc OOD detection methods that enhance the single-vision-modal networks (*e.g.*, ResNet [9] and ViT [6]) have been widely studied [3, 24, 25, 39, 5]. In this section, we make a step towards combining vision-language models with previous post-hoc enhancements for better OOD performance. The results are shown in Tab. A6, where ReAct [24] achieves a remarkable improvement. It indicates that pruning the extreme feature values according to the unified distributional statistics may be more suitable for VLMs to reduce the overconfidence on OOD samples. We hope this can bring new insights to the community.

Table A6: Combination with post-hoc methods.

Cmobine	ImageNet-100		ImageNet-1K	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑
None	10.31	97.82	29.66	93.48
ReAct [24]	**10.06**	97.82	**27.56**	**93.77**
BATS [39]	10.16	**97.84**	29.37	93.59
ASH [5]	10.19	97.81	29.14	93.27

D Additional Analysis

Performance improvement. To further evaluate the improvement brought by our method (*e.g.*, 8% decrease of FPR95 against NPOS), we conduct a comparative experiment on ImageNet-1K. To provide a unified analysis across two models, we take a third-party ResNet-50 model [32] (pre-trained on ImageNet-1K classification only) to produce the Maximum SoftMax Probability for each OOD sample that is correctly detected by our CATEX while wrongly viewed as ID samples by NPOS. According to Fig. A3, our method consistently improves the OOD detection on each interval, where the high-probability OOD (generally hard samples) detection is significantly enhanced. It indicates that properly leveraging the prior knowledge from pre-trained VLMs can alleviate the OOD problem when the fine-tuned visual features are indistinguishable, which is consistent with our motivation.

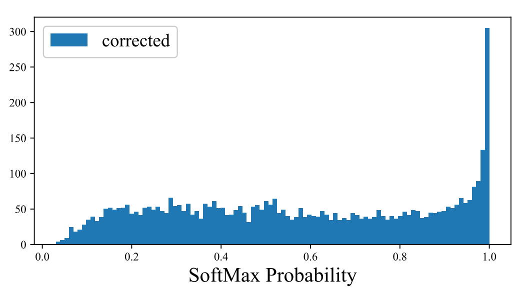


Figure A3: Corrected OOD detections compaired with NPOS. The softmax probability predictions on those OOD samples are produced by another pre-trained ResNet-50 [32] classifier.

Failure cases. As our method still gets 29% FPR95 on ImageNet-1K, we provide some failure cases in Fig. A4, which can be summarized into three kinds:

- Noisy label, where the ID objects (*e.g.*, *dam*) also exists in some OOD images from the test set. And the dataset composition may need a further examination.
- Similar texture, shared by some OOD samples (*e.g.*, *flower*) against ID images (*e.g.*, *starfish*), and the pre-trained encoders of CLIP are unable to distinguish their features. Applying image-level spurious OOD syntheses (*e.g.*, image attacks [17, 38]) may reduce the texture-bias.
- Same background (*e.g.*, *sky*) that seizes a large proportion of the image may lead to similar feature representations. Adopting image-level automatic masking techniques [1, 35] to synthesize spurious OOD samples may alleviate such problem.

Similar failure cases are also observed in recent SOTA methods, which reveal the unsolved challenges of OOD detection and suggest the potential directions for future works.

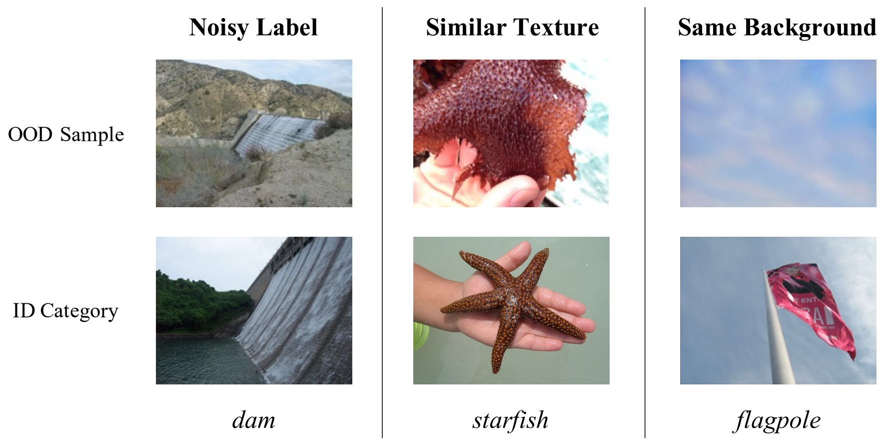


Figure A4: Failed OOD detections of our CATEX

E Datasets and Baselines

For reproducibility, we present the details of datasets and baselines as follows.

ImageNet-100 (I). Following MCM [18], we take the randomly-sampled 100 classes from ImageNet-1K [4] as the ImageNet-100 (I) subset, which contains the following categories: n03877845, n03000684, n03110669, n03710721, n02825657, n02113186, n01817953, n04239074, n02002556, n04356056, n03187595, n03355925, n03125729, n02058221, n01580077, n03016953, n02843684, n04371430, n01944390, n03887697, n04037443, n02493793, n01518878, n03840681, n04179913, n01871265, n03866082, n03180011, n01910747, n03388549, n03908714, n01855032, n02134084, n03400231, n04483307, n03721384, n02033041, n01775062, n02808304, n13052670, n01601694, n04136333, n03272562, n03895866, n03995372, n06785654, n02111889, n03447721, n03666591, n04376876, n03929855, n02128757, n02326432, n07614500, n01695060, n02484975, n02105412, n04090263, n03127925, n04550184, n04606251, n02488702, n03404251, n03633091, n02091635, n03457902, n02233338, n02483362, n04461696, n02871525, n01689811, n01498041, n02107312, n01632458, n03394916, n04147183, n04418357, n03218198, n01917289, n02102318, n02088364, n09835506, n02095570, n03982430, n04041544, n04562935, n03933933, n01843065, n02128925, n02480495, n03425413, n03935335, n02971356, n02124075, n07714571, n03133878, n02097130, n02113799, n09399592, n03594945.

ImageNet-100 (II). Disjoint from ImageNet-100 (I), ImageNet-100 (II) contains another 100 classes randomly sampled from ImageNet-1K: n02096177, n03769881, n01629819, n04033995, n04357314, n02101388, n02328150, n03729826, n02655020, n01985128, n02109525, n07715103, n02099429, n04517823, n02088632, n03207743, n03657121, n02948072, n02106662, n01631663, n09229709, n03793489, n03776460, n07860988, n02129604, n03483316, n02107574, n07716358, n04208210, n02107908, n04372370, n02119022, n12144580, n01693334, n04548280, n03785016, n03535780, n03599486, n02859443, n04335435, n02110341, n03902125, n04146614, n01774750, n03314780, n03045698, n01697457, n02869837, n02276258, n04081281, n03956157, n02487347, n04311174, n02094114, n04409515, n03028079, n03384352, n04532106, n02087394, n04612504, n02100583, n11939491, n02107142, n01669191, n12998815, n04522168, n02894605, n03529860, n10148035, n01677366, n03775071, n03208938, n04238763, n02363005, n02804414, n02106382, n03950228, n02128385, n02028035, n04099969, n02481823, n01729322, n02939185, n02483708, n04162706, n03857828, n02093647, n02927161, n03160309, n02840245, n03920288, n07871810, n04404412, n03947888, n04509417, n02086910, n02256656, n02412080, n02410509, n03584829.

ImageNet-21K. The ImageNet-21K dataset on which we conduct the category-extended experiment is the official winter 2021 released version [1]. For pre-processing, we follow Ridnik *et al* [21] to clean invalid classes, allocating 50 images per class for validation, and crop-resizing all the images to 224 resolution. Training settings are the same as Sec.4 in our manuscript.

OOD datasets. Following the literature [30, 26, 31, 18], we mainly consider subsets of `iNaturalist` [28], `SUN` [33], `Places` [36], and `Texture` [2] as the OOD datasets, which contains 35640 images in total.

Baselines. To evaluate the baselines on our experiment settings, we re-implement the most representative and relevant methods, including MSP [11, 8], Energy [16], VOS [7], and NPOS [26]. For a fair comparison, we train all the baselines with NPOS's codebase [2], and only fine-tune the last two transformer blocks of image encoder [26].

- For MSP and Energy, we train a single model with standard cross-entropy loss function for ID classification only, and infer with respective OOD metrics.
- For VOS, we take the likelihood-based sampling strategy to generate spurious OOD syntheses, and train the model with uncertainty regularization as suggested [7].
- For NPOS, we take the non-parametric distance-based sampling strategy to generate spurious OOD syntheses, and train the model with open-set ERM as suggested [26].

References

[1] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023.

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

[3] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR, 2019.

[4] 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. IEEE, 2009.

[5] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *International Conference on Learning Representations*, 2023.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Machine Learning*, 2021.

[7] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. In *International Conference on Learning Representations*, 2022.

[8] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34:7068–7081, 2021.

[9] 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 (CVPR)*, June 2016.

[10] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.

[11] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

[1] https://image-net.org/

[2] https://github.com/deeplearning-wisc/npos

[12] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *Proceedings of the International Conference on Learning Representations*, 2019.

[13] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.

[14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[15] 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.

[16] 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.

[17] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[18] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *Advances in Neural Information Processing Systems*, 2022.

[19] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[20] 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.

[21] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.

[22] 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.

[23] 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.

[24] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems*, 2021.

[25] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 691–708. Springer, 2022.

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

[27] Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.

[28] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[29] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019.

[30] Haotao Wang, Aston Zhang, Yi Zhu, Shuai Zheng, Mu Li, Alex J Smola, and Zhangyang Wang. Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition. In *International Conference on Machine Learning*, pages 23446–23458. PMLR, 2022.

[31] Qizhou Wang, Junjie Ye, Feng Liu, Quanyu Dai, Marcus Kalander, Tongliang Liu, HAO Jianye, and Bo Han. Out-of-distribution detection with implicit outlier transformation. In *The Eleventh International Conference on Learning Representations*, 2023.

[32] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

[33] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

[34] Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8301–8309, 2021.

[35] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023.

[36] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

[37] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[38] Yao Zhu, Yuefeng Chen, Xiaodan Li, Kejiang Chen, Yuan He, Xiang Tian, Bolun Zheng, Yaowu Chen, and Qingming Huang. Toward understanding and boosting adversarial transferability from a distribution perspective. *IEEE Transactions on Image Processing*, 31:6487–6501, 2022.

[39] Yao Zhu, YueFeng Chen, Chuanlong Xie, Xiaodan Li, Rong Zhang, Xiang Tian, Yaowu Chen, et al. Boosting out-of-distribution detection with typical features. In *Advances in Neural Information Processing Systems*, 2022.