A. Related Works

Another related topic is **Meta-learning** [17], also called *learning-to-learn*, which uses a unique paradigm where a model is trained over a variety of related learning episodes for the benefits of future learning tasks. Such learning-to-learn has been argued to be better aligned with human and animal learning by improving learning on an evolutionary scale [7]. Meta-learning primarily involves learning at two different levels [11]: one of an *inner* level where a new task is presented for quick learning of the training agent, and another of an *outer* level where knowledge has accumulated across earlier inner tasks. Concerned with different learning strategies of inner and outer tasks, Meta-learning approaches can be categorized into three kinds [9]: optimization-based [6], metric-based [10] or model-based [15], with a wide range of applications on few-short learning [4, 10], incremental learning [13], reinforcement learning [21] and domain generalization [2].

Model Agnostic Meta-Learning (MAML) [5] is an optimization-based meta-learning method proposed for fast adaptation of deep neural networks on a small amount of training samples. In OOD detection, one recent work [12] few short *static* OOD detection by directly applying MAML, which is fundamentally different from our proposal in solving CAOOD. Inspired by the idea of learning-to-learn, we link the CAOOD problem to meta-learning by regarding dynamic adaptations over continuously shifting distributions as the inner tasks.

B. CAOOD Benchmark Details

Rotation MNIST and Corruption datasets are frequently used in continuous domain adaptation [18, 19]. In practical applications, rotation can replicate shifts arising from camera positions, while corruptions mimic various weather conditions (e.g., fog, snow) and movements (e.g., motion). In CAOOD, we evaluate both ID accuracy and OOD detection effectiveness.
Rotation MNIST [1]. In this benchmark, we use R-MNIST as the ID dataset and use R-Cifar10bw and R-NOTMNIST as the OOD dataset for CAOOD evaluation. Three rotation datasets are derived from their original datasets (i.e., MNIST, cifar10bw, NOTMNIST) by applying the same rotations $0 - 180^{\circ}$, i.e., $T = [0, 180°]$. MNIST dataset consists of 60000 training samples, and 10000 test samples describing handwritten digits 0-9. NOTMINIST is a near OOD dataset that contains 19000 test samples describing handwritten alphabets A-Z. Cifar10bw is a black-and-white version of Cifar10 with 10000 test samples, which is closer to MNIST.

The original 60000 labeled ID training samples of size 28×28 are used as S. Then we take the images from rotation $(0°, 60°]$ as the training shifting samples, i.e., $\{S_{T_k}\}$. In meta-training, we randomly sample $\mathbf{S}_{\text{spt}}$ and $\mathbf{S}_{\text{qry}}$ of length 10, i.e., $|T_k'| = 10$ from $\{S_{T_k}\}$. In meta-testing, for efficient evaluation, we adapt the model sequentially to shifting distributions on rotation $T_K^- = (120°, 126°, \cdots, 174°)$. Note that for each rotation in meta-testing, we have access to only 100 samples. Lastly, we evaluate the ID classification performance on the Rotation MNIST test set, and OOD detection performance on the R-Cifar10bw, and R-NOTMNIST test set. Each test set contains 10000 test samples for every rotation.

Cifar10C [8] This dataset was initially released for evaluating the robustness of DL models by applying 15 common types of corruptions to Cifar10 [14], and each type has 5 levels of severities. Concretely it leads to 15×5 distributions concerned with various corruptions. We use Cifar10C as the ID data and test on two near OOD datasets applying the same shifting distribution: TinyimageNetC, and Cifar100C. Following the OOD benchmark literature [20], we create TinyimageNetC from a subset of the Tinyimagenet [16] where 1207 images overlap semantic labels with Cifar10 are removed.

In our protocol, we take clean images from Cifar10 as original labeled training samples S, and consider images with various corruptions as shifting samples $\{S^t\}_T$. Specifically, we design the continuous shifting distributions by gradually changing the severity across all corruption types, for example:

$$\underbrace{\cdots, \mathrm{C}_{t-1}^5}_{\text{t-1 and before}} \rightarrow \underbrace{\mathrm{C}_t^1, \mathrm{C}_t^2, \mathrm{C}_t^3, \mathrm{C}_t^4, \mathrm{C}_t^5, \mathrm{C}_t^4, \mathrm{C}_t^3, \mathrm{C}_t^2, \mathrm{C}_t^1}_{\text{corruption type t, changing gradually}} \rightarrow \underbrace{\mathrm{C}_{t+1}^1, \mathrm{C}_{t+1}^2, \cdots}_{\text{t+1 and on}}$$

By doing so each corruption type covers 9 shifting distributions with overlaps, resulting in 135 continuously shifting distributions in total, i.e., $|T = 135|$. This reflects a realistic scenario when test samples come from recursive shifting distributions.

In meta-training, we randomly sample $\mathbf{S}_{\text{spt}}$, $\mathbf{S}_{\text{qry}}$ of length 10 (i.e., $|T_k^{'}| = 10$), from the first 7 corruptions of 63 continuously shifting distributions, i.e., $\{S^t\}_{t \in T_k}, |T_k| = 63$. In meta-testing, we adapt our model to a trajectory of length 10 $\left(|T_K^-| = 10\right)$ that is randomly sampled from the last 7 corruptions. Similarly, we only had access to 100 samples in meta-testing. Lastly, we evaluate the ID performance on a trajectory of length 10 from the last 7 corruptions (i.e., frost, fog, brightness, contrast, elastic transform, pixelate, jpeg compression) on the Cifar10C test image, and test the OOD detection performance on TinyimagenetC and Cifar100C test sets.

Cifar100C [8]. This dataset is similar to Cifar10C by applying the same corruptions to Cifar100. We use Cifar100C as the ID dataset and Cifar10C and TinyimageNetC as the OOD datasets. We re-create TinyimageNetC after 2505 images sharing the same classes with Cifar100 have been removed [20]. We applied the same training and testing fashion as used in Cifar10C.

C. Virtual OOD Generation Details

We generate virtual OOD samples in each inner task. Following VOS [3], we maintain an ID class-conditional queue $|Q_y|$ for each class $y \in \mathcal{Y}$ for continuous online estimation of $\hat{\mu}_c^t$ and $\hat{\Sigma}^t$. Specifically, during the starting stage of the training, we en-queue the embeddings of ID features for each class until the queues are filled up. Then, the queues are kept updated by adding new features and deleting the oldest features dynamically. In our experiment, we set $|Q_y|$ to 500 for all three datasets. An effect of $|Q_y|$ on the R-MNIST benchmark is provided in Section D. For the $\delta - \text{likelihood}$ region, considering that δ can be infinitely small, we instead implement by selecting the $p - \text{th}$ smallest likelihood in a pool of 1000 samples generated from the estimated Gaussian distribution (per class) [3]. Intuitively, a larger p resembles a larger threshold. We set $p = 1$ for all experiments.

D. Detailed Ablation Study

Effect on Uncertainty Regularization Weight λ. In Table 1 we reported the average OOD detection performance on R-NOTMNIST and R-Cifar10bw when the model was trained on R-MNIST. Generally, a mild uncertainty weight λ on $\mathcal{L}^{\text{ood}}$ returns an optimal performance, while larger ones over-regularize the model leading to poor ID and OOD performance.
Effect on Queue Size $|Q_k|$. In Table 2 we investigate the effect of ID queue size $|Q_y|$ by varying $|Q_y| = \{100, 300, 500, 800\}$. Overall, a larger queue size encourages a more accurate estimation of the ID Gaussian distribution parameters thus benefiting the virtual OOD generation hence improving OOD detection.
Effect on Size of Sampling Pool. Our method can quickly adapt to newly arriving testing samples for continuously adaptive OOD detection. Note that the size of the sampling pool during testing time could affect the adaptation speed. In Table 3 we show that a fair-size of sampling pool is enough for fast adaptive OOD detection. A larger size of sampling scale takes more time to finish the adaptation but does not necessarily improve the performance.

Table 1: Ablation study on uncertainty regularization weight λ. Bold marks the chosen parameter.

λ	ID Accuracy↑	AUROC ↑	AUPR ↑	FPR95 ↓
0.01	35.61	97.50	98.39	11.91
0.015	**35.60**	**97.70**	**98.42**	**11.53**
0.02	33.73	97.99	99.01	10.32
0.05	31.33	89.70	87.65	22.38
0.10	28.51	47.13	76.07	99.99

Table 2: Ablation study on the size of queue $|Q_y|$. Bold marks the chosen parameter.

$\lvert Q_y \rvert$	ID Accuracy↑	AUROC ↑	AUPR ↑	FPR95 ↓
800	35.60	97.98	98.01	11.31
500	**35.60**	**97.70**	**98.42**	**11.53**
300	33.97	95.66	96.62	14.85
100	32.67	93.42	94.63	16.17

E. Comparisons on Training Time

For fair comparisons, we used the same network backbones across all experiments. Table 4 compares the training time (on 1x A100 GPU). Note that despite a longer training duration, our method achieved significant performance gains, which is particularly crucial for safety-critical scenarios. Such improvements were attained with a brief adaptation period (Table 3) in testing, which is practically allowed in terms of gained ID accuracy and OOD detection performance advantages.

Table 3: Ablation study on the size of sampling pool during the testing process. Bold marks the chosen parameter.

Size of Sampling Pool	ID Accuracy↑	AUROC ↑	AUPR ↑	FPR95 ↓	Time Required (seconds)
200	34.79	95.83	97.00	17.24	379.40
500	35.68	95.78	95.95	15.08	422.24
1000	**35.60**	**97.70**	**98.42**	**11.53**	435.87
1500	35.18	96.72	97.92	12.36	500.77
5000	32.60	96.04	96.27	14.19	561.94

Table 4: Training time on Energy, VOS, LogitNorm and our method, with ID Accuracy and FPR gaps compared to ours.

Dataset	Energy		VOS		LogitNorm		Ours
	Time (sec.)	Acc. / FPR	Time (sec.)	Acc. / FPR	Time (sec.)	Acc. / FPR	Time (sec.)
R-MNIST	4.7k	-10.0 / -18.3	6.8k	-7.9 / -10.6	3.4k	-10.9 / -16.8	15.1k

F. OOD Detection Results on Standard OOD Datasets

In Table 5,6 and 7, we evaluate our methods on commonly used standard OOD datasets from *static* distributions. In this experiment, the test samples consist of continuously shifted ID samples and *static* OOD samples. The results show that existing OOD baselines obtained poor results on standard OOD datasets when the arriving testing ID samples come from continuously shifting distributions. This highlights that distribution shifts on testing ID samples can cause severe damage to OOD detection even if OOD samples are drawn from a static distribution. This may question the generalization ability of current OOD detection methods.

A simple adaptive adaptation strategy could potentially relieve such impact by adapting the model sequentially to the arriving ID distributions, leading to an improved OOD detection performance. We note that our method outperforms competitive OOD baselines on challenging near-OOD datasets when the OOD samples show similar semantics with the ID datasets.

Table 5: **OOD detection** on standard OOD datasets (ID dataset: R-MNIST).

Dataset	Energy			MSP			Simple Adaptive Energy			Simple Adaptive MSP			**Ours**		
	AUROC	AUPR	FPR95	AUROC	AUPR	FPR95	AUROC	AUPR	FPR95	AUROC	AUPR	FPR95	AUROC	AUPR	FPR95
NOTMNIST	89.91	91.05	45.28	88.00	89.60	50.84	97.98	98.13	25.65	97.13	95.07	20.52	99.99	99.99	0.00
Cifar10bw	97.99	97.99	30.33	99.10	99.21	41.93	99.11	99.08	10.29	99.99	99.99	27.09	100.00	100.00	0.00
Average	93.95	94.52	37.81	93.55	94.41	46.39	98.55	98.61	17.97	98.56	97.53	23.81	**100.00**	**100.00**	**0.00**

Table 6: **OOD detection** on standard OOD datasets (ID dataset: Cifar10C).

Dataset	Energy			MSP			Simple Adaptive Energy			Simple Adaptive MSP			**Ours**		
	AUROC	AUPR	FPR95	AUROC	AUPR	FPR95	AUROC	AUPR	FPR95	AUROC	AUPR	FPR95	AUROC	AUPR	FPR95
Textures	43.21	43.21	57.24	45.54	58.46	95.06	64.17	71.66	78.77	65.13	73.14	83.98	69.57	69.57	77.13
LSUN-C	71.46	71.46	69.43	66.03	65.92	86.69	87.86	87.31	44.98	81.88	83.24	65.62	85.73	85.73	84.44
LSUN-R	54.12	54.12	50.99	54.72	51.55	92.01	87.25	85.82	45.25	81.77	82.22	66.82	87.78	85.73	87.13
iSUN	52.69	52.69	53.04	53.87	53.85	92.90	85.50	84.58	48.63	81.08	83.18	69.01	86.47	86.47	86.95
Places365	52.43	52.98	65.21	53.08	54.98	92.50	76.98	77.94	61.91	74.80	76.75	74.61	79.08	78.67	83.20
Average Far OOD	54.78	54.89	59.18	54.65	56.95	91.83	80.35	**81.46**	**55.91**	76.93	79.71	72.01	**81.73**	81.23	83.77
Cifar100	40.67	43.43	95.35	45.22	45.13	95.86	60.14	60.34	91.90	64.12	61.98	87.60	65.87	65.87	80.37
TinyImageNet	59.31	70.93	91.18	60.12	70.77	92.32	68.51	64.04	74.40	68.91	67.57	81.45	67.05	66.04	80.01
Average Near OOD	49.99	57.18	93.26	52.67	57.95	94.09	64.32	62.19	83.15	66.52	64.78	84.53	**66.46**	**65.96**	**80.19**

G. Detailed Experimental Results

Tables 8 to 16 provide detailed ID classification (Accuracy) and OOD detection results (AUROC, AUPR, FPR95) over continuously shifting distributions for each method.

Table 7: **OOD detection** on standard OOD datasets (ID dataset: Cifar100C).

Dataset	Energy			MSP			Simple Adaptive Energy			Simple Adaptive MSP			**Ours**		
	AUROC	AUPR	FPR95	AUROC	AUPR	FPR95	AUROC	AUPR	FPR95	AUROC	AUPR	FPR95	AUROC	AUPR	FPR95
Textures	50.01	57.17	94.19	48.24	62.65	93.85	58.72	69.72	87.39	57.67	69.82	91.08	52.58	66.51	91.63
LSUN-C	60.30	68.18	96.97	38.82	42.19	99.16	76.32	75.91	69.32	71.08	72.07	79.42	75.98	78.53	77.21
LSUN-R	45.97	46.75	89.11	55.05	52.19	90.86	62.29	60.22	80.97	59.38	58.53	87.18	65.75	64.53	84.11
iSUN	52.18	52.01	85.07	57.14	56.44	88.36	59.01	60.03	86.71	56.34	58.65	90.19	64.13	66.18	84.48
Average Far OOD	52.12	56.03	91.34	49.81	53.37	93.06	64.09	66.47	**81.10**	61.12	64.77	86.97	**64.61**	**68.94**	84.36
Cifar10	49.11	46.31	98.18	45.35	47.03	97.06	69.95	73.24	78.51	70.92	91.93	84.63	70.77	76.85	73.94
TinyImageNet	47.21	40.09	94.97	42.55	44.68	95.95	70.09	69.25	72.02	66.39	65.51	84.09	71.25	66.99	71.75
Average Near OOD	48.16	43.20	96.58	43.95	45.86	96.51	70.02	71.25	75.27	68.66	**78.72**	84.36	**71.01**	71.92	**72.85**

Table 8: **ID accuracy** on R-MNIST. Post-hoc methods include MSP/Odin/Mahalanobis/Energy/Gram.

Strategy	**Method**	120°	126°	132°	138°	144°	150°	156°	162°	168°	172°	**Average**
Direct Test	Post-hoc methods	20.32	21.66	22.92	24.15	25.28	26.36	27.44	28.38	29.24	30.04	25.58
	VOS	20.98	21.99	22.86	24.73	26.98	29.17	30.97	31.51	32.98	34.77	27.69
Simple Adaptive	Post-hoc methods	17.85	19.86	22.28	25.15	28.04	29.76	32.14	34.66	36.14	38.26	28.41
	VOS	18.47	21.25	23.92	27.23	29.91	33.18	36.03	38.17	39.44	41.38	30.90
Domain Adaptation	Post-hoc methods	19.44	19.28	21.42	23.16	25.43	28.78	31.09	33.05	34.47	36.31	27.24
MOL		25.24	28.28	31.05	33.81	36.28	38.06	39.74	40.38	41.23	41.93	35.60

Table 9: **OOD detection** performance R-NOTMNIST (OOD) when the model was trained on R-MNIST. For each method, we report AUROC↑, AUPR ↑, and FPR95 ↓ in order.

Strategy	**Method**	120°	126°	132°	138°	144°	150°	156°	162°	168°	172°	**Average**
Direct Test	MSP	74.08	74.80	75.19	76.62	77.22	78.58	78.85	77.96	77.37	78.44	76.91
		76.27	7.14	77.55	78.57	79.34	80.80	80.95	79.69	78.33	79.05	78.77
		76.64	74.91	73.45	71.85	73.58	71.62	70.66	70.46	68.53	66.96	71.87
	Odin	63.07	62.90	63.26	63.24	63.88	65.00	64.75	64.48	62.64	62.83	63.61
		63.26	63.51	64.26	64.45	64.62	65.20	65.13	65.19	63.55	64.02	64.32
		84.32	86.78	85.49	86.06	85.79	84.79	83.74	83.82	85.42	85.40	85.16
	Energy	90.26	91.88	92.62	93.06	93.07	93.51	93.27	92.74	92.38	93.07	92.59
		90.52	92.30	93.11	93.48	93.54	94.07	94.00	93.50	92.96	93.51	93.10
		41.87	39.42	37.17	33.30	34.26	33.74	36.22	37.06	35.75	31.90	36.07
	Mahalanobis	79.79	85.23	85.35	89.12	88.94	90.02	91.82	92.20	91.40	92.70	88.66
		84.37	89.84	89.17	91.40	91.90	91.94	92.43	92.22	90.56	91.19	90.50
		85.00	71.00	65.00	62.00	71.00	60.00	36.00	28.00	40.00	23.00	54.10
	VOS	90.37	88.73	87.89	87.69	88.30	87.04	85.72	83.26	83.30	85.15	86.74
		91.06	89.46	89.01	88.94	89.66	88.63	87.36	84.99	84.67	85.91	87.97
		44.64	50.05	53.19	55.34	57.18	62.25	62.63	66.39	65.87	58.22	57.58
	Gram	96.14	95.55	96.30	97.09	95.06	95.64	97.50	96.90	96.31	96.08	96.26
		96.79	96.10	95.98	98.13	97.67	96.43	97.01	96.33	97.99	97.47	96.99
		5.09	8.48	9.22	11.74	7.27	15.59	12.81	13.21	12.68	9.91	10.60
Simple Adaptive	MSP	71.40	72.98	74.65	76.84	77.51	77.68	77.11	77.05	76.06	77.48	75.88
		71.10	73.31	75.74	77.68	78.34	79.32	78.69	78.05	76.83	78.34	76.74
		77.62	75.69	72.95	69.56	68.09	69.33	69.09	66.82	68.00	67.17	70.43
	Odin	72.33	74.13	77.43	79.84	80.90	79.97	77.55	75.90	74.50	74.83	76.74
		72.84	74.64	77.82	80.70	82.52	82.08	79.45	77.36	75.36	75.17	77.79
		80.09	77.13	72.26	70.18	67.64	70.13	72.14	72.61	72.11	71.60	72.59
	Energy	88.75	90.49	92.09	92.70	93.17	93.75	93.13	92.16	90.45	90.76	91.75
		88.87	90.96	92.68	93.38	93.88	94.50	93.91	92.84	90.89	91.12	92.30
		49.26	45.65	39.87	37.18	35.85	34.80	37.64	37.02	40.60	39.07	39.69
	Mahalanobis	88.75	90.49	92.09	92.70	93.17	93.75	93.13	92.16	90.45	90.76	91.75
		88.87	90.96	92.68	93.38	93.88	94.50	93.91	92.84	90.89	91.12	92.30
		49.26	45.65	39.87	37.18	35.85	34.80	37.64	37.02	40.60	39.07	39.69
	VOS	90.99	91.82	93.03	94.33	95.36	96.04	95.59	94.26	92.02	91.98	93.54
		90.14	91.62	93.27	94.72	95.80	96.52	96.12	94.83	92.57	92.62	93.82
		34.00	36.07	32.91	29.76	25.95	24.20	27.61	33.34	39.39	40.05	32.33
MOL		94.67	95.37	96.07	96.19	96.62	96.70	97.15	97.21	97.13	97.59	96.47
		96.73	96.77	97.38	97.98	98.09	98.21	98.27	98.28	98.24	98.66	97.86
		21.65	19.12	17.32	15.05	12.99	11.75	11.24	9.22	10.31	9.83	13.85

Table 10: **OOD detection** performance on R-Cifar10bw (OOD) when the model was trained on R-MNIST (ID). For each method, we report AUROC↑, AUPR ↑, and FPR95 ↓ in order.

Strategy	Method	120°	126°	132°	138°	144°	150°	156°	162°	168°	172°	Average
Direct Test	MSP	94.64	93.98	93.03	91.61	90.44	89.28	88.05	87.42	86.38	86.28	90.11
		95.67	95.09	94.41	93.36	92.44	91.49	90.46	89.85	88.83	88.44	92.00
		34.63	37.22	44.57	54.79	58.27	61.81	65.20	66.77	67.19	65.24	55.57
	Odin	86.81	86.74	86.42	86.76	87.02	88.55	88.65	88.23	87.22	86.08	87.25
		88.85	88.62	88.34	88.55	88.72	89.97	90.09	89.80	89.02	87.78	88.98
		60.52	59.77	60.28	58.33	58.82	53.71	55.33	56.70	58.94	59.81	58.22
	Energy	99.11	99.15	99.16	98.95	98.66	97.70	96.76	95.72	94.73	95.10	97.50
		99.10	99.13	99.15	98.96	98.70	97.87	97.03	95.89	94.90	94.87	97.56
		4.04	4.09	4.04	4.92	6.33	11.98	18.08	23.08	27.12	22.65	12.63
	Mahalanobis	89.75	89.52	89.24	89.45	89.62	90.87	90.99	90.70	89.92	83.23	89.33
		91.08	90.86	90.49	89.90	90.31	91.50	92.51	93.22	93.60	87.72	91.12
		60.21	59.46	59.97	58.02	58.51	53.40	55.02	56.39	58.63	61.39	58.10
	VOS	97.55	96.76	95.88	94.60	93.38	91.40	90.46	88.96	88.76	89.66	92.74
		97.51	96.69	95.84	94.51	93.21	91.06	89.98	88.60	88.41	89.15	92.50
		12.03	16.42	20.59	25.41	29.13	35.89	38.33	44.61	46.31	41.72	31.04
	Gram	99.37	99.38	99.07	99.09	98.42	98.50	98.71	98.84	98.09	99.15	98.86
		99.33	99.40	99.13	98.53	98.59	98.85	99.05	98.34	98.11	99.12	98.85
		2.91	3.48	4.65	4.27	7.25	6.47	5.22	3.93	4.96	2.86	4.60
Simple Adaptive	MSP	87.69	88.23	88.67	88.66	89.20	89.98	90.25	90.26	90.20	90.49	89.36
		88.45	89.09	89.55	89.56	90.14	90.98	91.25	91.34	91.19	91.55	90.31
		49.23	49.66	48.68	48.08	47.80	48.02	47.59	48.78	48.76	48.75	48.54
	Odin	88.52	87.89	88.04	87.69	87.84	88.08	88.09	88.15	87.95	88.29	88.05
		89.29	88.77	88.83	88.53	88.57	88.79	88.76	88.67	88.31	88.76	88.73
		48.32	50.25	50.10	51.18	51.84	49.75	51.80	51.77	52.08	52.20	50.93
	Energy	96.88	96.66	96.45	96.17	95.95	96.15	95.80	95.67	94.87	94.81	95.94
		96.82	96.57	96.31	95.99	95.73	95.95	95.66	95.60	94.85	94.90	95.84
		15.79	16.55	17.15	18.04	19.23	18.88	20.58	21.40	24.95	26.26	19.88
	Mahalanobis	90.75	90.50	90.12	89.28	89.55	90.60	91.72	92.33	92.83	83.32	90.10
		92.78	92.56	92.19	91.60	92.01	93.20	94.21	94.92	95.30	89.92	92.87
		51.62	52.19	49.13	50.71	48.34	50.65	48.12	50.14	50.39	43.71	49.50
	VOS	97.00	96.61	96.29	96.28	96.35	96.94	97.14	97.32	96.84	96.66	96.74
		97.01	96.62	96.40	96.43	96.55	97.12	97.29	97.50	97.06	96.90	96.89
		14.60	16.41	17.80	18.60	17.92	15.67	14.62	14.23	16.62	17.65	16.41
MOL		99.28	99.27	99.21	99.05	99.00	98.92	98.88	98.73	98.53	98.47	98.93
		99.27	99.25	99.19	99.03	98.97	98.91	98.87	98.74	98.56	98.50	98.93
		4.94	4.67	4.99	5.68	6.40	7.66	9.69	11.85	16.87	19.25	9.20

Table 11: **ID accuracy** on Cifar10C. Post-hoc methods include MSP/Odin/Mahalanobis/Energy/Gram/Gradnorm. The test corruption types cover frost, fog, brightness, contrast, elastic transform, pixelate, and jpeg compression.

Strategy	Method	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	Average
Direct Test	Post-hoc methods	46.85	41.23	40.07	40.96	37.76	36.00	36.40	34.20	32.85	34.84	38.48
	VOS	28.42	27.22	26.96	26.77	26.07	25.62	25.26	24.17	23.45	23.82	25.77
Simple Adaptive	Post-hoc methods	70.38	53.87	57.49	68.44	42.06	46.06	69.24	32.41	41.91	79.67	56.15
	VOS	37.04	32.51	31.22	39.08	30.51	30.82	35.27	22.85	25.25	41.98	31.62
Domain Adaptation	Post-hoc methods	42.81	33.98	36.06	44.08	28.91	30.05	40.84	23.85	26.96	48.62	35.62
MOL		72.34	60.74	63.05	74.98	51.42	55.10	76.63	47.29	66.05	81.22	64.18

Table 12: **OOD detection** performance on Cifar100C (OOD) when the model was trained on Cifar10C (ID). For each method, we report AUROC↑, AUPR ↑, and FPR95 ↓ in order. The test corruption types cover frost, fog, brightness, contrast, elastic transform, pixelate, and jpeg.

Strategy	**Method**	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	**Average**
Direct Test	MSP	57.68	53.34	53.94	55.75	48.97	49.93	56.00	51.14	52.24	60.24	53.22
		58.48	53.92	54.34	56.61	49.61	50.20	56.21	50.13	51.07	61.50	53.40
		93.57	94.32	94.23	93.84	95.83	95.09	93.40	94.10	93.93	92.49	94.26
	Odin	42.32	46.66	46.06	44.25	51.03	50.07	44.00	48.86	47.76	39.76	46.08
		45.55	47.93	47.79	46.41	51.12	50.12	46.12	48.71	48.10	43.93	47.58
		98.00	97.10	96.94	97.82	95.22	95.45	97.39	94.25	94.57	98.71	96.55
	Energy	57.29	52.73	53.65	55.23	48.15	49.09	55.80	50.84	51.89	60.31	52.74
		58.16	53.32	53.90	56.08	49.11	49.69	55.77	49.87	50.71	61.25	52.96
		93.95	94.61	94.70	93.58	95.64	95.60	93.21	94.27	94.08	91.67	94.40
	Mahalanobis	43.69	47.01	46.80	45.62	49.65	49.32	45.83	49.15	48.52	42.03	46.76
		46.00	48.19	48.21	47.26	49.57	49.36	46.85	49.51	48.75	44.74	47.84
		97.69	97.28	97.21	97.92	95.56	95.84	96.88	95.37	95.24	98.32	96.73
	VOS	53.89	52.66	52.60	54.46	53.59	53.90	52.78	48.64	49.80	54.05	52.64
		54.84	53.13	53.21	55.64	54.36	54.27	53.66	49.05	50.38	55.11	53.37
		96.07	95.92	95.85	95.33	94.88	94.16	95.67	96.73	96.09	96.00	95.67
	Gradnorm	46.32	13.24	17.31	26.47	6.21	5.19	36.31	32.96	40.77	70.42	29.52
		48.32	32.92	33.91	36.54	31.63	31.48	41.03	41.82	100.00	78.00	47.56
		97.00	100.00	100.00	98.00	100.00	100.00	100.00	99.00	98.68	79.69	97.24
	Gram	56.65	54.46	54.46	55.75	52.66	53.56	57.03	53.43	55.43	58.34	55.18
		50.96	48.88	48.81	49.20	46.54	47.36	50.51	46.62	47.85	71.74	50.85
		88.27	89.93	90.46	88.29	91.83	91.34	88.49	92.54	90.06	86.02	89.72
Simple Adaptive	MSP	66.71	60.55	60.94	66.97	57.09	58.37	67.63	54.24	56.78	72.63	62.19
		66.15	60.33	60.35	67.17	57.91	59.08	68.45	53.97	57.19	72.42	62.30
		89.66	92.50	91.49	90.62	93.71	93.58	90.50	94.11	93.48	86.18	91.58
	Odin	66.78	59.89	60.75	66.21	56.28	57.08	66.10	52.87	55.30	71.87	61.31
		66.30	59.94	60.64	65.93	56.45	57.42	66.40	52.30	54.98	71.32	61.17
		89.74	92.73	92.26	90.22	93.83	93.82	90.81	93.93	94.11	87.17	91.86
	Energy	68.07	60.77	61.69	67.22	56.41	57.72	67.29	52.75	55.61	72.93	60.84
		68.36	62.34	62.88	68.42	58.62	58.54	68.17	55.98	58.00	73.26	62.37
		88.52	91.96	91.35	89.37	93.29	92.87	89.32	93.37	92.82	84.79	91.43
	Mahalanobis	68.45	62.36	62.90	68.31	57.50	58.82	68.24	55.27	57.25	73.53	63.26
		69.36	63.81	63.63	69.69	59.13	60.03	69.98	56.06	58.82	74.19	64.47
		89.76	92.10	91.32	89.73	92.99	93.39	89.66	93.75	93.35	86.83	91.29
	VOS	54.90	52.41	52.31	56.71	54.09	54.31	54.65	51.05	51.93	57.00	53.60
		57.11	54.32	54.06	59.05	55.63	55.66	56.50	52.48	53.36	59.51	55.35
		96.47	96.63	96.28	96.24	95.91	95.56	96.91	96.94	97.09	96.63	96.45
MOL		73.39	66.73	68.69	74.99	61.67	62.54	79.20	61.36	65.41	82.55	69.65
		73.30	67.89	68.28	75.09	61.50	63.30	77.46	60.74	66.06	81.27	69.49
		84.49	88.58	87.51	83.08	91.57	90.58	81.27	91.12	89.70	75.48	86.34

Table 13: **OOD detection** performance on TinyImageNetC (OOD) when the model was trained on Cifar10C (ID). For each method, we report AUROC↑, AUPR ↑, and FPR95 ↓ in order. The test corruption types cover frost, fog, brightness, contrast, elastic transform, pixelate, and jpeg compression.

Strategy	**Method**	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	**Average**
Direct Test	MSP	61.50	60.65	59.63	63.68	60.26	59.55	61.66	54.60	56.04	63.65	59.73
		72.45	71.08	70.51	73.71	70.93	70.32	72.53	65.78	66.49	73.89	70.42
		92.60	92.06	92.35	91.06	92.60	92.82	91.75	93.81	93.10	91.00	92.46
	Odin	38.50	39.35	40.37	36.32	39.74	40.45	38.34	45.40	43.96	36.35	40.27
		53.24	53.08	53.83	51.53	53.40	53.93	52.84	56.92	55.84	51.70	53.85
		99.51	99.26	99.34	99.50	99.28	99.24	99.54	98.24	98.34	99.57	99.14
	Energy	62.86	61.95	61.68	64.87	62.49	61.33	63.56	56.56	57.80	64.59	61.46
		73.16	71.57	71.44	74.40	72.19	71.41	73.42	66.87	67.40	74.29	71.32
		91.57	89.84	89.41	90.72	91.57	92.37	91.21	94.75	93.59	90.15	91.67
	Mahalanobis	42.56	46.93	47.90	42.33	60.17	58.32	38.74	42.51	43.22	39.04	46.97
		55.02	58.35	58.67	54.95	71.23	69.13	52.82	57.45	56.32	52.45	59.33
		98.79	98.10	97.99	98.59	96.04	96.68	99.13	99.13	98.72	98.94	98.13
	VOS	57.66	58.70	61.18	58.56	61.53	60.93	56.82	50.10	50.57	54.95	57.34
		68.35	68.23	71.05	69.16	72.14	71.52	68.66	62.87	62.91	65.71	68.32
		97.71	97.28	97.16	96.78	97.03	96.69	97.29	97.41	97.09	97.16	97.16
	Gradnorm	26.41	27.57	33.37	63.24	61.16	49.15	20.43	21.09	41.24	69.79	41.34
		45.91	46.12	50.03	69.90	74.68	63.72	43.86	44.71	32.85	34.84	50.66
		100.00	100.00	95.59	86.76	92.65	98.53	100.00	100.00	100.00	91.18	96.47
	Gram	69.01	59.28	67.12	61.87	89.55	88.75	47.13	46.09	46.73	89.04	66.46
		71.73	62.76	70.09	65.73	92.55	91.75	53.05	51.49	61.59	58.16	67.89
		79.34	88.12	83.81	86.04	57.95	58.09	93.28	95.33	94.31	89.04	82.53
Simple Adaptive	MSP	69.42	64.66	64.81	69.57	59.58	60.25	70.65	57.43	60.03	73.71	64.04
		78.15	74.25	74.13	78.49	71.25	71.67	80.07	68.83	71.30	80.41	74.24
		89.18	90.24	90.31	90.00	92.51	92.47	89.51	93.68	93.25	85.79	91.24
	Odin	68.91	63.93	64.00	69.17	59.03	59.13	69.61	54.67	58.50	73.27	62.99
		77.20	73.10	72.97	77.62	69.95	70.12	78.99	66.40	69.51	80.00	72.87
		89.29	90.97	90.03	89.59	92.91	93.31	89.84	93.99	93.38	86.96	91.48
	Energy	70.69	65.06	65.68	70.33	61.03	60.86	72.04	58.32	61.18	74.78	65.02
		77.85	73.33	73.63	77.90	70.71	70.68	80.22	68.97	71.49	80.27	73.86
		86.01	88.56	87.03	87.18	90.00	90.90	86.84	92.10	92.18	81.97	88.98
	Mahalanobis	71.47	67.02	67.57	70.03	59.72	60.17	72.64	62.56	62.68	75.39	65.98
		80.65	77.25	77.42	79.38	71.13	71.52	82.24	73.81	74.09	82.94	76.39
		89.12	90.03	88.25	89.03	92.93	92.96	88.21	92.32	92.22	86.06	90.56
	VOS	61.85	62.55	65.85	60.85	61.64	61.06	59.24	52.26	53.77	59.04	59.90
		72.32	72.82	75.40	70.99	70.79	70.26	70.51	66.14	66.42	70.07	70.63
		96.53	95.84	94.10	96.84	96.75	96.47	98.63	99.53	98.99	97.29	97.08
MOL		76.44	71.69	72.82	75.33	66.44	66.07	79.67	65.25	68.89	81.36	71.40
		83.88	80.83	81.28	82.75	74.62	74.65	86.65	75.97	78.78	87.19	79.93
		82.46	87.49	85.40	82.49	86.53	86.93	78.44	92.18	88.53	74.09	85.60

Table 14: **ID accuracy** on Cifar100C. Post-hoc methods include MSP/Odin/Mahalanobis/Energy/Gram/Gradnorm. The test corruption types cover frost, fog, brightness, contrast, elastic transform, pixelate, and jpeg compression.

Strategy	**Method**	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	**Average**
Direct Test	Post-hoc methods	33.69	28.65	27.67	29.55	27.25	25.95	27.19	25.89	25.31	27.22	27.84
	VOS	30.23	27.59	26.18	27.32	24.08	24.99	25.31	24.18	24.67	26.76	26.13
Simple Adaptive	Post-hoc methods	50.24	36.62	39.75	51.21	27.82	30.19	53.94	27.84	34.73	61.79	41.41
	VOS	51.63	39.34	42.30	52.74	29.71	31.69	55.47	30.81	36.37	61.56	43.16
Domain Adaptation	Post-hoc methods	45.41	33.92	36.31	47.71	27.57	29.74	47.20	23.13	29.63	56.34	37.70
MOL		66.63	54.34	57.30	66.74	43.71	45.69	69.47	44.81	50.37	75.56	57.46

Table 15: **OOD detection** performance on Cifar10C (OOD) when the model was trained on Cifar100C (ID). For each method, we report AUROC↑, AUPR ↑, and FPR95 ↓ in order. The test corruption types cover frost, fog, brightness, contrast, elastic transform, pixelate, and jpeg compression.

Strategy	**Method**	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	**Average**
Direct Test	MSP	60.47	57.11	57.68	61.22	54.62	55.37	61.76	52.90	55.46	65.42	58.20
		62.10	57.13	58.49	62.64	55.12	55.92	63.36	52.47	55.24	67.92	59.04
		92.69	92.92	93.18	91.93	93.70	93.95	92.22	94.23	93.17	90.51	92.85
	Odin	39.53	42.89	42.32	38.78	45.38	44.63	38.24	47.10	44.54	34.58	41.80
		43.84	45.38	45.34	43.29	46.66	46.65	43.17	47.93	46.49	41.07	44.98
		98.91	97.67	98.19	98.86	97.37	97.74	99.06	95.87	97.00	99.52	98.02
	Energy	61.18	57.88	58.33	62.06	55.02	55.94	62.45	52.65	55.21	66.48	58.72
		62.20	57.20	58.58	62.90	55.16	56.10	63.33	51.83	54.61	68.45	59.04
		92.60	91.99	93.41	91.94	93.32	93.40	92.66	94.73	93.99	90.89	92.89
	Mahalanobis	38.54	41.71	41.41	37.76	43.96	43.51	36.65	44.46	42.26	33.49	40.38
		43.87	45.13	45.07	43.12	46.14	46.04	42.49	46.32	45.20	40.93	44.43
		99.70	98.36	98.51	99.71	98.23	98.50	99.80	98.33	98.81	99.89	98.98
	VOS	57.66	58.70	61.18	58.56	61.53	60.93	56.82	50.10	50.57	54.95	57.34
		68.35	68.23	71.05	69.16	72.14	71.52	68.66	62.87	62.91	65.71	68.32
		97.71	97.28	97.16	96.78	97.03	96.69	97.29	97.41	97.09	97.16	97.16
	Gram	50.01	49.81	49.85	49.99	49.77	49.66	49.99	48.30	48.09	50.02	49.55
		47.38	47.48	47.53	47.36	47.26	47.31	47.35	46.85	47.19	47.38	47.31
		95.00	95.00	94.99	95.03	95.03	95.32	95.03	97.44	98.17	95.04	95.60
Simple Adaptive	MSP	66.62	61.82	62.26	66.87	58.79	59.02	69.14	60.76	62.87	72.11	64.03
		69.25	64.28	65.32	69.49	59.92	60.65	72.11	62.92	65.44	75.01	66.44
		90.44	92.38	92.40	89.79	92.13	92.84	90.40	92.34	91.64	87.88	91.22
	Odin	65.85	60.69	61.75	66.27	57.94	58.47	69.33	59.77	62.98	71.94	63.50
		68.24	62.83	64.38	68.55	59.03	59.72	72.00	61.10	64.82	74.70	65.54
		89.18	92.32	92.24	89.91	91.92	93.13	88.67	92.05	91.17	87.06	90.77
	Energy	66.73	60.80	61.70	66.86	58.42	59.35	69.35	58.60	61.93	72.64	63.64
		69.10	62.59	64.27	69.03	59.12	60.22	71.69	58.81	62.93	75.13	65.29
		90.95	92.69	93.00	89.90	92.01	92.72	89.28	93.20	92.73	87.12	91.36
	VOS	57.34	52.06	52.61	57.62	50.86	51.49	59.48	51.53	53.95	61.06	54.80
		61.33	56.69	57.20	61.62	55.07	55.50	63.44	55.36	57.96	73.93	59.81
		98.01	100.00	100.28	98.06	100.43	99.95	97.95	100.24	99.66	85.42	98.00
MOL		67.28	62.82	64.63	66.91	63.22	63.42	66.13	50.57	56.04	69.99	63.10
		79.65	75.17	76.99	79.12	75.32	75.32	79.68	67.10	71.45	84.09	76.39
		87.49	89.24	88.99	88.76	88.12	87.96	90.01	95.66	93.51	80.56	89.03

Table 16: **OOD detection** performance on TinyImageNetC (OOD) when the model was trained on Cifar100C (ID). For each method, we report AUROC↑, AUPR ↑, and FPR95 ↓ in order. The test corruption types cover frost, fog, brightness, contrast, elastic transform, pixelate, and jpeg compression.

Strategy	**Method**	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	**Average**
Direct Test	MSP	55.16	47.68	51.04	56.20	44.65	47.26	54.45	43.23	47.67	61.44	50.88
		65.07	57.53	60.61	65.47	55.69	57.66	64.53	55.15	58.23	70.83	61.08
		93.96	94.65	94.43	93.25	96.01	95.49	94.65	96.81	95.81	92.57	94.76
	Odin	44.84	52.32	48.96	43.80	55.35	52.74	45.55	56.77	52.33	38.56	49.12
		56.30	60.42	58.66	55.59	62.92	61.13	57.18	64.85	61.55	52.83	59.14
		97.81	93.41	95.79	97.44	92.49	93.85	97.49	93.01	94.69	98.72	95.47
	Energy	56.12	47.69	51.49	55.55	43.49	46.40	54.12	42.86	47.33	61.50	50.66
		64.99	57.08	60.26	64.61	54.60	56.69	63.75	54.93	57.80	70.46	60.52
		93.00	94.65	94.40	93.93	96.07	95.44	95.41	97.21	96.21	92.69	94.90
	Mahalanobis	46.37	50.48	49.28	42.74	57.86	55.42	42.56	54.30	50.14	36.19	48.53
		59.34	59.82	60.07	55.32	66.16	64.80	54.97	63.25	59.89	51.44	59.51
		98.50	95.90	96.74	98.66	93.15	94.59	99.19	96.57	97.79	99.40	97.05
	VOS	57.13	49.07	50.31	51.06	46.97	43.44	50.74	45.77	48.91	55.60	49.90
		64.80	55.17	56.93	59.71	53.04	55.58	65.61	52.79	64.03	69.44	59.71
		95.11	94.96	95.41	94.90	98.17	96.28	97.01	98.99	97.45	92.02	96.03
	Gram	54.06	39.78	46.81	49.09	65.09	63.81	37.37	32.52	35.44	47.12	47.11
		60.99	51.22	56.58	56.79	75.19	73.46	48.04	46.93	48.51	54.21	57.19
		94.47	97.33	95.64	95.38	91.15	92.15	97.92	99.68	99.58	95.80	95.91
Simple Adaptive	MSP	65.52	60.19	61.22	66.10	55.19	56.59	67.11	52.01	57.62	71.34	61.29
		74.70	69.78	71.29	74.92	65.21	66.42	76.46	63.54	68.45	79.38	71.01
		89.49	91.69	91.94	88.90	92.57	92.81	90.46	95.37	93.37	86.62	91.32
	Odin	64.89	57.61	60.27	65.25	52.64	54.68	66.76	48.12	55.79	71.12	59.71
		74.15	66.78	69.87	73.92	62.86	64.54	75.73	59.27	65.99	79.08	69.22
		89.56	91.76	91.26	89.65	93.10	93.29	89.31	95.85	93.71	86.47	91.40
	Energy	64.72	55.19	58.77	63.98	51.56	54.24	64.47	44.23	52.49	70.62	58.03
		73.19	63.72	67.65	72.34	61.18	63.46	73.35	55.94	62.62	78.13	67.16
		89.44	92.72	91.91	90.15	92.56	92.85	91.47	97.59	95.93	86.74	92.14
	VOS	56.81	50.98	52.62	58.09	48.61	50.57	57.01	39.92	47.66	58.72	52.10
		66.20	59.18	61.49	66.98	58.33	59.92	65.97	50.74	56.97	71.31	61.71
		95.71	97.99	97.03	94.51	98.78	97.63	96.19	100.65	98.99	85.53	96.30
MOL		70.28	65.82	67.63	71.91	68.22	72.42	72.13	63.57	69.04	73.77	69.48
		82.65	78.17	79.99	82.12	78.32	78.32	82.68	70.10	74.45	85.85	79.27
		86.49	88.24	87.99	87.76	87.12	86.96	89.01	94.66	92.51	86.94	88.77

References

[1] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. 1

[2] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[3] 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. 2

[4] Thomas Elsken, Benedikt Staffler, Jan Hendrik Metzen, and Frank Hutter. Meta-learning of neural architectures for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12365–12375, 2020. 1

[5] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017. 1

[6] Chelsea Finn and Sergey Levine. Meta-learning and universality: Deep representations and gradient descent can approximate any learning algorithm. *arXiv preprint arXiv:1710.11622*, 2017. 1

[7] Harry F Harlow. The formation of learning sets. *Psychological review*, 56(1):51, 1949. 1

[8] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. 1, 2

[9] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5149–5169, 2021. 1

[10] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[11] Mike Huisman, Jan N Van Rijn, and Aske Plaat. A survey of deep meta-learning. *Artificial Intelligence Review*, 54(6):4483–4541, 2021. 1

[12] Taewon Jeong and Heeyoung Kim. Ood-maml: Meta-learning for few-shot out-of-distribution detection and classification. *Advances in Neural Information Processing Systems*, 33:3907–3916, 2020. 1

[13] KJ Joseph, Jathushan Rajasegaran, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Incremental object detection via meta-learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9209–9216, 2021. 1

[14] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[15] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *International Conference on Machine Learning*, pages 2554–2563. PMLR, 2017. 1

[16] 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. 1

[17] Ricardo Vilalta and Youssef Drissi. A perspective view and survey of meta-learning. *Artificial intelligence review*, 18:77–95, 2002. 1

[18] Hao Wang, Hao He, and Dina Katabi. Continuously indexed domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9898–9907, 2020. 1

[19] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022. 1

[20] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:32598–32611, 2022. 1, 2

[21] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pages 1094–1100. PMLR, 2020. 1