ViM: Out-Of-Distribution with Virtual-logit Matching Supplementary Material

Haoqi Wang* Zhizhong Li* Litong Feng Wayne Zhang

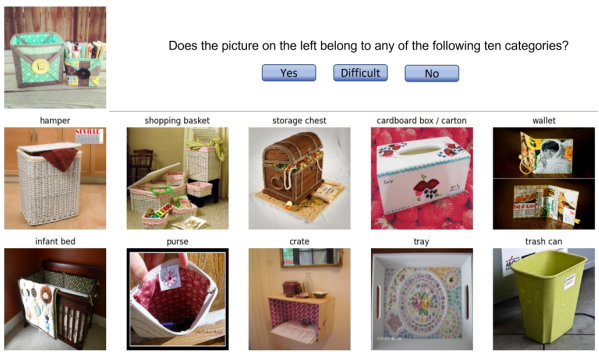

Figure 6. A demonstrative UI for the labelers. The image on the left-top corner is the candidate OOD image to be labeled. The two rows of images below are from the 10 most similar ID classes. Labelers choose from *yes/difficult/no* according to these information.

A. Detailed Information of Models (Sec. 6)

In the experiment, we benchmarked a collection of deep classification models. Their detailed information, including the specification, the architecture, the pre-train information, and the top-1 accuracy, is listed in Tab. 5. To summarize, half of them are CNN-based, and half are transformer-based. Vision Transformer and Swin Transformer are pre-trained on ImageNet-21K before training on the ImageNet-1K.

B. Detailed Results of Four Models (Sec. 6.3)

In Sec. 6.3 we gave the average AUROC and FPR95 for RepVGG, ResNet-50d, Swin Transformer and DeiT. We provide the detailed AUROC and FPR95 on OpenImage-O, Texture, iNaturalist, and ImageNet-O in Tab. 6.

C. Details on OpenImage-O (Sec. 5)

An illustrative software interface for labelers is shown in Fig. 6. For each candidate OOD image to be labeled, we find the top 10 classes in ImageNet-1K predicted by a classification model. Then we gather the most similar images in those top 10 classes by cosine similarity in the feature space. Next, we patch them as well as their labels with the corresponding OpenImage samples, and let the labelers distinguish whether the OpenImage sample belongs to any of the top 10 categories. We also set a choice called difficult, so that labelers can put the undistinguishable hard samples into the difficult category. To reduce annotation noises, each image is labeled twice from different group of labelers. Then we take the set of OOD images having consensus from the two groups, resulting in an OOD dataset with 17,632 unique images. In the end, a random inspection process is performed to guarantee the quality of the OOD dataset.

The OpenImage-O follows a natural image distribution as both the source dataset and the labeling process do not involve any filtration based on pre-defined list of labels. To get a sense of its distribution, we use the BiT model to find the most similar ID class in ImageNet for each OOD image. Then the histogram is illustrated in Fig. 7. It shows that the coverage of OpenImage-O is broader compared to the other three OOD datasets.

D. Details on Grouping (Sec. 6.5)

MOS [7] is trained using the officially released code and its default parameter setting. For all experiments in Sec. 6.5, the grouping strategy follows the taxonomy grouping defined in [7].

Grouping Results on ViT The grouping strategy is less effective for the ViT model, as seen from results in Tab. 7. Comparing MSP with its group version, MaxGroup, we can see that the improvement on AUROC is very small, while FPRs become even worse. Examining ViM with its group variant ViM+Group, we can see that their difference is very small, and the original version of ViM is slightly better than ViM+Group.

E. Details on Baselines (Sec. 6)

Mahalanobis On the BiT model, when including lower level features, the performance of Mahalanobis degrades a lot. The average AUROC on the four OOD datasets is 56%, which is much worse than the baseline MSP. Similar results is also found in [7, Table 1]. In this paper, we implement the

Model	Specification	Architecture	Pre-Trained Dataset	Top1 (%)
BiT [8]	BiT-S-R101x1	CNN	—	81.30
ViT [2]	ViT-B/16	Transformer	ImageNet-21K	85.43
RepVGG [1]	RepVGG-b3	CNN	—	80.52
Res50d [4]	ResNet-50d	CNN	—	80.52
Swin [12]	Swin-base-patch4-window7-224	Transformer	ImageNet-21K	85.27
DeiT [14]	DeiT-base-patch16-224	Transformer	—	81.98

Table 5. Detailed information on the used models. The detailed specification and the top-1 accuracy of the model are provided. Three of them are CNN-based, and the other three are transformer-based. Both ViT and Swin Transformer are pre-trained on ImageNet-21K before training on ImageNet-1K, so their general OOD performances are much better than alternatives.

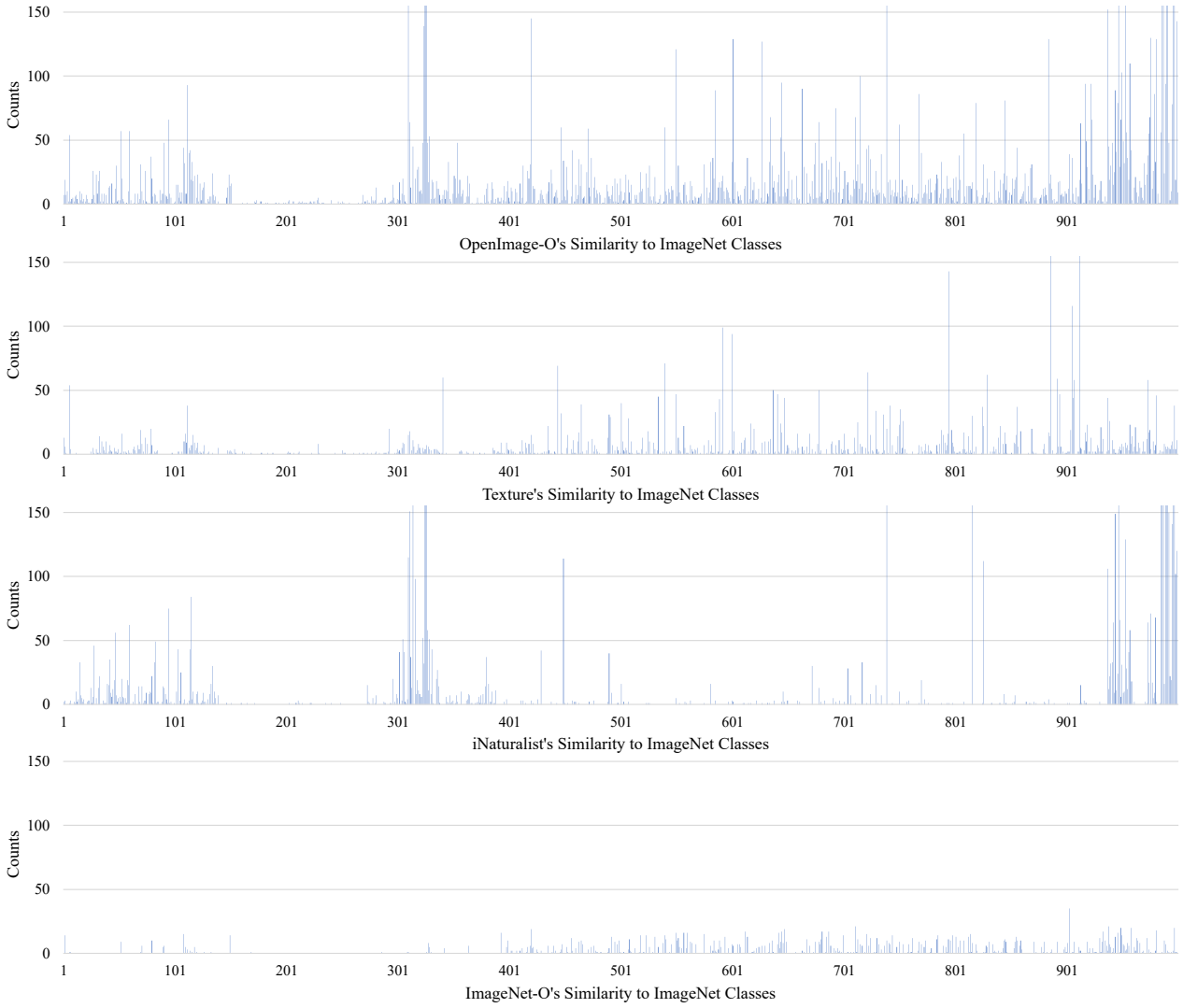

Figure 7. The diversity of four OOD datasets shown by how they look similar to the ImageNet-1K classes. We use the BiT model to predict which ID class the image most resembles, and count the number of such OOD images for each class. Results are shown above. Due to space limitation, the y-axis is clipped at 155. Our newly created OpenImage-O has a wider coverage on ImageNet ID classes.

Model	Method	Source	OpenImage-O		Texture		iNaturalist		ImageNet-O		Average	
			AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
RepVGG [1]	MSP [6]	prob	85.06	63.36	78.58	72.62	87.11	54.93	61.65	91.30	78.10	70.55
	Energy [11]	logit	83.64	69.92	74.53	82.97	83.92	75.31	63.36	87.75	76.36	78.99
	ODIN [10]	prob+grad	85.22	63.48	76.77	76.14	86.37	61.40	62.50	89.70	77.72	72.68
	MaxLogit [5]	logit	84.81	65.04	76.33	76.86	86.22	62.20	62.87	89.90	77.56	73.50
	KL Matching [5]	prob	86.80	57.48	83.18	62.09	89.06	**42.07**	66.36	84.95	81.35	61.65
	Residual†	feat	82.51	65.13	93.05	28.66	86.09	62.40	75.11	79.80	84.19	59.00
	ReAct [13]	feat	46.08	99.65	54.56	97.66	47.18	99.88	48.76	98.65	49.14	98.96
	Mahalanobis [9]	feat+label	85.71	64.93	92.71	32.03	89.17	58.79	76.68	81.80	86.07	59.39
	ViM (Ours)	feat+logit	**89.27**	**52.40**	**93.69**	**23.76**	**91.35**	46.79	**76.93**	**79.05**	**87.81**	**50.50**
Res50d [4]	MSP [6]	prob	84.50	63.53	82.75	64.40	88.58	50.05	56.13	93.85	77.99	67.96
	Energy [11]	logit	75.95	76.83	73.93	75.31	80.50	71.32	53.95	90.10	71.08	78.39
	ODIN [10]	prob+grad	81.53	64.49	80.21	63.93	86.48	52.58	52.87	93.25	75.27	68.56
	MaxLogit [5]	logit	81.50	65.50	79.25	66.20	86.42	53.00	54.39	92.65	75.39	69.34
	KL Matching [5]	prob	87.31	60.58	86.07	61.36	**90.48**	**47.22**	67.00	88.50	82.72	64.41
	Residual†	feat	87.64	59.65	94.62	25.89	84.63	75.81	**81.15**	**72.85**	87.01	58.55
	ReAct [13]	feat	85.30	60.79	91.12	39.26	87.27	56.03	68.02	78.45	82.93	58.63
	Mahalanobis [9]	feat+label	89.52	55.91	94.15	28.22	89.48	62.69	80.15	76.00	88.33	55.70
	ViM (Ours)	feat+logit	**90.76**	**50.45**	**95.84**	**20.58**	89.26	64.59	81.02	74.80	**89.22**	**52.61**
Swin [12]	MSP [6]	prob	91.35	34.96	85.21	51.90	94.76	23.19	78.97	63.70	87.57	43.44
	Energy [11]	logit	90.93	27.58	82.62	51.57	95.22	15.47	82.29	45.70	87.77	35.08
	ODIN [10]	prob+grad	91.38	28.42	85.74	44.59	94.24	19.65	80.62	53.65	88.00	36.58
	MaxLogit [5]	logit	91.91	26.79	84.67	47.42	95.72	15.41	81.28	51.50	88.40	35.28
	KL Matching [5]	prob	91.92	40.05	86.89	52.93	94.77	27.62	81.91	67.35	88.87	46.99
	Residual†	feat	94.64	32.19	91.31	43.97	98.89	4.81	86.68	68.55	92.88	37.38
	ReAct [13]	feat	93.58	**23.07**	85.51	47.91	97.51	9.98	84.09	**44.50**	90.17	31.36
	Mahalanobis [9]	feat+label	94.57	33.41	89.92	49.17	98.69	5.43	85.46	73.55	92.16	40.39
	ViM (Ours)	feat+logit	**96.04**	23.88	**92.34**	**38.49**	**99.28**	**2.60**	**88.78**	59.20	**94.11**	**31.04**
DeiT [14]	MSP [6]	prob	84.04	62.03	81.99	64.48	88.25	52.00	63.65	87.20	79.48	66.43
	Energy [11]	logit	74.50	67.21	77.47	64.77	78.63	65.82	60.60	82.75	72.80	70.14
	ODIN [10]	prob+grad	80.19	**59.53**	81.26	**59.38**	85.36	51.81	61.70	84.95	77.13	**63.92**
	MaxLogit [5]	logit	80.11	60.83	80.45	60.89	85.22	52.54	61.38	83.70	76.79	64.49
	KL Matching [5]	prob	87.49	60.66	**84.89**	63.47	90.54	**50.47**	71.05	84.60	83.49	64.80
	Residual†	feat	88.07	69.21	82.68	77.75	91.32	58.30	74.54	91.25	84.15	74.13
	ReAct [13]	feat	80.29	63.11	80.45	63.99	84.43	59.07	64.32	**81.85**	77.37	67.00
	Mahalanobis [9]	feat+label	89.03	66.51	83.58	77.31	91.56	58.67	75.95	90.25	85.03	73.18
	ViM (Ours)	feat+logit	**89.13**	64.58	84.42	73.02	**92.15**	52.79	**95.30**	89.40	**85.25**	69.95

Table 6. OOD detection for ViM and baseline methods on RepVGG, ResNet-50d, Swin Transformer, and DeiT. Their pre-trained weights are used. The ID dataset is ImageNet-1K, and OOD datasets are OpenImage-O, Texture, iNaturalist, and ImageNet-O. Both metrics AUROC and FPR95 are in percentage. The best performing item is bolded, and the second and the third places are underlined. The proposed ViM has the largest AUROC and the lowest FPR in most cases. †: Residual is defined in Equ. (4).

Method	OpenImage-O		Texture		iNaturalist		ImageNet-O	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
MSP	92.53	34.18	87.10	48.55	96.11	19.04	81.86	64.85
MaxGroup	92.60	48.08	87.84	60.08	95.39	31.40	84.45	71.90
ViM	97.61	12.61	**95.34**	**20.31**	**99.41**	**2.60**	**92.55**	**36.75**
ViM+Group	**97.64**	**12.51**	95.29	20.41	99.40	2.70	92.50	37.05

Table 7. Comparison of effect of grouping on ViT. All numbers are in percentage. The grouping is defined in [7] based on taxonomy. MaxGroup is the group version of MSP and ViM+Group is the group version of ViM.

Method	OpenImage-O	Texture	iNaturalist	ImageNet-O
Residual	1.70s	0.56s	1.00s	0.19s
KL Matching	249.97s	78.65s	141.63s	33.51s
Mahalanobis	2135.13s	626.80s	1210.82s	243.69s
ViM	1.49s	0.51s	0.86s	0.18s

Table 8. Score computation time for four methods on four OOD datasets. We assume that the features have been extracted, so the network forward time is not included. The implementation uses numpy and runs on Intel Xeon (Skylake) 23.20GHz CPU.

Mahalanobis score using the feature vector before the final classification fc layer, as in [3]. The precision matrix and the class-wise average vector are estimated using 200,000 random training samples. The ground-truth class label is used during computation.

KL Matching We estimate the class-wise average probability using 200,000 random training samples. Following the practice of [5], the predicted class is used instead of ground-truth labels. We would like to note that the hyperparameter selection for OOD methods should not base on the ID set that is used for computing FPR95 and AUROC (in our case, its the validation set of ImageNet), because once the OOD method overfits the validation set, the evaluation result can be higher than the actual performance.

ReAct For ReAct, we use the Energy+ReAct setting, which is the most effective settings in [13]. In the original paper, they recommended the 90-th percentile of activations estimated on the ID data for the clipping threshold. However, for BiT and ViT, we found that the rectification percentile $p = 99$ works much better than 90. So we report results using $p = 99$.

F. OOD Examples Detected by KL Matching and Residual (Sec. 3)

In Sec. 3, we showed that feature-based OOD scores (*e.g.* Residual) and logit/softmax-based OOD scores (*e.g.* KL Matching) have different performances on the Texture OOD dataset. Here we visualize the OOD examples found by the two methods in Fig. 8.

G. Running Time of Four Methods (Sec. 6.2)

From Tab. 2 and Tab. 6, it is clear that the four most competitive methods are ViM, Mahalanobis, KL Matching, and Residual. Our ViM is the fastest among all four methods. We show their inference time on the four datasets in Tab. 8.

References

[1] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. RepVGG: Making VGG-style ConvNets great again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13733–13742, 2021. 2, 3

[2] 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. In *International Conference on Learning Representations*, 2021. 2

[3] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34, 2021. 4

[4] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 558–567, 2019. 2, 3

[5] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*, 2019. 3, 4

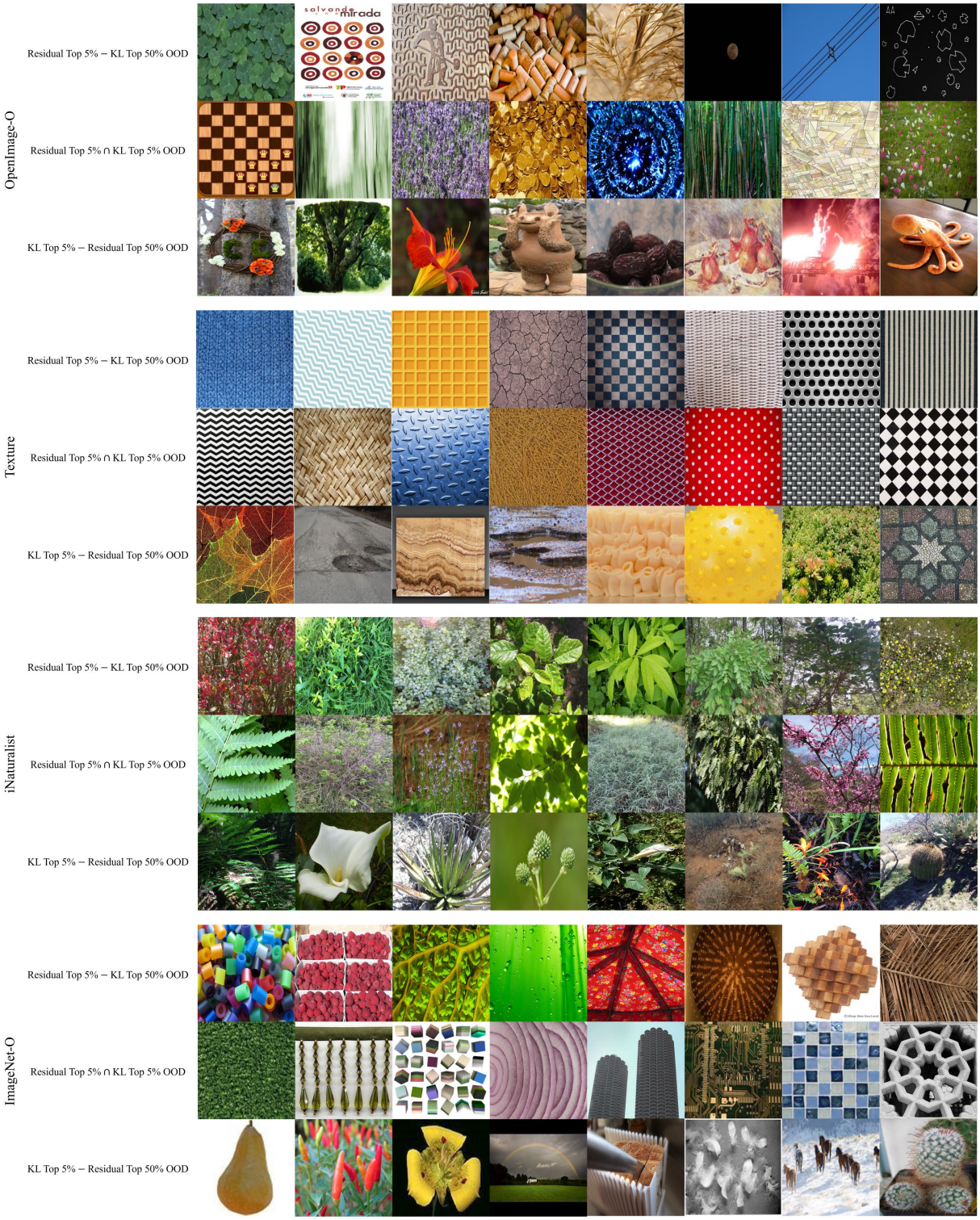

Figure 8. OOD examples detected by Residual and KL Matching. There are three rows for each OOD dataset. The first row shows images from the top 5% OODs detected by Residual, with overlapping images in the top 50% list of KL Matching removed. The second row displays images from the intersection of the top 5% OODs detected by Residual and the top 5% OODs detected by KL Matching. The third row shows images from the top 5% OODs detected by KL Matching, with overlapping ones in the top 50% list of Residual removed.

[6] 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. 3

[7] Rui Huang and Yixuan Li. MOS: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8710–8719, 2021. 1, 4

[8] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (BiT): General visual representation learning. In *European Conference on Computer Vision*, pages 491–507. Springer, 2020. 2

[9] 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. 3

[10] 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. 3

[11] 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. 3

[12] 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. 2, 3

[13] Yiyou Sun, Chuan Guo, and Yixuan Li. ReAct: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34, 2021. 3, 4

[14] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 2, 3