MOS: Towards Scaling Out-of-distribution Detection for Large Semantic Space (Supplementary Material)

A. Selected Categories in OOD Datasets

To evaluate our approach, we consider a diverse collection of OOD test datasets, spanning various domains. We carefully curate the OOD evaluation benchmarks to make sure concepts in these datasets do not overlap with ImageNet-1k [8]. Below we provide the list of concepts chosen for each OOD dataset, including iNaturalist [46], SUN [48], Places365 [50], and Textures [6]. We hope the information would allow future research to reproduce our results.

iNaturalist *Coprosma lucida*, *Cucurbita foetidissima*, *Mitella diphylla*, *Selaginella bigelovii*, *Toxicodendron vernix*, *Rumex obtusifolius*, *Ceratophyllum demersum*, *Streptopus amplexifolius*, *Portulaca oleracea*, *Cynodon dactylon*, *Agave lechuguilla*, *Pennantia corymbosa*, *Sapindus saponaria*, *Prunus serotina*, *Chondracanthus exasperatus*, *Sambucus racemosa*, *Polypodium vulgare*, *Rhus integrifolia*, *Woodwardia areolata*, *Epifagus virginiana*, *Rubus idaeus*, *Croton setiger*, *Mammillaria dioica*, *Opuntia littoralis*, *Cercis canadensis*, *Psidium guajava*, *Asclepias exaltata*, *Linaria purpurea*, *Ferocactus wislizeni*, *Briza minor*, *Arbutus menziesii*, *Corylus americana*, *Pleopeltis polypodioides*, *Myoporum laetum*, *Persea americana*, *Avena fatua*, *Blechnum discolor*, *Physocarpus capitatus*, *Ungnadia speciosa*, *Cercocarpus betuloides*, *Arisaema dracontium*, *Juniperus californica*, *Euphorbia prostrata*, *Leptopteris hymenophylloides*, *Arum italicum*, *Raphanus sativus*, *Myrsine australis*, *Lupinus stiversii*, *Pinus echinata*, *Geum macrophyllum*, *Ripogonum scandens*, *Echinocereus triglochidiatus*, *Cupressus macrocarpa*, *Ulmus crassifolia*, *Phormium tenax*, *Aptenia cordifolia*, *Osmunda claytoniana*, *Datura wrightii*, *Solanum rostratum*, *Viola adunca*, *Toxicodendron diversilobum*, *Viola sororia*, *Uropappus lindleyi*, *Veronica chamaedrys*, *Adenocaulon bicolor*, *Clintonia uniflora*, *Cirsium scariosum*, *Arum maculatum*, *Taraxacum officinale officinale*, *Orthilia secunda*, *Eryngium yuccifolium*, *Diodia virginiana*, *Cuscuta gronovii*, *Sisyrinchium montanum*, *Lotus corniculatus*, *Lamium purpureum*, *Ranunculus repens*, *Hirschfeldia incana*, *Phlox divaricata laphamii*, *Lilium martagon*, *Clarkia purpurea*, *Hibiscus moscheutos*, *Polanisia dodecandra*, *Fallugia paradoxa*, *Oenothera rosea*, *Proboscidea louisianica*, *Packera glabella*, *Impatiens parviflora*, *Glaucium flavum*, *Cirsium andersonii*, *Heliopsis helianthoides*, *Hesperis matronalis*, *Callirhoe pedata*, *Crocosmia × crocosmiiflora*, *Calochortus albus*, *Nuttallanthus canadensis*, *Argemone albiflora*, *Eriogonum fasciculatum*, *Pyrrhopappus pauciflorus*, *Zantedeschia aethiopica*, *Melilotus officinalis*, *Peritoma arborea*, *Sisyrinchium bellum*, *Lobelia siphilitica*, *Sorghastrum nutans*, *Typha domingensis*, *Rubus laciniatus*, *Dichelostemma congestum*, *Chimaphila maculata*, *Echinocactus texensis*

SUN *badlands*, *bamboo forest*, *bayou*, *botanical garden*, *canal (natural)*, *canal (urban)*, *catacomb*, *cavern (indoor)*, *corn field*, *creek*, *crevasse*, *desert (sand)*, *desert (vegetation)*, *field (cultivated)*, *field (wild)*, *fishpond*, *forest (broadleaf)*, *forest (needleleaf)*, *forest path*, *forest road*, *hayfield*, *ice floe*, *ice shelf*, *iceberg*, *islet*, *marsh*, *ocean*, *orchard*, *pond*, *rainforest*, *rice paddy*, *river*, *rock arch*, *sky*, *snowfield*, *swamp*, *tree farm*, *trench*, *vineyard*, *waterfall (block)*, *waterfall (fan)*, *waterfall (plunge)*, *wave*, *wheat field*, *herb garden*, *putting green*, *ski slope*, *topiary garden*, *vegetable garden*, *formal garden*

Places *badlands*, *bamboo forest*, *canal (natural)*, *canal (urban)*, *corn field*, *creek*, *crevasse*, *desert (sand)*, *desert (vegetation)*, *desert road*, *field (cultivated)*, *field (wild)*, *field road*, *forest (broadleaf)*, *forest path*, *forest road*, *formal garden*, *glacier*, *grotto*, *hayfield*, *ice floe*, *ice shelf*, *iceberg*, *igloo*, *islet*, *japanese garden*, *lagoon*, *lawn*, *marsh*, *ocean*, *orchard*, *pond*, *rainforest*, *rice paddy*, *river*, *rock arch*, *ski slope*, *sky*, *snowfield*, *swamp*, *swimming hole*, *topiary garden*, *tree farm*, *trench*, *tundra*, *underwater (ocean deep)*, *vegetable garden*, *waterfall*, *wave*, *wheat field*

Textures all images in this dataset

B. More Ablation Studies

B.1. MOS with Increasing Numbers of Classes (A More Challenging Setting)

In Section 4.3.2, we increase the number of in-distribution classes while fixing the number of *training images in each class* and observe the degradation of OOD detection performance. Here we investigate an alternative setting where we fix the number of *total training images* to be $35,000$, as we increase the number of classes $C \in \{50, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}$. For each C, we create training data by first randomly sampling C labels from the entire 1,000 classes in ImageNet-1k, and then sampling 35,000 / C images for each chosen label. In Figure 9, we show the OOD detection performance with varying numbers of in-distribution classes C.

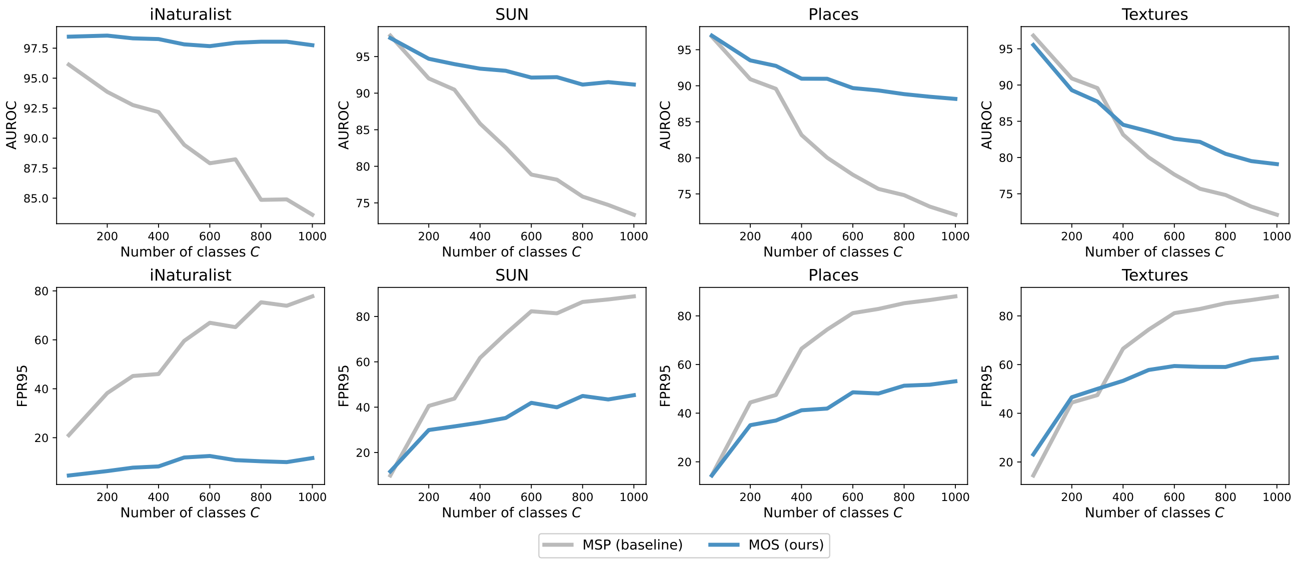

Figure 9: OOD detection performance of MOS (blue) and the MSP baseline (gray). MOS exhibits more stabilized performance as the number of in-distribution classes increases. For each OOD dataset, we show AUROC (*top*) and FPR95 (*bottom*). Different from Figure 5, we fix the number of *total training images* instead of the number of *training images in each category* in this experiment.

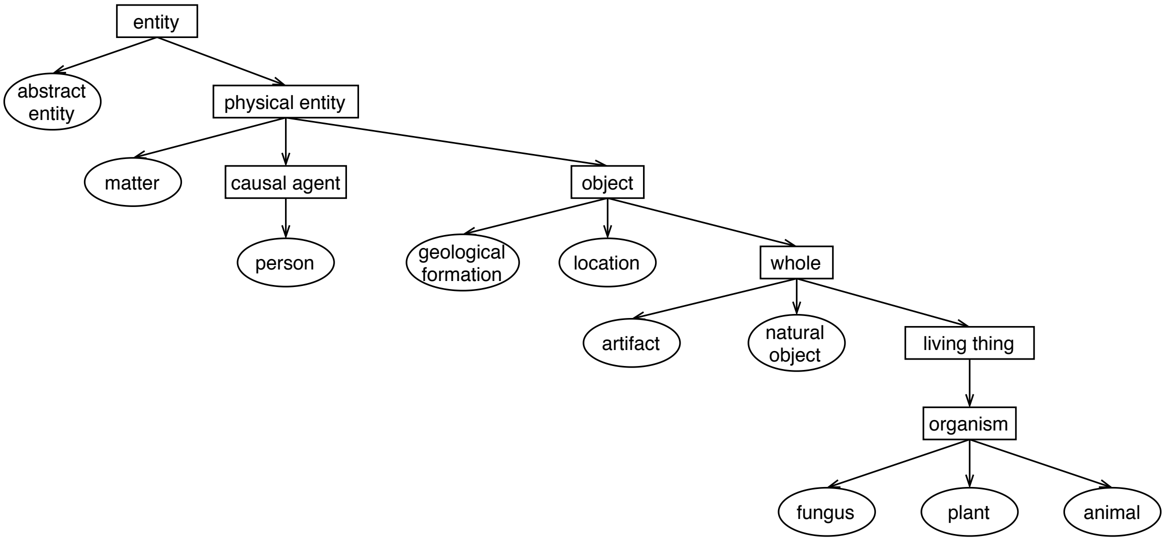

Figure 10: WordNet hierarchy. Super-classes of ImageNet-1k are based on the leaf nodes (in ellipses) except for `misc`. The super-class of `misc` contains 3 leaf nodes: `abstract entity`, `matter`, and `location`.

This is a more challenging setting because when the number of classes increases, there will be fewer training images for each class. Unsurprisingly, the performance degradation of OOD detection is more severe in this setting. For instance, on the iNaturalist OOD dataset, the FPR95 performance of MSP [16] degrades by 56.74% when the number of classes increases from 50 to 1,000, while the corresponding degradation is only 42.34% in the previous setting. Importantly, MOS remains much less sensitive to the change of the number of in-distribution classes compared to the MSP baseline (without grouping). In particular, on the Places OOD dataset, the FPR95 performance drops from 14.45% to 88.02% using MSP, while MOS degrades by only 38.73%.

B.2. MOS with Varying Numbers of Groups

In this ablation we investigate how different numbers of groups K affect the OOD detection performance of MOS under three grouping strategies: (1) taxonomy, (2) feature clustering, and (3) random grouping. For taxonomy-based grouping, in

Level	**Number of Groups**	**Grouping Strategy**	**iNaturalist**		**SUN**		**Places**		**Textures**		**Average**	
			AUROC ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓	**AUROC** ↑	**FPR95** ↓
-3	2	taxonomy	**97.66**	**11.20**	85.27	65.86	82.21	70.31	79.06	63.88	86.05	**52.81**
		feature clustering	94.91	33.12	**87.87**	**57.02**	**84.31**	**66.23**	**84.58**	**56.06**	**87.92**	53.11
		random grouping	91.12	46.82	79.28	78.73	78.64	76.50	79.02	63.53	82.02	66.40
-2	3	taxonomy	**97.21**	**15.78**	**92.28**	**40.08**	**89.35**	**49.74**	81.00	60.64	**89.96**	**41.56**
		feature clustering	94.57	33.58	87.23	57.18	83.60	65.34	**83.06**	**57.23**	87.12	53.33
		random grouping	90.75	47.55	76.57	83.00	75.89	81.33	80.40	61.93	80.90	68.45
-1	6	taxonomy	**97.16**	**16.16**	**92.07**	**40.28**	**89.12**	**49.53**	81.34	**60.27**	**89.92**	**41.56**
		feature clustering	90.68	53.49	82.95	72.24	79.48	75.06	**81.78**	62.50	83.72	65.82
		random grouping	91.55	44.73	79.11	78.99	78.31	76.17	80.93	62.30	82.48	65.55
0	8	taxonomy	**98.15**	**9.28**	**92.01**	**40.63**	**89.06**	**49.54**	**81.23**	**60.43**	**90.11**	**39.97**
		feature clustering	93.29	41.13	84.84	63.15	81.84	66.33	80.62	64.61	85.15	58.81
		random grouping	90.63	47.47	76.95	81.89	76.65	78.47	79.02	65.25	80.81	68.27
1	36	taxonomy	**95.85**	**23.73**	**90.51**	**47.53**	**87.74**	**52.51**	**88.47**	**45.55**	**90.64**	**42.33**
		feature clustering	91.22	47.73	79.25	79.48	76.53	79.00	82.81	61.72	82.45	66.98
		random grouping	91.01	46.96	79.66	77.79	79.36	74.35	78.91	68.72	82.24	66.96
2	85	taxonomy	92.22	46.19	**88.07**	**59.41**	**86.02**	**60.28**	**85.40**	**57.70**	**87.93**	**55.90**
		feature clustering	**93.13**	**40.07**	81.05	77.06	78.33	76.44	82.28	64.57	83.70	64.54
		random grouping	90.58	50.04	78.75	81.21	78.81	77.23	76.95	76.17	81.27	71.16
3	225	taxonomy	90.35	57.38	**85.19**	**71.72**	**83.57**	**69.99**	**81.40**	**72.27**	**85.13**	**67.84**
		feature clustering	**91.49**	**48.90**	79.59	82.16	78.06	79.97	79.40	75.09	82.14	71.53
		random grouping	89.66	56.81	77.55	84.73	78.16	79.61	75.07	82.43	80.11	75.90
4	416	taxonomy	89.18	63.48	**82.34**	80.60	**81.30**	**76.88**	**78.37**	81.17	**82.80**	75.53
		feature clustering	**91.66**	**47.91**	80.40	**79.40**	79.12	77.26	78.24	**80.00**	82.36	**71.14**
		random grouping	88.68	61.29	76.94	86.01	77.67	81.41	73.24	86.91	79.13	78.91
5	642	taxonomy	88.10	67.11	**80.07**	84.04	**79.65**	79.89	75.17	87.22	80.75	79.57
		feature clustering	**90.45**	**55.74**	79.83	**82.89**	79.22	**79.18**	**75.77**	**86.01**	**81.32**	**75.96**
		random grouping	88.31	63.92	77.09	85.94	77.60	81.69	72.16	89.08	78.79	80.16
6	789	taxonomy	87.39	69.61	78.57	85.73	**78.64**	81.04	73.57	89.29	79.54	81.42
		feature clustering	**89.81**	**59.68**	**78.78**	**84.75**	78.31	**80.82**	**74.66**	**88.65**	**80.39**	**78.48**
		random grouping	88.07	65.12	76.91	86.65	77.52	82.15	71.84	89.84	78.59	80.94

Table 2: Effect of different numbers of groups on OOD detection performance for 3 grouping strategies (taxonomy, feature clustering, and random grouping). Level 0 represents the level of super-classes in the taxonomy tree (main setting). Positive levels indicate splitting the super-classes into more groups (tracing down the taxonomy tree), while negative levels indicate merging the super-classes into fewer groups (tracing up the taxonomy tree).

order to increase the number of groups, we split the nodes of each super-class into their descendants in the label tree and map the 1,000 classes into one of the descendants instead of the super-classes themselves; in order to decrease the number of groups, we merge some of the super-classes into one group based on Figure 10. Specifically, we construct 10 taxonomy levels with increasing numbers of groups based on the label tree in the following way:

Level -3 There are 2 groups in Level -3: {`animal, plant, fungus`}, {`artifact, natural object, geological formation, person, misc`}.

Level -2 There are 3 groups in Level -2: {`animal, plant, fungus`}, {`artifact, natural object`}, {`geological formation, person, misc`}.

Level -1 There are 6 groups in Level -1: {`animal, plant, fungus`}, {`artifact`}, {`natural object`}, {`geological formation`}, {`person`}, {`misc`}.

Level 0 This is the level of 8 super-classes (main setting).

Level 1∼6 Groups in Level i are direct children of the nodes in Level $(i-1)$.

For feature clustering and random grouping, we set the numbers of groups to be equal to the corresponding numbers at each of the taxonomy levels for fair comparisons.

As shown in Table 2, for taxonomy-based grouping, the performance of OOD detection is almost optimal when the number of groups is 8 (Level 0), and further increasing or decreasing the number of groups will not lead to improved performance. Moreover, taxonomy-based grouping outperforms feature clustering and random grouping when K is small and mildly large. However, feature clustering surpasses taxonomy-based grouping when the number of groups is sufficiently large. We hypothesize that as we trace down the label tree, the numbers of categories in each group become more imbalanced, which could adversely impact the performance of OOD detection using taxonomy-based grouping.

C. AUROC Curves

Figure 11 shows the AUROC curves of MOS and MSP for OOD detection. All settings and training details are the same as in Table 1. The gray curve corresponds to the MSP baseline [16], while the blue curve corresponds to MOS with taxonomy-based grouping. We observe huge gaps between the gray and the blue AUROC curves on all OOD datasets. For instance, when TPR = 95%, the FPR can be reduced from 63.69% to 9.28% on the iNaturalist OOD dataset.

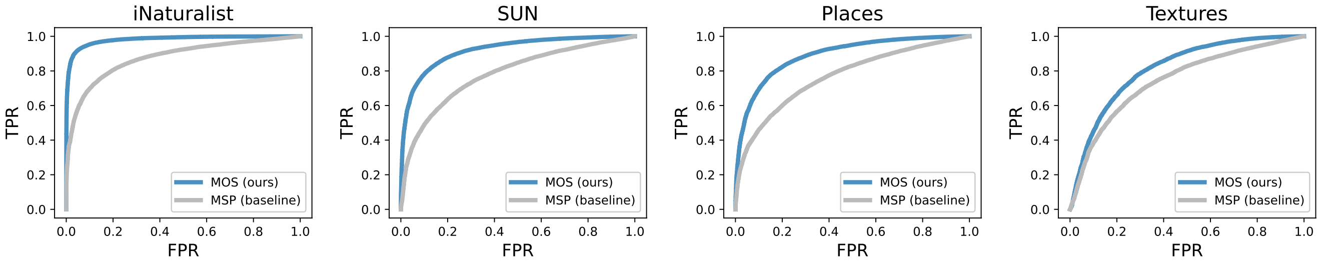

Figure 11: AUROC curves of MOS (blue) and MSP (gray) on four OOD datasets.

D. AUPR Results

OOD Dataset	Method	AUROC ↑	FPR95 ↓	AUPR ↑
iNaturalist	MSP	87.59	63.69	97.23
	ODIN	89.36	62.69	97.76
	Mahalanobis	46.33	96.34	81.14
	Energy	88.48	64.91	97.58
	KL Matching	93.00	27.36	97.93
	MOS (ours)	**98.15**	**9.28**	**99.62**
SUN	MSP	78.34	79.98	94.45
	ODIN	83.92	71.67	96.26
	Mahalanobis	65.20	88.43	88.81
	Energy	85.32	65.33	96.57
	KL Matching	78.72	67.52	94.10
	MOS (ours)	**92.01**	**40.63**	**98.17**
Places	MSP	76.76	81.44	94.15
	ODIN	80.67	76.27	95.35
	Mahalanobis	64.46	89.75	88.85
	Energy	81.37	73.02	95.49
	KL Matching	76.49	72.61	93.61
	MOS (ours)	**89.06**	**49.54**	**97.36**
Textures	MSP	74.45	82.73	95.65
	ODIN	76.30	81.31	96.12
	Mahalanobis	72.10	52.23	91.89
	Energy	75.79	80.87	96.05
	KL Matching	**87.07**	**49.70**	**97.97**
	MOS (ours)	81.23	60.43	96.65
Average	MSP	79.29	76.96	95.37
	ODIN	82.56	72.99	96.37
	Mahalanobis	62.02	81.69	87.67
	Energy	82.74	71.03	96.42
	KL Matching	83.82	54.30	95.90
	MOS (ours)	**90.11**	**39.97**	**97.95**

Table 3: Main results with AUPR. Experimental setups are the same as in Table 1.

In Table 3 we report the area under the precision-recall curve (AUPR) complementing the AUROC and FPR95 results in Table 1. AUPR is an informative metric in the presence of class imbalance, which is common in OOD detection. Again, MOS demonstrates state-of-the-art performance in terms of AUPR.

E. Others Scores for All In-distribution Groups and OOD Datasets

Figure 12 and Figure 13 show average others scores for 8 in-distribution groups and 4 OOD datasets, respectively. For in-distribution groups, others scores are averaged among all validation images in each group in ImageNet-1k; for OOD datasets, others scores are averaged among all sampled images in the curated datasets.

These histograms provide visual justifications for our method MOS: in-distribution images will have low others scores in at least one group (shown in red boxes), while out-of-distribution images will have high others scores in all 8 groups. Therefore, MOS is effective in distinguishing between in- vs. out-of-distribution data.

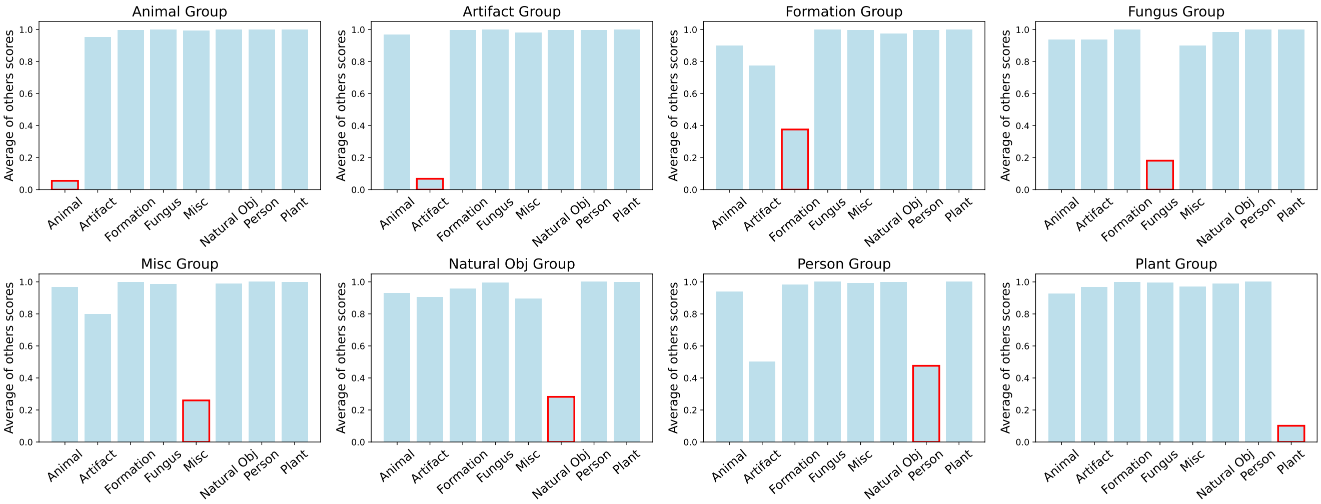

Figure 12: Average others scores for all in-distribution groups. Red boxes indicate the corresponding groups these images belong to.

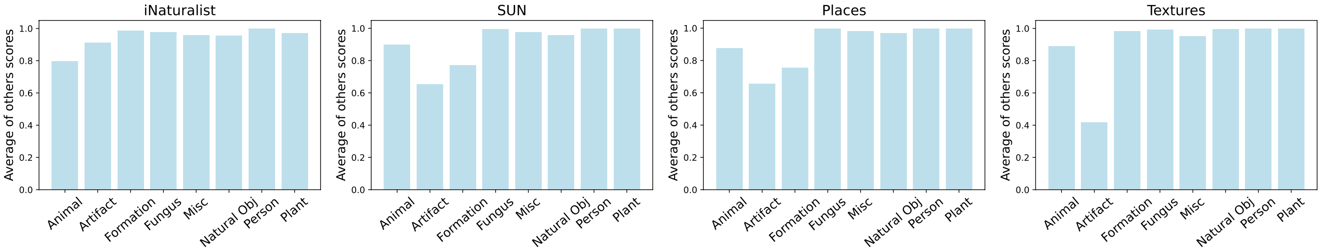

Figure 13: Average others scores for all OOD datasets