Table 3. Feature selection methods.

Selection Method	**COCOShift_balanced**			**COCOShift75**			**COCOShift95**		
	ROC-AUC ↑		% selected feat.	ROC-AUC ↑		% selected feat.	ROC-AUC ↑		% selected feat.
	all feat.	**kept** feat.		all feat.	**kept** feat.		all feat.	**kept** feat.	
Stylist (ours)	80.9	83.5 (+2.6)	40	80.4	**84.7 (+4.3)**	10	79.8	**85.1 (+5.3)**	30
InfoGain	80.9	81.0 (+0.0)	95	80.4	80.7 (+0.3)	90	79.8	79.9 (+0.1)	90
FisherScore	80.9	80.9 (+0.0)	100	80.4	80.4 (+0.0)	100	79.8	79.8 (+0.0)	100
MAD	80.9	80.9 (+0.0)	100	80.4	80.4 (+0.0)	100	79.8	79.8 (+0.0)	100
Dispersion	80.9	84.1 (+3.2)	35	80.4	83.0 (+2.6)	45	79.8	82.2 (+2.4)	50
Variance	80.9	80.9 (+0.0)	100	80.4	80.4 (+0.0)	100	79.8	79.8 (+0.0)	100
PCA Loadings	80.9	**84.6 (+3.7)**	35	80.4	**84.7 (+4.3)**	30	79.8	83.7 (+3.9)	35

A. Feature selection methods - detailed

We show in Tab. 3 individual results for multiple feature selection algorithms, grouped into env-aware ones and algorithms that are not env-aware. Please note that we adapt basic algorithms for feature selection to make them env-aware.

Considered feature selection methods:

- env-aware
  - InfoGain infogain: We adapt the method to the env-ware setup. We compute the mutual information between each feature and the style labels. High scores indicate a higher dependency between feature and style labels → environment-biased feature.
  - FisherScore fisher: We adapt the method to the env-aware setup. We rank the features based on their relevance for the classification of style categories.
- non environment-aware
  - MAD: For one feature, it computes the average of absolute differences between each sample value and the mean value. High MAD values indicate high discriminatory power.
  - Dispersion: This is computed as the ratio of arithmetic and geometric means. High dispersion implies a higher discriminatory power.
  - Variance: Generally, you can use variance to discard zero variance features as being completely uninformative. We have ranked the features based on their variance, considering high-variance features as being more informative.
  - PCA Loadings: We compute the contribution of each feature to the set of principal components identified by PCA.

B. Different shift robustness

We analyzed in Fig. 7 what is the impact of *Stylist* when we consider *Sub-Population shifts*, in extension to *Covariate shifts*, presented in the main paper. The testing dataset in this case is balanced, ID with the first plot (with no correlation between style and content). Our method manages to improve its performance when compared with other feature selectors, as training and testing become more OOD. The observations are similar in both kinds of shifts.

C. Stylist ablation distances

In Fig. 8 and Tab. 4 we show results when using symmetric KL or Wasserstein to measure per feature distribution distances between any two training environments. We combine the score or ranking obtained, over all pairs of training envs, using mean, median, median ranking or weighted mean ranking.

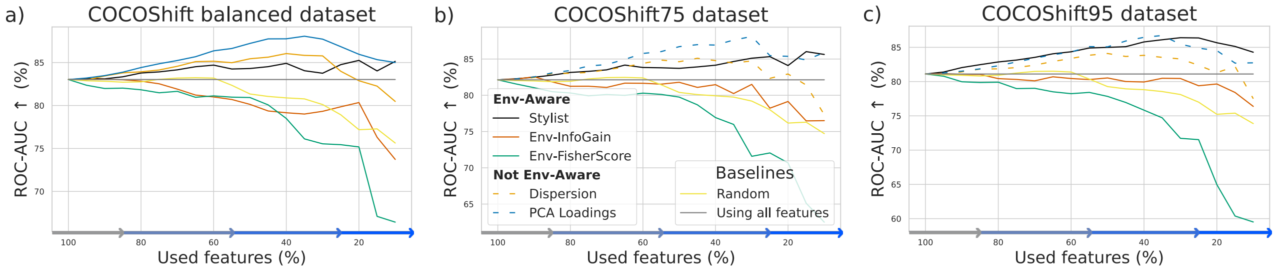


Figure 7. Feature selection algorithms. The reported ROC-AUC performance is for a testing dataset coming from the same (ID) distribution with a), showing that similar observations remain for sub-population OOD shifts. TODO: a) corresponds to ID setting while c) corresponds to a strong OOD setting

Table 4. Ablation distance and ranking combining. We notice here that all variants of distances and feature ranking combinations manage to improve. Depending on the dataset, there are different chooses to make as hyper-parameters. We use for Stylist the mean of Wasserstein distances over all training pairs.

Distance	**Method**	**fMoW**			**DomainNet**			**COCOShift95**		
		ROC-AUC ↑		% selected feat.	ROC-AUC ↑		% selected feat.	ROC-AUC ↑		% selected feat.
		all feat.	**Stylist** feat.		all feat.	**Stylist** feat.		all feat.	**Stylist** feat.	
Wasserstein	**mean**	59.0	60.3 (+1.3)	20	50.6	50.8 (+0.2)	40	79.8	85.1 (+5.3)	30
	median	59.0	60.6 (+1.5)	15	50.6	50.8 (+0.2)	75	79.8	84.2 (+4.4)	20
	median ranking	59.0	60.9 (+1.9)	15	50.6	50.8 (+0.2)	55	79.8	85.5 (+5.7)	30
	weighted mean ranking	59.0	60.7 (+1.7)	15	50.6	50.8 (+0.2)	45	79.8	85.2 (+5.4)	25
KL symmetric	mean	59.0	61.3 (+2.2)	30	50.6	50.8 (+0.2)	50	79.8	83.0 (+3.1)	15
	median	59.0	60.7 (+1.6)	50	50.6	50.8 (+0.2)	45	79.8	83.7 (+3.8)	25
	median ranking	59.0	61.6 (+2.6)	30	50.6	50.7 (+0.1)	45	79.8	83.3 (+3.5)	10
	weighted mean ranking	59.0	60.9 (+1.9)	25	50.6	50.8 (+0.2)	35	79.8	83.4 (+3.6)	30

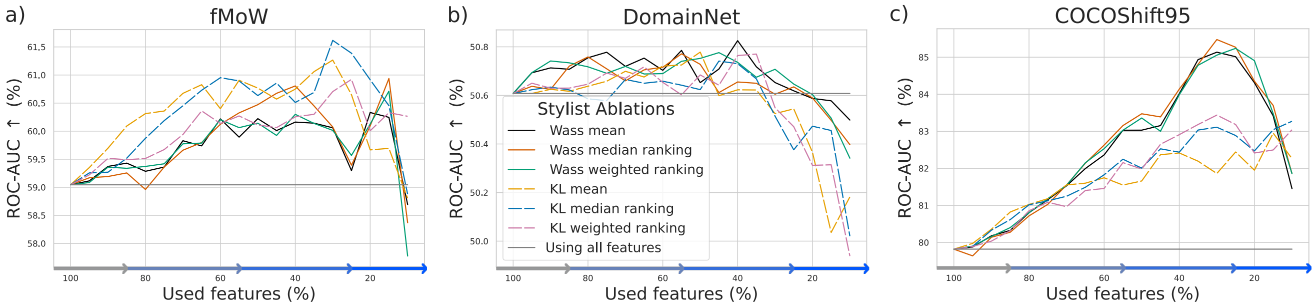


Figure 8. **Stylist ablations.** We vary both the distance metrics (based on Wasserstein or symmetric KL) and the ranking combination approaches for pairs of environments (mean, median ranking, weighted ranking). Notice how usually Wasserstein distances perform better (except for a) fMoW). Also, see how the ranking combination does not have a high influence on the result.

D. Qualitative analysis

In this section, we analyze the top-ranked features identified by Stylist. Specifically, we extract samples with the highest activation for a given feature and investigate their common characteristics. To this end, Figure 9 showcases the top six samples with the highest activations for a selection of features from the top of the Stylist ranking. Each subfigure reveals that the common traits among the images are related to style, particularly centered around environments (sketches/quickdraw for DomainNet and forest/field for COCOShift95). Table 5 presents the features alongside their positions in Stylist's ranking, as well as in other ranking algorithms. The table illustrates that some features, highly ranked by Stylist, receive lower ratings from other algorithms. However, the images in Figure 9 suggest that these features are indeed related to style.

Table 5. The proportion within which features are located for different rankings and training datasets. Features top-ranked by Stylist are ranked lower by InfoGain and PCA Loadings ranking algorithms.

Training Dataset	Feature Index	Feature in the Top % of Rankings		
		Stylist	InfoGain	PCA Loadings
DomainNet	189	3.12%	17.58%	6.25%
DomainNet	361	3.71%	29.30%	3.91%
DomainNet	154	3.91%	8.59%	54.49%
DomainNet	58	4.10%	10.55%	21.09%
COCOShift95	297	0.78%	2.15%	23.83%
COCOShift95	435	2.34%	31.45%	3.91%
COCOShift95	252	2.73%	20.31%	54.69%

E. Determine the proportion of selected features

We can determine the optimal proportion of features following a validation methodology on the ID validation set or by employing an oracle strategy and interrogating the OOD test set. In Tab. 6 we compare these approaches, highlighting that the optimal number of features is stable between the two setups. The variance between the two approaches is less than 0.015. Also, we emphasize that Stylist always improves over the baseline considering 100% of the features, in all the considered configurations.

F. Future Work

We leave here several unexplored directions we consider valuable for further investigations:

1. Analyze the impact of the pretrained feature extractor, looking after different axes of variation: supervised/unsupervised pretraining, high/low disentanglement. And going even further, find an unsupervised way to choose the best feature extractor, given a dataset. Also, target methods that promise to disentangle the features, like Sparse AE.

2. Explore more complex algorithms for ranking, based on the same principle of emphasizing the intra and inter environment distances. More related to the algorithm, explore an unsupervised manner to choose the best percent of features to keep.

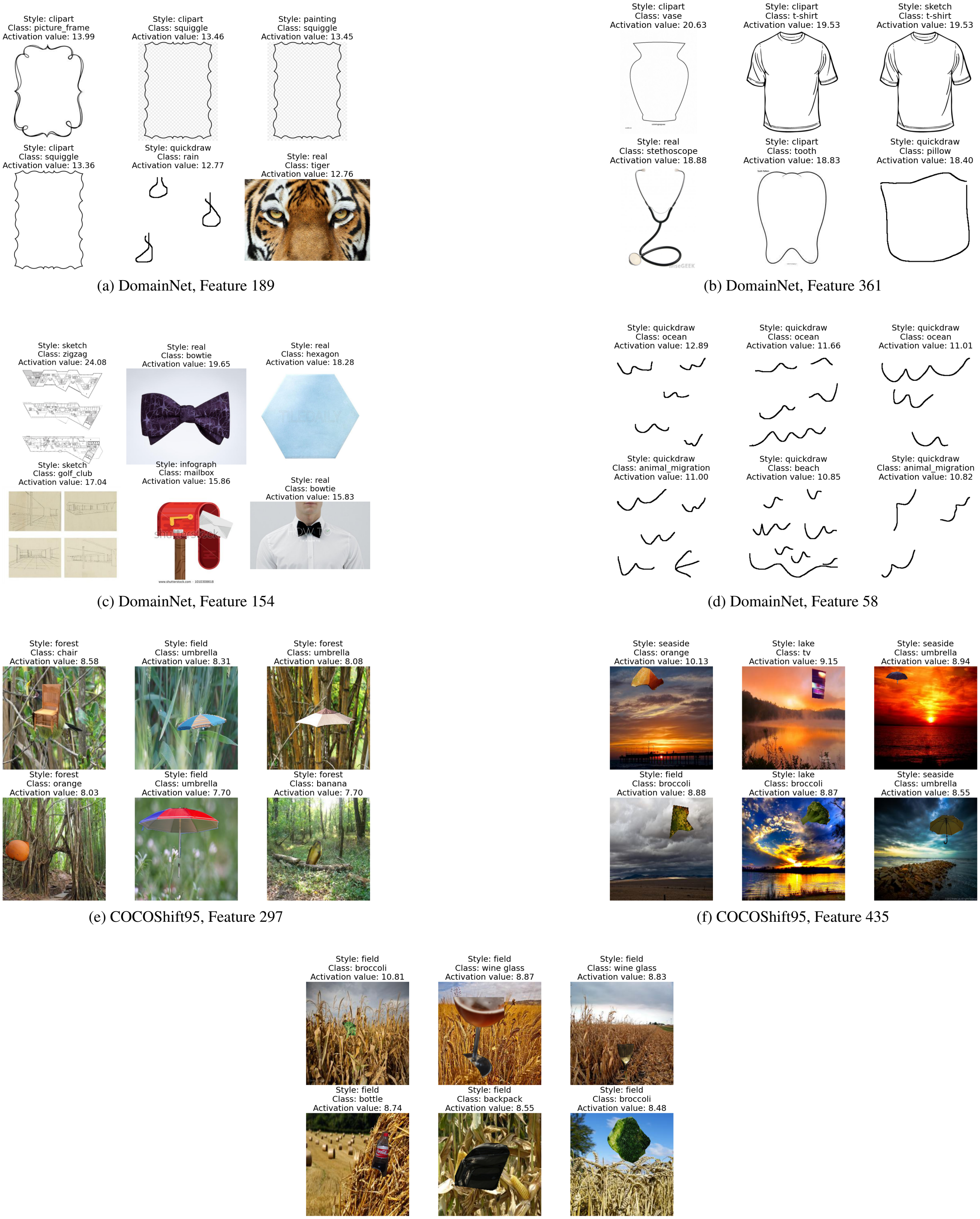
3. Take the approach beyond novelty detection, analyzing the performance improvement of feature ranking and selection w.r.t. other supervised approaches for OOD robustness.

4. Making the approach more automatic and less environment label dependent, by providing (or couple it with) an environment discovery solution.

G. Benchmarks

G.1. fMoW

- Content (functional purpose of buildings):



(a) DomainNet, Feature 189

(b) DomainNet, Feature 361

(c) DomainNet, Feature 154

(d) DomainNet, Feature 58

(e) COCOShift95, Feature 297

(f) COCOShift95, Feature 435

(g) COCOShift95, Feature 252

Figure 9. Images with the highest activation for a given feature and training dataset (ResNet-18 Embeddings)

Table 6. Feature selection approach: based on ID validset vs OOD testset. Notice extremely small variances in results, between the two approaches.

Selection Method	COCOShift_balanced			COCOShift75			COCOShift95		
	ROC-AUC ↑		% selected feat.	ROC-AUC ↑		% selected feat.	ROC-AUC ↑		% selected feat.
	all feat.	**kept** feat.		all feat.	**kept** feat.		all feat.	**kept** feat.	
ResNet-18									
Based on ID val set	80.93	83.24 (+2.31)	10	80.40	84.71 (+4.31)	10	79.82	85.01 (+5.19)	25
Based on OOD test set	80.93	83.49 (+2.56)	40	80.40	84.71 (+4.31)	10	79.82	85.13 (+5.31)	30
Variance		**0.015**			**0.00**			**0.003**	
CLIP									
Based on ID val set	95.06	95.40 (+0.34)	95	94.88	95.21 (+0.33)	95	94.53	94.92 (+0.39)	95
Based on OOD test set	95.06	95.40 (+0.34)	95	94.88	95.21 (+0.33)	95	94.53	94.92 (+0.39)	95
Variance		**0.00**			**0.00**			**0.00**	

  - **normal**: airport, airport terminal, barn, burial site, car dealership, dam, debris or rubble, educational institution, electric substation, fountain, gas station, golf course, hospital, interchange, multi-unit residential, parking lot or garage, police station, port, railway bridge, recreational facility, road bridge, runway, shipyard, shopping mall, solar farm, space facility, surface mine, swimming pool, waste disposal, water treatment facility, zoo
  - **novel**: airport hangar, amusement park, aquaculture, archaeological site, border checkpoint, construction site, crop field, factory or powerplant, fire station, flooded road, ground transportation station, helipad, impoverished settlement, lake or pond, lighthouse, military facility, nuclear powerplant, office building, oil or gas facility, park, place of worship, prison, race track, single-unit residential, smokestack, stadium, storage tank, toll booth, tower, tunnel opening, wind farm

- Style (geographical area):
  - **ID**: Europe, America, Asia, Africa
  - **OOD**: Australia

- Number of samples:
  - OOD test set: 3469
  - ID train set: 55288
  - ID test set: 13817
  - ID val set: 6911

G.2. DomainNet

- Content (object classes):
  - **normal**: aircraft carrier, angel, animal migration, apple, arm, backpack, barn, basketball, bed, belt, birthday cake, blackberry, blueberry, book, boomerang, bowtie, brain, bread, bucket, butterfly, cactus, cake, camouflage, cannon, carrot, cat, cello, chandelier, circle, cloud, coffee cup, computer, cookie, couch, crab, crayon, crocodile, cruise ship, diamond, diving board, dog, donut, door, dresser, drill, drums, duck, ear, elbow, envelope, eraser, fan, fence, flower, flying saucer, fork, frog, garden, guitar, hand, headphones, helicopter, helmet, hexagon, hockey stick, horse, hospital, hot air balloon, hot dog, hourglass, house, hurricane, jacket, jail, kangaroo, knife, laptop, leg, light bulb, lighter, lightning, lipstick, lobster, map, microphone, microwave, mountain, moustache, mug, mushroom, necklace, nose, owl, paint can, paintbrush, palm tree, parachute, parrot, peanut, pear, peas, piano, pig, pillow, pineapple, pizza, police car, pond, postcard, power outlet, radio, rain, rake, remote control, roller coaster, sailboat, saw, saxophone, screwdriver, sea turtle, see saw, shark, shorts, skull, sleeping bag, snail, snowman, soccer ball, spider, spoon, square,

stairs, star, stethoscope, stitches, stop sign, strawberry, streetlight, string bean, submarine, suitcase, sun, swan, sweater, swing set, syringe, table, teapot, teddy-bear, telephone, television, tennis racquet, tent, toaster, toe, tooth, traffic light, train, tree, trombone, truck, trumpet, umbrella, underwear, van, vase, violin, whale, wheel, wine bottle, wristwatch, zebra

  - **novel**: airplane, alarm clock, ambulance, ant, anvil, asparagus, axe, banana, bandage, baseball, baseball bat, basket, bat, bathtub, beach, bear, beard, bee, bench, bicycle, binoculars, bird, bottlecap, bracelet, bridge, broccoli, broom, bulldozer, bus, bush, calculator, calendar, camel, camera, campfire, candle, canoe, car, castle, ceiling fan, cell phone, chair, church, clarinet, clock, compass, cooler, cow, crown, cup, dishwasher, dolphin, dragon, dumbbell, elephant, eye, eyeglasses, face, feather, finger, fire hydrant, fireplace, firetruck, fish, flamingo, flashlight, flip flops, floor lamp, foot, frying pan, garden hose, giraffe, goatee, golf club, grapes, grass, hamburger, hammer, harp, hat, hedgehog, hockey puck, hot tub, house plant, ice cream, key, keyboard, knee, ladder, lantern, leaf, lighthouse, line, lion, lollipop, mailbox, marker, matches, megaphone, mermaid, monkey, moon, mosquito, motorbike, mouse, mouth, nail, ocean, octagon, octopus, onion, oven, panda, pants, paper clip, passport, pencil, penguin, pickup truck, picture frame, pliers, pool, popsicle, potato, purse, rabbit, raccoon, rainbow, rhinoceros, rifle, river, rollerskates, sandwich, school bus, scissors, scorpion, sheep, shoe, shovel, sink, skateboard, skyscraper, smiley face, snake, snorkel, snowflake, sock, speedboat, spreadsheet, squiggle, squirrel, steak, stereo, stove, sword, t-shirt, The Eiffel Tower, The Great Wall of China, The Mona Lisa, tiger, toilet, toothbrush, toothpaste, tornado, tractor, triangle, washing machine, watermelon, waterslide, windmill, wine glass, yoga, zigzag

- Style (manner of depiction):
  - **ID**: real, painting, clipart, infograph
  - **OOD**: sketch, quickdraw
- Number of samples:
  - OOD test set: 242886
  - ID train set: 142026
  - ID test set: 35313
  - ID val set: 17753

G.3. COCOShift

Each COCOShift environment is composed of 5 closely related categories of Places365 as follows:

- **forest**: forest, rainforest, bamboo_forest, forest_path, forest_road,
- **mountain**: mountain, mountain_snowy, glacier, mountain_path, crevasse
- **seaside**: beach, coast, ocean, boathouse, beach_house"
- **garden**: botanical_garden, formal_garden, japanese_garden, vegetable_garden, greenhouse
- **field**: field_cultivated, field_wild, wheat_field, corn_field, field_road
- **rock**: badlands, butte, canyon, cliff, grotto
- **lake**: lake, lagoon, swamp, marsh, hot_spring
- **farm**: orchard, vineyard, farm, rice_paddy, pasture
- **sport_field**: soccer_field, football_field, golf_course, baseball_field, athletic_field

On the other hand, we worked with superclasses from COCO COCO such that there would be a significant shift between classes.

To make sure that the content is identifiable from each image (as many COCO segmentations provide little content without it's context), we tested each generated image against CLIP clip. Specifically, we took a given merged picture into COCOShift dataset if CLIP could correctly identify the content between the selected COCO classes (not superclasses) listed below. The same test is effectuated on images of segmentations over white backgrounds.

- Content (object category):
  - **normal**: food (composed of classes: hot dog, cake, donut, carrot, sandwich, broccoli, banana, apple, pizza, orange)
  - **novel**: electronic (composed of classes: remote, laptop, tv, cell phone, keyboard), kitchen (composed of classes: bottle, cup, wine glass, knife, fork, bowl, spoon)
- Style (background area surrounding the object):
  - **ID**: forest, mountain, field, rock, farm
  - **OOD**: lake, seaside, garden, sport field
- Number of samples:
  - OOD test set: 13013
  - ID test set: 1623
  - COCOShift_balanced
    - ID train set: 7033
    - ID val set: 880
  - COCOShift75
    - ID train set: 4221
    - ID val set: 529
  - COCOShift90
    - ID train set: 3284
    - ID val set: 412
  - COCOShift95
    - ID train set: 3037
    - ID val set: 379
- **Spurious correlation** sets are variants of train sets used for the synthetic COCOShift dataset, where we eliminate samples to create spuriousity.
  - **COCOShift_balanced**: normal samples are uniformly distributed among the ID environments ($\approx 1.4k$ samples per environment)
  - **COCOShift75**: environments [farm, mountain] have $\approx 1.4k$ normal samples, while environments [rock, forest, field] have ≈ 400 normal samples
  - **COCOShift90**: environments [farm, mountain] have $\approx 1.4k$ normal samples, while environments [rock, forest, field] have ≈ 150 normal samples
  - **COCOShift95**: environments [farm, mountain] have $\approx 1.4k$ normal samples, while environments [rock, forest, field] have ≈ 70 normal samples