Supplemental Material for: What Ails One-Shot Image Segmentation: A Data Perspective

Abhinav Patel[*]
Adobe Inc.
abpatel@adobe.com

Anirudha Ramesh[*]
IIIT Hyderabad, India
anramesh@iiit.ac.in

Tejas Shimpi[*]
IIT BHU, India
tshimpi@iitbhu.ac.in

Mayur Hemani[*]
Adobe Inc.
mayur@adobe.com

Balaji K.
Adobe Inc.
kbalaji@adobe.com

1 Author Statement

The current dataset is a reorganization of the existing PASCAL 5^i dataset by Shaban et al. and the FSS-1000 dataset by Li et al. There are no new assets, but the existing assets are augmented at the time of test setup. For more instructions see the dataset web-page. `https://github.com/fewshotseg/toss`

2 Obtaining the TOSS Dataset

The Tiered One-shot Segmentation dataset is constructed for nuanced evaluation of one-shot segmentation solutions. It is based on the PASCAL 5^i and FSS-1000 datasets. URL: `https://github.com/fewshotseg/toss`
To setup the data for testing, follow the instructions in the README file.

3 Experiments supporting Class-Negative Bias: More details

As discussed in the paper in section 3.2, the *leave-out* experiment demonstrates that training with fewer examples can improve the results provided there are no *distracting* objects in them. Table 1 presents the class-wise results for each split of the PASCAL 5^i dataset on the baseline network and RPMM. It indicates a strong correlation between the percentage of examples with distracting pixels and the improvement in performance that can be obtained by dropping those examples in training. The drop in accuracy for other classes is expected and can be attributed to the fewer examples used for training.

4 Design Choices for the TOSS Dataset

4.1 Attributes for Input Complexity

As discussed in Section 4.1.1 (and Figure 3) of the paper, a number of different attributes can be considered to measure the complexity of the images (Figure 2). We analyze 15000 images from the PASCAL VOC dataset. The distribution of these attributes are depicted in Figure 3. The distributions are used to compute thresholds to convert these real-valued attributes into binary attributes for

---

[*] equal contribution

35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks.

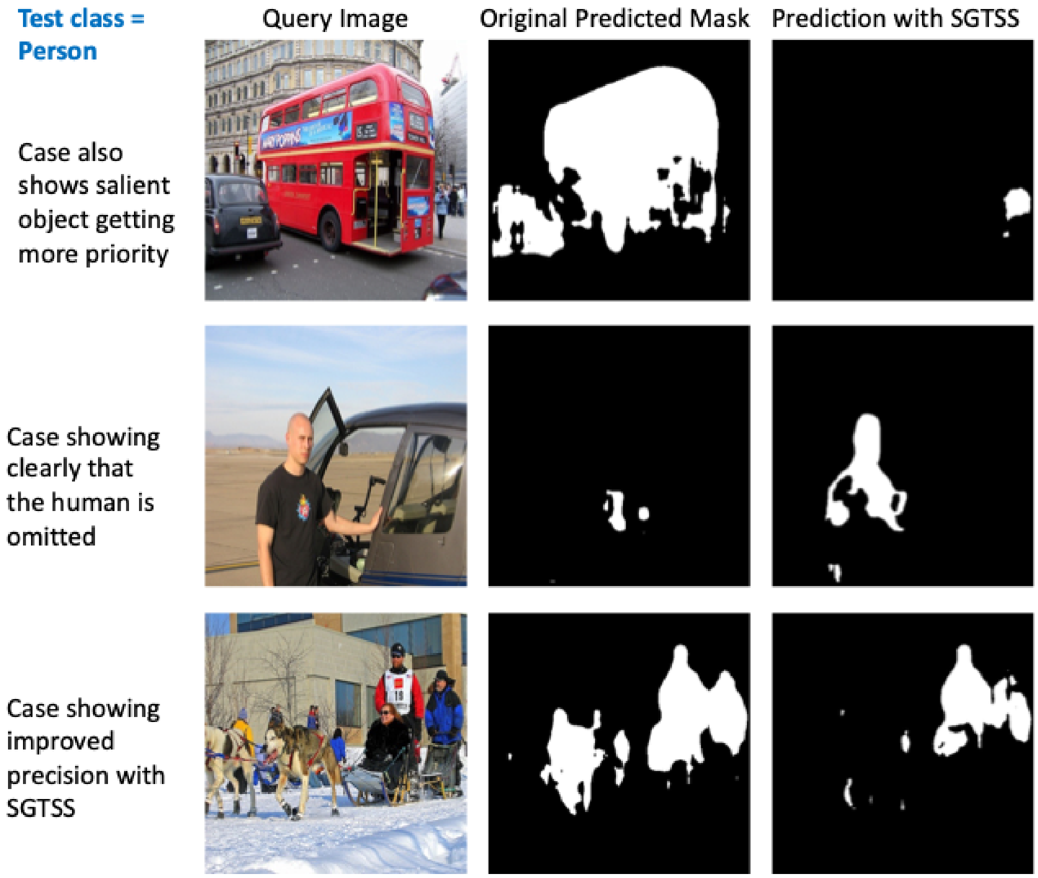
Test class = Person
Query Image
Original Predicted Mask
Prediction with SGTSS
Case also shows salient object getting more priority
Case showing clearly that the human is omitted
Case showing improved precision with SGTSS

Figure 1: Qualitative examples for the Person test class indicating clear cases of class-negative and salience bias at play.

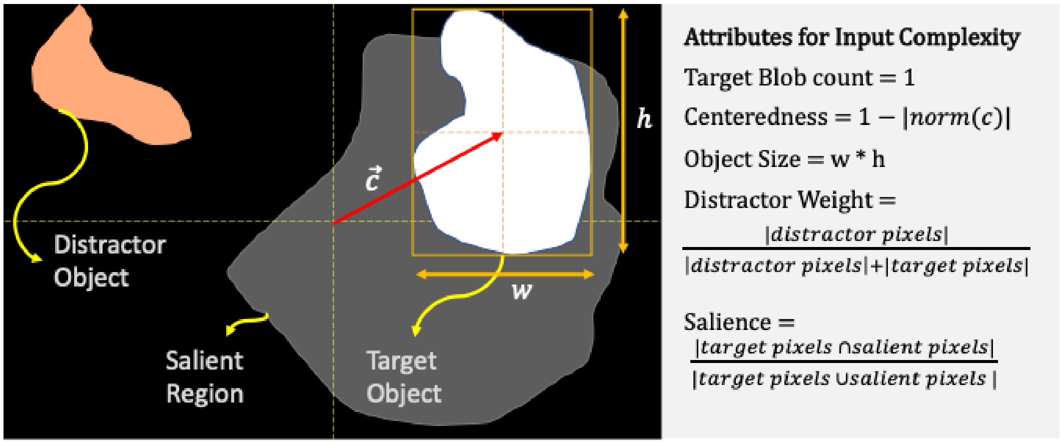
Attributes for Input Complexity
Target Blob count = 1
Centeredness = $1 - |norm(c)|$
Object Size = w * h
Distractor Weight = $\frac{|distractor\ pixels|}{|distractor\ pixels| + |target\ pixels|}$
Salience = $\frac{|target\ pixels \cap salient\ pixels|}{|target\ pixels \cup salient\ pixels|}$
h
$\vec{c}$
w
Distractor Object
Salient Region
Target Object

Figure 2: Attributes considered for computing the query image complexity. TOSS only uses salience to keep the number of splits manageable.

determining samples.Each binary attribute (high/low) adds a factor of 2 to the number of splits. To keep the number of splits within manageable limits, we only choose the attribute with the highest mutual information with respect to the accuracy of the predicted mask through the baseline network.

	Baseline Network		RPMM		
Class	Baseline	Leave-out	Baseline	Leave-out	Percent Distractors
aeroplane	67.16	65.71	64.29	71.08	0.73
bicycle	54.04	51.92	40.78	50.78	3.38
bird	65.70	66.50	60.99	67.05	0.42
boat	39.43	38.25	39.89	48.41	1.61
bottle	28.99	33.1	22.62	26.96	6.54
bus	81.49	82.09	76.99	81.74	1.68
car	53.28	58.89	45.06	55.78	6.29
cat	82.37	81.96	80.35	83.68	1.73
chair	22.6	26.59	22.93	26.53	12.24
cow	82.05	82.10	81.69	82.02	0.48
dining-table	26.27	21.71	10.07	21.31	7.29
dog	79.06	78.4	76.12	79.14	2.83
horse	75.19	73.98	76.88	74.96	0.47
motorbike	66.94	66.84	65.46	64.94	1.54
person	31.00	49.42	8.96	36.82	30.05
potted-plant	27.84	28.68	20.79	18.10	3.97
sheep	82.05	77.7	76.89	79.20	0.58
sofa	48.49	43.34	40.86	50.20	3.93
train	65.49	72.22	65.35	69.36	1.15
tv-monitor	28.96	31.35	17.68	20.48	3.83

Table 1: Leave-out Scores: Filtering out training images with test-class pixels produces better results for classes with high percentage of distractors. See Person class for instance.

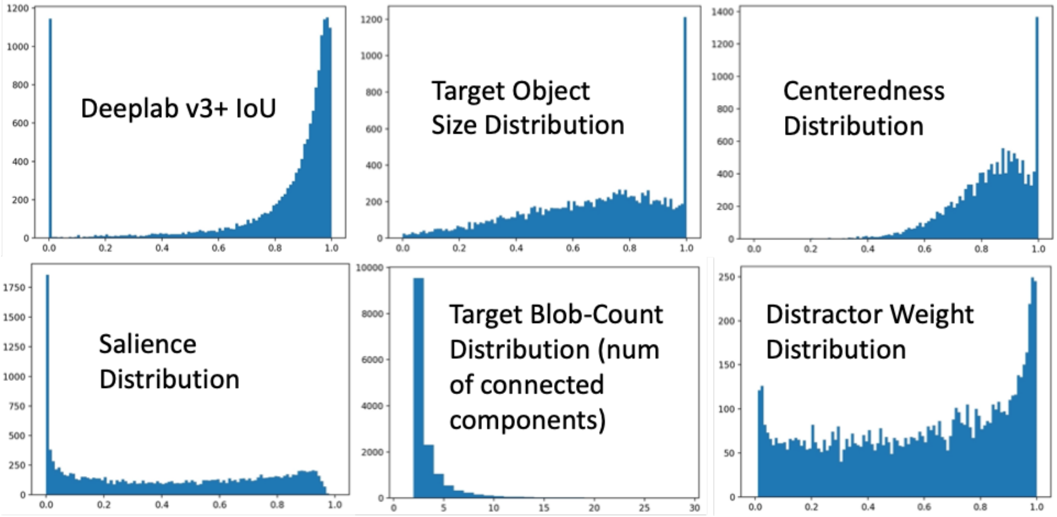

Figure 3: Distribution of different complexity-related attributes in the PASCAL VOC Dataset. These are used to determine threshold values to convert these real-valued attributes into binary attributes.

	DLv3+ IoU	Size	Centeredness	Salience	Blobs	Distractor weights
Mutual info	0.091	0.033	0.022	0.21	0.0023	0.16

The mutual information is measured against the binary criterion that determines if the prediction from the baseline network for a particular image (when treated as the query) results in a mask with higher than 0.5 mIoU. Multiple predictions for the same query image are used to average the baseline mIoU for the image.

The DeepLab v3+ iou is a composite factor because it entails several other latent/unconsidered factors for diminished performance. Also as presented in the paper, baseline one-shot performance follows the same trend as supervised accuray with DeepLab v3+. So the first attribute considered is IoU with DeepLab v3+.

Of the others, the maximum mutual information is obtained from Salience. Thus we chose Salience as the second attribute for dividing the test files.

4.2 Using training class images in the test set

We observe that the PASCAL 5i test set is quite small. We therefore include training class images as part of the test set. This is valid because we still differentiate between the training classes and the test classes. viz. a network trained for a fold with test classes c1, c2, c3, c4, c5 are only tested against annotations for these classes, albeit the images themselves maybe from the training set.

4.3 Scoring Weights for Query Complexity

The query complexity scores use weights derived from the distribution of salience and mIoU with a supervised method (DLv3+). The proportions are adjusted for computational convenience.

	μ(Salience)	mIoU (DLv3+)	Product/sum(Product)	Proportion	Adjusted Proportion
easy-sal	0.45	0.37	0.60	3.01	3
easy-nsal	0.15	0.36	0.19	0.97	0.75
hard-sal	0.32	0.15	0.17	0.84	0.75
hard-nsal	0.08	0.13	0.04	0.18	0.25

4.4 CLIP for similarity measurement

We use CLIP for similarity measurement because it is able to identify fine-grained differences between images [2]. We determine the cosine similarity between image features, and use the mean similarity value for each split as the threshold. We posit that this approach is aligned with other neural-network based means of measuring image similarity. To corroborate this, we compute similarity using pre-trained VGG-19 and ResNeXT models for 70000 pairs of images. We compute the cosine similarity between the features obtained from {CLIP, VGG19, ResNeXT} networks. Each one is converted to either high-similarity (1) and low-similarity (0) based on their mean values. Then we measure the agreement between CLIP similarity-classes with the other two networks. VGG19 agrees with CLIP 80.08% of the times, and ResNeXT agrees 86.25% with it. Thus, it is unlikely that the similarity splits obtained from using CLIP will be very different from those obtained using other image similarity measures.

4.5 Choice of U^2-Net for Salient Region Detection

We chose $U^2 - Net$ over other static/deep neural network-based salient region detection methods because it produces very crisp salient region boundaries whereas the other methods (Minimum Barrier (MBD),MR ,Spectral Residuals) produced diffused boundaries.

4.6 Removed classes in the Generalization Tier

To obtain classes that are semantically different from the PASCAL 5^i dataset, we use a subset of the FSS-1000 dataset. This dataset is constructed by removing classes from the FSS-1000 class set with any visual similarity to a class in the PASCAL 5^i classes. The following 222 classes were removed from the FSS-1000 dataset for use with the TOSS generalization split:

> afghan-hound, african-grey, air-strip, aircraft-carrier, airedale, airliner, airship, albatross, ambulance, american-staffordshire, andean-condor, angora, arctic-fox, astronaut, australian-terrier, baby, bald-eagle, banana-boat, baseball-player, basset, bat, beagle, bedlington-terrier, beer-bottle, beer-glass, bighorn-sheep, bison, bittern, black-grouse, black-stork, black-swan, blenheim-spaniel, bloodhound, bluetick, border-terrier, boston-bull, brambling, brasscica, briard, bulbul-bird, bullet-train, bus, bushtit, bustard, cactus, cactus-ball, canoe, car-wheel, cardoon, carriage, chickadee-bird, chicken, chihuahua, cocacola, condor, convertible, coucal, cougar,

---

[2] https://openai.com/blog/clip/

coyote, crane, crt-screen, cuckoo, curlew, daisy, dandie-dinmont, delta-wing, dhole, dingo, donkey, doublebus, dowitcher, downy-pitch, drake, eagle, egret, egyptian-cat, english-foxhound, english-setter, esport-chair, f1-racing, ferrari911, fire-engine, flamingo, flat-coated-retriever, flowerpot, folding-chair, fox, ganeva-chair, garbage-truck, gas-tank, germain-pointer, giant-schnauzer, golden-plover, golden-retriever, goldfinch, goose, grey-fox, groenendael, hang-glider, hawk, helicopter, hock, hornbill, housefinch, hummingbird, ibex, impala, jacamar, jay-bird, jet-aircraft, kinguin, kit-fox, laptop, lapwing, lark, leeks, lhasa-apso, lifeboat, little-blue-heron, lorikeet, lynx, macaque, macaw, magpie-bird, manx, maotai-bottle, meerkat, minicooper, mink, monitor, monocycle, motor-scooter, motorbike, muscle-car, narcissus, ocicat, oiltank-car, oriole, osprey, ostrich, owl, ox, park-bench, partridge, peacock, peregine-falcon, perfume, persian-cat, pheasant, pickup, police-car, potted-plant, prairie-chicken, ptarmigan, pteropus, quail, raft, rally-car, raven, recreational-vehicle, red-breasted-merganser, red-fox, red-wolf, redshank, rocket, rocking-chair, rose, ruddy-turnstone, ruffed-grouse, saluki, school-bus, schooner, seagull, shih-tzu, siamese-cat, skua, space-shuttle, sparrow, speedboat, spoonbill, sports-car, staffordshire, stealth-aircraft, steam-locomotive, stonechat, stork, streetcar, submarine, sulphur-crested, taxi, tebby-cat, television, tiger-cat, timber-wolf, toucan, tow-truck, trailer-truck, transport-helicopter, trimaran, trolleybus, tulip, vase, vulture, wagtail, wandering-albatross, warplane, water-bike, water-buffalo, water-ouzel, wheelchair, whippet, whiptail, white-stork, white-wolf, wine-bottle, wolf, wooden-boat, woodpecker, wreck, yawl, yorkshire-terrier, zebra

5 Uncombined Results for the TOSS Dataset

We acknowledge that aggregating results for computing scores can conceal details about the performance of an OSS solution. However, comparing the results directly from the different splits is not manageable. Therefore, we advocate using the aggregates for comparison and reporting, and using the raw results for analysis of one's own solution. Table 2, 3, and 4 present the mIoU values for each studied method, as well as the corresponding aggregate scores.

6 Person Class Images

As discussed in section 6 of the paper, there are no human face related images in the FSS-1000 dataset besides the two which are depicted in Figure 4

Tier 1: Query Complexity

	Fold-1	Fold-2	Fold-3	Fold-4	Mean
Baseline					
easy + salient	0.65	0.75	0.61	0.68	0.67
easy + non-salient	0.46	0.57	0.51	0.48	0.51
hard + salient	0.49	0.64	0.49	0.55	0.54
hard + non-salient	0.22	0.26	0.29	0.27	0.26
LCA	0.56	0.66	0.54	0.59	0.59
HCA	0.34	0.41	0.39	0.39	0.38
PFENet					
easy + salient	0.71	0.78	0.56	0.70	0.69
easy + non-salient	0.45	0.58	0.47	0.48	0.50
hard + salient	0.56	0.66	0.47	0.59	0.57
hard + non-salient	0.23	0.25	0.24	0.29	0.25
LCA	0.60	0.68	0.50	0.61	0.60
HCA	0.36	0.42	0.34	0.41	0.38
RPMM*					
easy + salient	0.62	0.73	0.54	0.62	0.63
easy + non-salient	0.40	0.54	0.45	0.41	0.45
hard + salient	0.49	0.66	0.43	0.53	0.53
hard + non-salient	0.18	0.23	0.25	0.23	0.22
LCA	0.52	0.64	0.48	0.53	0.54
HCA	0.30	0.39	0.33	0.34	0.34
RePRI					
easy + salient	0.67	0.72	0.59	0.57	0.64
easy + non-salient	0.44	0.52	0.43	0.39	0.45
hard + salient	0.48	0.58	0.45	0.39	0.48
hard + non-salient	0.18	0.22	0.22	0.19	0.20
LCA	0.56	0.62	0.51	0.48	0.54
HCA	0.32	0.37	0.33	0.29	0.33
HSNet					
easy + salient	0.66	0.74	0.63	0.69	0.68
easy + non-salient	0.48	0.55	0.52	0.51	0.51
hard + salient	0.52	0.65	0.47	0.53	0.54
hard + non-salient	0.22	0.27	0.26	0.30	0.26
LCA	0.57	0.65	0.55	0.60	0.59
HCA	0.34	0.42	0.37	0.40	0.38

Table 2: Raw scores on the Tier-1 splits of TOSS for the studied OSS networks.

Tier 2: Support Cognizance

Baseline

	Fold-1	Fold-2	Fold-3	Fold-4	Mean
L0	0.80	0.85	0.79	0.81	0.81
L1	0.67	0.79	0.74	0.73	0.73
L2 + L3	0.50	0.60	0.50	0.53	0.51
L4	0.46	0.46	0.37	0.52	0.45
L5	0.17	0.23	0.16	0.19	0.19
L6	0.35	0.51	0.37	0.42	0.41

PFENet

	Fold-1	Fold-2	Fold-3	Fold-4	Mean
L0	0.64	0.69	0.62	0.65	0.65
L1	0.57	0.70	0.58	0.57	0.60
L2 + L3	0.54	0.60	0.46	0.55	0.51
L4	0.54	0.41	0.35	0.50	0.45
L5	0.11	0.19	0.13	0.27	0.17
L6	0.28	0.30	0.33	0.41	0.33

RPMM

	Fold-1	Fold-2	Fold-3	Fold-4	Mean
L0	0.51	0.64	0.45	0.54	0.54
L1	0.48	0.64	0.49	0.49	0.53
L2 + L3	0.45	0.48	0.36	0.44	0.43
L4	0.45	0.43	0.31	0.46	0.41
L5	0.25	0.25	0.14	0.15	0.20
L6	0.33	0.37	0.29	0.28	0.32

RePRI

	Fold-1	Fold-2	Fold-3	Fold-4	Mean
L0	0.81	0.84	0.78	0.79	0.80
L1	0.68	0.75	0.70	0.66	0.70
L2 + L3	0.49	0.53	0.53	0.37	0.52
L4	0.73	0.68	0.54	0.52	0.62
L5	0.00	0.00	0.00	0.00	0.00
L6	0.17	0.33	0.33	0.17	0.25

HSNet

	Fold-1	Fold-2	Fold-3	Fold-4	Mean
L0	0.79	0.81	0.76	0.76	0.78
L1	0.68	0.75	0.69	0.64	0.69
L2 + L3	0.52	0.59	0.51	0.52	0.53
L4	0.41	0.48	0.37	0.44	0.54
L5	0.00	0.00	0.18	0.00	0.05
L6	0.08	0.21	0.18	0.14	0.15

Table 3: Raw mIoU numbers for the studied networks on Tier 2 of the TOSS dataset.

Tier 3: Generalization to unseen classes

	Fold-1	Fold-2	Fold-3	Fold-4	Mean
Baseline	0.82	0.81	0.82	0.82	0.82
PFENet	0.84	0.83	0.84	0.84	0.84
RPMM	0.78	0.79	0.78	0.79	0.79
RePRI	0.86	0.84	0.84	0.85	0.85
HSNet	0.88	0.88	0.87	0.87	0.88

Table 4: Raw mIoU numbers for the studied networks for Tier 3 of the TOSS dataset.

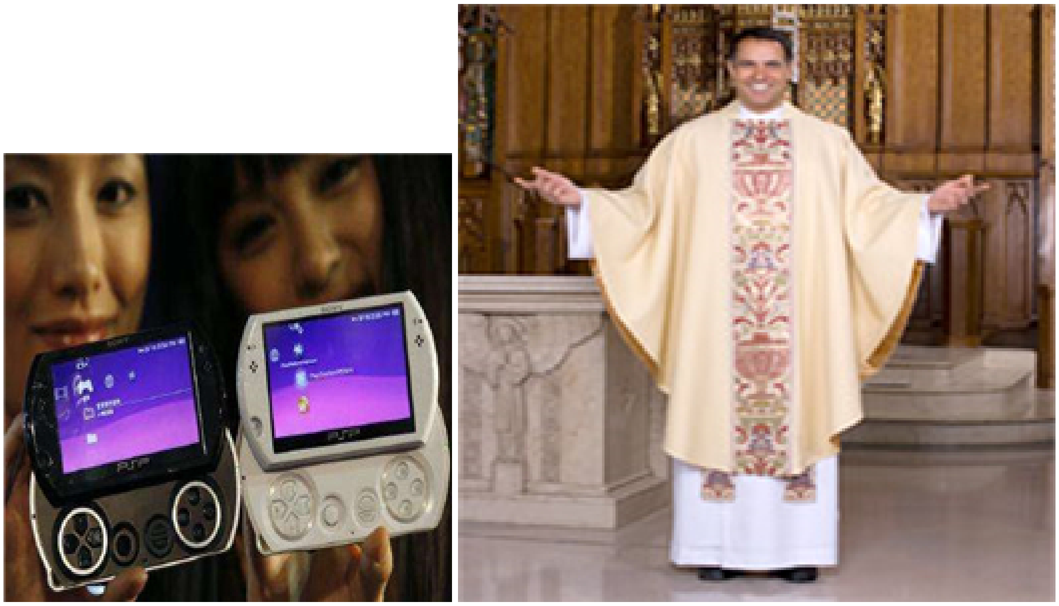

Figure 4: Images in the FSS-1000 subset used for the TOSS Dataset containing human faces