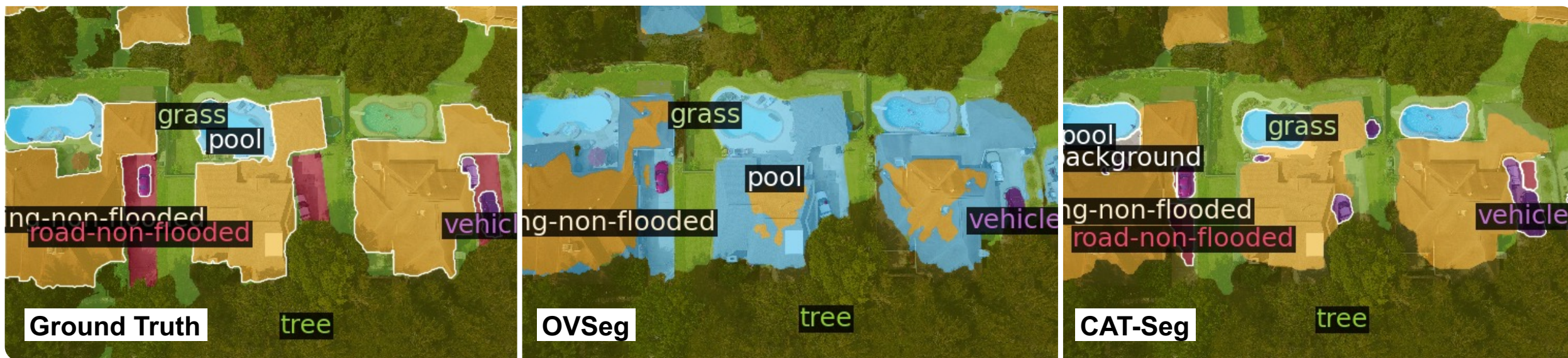


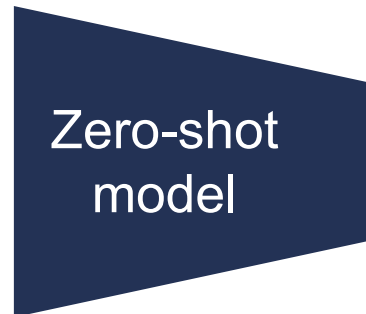
What a MESS: Multi-Domain Evaluation of Zero-Shot Semantic Segmentation

Benedikt Blumenstiel, Johannes Jakubik, Hilde Kühne, Michael Vössing



Zero-shot semantic segmentation models can process varying class names during inference.

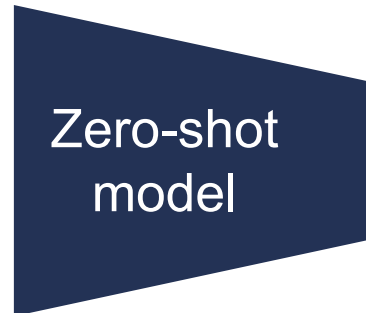
Variable classes: *road, building, sky, ...*



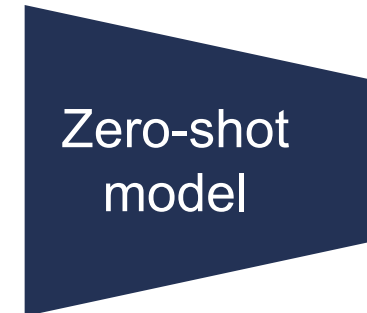
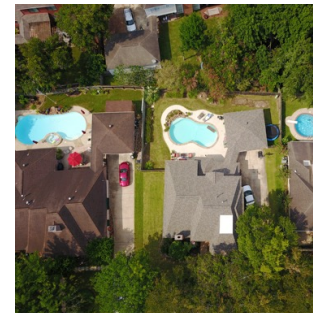
Dataset examples from BDD100K and FloodNet.

Zero-shot semantic segmentation models can process varying class names during inference.

Variable classes: *road, building, sky, ...*



Variable classes: *grass, tree, pool, ...*



Dataset examples from BDD100K and FloodNet.

Previous zero-shot evaluation focused on in-domain while MESS evaluates the generalizability across domains.



person, dog, car, chair, ...



person, dog, ... building, wall

Dataset examples from COCO-Stuff, ADE20K, Pascal Context, CHASE DB1, WorldFloods, UAVid, and SUIM.

Previous zero-shot evaluation focused on in-domain while MESS evaluates the generalizability across domains.



person, dog, car, chair, ...



person, dog, ... building, wall



blood vessels, flood, ...

Dataset examples from COCO-Stuff, ADE20K, Pascal Context, CHASE DB1, WorldFloods, UAVid, and SUIM.

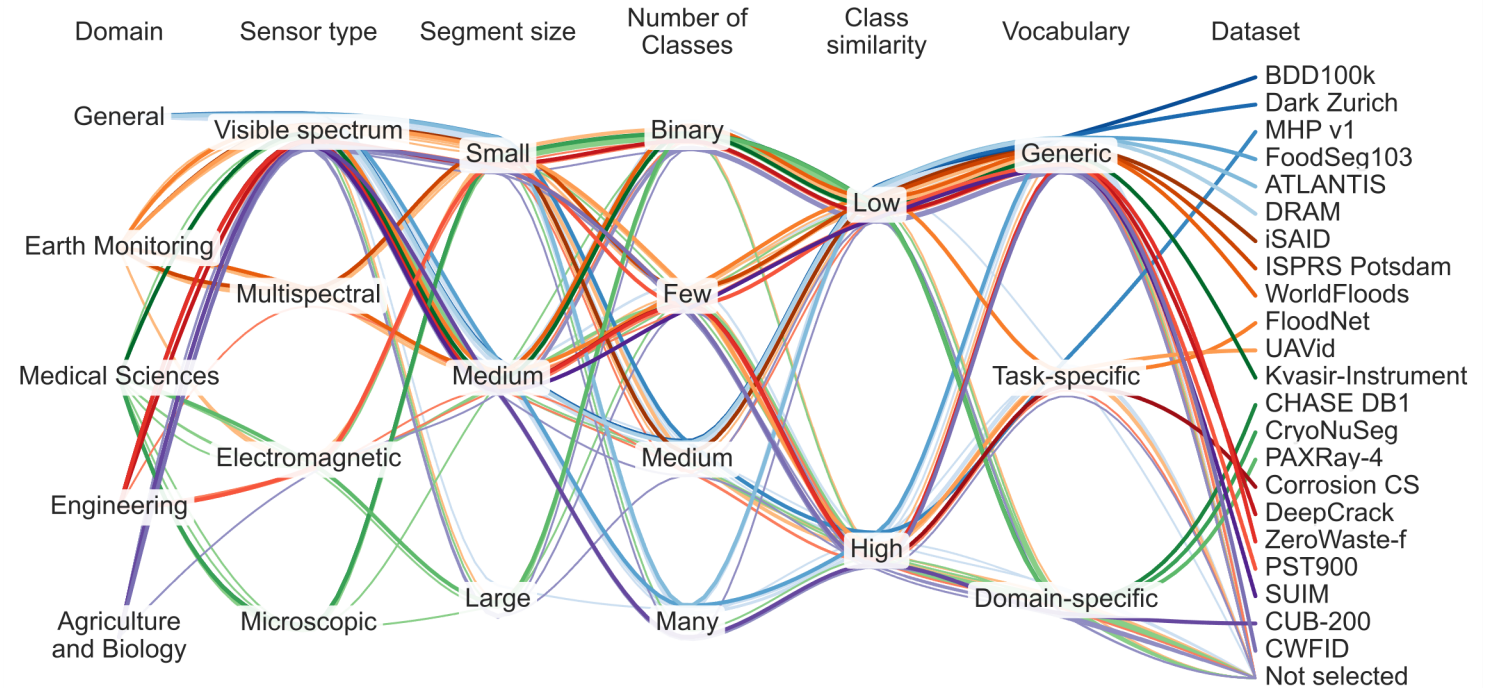
The MESS benchmark includes 22 datasets and covers a variety of visual and language characteristics.

>500
reviewed
datasets

120
classified
datasets



Visual and language characteristics of segmentation tasks



Taxonomy development based on from Nickerson, R. C. et al. (2013). A method for taxonomy development and its application in information systems. *European Journal of Information Systems*.

The domain-averaged mIoU results show differences in the generalizability between models and domains.



Model	General	Earth Monitoring	Medical Sciences	Engineering	Agri. and Biology	Mean
<i>Random (LB)</i>	<i>1.17</i>	<i>7.11</i>	<i>29.51</i>	<i>11.71</i>	<i>6.14</i>	<i>10.27</i>
<i>Best supervised (UB)</i>	<i>48.62</i>	<i>79.12</i>	<i>89.49</i>	<i>67.66</i>	<i>81.94</i>	<i>70.99</i>
ZSSeg-B	19.98	17.98	<u>41.82</u>	14.0	22.32	22.73
ZegFormer-B	13.57	17.25	17.47	17.92	<u>25.78</u>	17.57
X-Decoder-T	22.01	18.92	23.28	15.31	18.17	19.8
SAN-B	29.35	<u>30.64</u>	29.85	23.58	15.07	26.74
OpenSeeD-T	22.49	25.11	44.44	16.5	10.35	24.33
CAT-Seg-B	34.96	34.57	41.65	<u>26.26</u>	29.32	33.74
Grounded-SAM-B	<u>29.51</u>	25.97	37.38	29.51	17.66	<u>28.52</u>
OVSeg-L	29.54	29.04	31.9	14.16	28.64	26.94
SAN-L	36.18	<u>38.83</u>	30.27	16.95	20.41	30.06
CAT-Seg-L	39.93	39.85	48.49	26.04	<u>34.06</u>	38.14
Grounded-SAM-L	30.32	26.44	<u>38.69</u>	29.25	17.73	29.05
CAT-Seg-H	<u>37.98</u>	37.74	<u>34.65</u>	<u>29.04</u>	37.76	<u>35.66</u>
Grounded-SAM-H	30.27	26.44	38.45	28.16	17.67	28.78

Bold = best model; underlined = second-best model; Random = randomly expected mIoU with uniformly distributed predictions; Best supervised = SOTA separately selected for each dataset.

All results are available at <https://blumenstiel.github.io/mess-benchmark/leaderboard/>.

The domain-averaged mIoU results show differences in the generalizability between models and domains.



Model	General	Earth Monitoring	Medical Sciences	Engineering	Agri. and Biology	Mean
<i>Random (LB)</i>	<i>1.17</i>	<i>7.11</i>	<i>29.51</i>	<i>11.71</i>	<i>6.14</i>	<i>10.27</i>
<i>Best supervised (UB)</i>	<i>48.62</i>	<i>79.12</i>	<i>89.49</i>	<i>67.66</i>	<i>81.94</i>	<i>70.99</i>
ZSSeg-B	19.98	17.98	<u>41.82</u>	14.0	22.32	22.73
ZegFormer-B	13.57	17.25	17.47	17.92	<u>25.78</u>	17.57
X-Decoder-T	22.01	18.92	23.28	15.31	18.17	19.8
SAN-B	29.35	<u>30.64</u>	29.85	23.58	15.07	26.74
OpenSeeD-T	22.49	25.11	44.44	16.5	10.35	24.33
CAT-Seg-B	34.96	34.57	41.65	<u>26.26</u>	29.32	33.74
Grounded-SAM-B	<u>29.51</u>	25.97	37.38	29.51	17.66	<u>28.52</u>
OVSeg-L	29.54	29.04	31.9	14.16	28.64	26.94
SAN-L	36.18	<u>38.83</u>	30.27	16.95	20.41	30.06
CAT-Seg-L	39.93	39.85	48.49	26.04	<u>34.06</u>	38.14
Grounded-SAM-L	30.32	26.44	<u>38.69</u>	29.25	17.73	29.05
CAT-Seg-H	<u>37.98</u>	37.74	34.65	<u>29.04</u>	37.76	<u>35.66</u>
Grounded-SAM-H	30.27	26.44	38.45	28.16	17.67	28.78

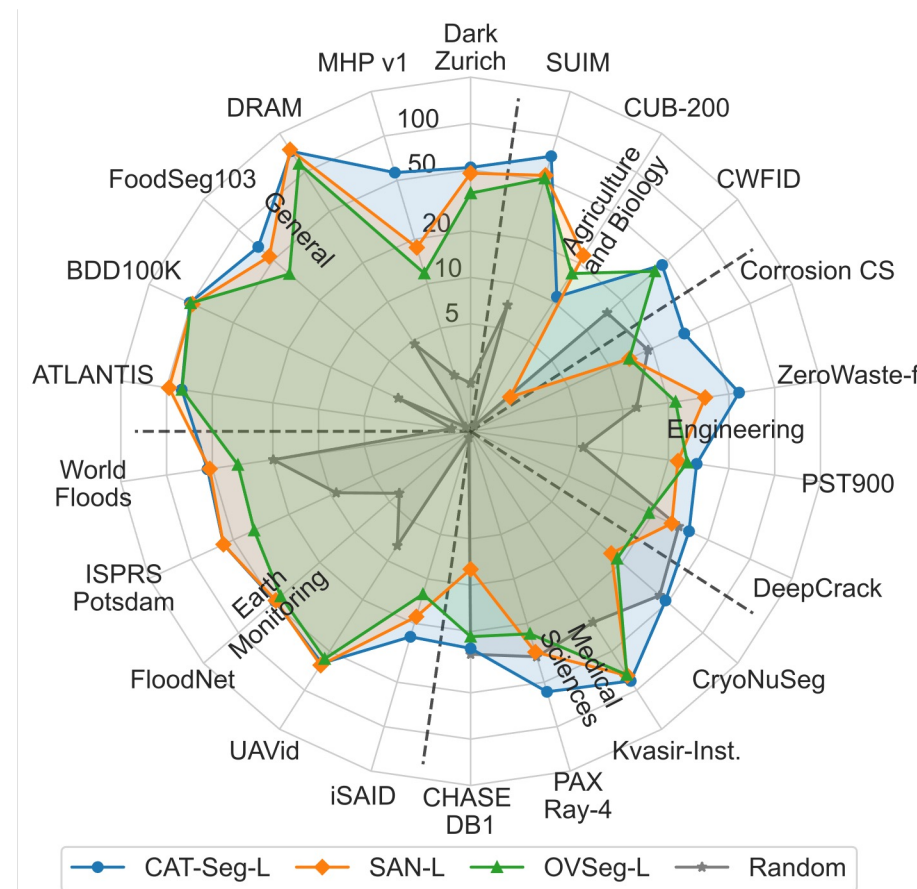
- CAT-Seg performs best, followed by SAN.
- Limited generalizability of two-stage or not CLIP based models.
- The performance varies between domains and datasets.

Bold = best model; underlined = second-best model; Random = randomly expected mIoU with uniformly distributed predictions; Best supervised = SOTA separately selected for each dataset.

All results are available at <https://blumenstiel.github.io/mess-benchmark/leaderboard/>.

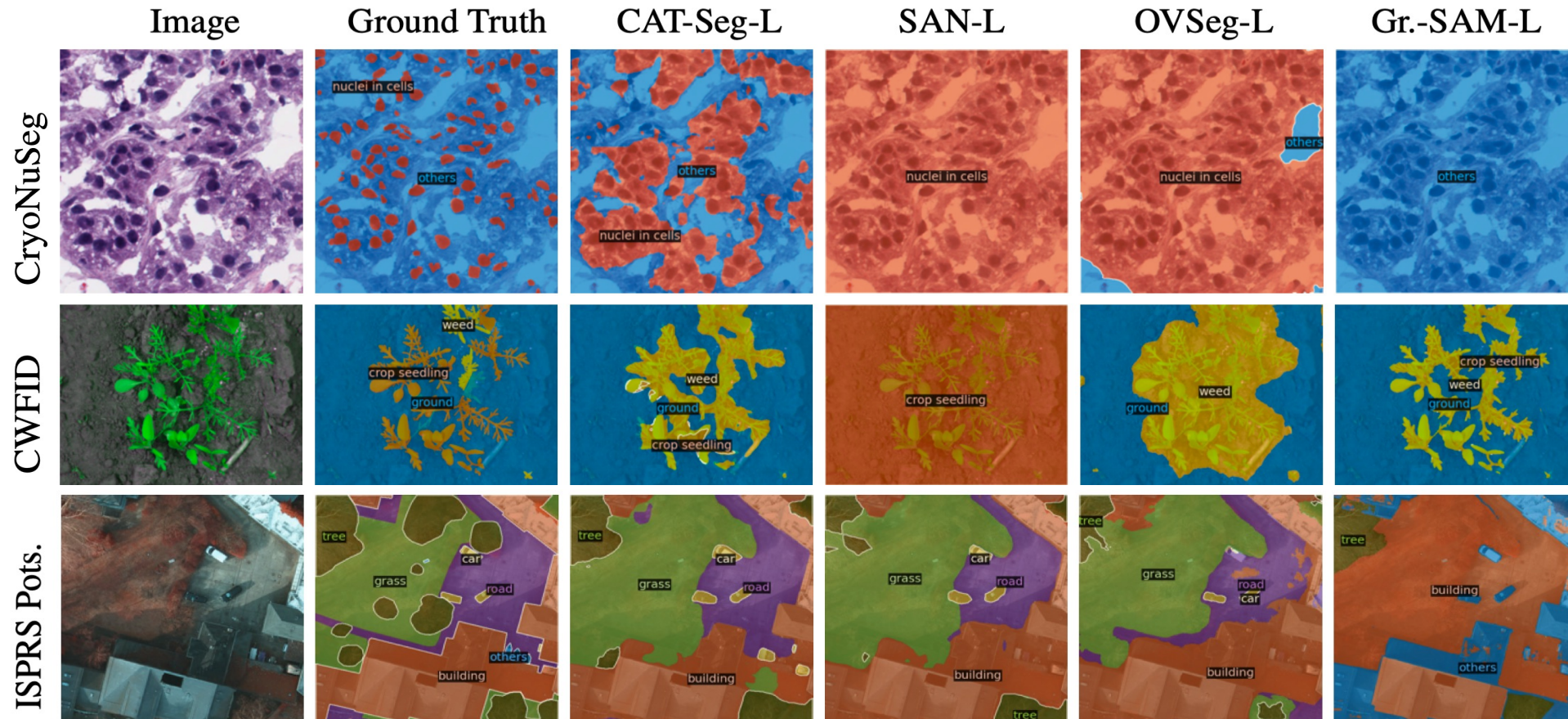
The visual and language characteristics from the datasets can influence the results of zero-shot models.

- Zero-shot transfer models **match supervised models** on everyday classes and images (DRAM, BDD100K).
- **Common classes** are correctly predicted in other domains (e.g., FloodNet, Kvasir-Instrument, SUIM).
- **Medical and engineering** datasets are very challenging for zero-shot transfer.
- The applicability is influenced by **visual and language characteristics** of the domain tasks.



Relative performance to the supervised results, visualized on a log scale (100 = supervised).

Small segments and similar class names are challenging but common classes can be detected in domain tasks.





Benedikt
Blumenstiel



Johannes
Jakubik



Hilde
Kühne



Michael
Vössing

✉ benedikt.blumenstiel@kit.edu

🏠 Karlsruhe Institute of Technology
University of Bonn
IBM Research Europe
MIT-IBM Watson AI Lab
IBM Germany

▶ <https://blumenstiel.github.io/mess-benchmark/>

MESS Benchmark

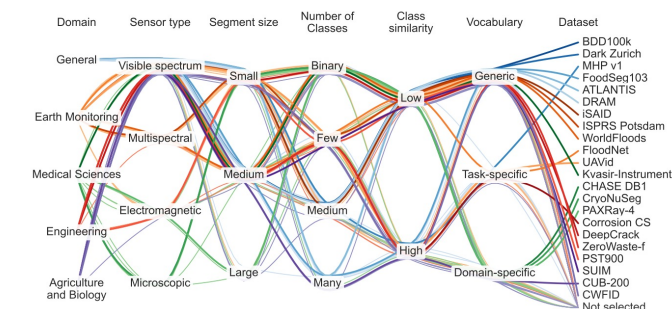
Leaderboard

Paper

Toolkit

MESS – Multi-Domain Evaluation of Semantic Segmentation

The MESS benchmark enables a holistic evaluation of semantic segmentation models on a variety of domains and datasets. The collection is based on a developed taxonomy which describes the semantic segmentation task space along six dimensions. We classified 120 datasets according to the taxonomy, visualized in the following figure. The 22 selected MESS datasets, highlighted with bold lines, cover all identified characteristics and are various domains such as earth monitoring, medicine, or engineering. You find details of the datasets, including links and licences, in this [overview](#).



Why MESS?

Zero-shot semantic segmentation models are regularly trained on COCO Stuff, a dataset including common scenes and classes. The standard evaluation datasets of these models are ADE20K, Pascal Context, and Pascal VOC. These datasets are from the same domain as the training data and therefore only test an in-domain transfer setting. With MESS, we provide a benchmark that enables a holistic evaluation of the generalization capabilities in a variety of other domains and use cases. The MESS evaluation is currently focused on zero-shot transfer – with potential extensions for few-shot or many-shot settings.

Toolkit

The MESS benchmark is available as a Python package on GitHub: <https://github.com/blumenstiel/mess-benchmark>