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

Benedikt Blumenstiel* **Johannes Jakubik*** **Hilde Kühne** **Michael Vössing**

The supplementary material is organized as follows:

- We detail the taxonomy development in Section A.
- The benchmark datasets are analyzed in Section B.
- We provide details about the evaluated models in Section C.
- Additional experiments are presented in Section D.
- Exemplary predictions are included in Section E.

A Taxonomy development

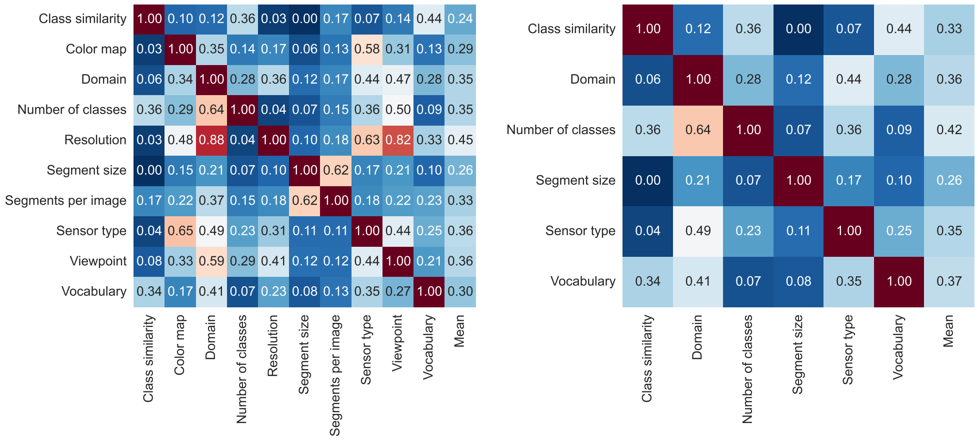
The taxonomy and the characterized datasets serve as a basis for the selection of the benchmark datasets. Therefore, we describe the taxonomy development in this section in detail. We applied the taxonomy development method proposed by Nickerson et al. [94] to analyze the task space of semantic segmentation. The method aims to develop a framework based on deduction and induction rather than *ad-hoc* decisions. We initialize the development by selecting our meta-characteristic (i.e., the goal): *identify visual and language characteristics of downstream tasks influencing the performance of zero-shot semantic segmentation models*.

We apply multiple empirical-to-conceptual or conceptual-to-empirical cycles until the ending conditions are reached. In an empirical-to-conceptual iteration, new objects are examined, and common characteristics are identified. The characteristics must derive from the meta-characteristic and discriminate between the objects to be of use for the taxonomy. In the conceptual step, the characteristics are grouped into dimensions. In contrast, a conceptual-to-empirical cycle starts by deducting potential dimensions and characteristics for the meta-characteristic based on prior knowledge. Next, the concept is evaluated by classifying objects. If a dimension does not differentiate between the objects or a characteristic has no real examples, it might not be appropriate. To fulfill the subjective ending conditions, the taxonomy must be concise, robust, comprehensive, extendible, and explanatory. Further, the objective ending conditions include, among others, dimension uniqueness and characteristic uniqueness within the dimension. We refer to Nickerson et al. [94] for more detailed information.

Starting with a first conceptual-to-empirical cycle, we analyzed other benchmarks and literature to initialize the taxonomy. The RF100 object detection benchmark [26] clusters the datasets into seven categories, representing different image types. We analyzed the literature of the RF100 categories to identify visual or language features relevant to these images. Aerial and satellite imagery has many important characteristics, such as the *sensor type* with different bands that can be mapped to true or false *color maps*, the *spatial resolution*, and *metadata* like time and location [97]. Electromagnetic images are often medical modalities using different sensor types, different *sensor directions* and having a domain-specific *vocabulary* that describes the anatomy [71]. Various sensors are also used in underwater imagery, and *preprocessing* plays an essential role in this image domain [81]. The spatial resolution is an important factor in microscopy, besides sensor types and specific hardware such as

*Equal contributions

37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks.



(a) Interim dimensions

(b) Final dimension

Figure 1: Peason correlation between dimensions based on the classified datasets for an interim status (a) and the final taxonomy (b).

phase contrast or fluorescence [80]. Images from documents or video games are often synthetic but use the same visual spectrum as real-world images. Segmentation in documents does include very fine *segment sizes*. We further added a *domain* dimension because the image and labels are often domain-specific. Therefore, we select domains inspired by the major subject areas from Scopus[2] but shorten the labels to improve the usability of the taxonomy. We added a domain "General" for everyday images, which we did not associate with any subject area. Based on this research, we used all identified dimensions relevant to at least two domains as domain-specific dimensions lead to redundancy in the taxonomy. The initial characteristics of each dimension are selected based on the literature review and complemented through the following empirical step.

We refine the taxonomy in multiple empirical-to-conceptual iterations. Therefore, we reviewed overall 500 datasets, including all semantic segmentation datasets on Papers with Code[3], Kaggle[4], and test datasets from other segmentation models like SAM [66]. We classified 120 datasets within our taxonomy which are presented in Table 1. Note, that different versions of a dataset are classified if they lead to varying characteristics. We did not classify all datasets of similar use cases as we aim for a diverse collection of datasets (e.g., seven driving datasets out of 30+). Other criteria for exclusion are deviating tasks (e.g., 3D data) and missing data availability. Also, we discarded use cases that seem to be very unique, like galaxy segmentation. If only a few reviewed datasets covered specific use cases, e.g., crack segmentation, we analyzed additional datasets from other sources. Based on the datasets, we added dimensions regarding the segmentation mask and language-related dimensions like the *class similarity*. We repeatedly reduced the dimensions in conceptual phases to the most meaningful ones for the meta-characteristic.

Finally, we utilized a statistical analysis to identify similar dimensions, specifically, the Pearson correlation between each pair of dimensions using the empirical data from the classified datasets. We applied one-hot encoding for categorical dimensions and scaled each ordinal dimension by the number of characteristics. Figure 1 visualize multiple pairs with high correlation, e.g., *segment size* and *number of segments*. We reduced the interim dimensions based on the statistical analysis and the meta-characteristic. The final taxonomy passes all ending conditions in [94].

[2]List of subject areas: https://www.scopus.com/sources

[3]Semantic segmentation datasets: https://paperswithcode.com/datasets?task=semantic-segmentation

[4]Seach results for "semantic segmentation": https://www.kaggle.com/datasets?search=semantic+segmentation

Table 1: All 120 classified semantic segmentation datasets within the taxonomy.

Dataset	Task	Domain	Sensor type	Segment size	Number of classes	Class similarity	Vocabulary
COCO Stuff [77]	Common	General	Visible spectrum	Medium	171 (Many)	Low	Generic
Pascal VOC 2012 [37]	Common	General	Visible spectrum	Medium	20 (Medium)	Low	Generic
ADE20K-150 [147]	Common	General	Visible spectrum	Medium	150 (Many)	Low	Generic
ADE20K-847 [147]	Common	General	Visible spectrum	Medium	847 (Many)	High	Generic
Pascal Context-59 [92]	Common	General	Visible spectrum	Medium	59 (Many)	Low	Generic
Pascal Context-459 [92]	Common	General	Visible spectrum	Medium	459 (Many)	High	Generic
LVIS [48]	Common	General	Visible spectrum	Small	1203 (Many)	High	Generic
FSS-1000 [74]	Common	General	Visible spectrum	Large	1000 (Many)	High	Generic
Mapillary Vistas v1 [93]	Driving	General	Visible spectrum	Small	66 (Many)	Low	Generic
Mapillary Vistas v2 [93]	Driving	General	Visible spectrum	Small	124 (Many)	Low	Task-spec.
Cityscapes [29]	Driving	General	Visible spectrum	Small	30 (Medium)	Low	Generic
BDD100K [140]	Driving	General	Visible spectrum	Medium	19 (Medium)	Low	Generic
Dark Zurich [105]	Driving	General	Visible spectrum	Medium	20 (Medium)	Low	Generic
SYNTHIA [102]	Driving	General	Visible spectrum	Small	13 (Medium)	Low	Generic
WoodScape [139]	Driving	General	Visible spectrum	Small	40 (Medium)	High	Generic
MVTec D2S [40]	Checkout	General	Visible spectrum	Medium	60 (Many)	Low	Generic
EgoHands [6]	Ego hands	General	Visible spectrum	Medium	5 (Few)	High	Task-spec.
WorkingHands [111]	Ego hands	General	Visible spectrum	Medium	16 (Medium)	Low	Generic
EgoHOS [143]	Ego hands	General	Visible spectrum	Medium	8 (Few)	High	Task-spec.
EYTH [124]	Ego hands	General	Visible spectrum	Medium	2 (Binary)	Low	Generic
VISOR [30]	Ego hands	General	Visible spectrum	Small	257 (Many)	High	Generic
Open Surfaces [9]	Materials	General	Visible spectrum	Medium	37 (Medium)	High	Domain-spec.
MINC [10]	Materials	General	Visible spectrum	Medium	23 (Medium)	Low	Generic
DMS [123]	Materials	General	Visible spectrum	Small	52 (Many)	High	Generic
DeepFashion2 [45]	Clothing	General	Visible spectrum	Small	13 (Medium)	Low	Generic
ModaNet [146]	Clothing	General	Visible spectrum	Small	13 (Medium)	Low	Generic
MHP v1 [72]	Body parts	General	Visible spectrum	Small	18 (Medium)	High	Task-spec.
MHP v2 [72]	Body parts	General	Visible spectrum	Small	58 (Many)	High	Task-spec.
FoodSeg103 [135]	Ingredients	General	Visible spectrum	Medium	103 (Many)	High	Generic
TACO [98]	Trash	General	Visible spectrum	Medium	60 (Many)	High	Domain-spec.
RailSem19 [141]	Rail	General	Visible spectrum	Small	11 (Medium)	High	Task-spec.
ATLANTIS [36]	Maritime	General	Visible spectrum	Small	56 (Many)	Low	Generic
Aircraft Context [115]	Aerial vehicles	General	Visible spectrum	Medium	8 (Few)	Low	Generic
RELLIS-3D [64]	Robotics	General	Visible spectrum	Small	20 (Medium)	Low	Generic
SketchyScene-7k [148]	Sketches	General	Visible spectrum	Small	45 (Medium)	Low	Generic
DRAM [28]	Paintings	General	Visible spectrum	Medium	12 (Medium)	Low	Generic
iSAID [134]	Objects	Earth Monitoring	Visible spectrum	Small	15 (Medium)	Low	Generic
DSTL Satellite [59]	Objects	Earth Monitoring	Multispectral	Small	10 (Medium)	High	Generic
ISPRS Potsdam [16]	Land use	Earth Monitoring	Multispectral	Small	6 (Few)	Low	Generic
LandCoverNet [4]	Land use	Earth Monitoring	Multispectral	Medium	7 (Few)	Low	Generic
LoveDA [129]	Land use	Earth Monitoring	Visible spectrum	Small	7 (Few)	Low	Generic
Deep Globe [32]	Land use	Earth Monitoring	Visible spectrum	Medium	7 (Few)	Low	Generic
GID-5 [120]	Land use	Earth Monitoring	Multispectral	Small	5 (Few)	Low	Generic
GID-15 [120]	Land use	Earth Monitoring	Multispectral	Small	16 (Medium)	High	Task-spec.
Dubai [57]	Land use	Earth Monitoring	Visible spectrum	Small	6 (Few)	Low	Generic
Sen1Floods11 [13]	Floods	Earth Monitoring	Electromagnetic	Small	2 (Binary)	Low	Generic
WorldFloods [84]	Floods	Earth Monitoring	Multispectral	Medium	3 (Binary)	Low	Generic
HR-GLDD [87]	Landslides	Earth Monitoring	Multispectral	Medium	2 (Binary)	Low	Generic
Antarctic fracture [67]	Ice fractures	Earth Monitoring	Multispectral	Small	2 (Binary)	Low	Generic
Active fire [31]	Wildfires	Earth Monitoring	Multispectral	Small	2 (Binary)	Low	Generic
xBD [49]	Buildings	Earth Monitoring	Visible spectrum	Small	5 (Few)	High	Task-spec.
MSAW [110]	Buildings	Earth Monitoring	Electromagnetic	Small	2 (Binary)	Low	Generic
3D PV Locator [85]	PV	Earth Monitoring	Visible spectrum	Small	2 (Binary)	Low	Generic
AgricultureVision [24]	Agriculture	Earth Monitoring	Multispectral	Medium	9 (Few)	Low	Domain-spec.
PASTIS [44]	Agriculture	Earth Monitoring	Multispectral	Small	18 (Medium)	High	Domain-spec.
CalCROP21 [47]	Agriculture	Earth Monitoring	Multispectral	Small	29 (Medium)	High	Domain-spec.
Arctic Sea Ice [117]	Sea ice	Earth Monitoring	Multispectral	Medium	8 (Few)	High	Task-spec.
ELAI Dust Storm [7]	Dust storm	Earth Monitoring	Visible spectrum	Large	2 (Binary)	Low	Generic
FloodNet [101]	Floods	Earth Monitoring	Visible spectrum	Medium	10 (Few)	Low	Task-spec.
SDD [60]	Objects	Earth Monitoring	Visible spectrum	Small	21 (Medium)	Low	Generic
UDD [22]	Objects	Earth Monitoring	Visible spectrum	Medium	6 (Few)	Low	Generic
UAVid [82]	Objects	Earth Monitoring	Visible spectrum	Small	6 (Few)	High	Task-spec.
PV thermography [132]	PV	Earth Monitoring	Electromagnetic	Small	6 (Binary)	High	Domain-spec.
CholecSeg8k [122]	Surgery	Medical Sciences	Visible spectrum	Medium	13 (Medium)	High	Domain-spec.
RoboTool [43]	Surgery	Medical Sciences	Visible spectrum	Medium	2 (Binary)	Low	Generic
Kvasir-Instrument [62]	Surgery	Medical Sciences	Visible spectrum	Medium	2 (Binary)	Low	Generic
ROBUST-MIS 2019 [103]	Surgery	Medical Sciences	Visible spectrum	Medium	2 (Binary)	Low	Generic
Kvasir SEG [63]	Surgery	Medical Sciences	Visible spectrum	Medium	2 (Binary)	Low	Domain-spec.
Vocalfolds [68]	Surgery	Medical Sciences	Visible spectrum	Medium	7 (Few)	Low	Domain-spec.
CHASE DB1 [41]	Retina scan	Medical Sciences	Microscopic	Small	2 (Binary)	Low	Domain-spec.
HRF [69]	Retina scan	Medical Sciences	Microscopic	Small	2 (Binary)	Low	Domain-spec.
STARE [55]	Retina scan	Medical Sciences	Microscopic	Small	2 (Binary)	Low	Domain-spec.
Intraretinal C. Fluid [2]	Retinal OCT	Medical Sciences	Microscopic	Small	2 (Binary)	Low	Domain-spec.

Dataset	Task	Domain	Sensor type	Segment size	Number of classes	Class similarity	Vocabulary
GLaS [113]	WSI	Medical Sciences	Microscopic	Medium	2 (Binary)	High	Domain-spec.
Gleason [95]	WSI		Microscopic	Large	6 (Few)	High	Domain-spec.
CryoNuSeg [83]	WSI		Microscopic	Small	2 (Binary)	Low	Domain-spec.
BBBC038v1 [17]	WSI		Microscopic	Small	2 (Binary)	Low	Domain-spec.
Vector-LabPics [35]	Lab vessels		Visible spectrum	Medium	58 (Medium)	High	Domain-spec.
vesselNN [118]	Brain vessel		Microscopic	Small	2 (Binary)	Low	Domain-spec.
MTNeuro [99]	Brain vessel		Microscopic	Small	3 (Few)	High	Domain-spec.
Neuronal Cells [53]	Brain cells		Microscopic	Small	2 (Binary)	Low	Domain-spec.
BraTS 2015 [88]	Brain tumor		Electromagnetic	Medium	5 (Few)	High	Domain-spec.
ISIC2018 Task1 [27]	Lesions		Visible spectrum	Large	2 (Binary)	Low	Domain-spec.
PAXRay-166 [108]	X-Ray		Electromagnetic	Small	166x2 (Binary)	High	Domain-spec.
PAXRay-4 [108]	X-Ray		Electromagnetic	Large	4x2 (Binary)	Low	Domain-spec.
Pulmonary Chest [18]	X-Ray		Electromagnetic	Large	2 (Binary)	Low	Generic
US segmentation [125]	Ultrasound		Electromagnetic	Medium	9 (Few)	High	Domain-spec.
Severstal [109]	Surface defect	Engineering	Visible spectrum	Medium	4 (Few)	High	Domain-spec.
KolektorSDD2 [14]	Surface defect		Visible spectrum	Medium	2 (Binary)	Low	Generic
EMPS [138]	Particles		Electromagnetic	Small	2 (Binary)	Low	Generic
LIB-HSI [50]	Building fasade		Multispectral	Medium	44 (Medium)	High	Generic
Corrosion CS [11]	Corrosion		Visible spectrum	Medium	4 (Few)	High	Task-spec.
LCW [12]	Cracks		Visible spectrum	Small	2 (Binary)	Low	Generic
DeepCrack [79]	Cracks		Visible spectrum	Small	2 (Binary)	Low	Generic
ZeroWaste-f [8]	Conveyor		Visible spectrum	Medium	4 (Few)	High	Generic
Thermal Dog [104]	Thermal		Electromagnetic	Medium	3 (Few)	Low	Generic
PST900 [112]	Thermal		Electromagnetic	Small	5 (Few)	Low	Generic
TAS-NIR [91]	Thermal		Electromagnetic	Medium	22 (Medium)	High	Generic
PIDRay [127]	Security		Electromagnetic	Small	12 (Medium)	Low	Generic
TTPLA [1]	Powerlines		Visible spectrum	Small	5 (Few)	High	Generic
Vale [56]	Terrain		Visible spectrum	Medium	5 (Few)	High	Task-spec.
AI4MARS [116]	Terrain		Visible spectrum	Small	4 (Few)	High	Generic
TrashCan [54]	Trash	Agriculture and Biology	Visible spectrum	Medium	4 (Few)	Low	Generic
SUIM [61]	Underwater		Visible spectrum	Medium	8 (Few)	Low	Generic
DeepFish [106]	Fish		Visible spectrum	Medium	2 (Binary)	Low	Generic
NDD20 [121]	Fish		Visible spectrum	Medium	2 (Binary)	Low	Generic
Ciona17 [42]	Maritime species		Visible spectrum	Large	4 (Few)	High	Domain-spec.
CUB-200 [126]	Bird species		Visible spectrum	Medium	201 (Many)	High	Domain-spec.
Oxford-IIIT Pet [96]	Animal species		Visible spectrum	Large	28 (Medium)	High	Domain-spec.
Plittersdorf [51]	Animals		Electromagnetic	Medium	2 (Binary)	Low	Generic
CAMO [70]	Animals		Visible spectrum	Medium	2 (Medium)	Low	Domain-spec.
COD [38]	Animals		Visible spectrum	Medium	78 (Many)	Low	Domain-spec.
CropAndWeed [114]	Plants		Visible spectrum	Small	100 (Many)	High	Domain-spec.
WGISD [107]	Plants		Visible spectrum	Medium	2 (Binary)	Low	Generic
PPDPS [90]	Plants		Visible spectrum	Large	2 (Binary)	Low	Generic
Plant seg. [34]	Plants		Visible spectrum	Small	3 (Few)	High	Task-spec.
CWFID [52]	Crops		Visible spectrum	Small	3 (Few)	High	Generic
PPDLS [90]	Leefs		Visible spectrum	Medium	2 (Binary)	Low	Generic
Leaf disease [3]	Leef disease		Visible spectrum	Small	2 (Binary)	Low	Generic
Rice Leaf dis. [75]	Leef disease		Visible spectrum	Small	5 (Few)	High	Domain-spec.

B Benchmark datasets

B.1 Overview

We selected 22 out of the 120 classified datasets for the MESS benchmark. The links, licenses, selected splits, and a sample of the class labels of the datasets are provided in Table 2. We specified some label names for better performances of the models similar to [73]. E.g., we use *crop seedling* instead of *crop* for the CWFID dataset. We refer to our implementation for all class labels.

We shortly introduce each dataset in the following: The general datasets include datasets with everyday scenes but more specific use cases and niche image themes in comparison to the standard evaluation datasets. Specifically, the use cases include two driving datasets with one covering nighttime images. Further, MHP v1 covers classes of body parts and clothes while FoodSeg103 requires the segmentation of different ingredients. The ATLANTIS dataset focuses on classes related to maritime environments and DRAM covers common classes in paintings. The selected earth monitoring datasets include iSAID, which requires the segmentation of 15 object categories in satellite images, e.g., a tennis court or a helicopter. ISPRS Potsdam and WorldFloods provide multispectral data, and our main evaluation uses an IRRG false color mapping. Near-infrared radiation is visualized in red and highlights vegetation. ISPRS Potsdam provides very high-resolution images of an urban area with multiple classes, while WorldFloods has a 10-meter resolution and focuses on

Table 2: Details for the 22 MESS datasets including the links and licenses. Nearly all datasets require attribution and many only allow non-commercial use.

Dataset	Link	Licence	Split	No. of classes	Classes
BDD100K [140]	berkeley.edu	custom	val	19	[road; sidewalk; building; wall; fence; pole; traffic light; traffic sign; ...]
Dark Zurich [105]	ethz.ch	custom	val	20	[unlabeled; road; sidewalk; building; wall; fence; pole; traffic light; ...]
MHP v1 [72]	github.com	custom	test	19	[others; hat; hair; sunglasses; upper clothes; skirt; pants; dress; ...]
FoodSeg103 [135]	github.io	Apache 2.0	test	104	[background; candy; egg tart; french fries; chocolate; biscuit; popcorn; ...]
ATLANTIS [36]	github.com	Flickr (images)	test	56	[bicycle; boat; breakwater; bridge; building; bus; canal; car; ...]
DRAM [28]	ac.il	custom (in download)	test	12	[bird; boat; bottle; cat; chair; cow; dog; horse; ...]
iSAID [134]	github.io	Google Earth (images)	val	16	[others; boat; storage tank; baseball diamond; tennis court; bridge; ...]
ISPRS Potsdam [16]	isprs.org	no licence provided[a]	test	6	[road; building; grass; tree; car; others]
WorldFloods [84]	github.com	CC NC 4.0	test	3	[land; water and flood; cloud]
FloodNet [101]	github.com	custom	test	10	[building-flooded; building-non-flooded; road-flooded; water; tree; ...]
UAVid [82]	uavid.nl	CC BY-NC-SA 4.0	val	8	[others; building; road; tree; grass; moving car; parked car; humans]
Kvasir-Inst. [62]	simula.no	custom	test	2	[others; tool]
CHASE DB1 [41]	kingston.ac.uk	CC BY 4.0	test	2	[others; blood vessels]
CryoNuSeg [83]	kaggle.com	CC BY-NC-SA 4.0	test	2	[others; nuclei in cells]
PAXRay-4 [108]	github.io	custom	test	4x2	[others, lungs], [others, bones], [others, mediastinum], [others, diaphragm]
Corrosion CS [11]	figshare.com	CC0	test	4	[others; steel with fair corrosion; ... poor corrosion; ... severe corrosion]
DeepCrack [79]	github.com	custom	test	2	[concrete or asphalt; crack]
PST900 [112]	github.com	GPL-3.0	test	5	[background; fire extinguisher; backpack; drill; human]
ZeroWaste-f [8]	ai.bu.edu	CC-BY-NC 4.0	test	5	[background or trash; rigid plastic; cardboard; metal; soft plastic]
SUIM [61]	umn.edu	MIT	test	8	[human diver; reefs and invertebrates; fish and vertebrates; ...]
CUB-200 [126]	caltech.edu	custom	test	201	[background; Laysan Albatross; Sooty Albatross; Crested Auklet; ...]
CWFID [52]	github.com	custom	test	3	[ground; crop seedling; weed]

[a]Upon request, the naming of the data provider and project is required.

water segmentation. We selected two drone datasets with similar use cases. UAVid includes urban scenes, and FloodNet covers flooded buildings and roads. The medical datasets cover four different modalities: Endoscopy (RGB images), retinal scans, whole slide imagery (WSI), and X-ray scans. Each binary segmentation task focuses on a specific object or anatomical structure, like blood vessels or lungs. The multi-label segmentation dataset PAXRay is a special case. We do not use each of the 166 annotated classes but only the four superclasses. Because of the mask overlay, each class is predicted in a binary setting, and we average the resulting metrics. Next, we selected four diverse engineering datasets. Corrosion CS includes images of corrosion on bridges and other infrastructure with four different condition states. DeepCrack consists of close-up images of crack. PST900 consists of thermal imagery with firefighter-related objects. We use a gray-scale color map in our main evaluation to visualize the thermal data. The Zero-Waste-f dataset includes images of a conveyor belt with annotations for four types of recyclable trash. The final three datasets cover biological-related datasets: SUIM is an underwater imagery dataset with fish, aquatic plants, and others. CUB-200 is a widely used dataset of 200 bird species. The images of CUB-200 are relatively easy to segment, but assigning the correct species is challenging. CWFID includes crop seedlings and weeds.

We looked up the current fully supervised performance to provide an upper threshold for each dataset and present them in Table 3. We did not find any mIoU results for the MHP v1 dataset as it is originally annotated for instance segmentation. Therefore, we trained MaskFormer [23] to provide a reference. We trained the model for 100K steps using the Swin-B ADE20K-150 settings and evaluated the best model based on the val mIoU.

B.2 Dataset analysis

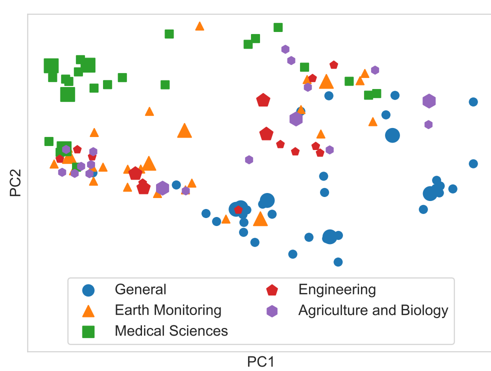
The classified datasets are visualized in Figure 2 by applying a Principal Component Analysis (PCA) along the taxonomy's dimensions. An analysis of the principal components reveals that, apart from the domain, mainly language-related features differentiate the datasets within the taxonomy. The PCA has two big clusters covering all domains – one cluster of datasets (top) with mostly domain-specific vocabulary and high class similarity and another one (bottom) with tasks of easily distinguishable generic classes. The PCA emphasizes the importance of these two dimensions for all domains. The datasets visualized in the center between these clusters have either a domain-specific vocabulary with low class similarly, which is often the case for medical datasets, or the opposite, often observed in

Table 3: Supervised mIoU results for the datasets.

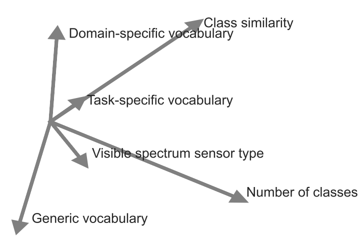
Dataset	Model	Year	mIoU
BDD100K	Two-branch Enet [89]	2023	44.8
Dark Zurich	Refign (HRDA) [15]	2023	63.9
MHP v1	MaskFormer (Swin-B) [23]	2021	53.18[a]
FoodSeg103	SeTR-MLA (ViT-16/B) [145]	2021	45.1
ATLANTIS	AQUANet [36]	2021	42.22
DRAM	DRAM model [28]	2022	45.71[b]
iSAID	IMP-ViTAEv2-S-UperNet [128]	2022	65.3
ISPRS Potsdam	DC-Swin [130]	2022	87.56
WorldFloods	UNet [84]	2021	92.71
FloodNet	SegFormer [5]	2023	82.22
UAVid	UNetFormer [131]	2022	67.8
Kvasir-Instrument	U-Net [62]	2021	93.7
CHASE DB1	RV-GAN [65]	2021	97.05
CryoNuSeg	TransUNet [21]	2022	73.45
PAXRay-4	Unet-R50 [108]	2022	93.77
Corrosion CS	DeeplabV3+ [11]	2021	49.92
DeepCrack	DeepCrack-GF [79]	2019	85.9
ZeroWaste-f	DeeplabV3+ [8]	2022	52.5
PST900	SpiderMesh [39]	2022	82.3
SUIM	LOCA [19]	2022	74.0
CUB-200	GFN [144]	2022	84.6
CWFID	Unet-Resnet-50 [119]	2022	87.23

[a] Own experiment because mIoU results are not reported in MHP v1 literature.
[b] The DRAM model is not trained on a labeled training set but self-supervised on generated images.

general datasets. Furthermore, medical datasets have few classes, while general use cases have many classes, with the other three domains in between.



(a) PCA



(b) Influence

Figure 2: PCA of the classified datasets, clustered by their domain (a), and the highest influencing factors apart from the domains (b). An increased size visualizes selected datasets. Some noise was added to visualize similar classified datasets.

Following Xu et al. [136], we conducted an analysis of the similarity between the labels of each dataset and the training labels from COCO-Stuff [77] which is used by most evaluated models. The similarity between two labels is computed using the cosine similarity between their CLIP text embeddings. Next, we select the maximum similarity value for each text label (i.e., the minimal distance of this label to the training labels). To calculate the similarity between a test set and the training set, we can select the minimum value among the test labels. This represents a Hausdoff

Distance between these two sets, i.e., the maximum distance in the embedding space [136]. However, this calculation is sensitive to outliers and we also report the mean similarity over all test labels.

The analysis in Figure 3 visualizes that most datasets do include classes with a low train similarity that are not related to the train labels. Some datasets have a high mean similarity (i.e., BDD100K, DRAM, ISPRS Potsdam, ZeroWaste-f). Therefore, most classes in these datasets are equal or similar to a training label from COCO-Stuff. The medical and engineering datasets often have a low mean train similarity and include labels that are not present in the training labels.

Additionally, Figure 3 includes the similarity values within each dataset. The similarity is calculated using the maximum cosine similarity for each label to the rest. Selecting the maximum value from all labels results in the inner max similarity, and a high value indicates that at least two labels in the task have very close embeddings. It corresponds to a high class similarity within our taxonomy. Therefore, these classes are challenging to differentiate, even without considering the image features (e.g., classes in MHP v1, Corrosion CS, and CUB-200).

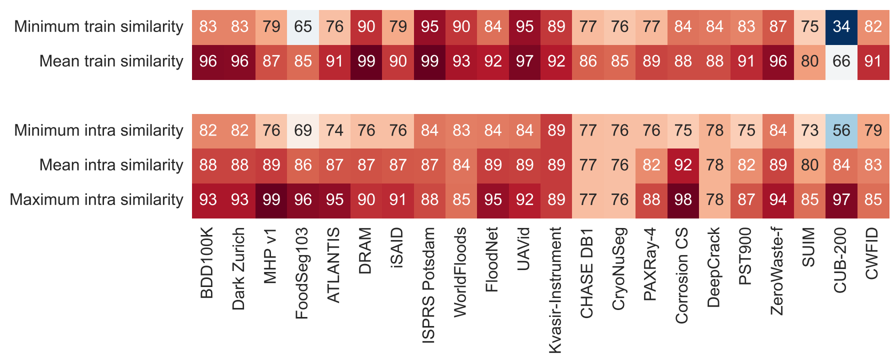


Figure 3: Class similarity to the COCO-Stuff training labels and within each dataset.

C Models

We provide an overview of the tested zero-shot semantic segmentation models in Table 4 including their modules and training datasets. We only include the datasets used for training the segmentation model and not the pre-training datasets of a utilized FM. We want to point out that the public versions of X-Decoder and OpenSeeD are using different FMs than the larger, non-public versions.

We utilize Grounded-SAM based on a re-implementation inspired by the demo code in [58]. To our knowledge, other implementations of Grounded-SAM are limited to demo scripts and do not apply semantic segmentation. The model combines bounding box predictions from Grounding DINO [78] with instance segmentations from SAM [66]. Grounding DINO is an open-vocabulary object detection model. The model predicts bounding boxes for all class labels in the label set. The labels also include the background class, as we noticed better results in prior experiments compared to discarding the background class. Next, SAM predicts one instance segmentation mask for each bounding box, and the pixel-wise confidence values are scaled by the confidence score of the bounding box. The instance masks of each class are combined by the maximum confidence value of each pixel, resulting in semantic masks. Negative values represent background predictions. Therefore, pixels with only negative values are predicted as background or no-object for datasets without a background class.

We noticed that Grounding DINO has a limited capability to predict non-general classes. SAM can also be combined with other open-vocabulary object detection models to improve performance. We refer to the oracle bounding box results for an upper bound.

D Additional results

We provide further experiments and detailed dataset-wise results in this section. Specifically, we analyze classes of interest, the similarity between evaluation and training classes, and the influence of the segment size.

Table 4: Overview of the evaluated models.

Name	Versions	Year	Modules	Training datasets
ZSSeg [137]	Base	2021	CLIP ViT-B & text encoder, Resnet 101	COCO-Stuff
ZegFormer [33]	Base	2022	CLIP ViT-B & text encoder, Resnet 101	COCO-Stuff-156
OVSeg [76]	Large	2022	CLIP ViT-L & text encoder, Swin-B	COCO-Stuff, COCO Caption
X-Decoder [149]	Tiny	2023	Focal-T/L, UniCL text encoder, ViT-decoder	COCO2017, 4M corpora
OpenSeeD [142]	Tiny	2023	Swin-T, UniCL text encoder, ViT-decoder	COCO2017, Objects365
SAN [136]	B/L	2023	CLIP ViT-B/L & text encoder, ViT-adapter	COCO-Stuff
CAT-Seg [25]	B/L/H	2023	CLIP ViT-B/L/H & text encoder, Swin-B, ViT-decoder	COCO-Stuff
Grounded-SAM [58]	B/L/H	2023	SAM ViT-B/L/H, DINO Swin-B, BERT-B	SA-1B, COCO, O365, GoldG, Cap4M, OpenImage, ODinW-35, RefCOCO

D.1 Classes of interest

For several datasets in the benchmark, a subset of the annotated classes is particularly relevant. We refer to this subset as Class(es) of Interest (CoI). E.g., binary segmentation tasks typically include a CoI (like pixels depicting a flood event) and a background class. In many cases, the model performance varies between CoI and background. To better understand the actual performance for these classes, we report the mIoU on the CoI subset (CoI-mIoU). With $CoI \subseteq \mathcal{C}$, and IoU_i being the intersection over union for class i, we calculate the metric $\textit{CoI-mIoU} = \frac{\sum_{i \in CoI} IoU_i}{|CoI|}$. In binary segmentation tasks, this is similar to the IoU_{pos} of the positive class [86].

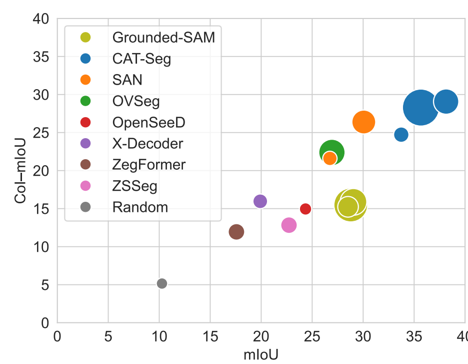
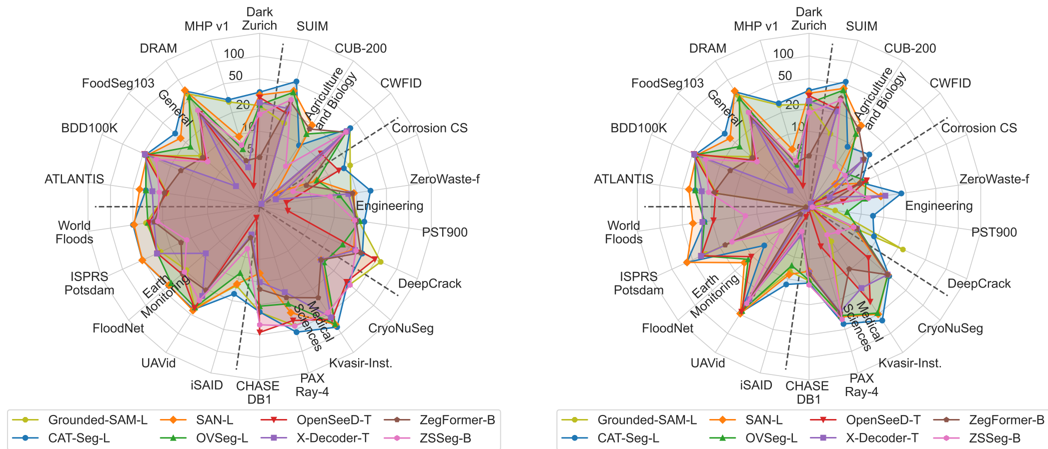


Figure 4: mIoU for Class(es) of Interest (CoI) in comparison to the mIoU of all classes. The size represents the parameter count of the models.

Figure 4 visualizes the CoI-mIoU compared to the mIoU of all predicted classes. On average, none of the models is able to segment classes of interest as well as all classes. Models like ZSSeg, OpenSeeD, and Grounded-SAM have a particularly strong bias toward misclassifying CoI than the other models. Also, CAT-Seg tends to misclassify classes of interest. For example, SAN-L has on average only a 3.18pp lower CoI-mIoU than the best-performing model CAT-Seg-L while the mIoU difference is 8.08pp. The differences between the mIoU and the CoI-mIoU vary between the datasets and domains. Figure 5 visualizes the mIoU and CoI-mIoU for all datasets and model architectures. The differences between both class sets are most evident for medical, engineering, and biological datasets (except for the datasets Kvasir-Instrument and SUIM). The CoI seem to be challenging for all models. These classes often include characteristics such as small segments, a high class similarity, or domain-specific labels. Furthermore, several model architectures tend to predict no or very few pixels as CoI, resulting in very low or zero scores. These architectures include X-Decoder, OpenSeeD, and Grounded-SAM. The models do not make use of CLIP, which may limit their capability to generalize to the domain-specific classes.



(a) mIoU (b) CoI-mIoU

Figure 5: mIoU (a) and CoI-mIoU (b) results for all model architectures on a log scale.

D.2 Similarity to training classes

The generalized zero-shot transfer setting does allow an overlap between the training labels and the evaluation labels. We analyze this overlap and the influences on the model performance by calculating the embedding similarity of each label to the training labels in COCO-Stuff. A high similarity corresponds to the concept being present in the training dataset. Figure 6 presents the correlation between the similarity and the class-wise IoU for the three large models which are trained on COCO-Stuff. The results indicate a positive correlation between the similarity of the training labels and the performance. We also observe a comparable correlation for all other model architectures (except ZegFormer) which are partly trained on more diverse datasets. The similarity of the training labels for the segmentation modules is not the only explanation. The correlation could be influenced by the open-vocabulary capabilities of the underlying FM. CLIP's understanding of common concepts, such as the training classes, is better than the understanding of domain-specific concepts [100].

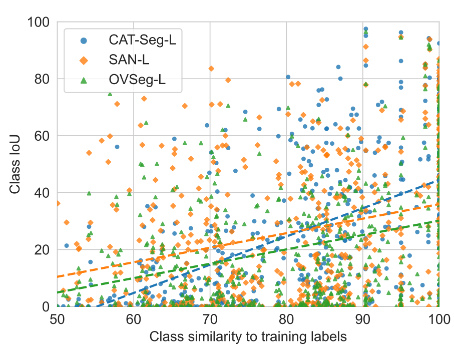


Figure 6: Class IoU in comparison to the class similarity with the labels in COCO-Stuff, represented by the maximum cosine similarity.

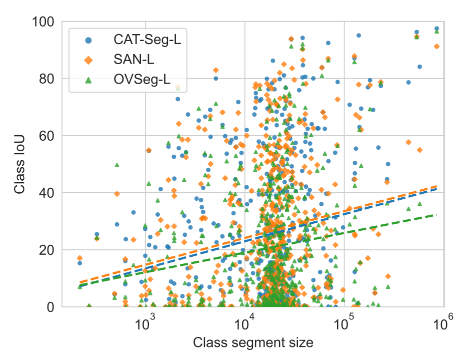


Figure 7: Class IoU in comparison to the class segment size on a log scale. The segment size is the class-wise average pixel count of a segment.

D.3 Segment size

Our benchmark includes multiple datasets with small segments, like WSI images with nuclei in cells or cars in satellite images. However, many models cannot correctly segment these small objects. We compare the average class segment size with the class IoU in Figure 7. The analysis considers all connected segments over 10 pixels to filter out potential annotation inaccuracies. Overall, all three large models have a positive correlation between segment size and mIoU—which also applies to other models. Therefore, the models have on average a lower performance on classes with small segments. We want to point out that the 200 CUB-200 classes are mostly correctly segmented but wrongly classified due to the challenging species labels. The correlation is higher without considering the CUB-200 classes. Visual inspection leads to a second insight: Some models, e.g., CAT-Seg-L and SAN-L, are able to locate small objects but fail to correctly segment the boundaries. Therefore, nearby instances, like cars in satellite images, are often included in one segment.

D.4 In-domain datasets

We present the results of the five commonly used in-domain evaluation datasets in Table 5. Some values differ from the officially reported performance, mostly within $\pm 1\%$, which may be due to repeated runs [136]. It is worth noting that we could not reproduce the results from CAT-Seg on Pascal Context-459 and report a 4.2% lower mIoU [25]. The results for Pascal VOC differ from the reported values in [25, 76, 136] because of a different evaluation setting. We included a 21st *background* class and did not ignore the background pixels during evaluation. We find it misleading to ignore wrong predictions in the background, even if some objects are potentially not annotated. Other works assign the Pascal Context-59 labels that are not in PASCAL VOC to the background class [25, 46]. This may lead to better results than using the uniform label *background*.

Grounded-SAM has a very strong performance on Pascal VOC and nearly matches the fully-supervised result. However, the predictions become very noisy with an increasing number of classes, resulting in low mIoU scores. The CAT-Seg and SAN architectures produce the best results for the ADE20K and Pascal Context datasets.

Table 5: mIoU results for all evaluated models on commonly used in-domain evaluation datasets.

Model	ADE20K-150	ADE20K-847	Pascal Context-59	Pascal Context-459	Pascal VOC	Mean
Random (LB)	*0.16*	*0.02*	*0.6*	*0.03*	*1.15*	*0.39*
Best supervised (UB) [a]	*62.9*	*17.4*	*70.3*	-	*84.56*	-
ZSSeg-B	19.85	4.91	47.5	8.81	42.27	24.67
ZegFormer-B	11.79	4.16	28.85	4.61	43.88	18.66
X-Decoder-T	25.13	6.37	54.19	9.72	38.13	26.71
SAN-B	**27.56**	**10.22**	54.07	<u>12.42</u>	44.21	<u>29.7</u>
OpenSeeD-T	23.85	6.08	<u>56.79</u>	12.19	39.17	27.61
CAT-Seg-B	<u>27.52</u>	<u>8.99</u>	**57.5**	**13.47**	<u>60.45</u>	**33.59**
Grounded-SAM-B	14.75	2.58	41.65	10.05	**77.19**	29.25
OVSeg-L	29.58	9.11	55.32	12.07	40.82	29.38
SAN-L	<u>31.93</u>	<u>12.92</u>	57.53	**16.31**	50.16	<u>33.77</u>
CAT-Seg-L	31.14	11.39	**61.97**	<u>16.2</u>	63.97	**36.93**
Grounded-SAM-L	15.18	2.58	44.02	10.75	**82.36**	30.98
CAT-Seg-H	**34.52**	**13.08**	<u>61.2</u>	16.03	43.53	33.67
Grounded-SAM-H	15.36	2.62	43.95	10.88	<u>81.51</u>	30.86

[a] The supervised models are InternImage-H [133] (ADE20K-150 and Pascal Context-59), MaskFormer [23] (ADE20K-847), and DeepLabv3+ (Xception-JFT) [20] (Pascal VOC). Pascal Context-459 is rarely used in supervised settings and has, to our knowledge, not been evaluated with recent models.

D.5 Dataset-wise results

Table 6 presents the mIoU results for all datasets. The best-performing model varies between the datasets. CAT-Seg is overall the best-performing model architecture, while SAN, Grounded-SAM, and OpenSeeD are better in some specific use cases. Table 7 presents the CoI-mIoU results for each dataset. As discussed above, models without using CLIP often predict background instead of domain-specific classes which leads to a very low or zero CoI-mIoU.

Table 6: mIoU results for all datasets grouped by their domain.

	General						Earth Monitoring					Medical Sciences				Engineering				Agri. and Biology			
	BDD100K	Dark Zurich	MHP v1	FoodSeg103	ATLANTIS	DRAM	iSAID	ISPRS Pots.	WorldFloods	FloodNet	UAVid	Kvasir-Inst.	CHASE DB1	CryoNuSeg	PAXRay-4	Corrosion CS	DeepCrack	PST900	ZeroWaste-f	SUIM	CUB-200	CWFID	Mean
Random (LB)	*1.48*	*1.31*	*1.27*	*0.23*	*0.56*	*2.16*	*0.56*	*8.02*	*18.43*	*3.39*	*5.18*	*27.99*	*27.25*	*31.25*	*31.53*	*9.3*	*26.52*	*4.52*	*6.49*	*5.3*	*0.06*	*13.08*	*10.27*
Best sup. (UB)	*44.8*	*63.9*	*50.0*	*45.1*	*42.22*	*45.71*	*65.3*	*87.56*	*92.71*	*82.22*	*67.8*	*93.7*	*97.05*	*73.45*	*93.77*	*49.92*	*85.9*	*82.3*	*52.5*	*74.0*	*84.6*	*87.23*	*70.99*
ZSSeg-B	32.36	16.86	7.08	8.17	22.19	33.19	3.8	11.57	23.25	20.98	30.27	46.93	37.0	**38.7**	44.66	3.06	25.39	18.76	8.78	30.16	4.35	32.46	22.73
ZegFormer-B	14.14	4.52	4.33	10.01	18.98	29.45	2.68	14.04	25.93	22.74	20.84	27.39	12.47	11.94	18.09	4.78	29.77	19.63	17.52	28.28	16.8	32.26	17.57
X-Decoder-T	47.29	24.16	3.54	2.61	27.51	26.95	2.43	31.47	26.23	8.83	25.65	55.77	10.16	11.94	15.23	1.72	24.65	19.44	15.44	24.75	0.51	29.25	19.8
SAN-B	37.4	24.35	8.87	19.27	**36.51**	49.68	4.77	37.56	31.75	**37.44**	**41.65**	69.88	17.85	11.95	19.73	3.13	50.27	19.67	**21.27**	22.64	**16.91**	5.67	26.74
OpenSeeD-T	**47.95**	**28.13**	2.06	9.0	18.55	29.23	1.45	31.07	30.11	23.14	39.78	59.69	**46.68**	33.76	37.64	13.38	47.84	2.5	2.28	19.45	0.13	11.47	24.33
CAT-Seg-B	44.58	27.36	20.79	**21.54**	33.08	**62.42**	**15.75**	**41.89**	**39.47**	35.12	40.62	**70.68**	25.38	25.63	**44.94**	13.76	49.14	21.32	20.83	**39.1**	3.4	**45.47**	**33.74**
Gr.-SAM-B	41.58	20.91	**29.38**	10.48	17.33	57.38	12.22	26.68	33.41	19.19	38.34	46.82	23.56	38.06	41.07	**20.88**	**59.02**	**21.39**	16.74	14.13	0.43	38.41	28.52
OVSeg-L	45.28	22.53	6.24	16.43	33.44	53.33	8.28	31.03	31.48	35.59	38.8	71.13	20.95	13.45	22.06	6.82	16.22	21.89	11.71	38.17	14.0	33.76	26.94
SAN-L	43.81	30.39	9.34	24.46	**40.66**	**68.44**	11.77	**51.45**	48.24	39.26	**43.41**	72.18	7.64	11.94	29.33	6.83	23.65	19.01	18.32	40.01	19.3	1.91	30.06
CAT-Seg-L	45.83	**33.1**	**30.03**	**30.47**	33.6	66.54	**16.09**	51.42	**49.86**	39.84	42.02	**79.4**	24.99	35.06	**54.5**	16.87	31.42	**25.26**	**30.62**	**53.94**	9.24	39.0	**38.14**
Gr.-SAM-L	42.69	21.92	28.11	10.76	17.63	60.8	12.38	27.76	33.4	19.28	39.37	47.32	**25.16**	**38.06**	44.22	**20.88**	**58.21**	21.23	16.67	14.3	0.43	38.47	29.05
CAT-Seg-H	**48.34**	29.72	23.53	29.06	40.43	56.78	9.04	49.37	47.92	**40.98**	41.36	70.7	13.37	12.82	41.72	12.17	57.69	19.61	26.71	47.8	**19.49**	**45.99**	35.66
Gr.-SAM-H	42.95	22.09	28.05	9.97	17.68	60.86	12.44	27.79	33.23	19.31	39.41	46.97	25.13	**38.06**	43.64	**20.88**	53.74	21.34	16.68	14.3	0.43	38.29	28.78

Table 7: CoI-mIoU results for all datasets grouped by their domain.

	General						Earth Monitoring					Medical Sciences				Engineering				Agri. and Biology			
	BDD100K	Dark Zurich	MHP v1	FoodSeg103	ATLANTIS	DRAM	iSAID	ISPRS Pots.	WorldFloods	FloodNet	UAVid	Kvasir-Inst.	CHASE DB1	CryoNuSeg	PAXRay-4	Corrosion CS	DeepCrack	PST900	ZeroWaste-f	SUIM	CUB-200	CWFID	Mean
Random (LB)	*1.48*	*1.28*	*1.06*	*0.22*	*0.56*	*1.62*	*0.18*	*8.87*	*15.35*	*1.83*	*4.84*	*8.38*	*6.22*	*19.28*	*21.58*	*4.46*	*4.15*	*0.67*	*3.33*	*4.53*	*0.06*	*3.38*	*5.15*
ZSSeg-B	32.36	17.75	4.33	8.16	22.19	30.71	2.2	13.35	7.13	3.12	33.74	2.77	**10.93**	3.25	36.3	3.92	4.49	0.93	6.24	29.63	4.35	4.29	12.83
ZegFormer-B	14.14	4.72	4.08	9.91	18.98	25.6	2.2	16.72	0.0	1.42	23.81	9.63	7.89	23.88	29.75	5.49	4.96	0.24	1.71	31.8	16.6	9.24	11.94
X-Decoder-T	47.29	25.3	2.98	2.13	27.51	22.55	2.54	37.71	**26.84**	0.77	28.95	19.25	7.54	23.88	28.73	2.0	4.98	0.0	**10.52**	22.28	0.07	7.96	15.99
SAN-B	37.4	25.63	6.32	19.16	**36.51**	47.7	4.55	45.0	20.01	**14.41**	**46.08**	45.69	8.86	23.89	30.18	3.48	6.5	1.35	7.0	25.52	**16.82**	3.17	21.6
OpenSeeD-T	**47.95**	**29.7**	2.03	8.81	18.55	29.62	1.41	37.28	19.26	10.32	45.46	31.38	0.0	8.97	3.69	**5.8**	0.0	0.17	2.85	22.16	0.13	1.19	14.85
CAT-Seg-B	44.58	28.8	17.05	**21.28**	33.08	**60.26**	**13.16**	**50.07**	5.74	6.74	45.09	**47.66**	10.35	**25.98**	**39.78**	5.12	17.63	2.38	7.84	**37.49**	2.93	**20.88**	**24.72**
Gr.-SAM-B	41.58	21.75	**26.7**	10.01	17.33	54.66	7.73	30.7	0.0	0.0	39.42	2.71	9.71	0.0	26.52	0.0	**23.72**	**2.42**	1.39	9.99	0.0	8.9	15.24
OVSeg-L	45.28	23.72	3.8	16.56	33.44	51.07	6.54	37.13	25.27	11.67	44.02	47.77	9.46	24.29	32.13	**6.75**	5.29	3.25	5.61	40.75	14.06	4.64	22.39
SAN-L	43.81	32.08	6.22	24.37	**40.66**	**66.81**	8.71	60.17	**36.03**	**13.65**	**48.67**	49.69	7.18	23.88	33.44	5.54	4.42	0.96	9.16	43.17	19.0	2.86	26.39
CAT-Seg-L	45.83	**34.84**	**26.91**	**30.26**	33.6	64.89	**11.92**	**60.53**	25.28	6.11	46.32	**62.54**	**10.33**	**25.49**	**41.91**	5.82	8.85	**7.19**	**17.24**	53.47	8.82	11.4	**29.07**
Gr.-SAM-L	42.69	22.8	25.44	10.28	17.63	58.18	7.89	32.0	0.0	0.0	40.35	3.52	9.63	0.0	32.92	0.0	**23.39**	2.24	1.34	10.18	0.0	8.99	15.89
CAT-Seg-H	**48.34**	31.29	20.61	28.92	40.43	55.02	8.31	58.91	26.92	11.49	45.88	46.67	8.04	23.74	33.47	4.09	19.4	1.27	14.08	**53.92**	**19.42**	**22.02**	28.28
Gr.-SAM-H	42.95	22.97	25.4	9.49	17.68	58.25	7.85	32.02	0.0	0.0	40.44	2.86	9.63	0.0	31.75	0.0	16.47	2.35	1.34	10.18	0.0	8.81	15.47

E Qualitative examples

An example of each dataset with predictions from the four large models is presented in Figure 8 and 9. CAT-Seg-L has visually the best predictions, which is in line with the quantitative results. The mask-based approaches SAN-L and OVSeg-L tend to segment very large areas with one class, e.g., in MHP v1, CryoNuSeg, and CWFID. Sometimes, they also fail to recognize the background as visualized in SUIM and CUB-200. This can happen when masks of the background include the predicted class itself. The prediction quality from Grounded-SAM-L varies the most. E.g., the model has a good prediction for UAVid but insufficient predictions for all other earth monitoring datasets.

Figure 8: Predictions for datasets of the domains general and earth monitoring.

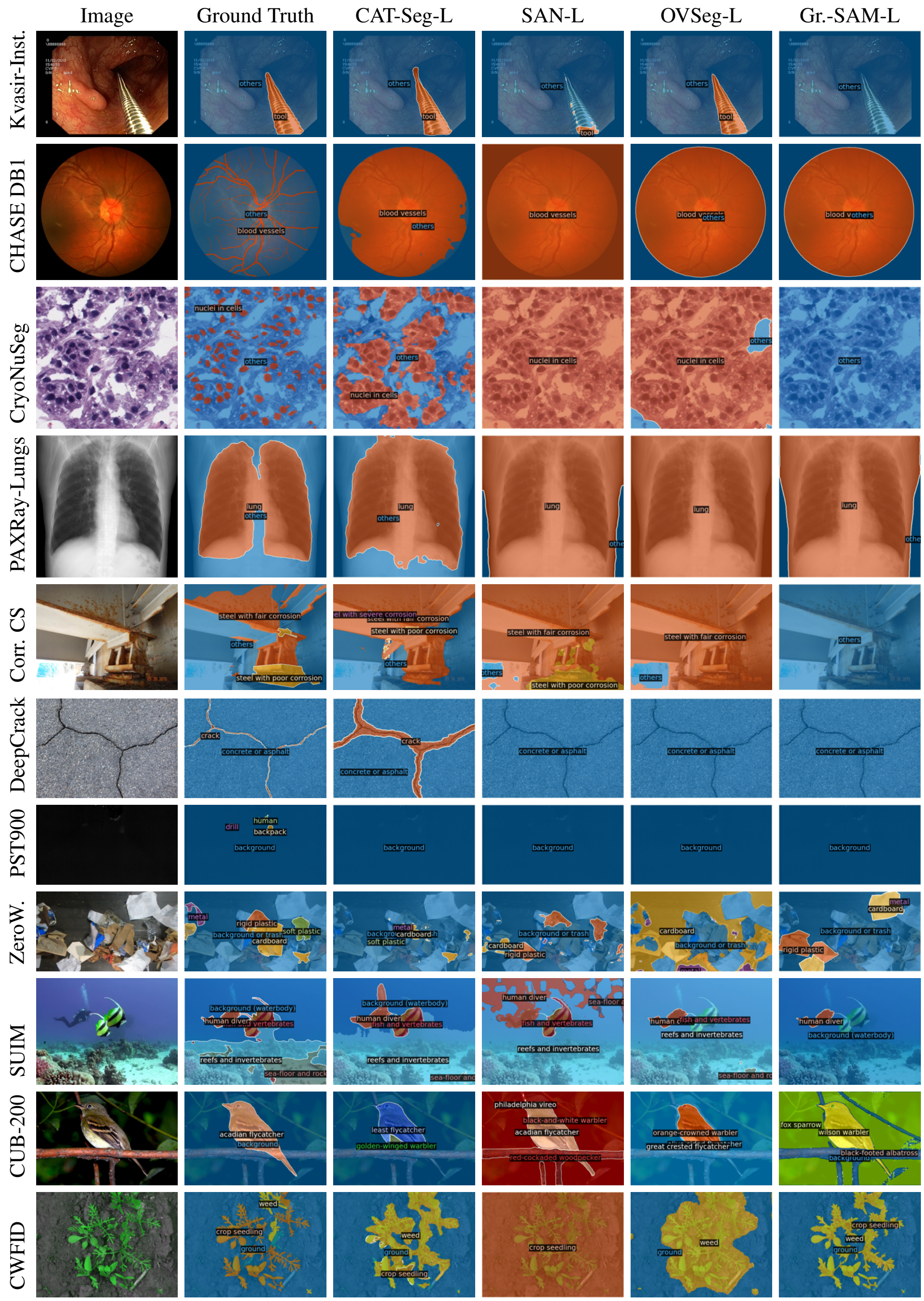


Figure 9: Predictions for datasets from medical sciences, engineering, and agriculture and biology.

References

[1] Abdelfattah, R., Wang, X., and Wang, S. (2020). Ttpla: An aerial-image dataset for detection and segmentation of transmission towers and power lines. In *Proceedings of the Asian Conference on Computer Vision*.

[2] Ahmed, Z., Panhwar, S. Q., Baqai, A., Umrani, F. A., Ahmed, M., and Khan, A. (2022). Deep learning based automated detection of intraretinal cystoid fluid. *International Journal of Imaging Systems and Technology*, 32(3):902–917.

[3] Alam, F. (2021). Leaf disease segmentation dataset. Kaggle. https://www.kaggle.com/datasets/fakhrealam9537/leaf-disease-segmentation-dataset.

[4] Alemohammad, H. and Booth, K. (2020). LandCoverNet: A global benchmark land cover classification training dataset. *arXiv preprint arXiv:2012.03111*.

[5] Asad, M. H., Asim, M. M., Awan, M. N. M., and Yousaf, M. H. (2023). Natural disaster damage assessment using semantic segmentation of uav imagery. In *2023 International Conference on Robotics and Automation in Industry (ICRAI)*, pages 1–7. IEEE.

[6] Bambach, S., Lee, S., Crandall, D. J., and Yu, C. (2015). Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *Proceedings of the IEEE international conference on computer vision*, pages 1949–1957.

[7] Bandara, N. (2022). Ensemble Deep Learning for Automated Dust Storm Detection Using Satellite Images. In *2022 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, volume 5, pages 178–183. IEEE.

[8] Bashkirova, D., Abdelfattah, M., Zhu, Z., Akl, J., Alladkani, F., Hu, P., Ablavsky, V., Calli, B., Bargal, S. A., and Saenko, K. (2022). Zerowaste dataset: Towards deformable object segmentation in cluttered scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21147–21157.

[9] Bell, S., Upchurch, P., Snavely, N., and Bala, K. (2013). OpenSurfaces: A richly annotated catalog of surface appearance. *ACM Transactions on graphics (TOG)*, 32(4):1–17.

[10] Bell, S., Upchurch, P., Snavely, N., and Bala, K. (2015). Material recognition in the wild with the materials in context database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3479–3487.

[11] Bianchi, E. and Hebdon, M. (2021a). Corrosion condition state semantic segmentation dataset. *University Libraries, Virginia Tech: Blacksburg, VA, USA*.

[12] Bianchi, E. and Hebdon, M. (2021b). Labeled Cracks in the Wild (LCW) Dataset. *University Libraries, Virginia Tech*.

[13] Bonafilia, D., Tellman, B., Anderson, T., and Issenberg, E. (2020). Sen1Floods11: A georeferenced dataset to train and test deep learning flood algorithms for sentinel-1. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 210–211.

[14] Božič, J., Tabernik, D., and Skočaj, D. (2021). Mixed supervision for surface-defect detection: From weakly to fully supervised learning. *Computers in Industry*, 129:103459.

[15] Brüggemann, D., Sakaridis, C., Truong, P., and Van Gool, L. (2023). Refign: Align and refine for adaptation of semantic segmentation to adverse conditions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3174–3184.

[16] BSF Swissphoto (2012). ISPRS Potsdam dataset within the ISPRS test project on urban classification, 3D building reconstruction and semantic labeling. https://www.isprs.org/education/benchmarks/UrbanSemLab/default.aspx.

[17] Caicedo, J. C., Goodman, A., Karhohs, K. W., Cimini, B. A., Ackerman, J., Haghighi, M., Heng, C., Becker, T., Doan, M., McQuin, C., et al. (2019). Nucleus segmentation across imaging experiments: the 2018 data science bowl. *Nature methods*, 16(12):1247–1253.

[18] Candemir, S., Jaeger, S., Palaniappan, K., Musco, J. P., Singh, R. K., Xue, Z., Karargyris, A., Antani, S., Thoma, G., and McDonald, C. J. (2013). Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE transactions on medical imaging*, 33(2):577–590.

[19] Caron, M., Houlsby, N., and Schmid, C. (2022). Location-aware self-supervised transformers. *arXiv preprint arXiv:2212.02400*.

[20] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018a). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, pages 801–818.

[21] Chen, X., Zhong, X., Li, T., An, Y., and Mo, L. (2022). Normtoraw: A style transfer based self-supervised learning approach for nuclei segmentation. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

[22] Chen, Y., Wang, Y., Lu, P., Chen, Y., and Wang, G. (2018b). Large-scale structure from motion with semantic constraints of aerial images. In *Pattern Recognition and Computer Vision: First Chinese Conference*, pages 347–359. Springer.

[23] Cheng, B., Schwing, A., and Kirillov, A. (2021). Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34.

[24] Chiu, M. T., Xu, X., Wei, Y., Huang, Z., Schwing, A. G., Brunner, R., Khachatrian, H., Karapetyan, H., Dozier, I., Rose, G., et al. (2020). Agriculture-vision: A large aerial image database for agricultural pattern analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2828–2838.

[25] Cho, S., Shin, H., Hong, S., An, S., Lee, S., Arnab, A., Seo, P. H., and Kim, S. (2023). CAT-Seg: Cost aggregation for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2303.11797v1*.

[26] Ciaglia, F., Zuppichini, F. S., Guerrie, P., McQuade, M., and Solawetz, J. (2022). Roboflow 100: A rich, multi-domain object detection benchmark. *arXiv preprint arXiv:2211.13523*.

[27] Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al. (2019). Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*.

[28] Cohen, N., Newman, Y., and Shamir, A. (2022). Semantic segmentation in art paintings. *Computer Graphics Forum*, 41(2):261–275.

[29] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223.

[30] Darkhalil, A., Shan, D., Zhu, B., Ma, J., Kar, A., Higgins, R., Fidler, S., Fouhey, D., and Damen, D. (2022). Epic-kitchens visor benchmark: Video segmentations and object relations. *Advances in Neural Information Processing Systems*, 35:13745–13758.

[31] de Almeida Pereira, G. H., Fusioka, A. M., Nassu, B. T., and Minetto, R. (2021). Active fire detection in Landsat-8 imagery: A large-scale dataset and a deep-learning study. *ISPRS Journal of Photogrammetry and Remote Sensing*, 178:171–186.

[32] Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., and Raskar, R. (2018). Deepglobe 2018: A challenge to parse the earth through satellite images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 172–181.

[33] Ding, J., Xue, N., Xia, G.-S., and Dai, D. (2022). Decoupling zero-shot semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592.

[34] Dobos, O., Horvath, P., Nagy, F., Danka, T., and Viczián, A. (2019). A deep learning-based approach for high-throughput hypocotyl phenotyping. *Plant physiology*, 181(4):1415–1424.

[35] Eppel, S., Xu, H., Bismuth, M., and Aspuru-Guzik, A. (2020). Computer vision for recognition of materials and vessels in chemistry lab settings and the vector-labpics data set. *ACS central science*, 6(10):1743–1752.

[36] Erfani, S. M. H., Wu, Z., Wu, X., Wang, S., and Goharian, E. (2022). Atlantis: A benchmark for semantic segmentation of waterbody images. *Environmental Modelling & Software*, 149:105333.

[37] Everingham, M., Gool, L. V., Williams, C. K. I., Winn, J. M., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338.

[38] Fan, D.-P., Ji, G.-P., Cheng, M.-M., and Shao, L. (2022). Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[39] Fan, S., Wang, Z., Wang, Y., and Liu, J. (2023). Spidermesh: Spatial-aware demand-guided recursive meshing for rgb-t semantic segmentation. *arXiv preprint arXiv:2303.08692*.

[40] Follmann, P., Bottger, T., Hartinger, P., Konig, R., and Ulrich, M. (2018). MVTec D2S: densely segmented supermarket dataset. In *European Conference on Computer Vision (ECCV)*, pages 569–585.

[41] Fraz, M. M., Remagnino, P., Hoppe, A., Uyyanonvara, B., Rudnicka, A. R., Owen, C. G., and Barman, S. A. (2012). An ensemble classification-based approach applied to retinal blood vessel segmentation. *IEEE Transactions on Biomedical Engineering*, 59(9):2538–2548.

[42] Galloway, A., Taylor, G. W., Ramsay, A., and Moussa, M. (2017). The ciona17 dataset for semantic segmentation of invasive species in a marine aquaculture environment. In *Conference on Computer and Robot Vision (CRV)*, pages 361–366. IEEE.

[43] Garcia-Peraza-Herrera, L. C., Fidon, L., D'Ettorre, C., Stoyanov, D., Vercauteren, T., and Ourselin, S. (2021). Image Compositing for Segmentation of Surgical Tools Without Manual Annotations. *IEEE Transactions on Medical Imaging*, 40(5):1450–1460.

[44] Garnot, V. S. F. and Landrieu, L. (2021). Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4872–4881.

[45] Ge, Y., Zhang, R., Wang, X., Tang, X., and Luo, P. (2019). Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5337–5345.

[46] Ghiasi, G., Gu, X., Cui, Y., and Lin, T.-Y. (2022). Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision (ECCV)*, pages 540–557. Springer.

[47] Ghosh, R., Ravirathinam, P., Jia, X., Khandelwal, A., Mulla, D., and Kumar, V. (2021). Calcrop21: A georeferenced multi-spectral dataset of satellite imagery and crop labels. In *IEEE International Conference on Big Data*, pages 1625–1632. IEEE.

[48] Gupta, A., Dollar, P., and Girshick, R. (2019a). Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364.

[49] Gupta, R., Hosfelt, R., Sajeev, S., Patel, N., Goodman, B., Doshi, J., Heim, E., Choset, H., and Gaston, M. (2019b). xbd: A dataset for assessing building damage from satellite imagery. *arXiv preprint arXiv:1911.09296*.

[50] Habili, N., Kwan, E., Li, W., Webers, C., Oorloff, J., Armin, M. A., and Petersson, L. (2022). A Hyperspectral and RGB Dataset for Building Facade Segmentation. *arXiv preprint arXiv:2212.02749*.

[51] Haucke, T., K¨uhl, H. S., and Steinhage, V. (2022). SOCRATES: Introducing Depth in Visual Wildlife Monitoring Using Stereo Vision. *Sensors*, 22(23):9082.

[52] Haug, S. and Ostermann, J. (2015). A crop/weed field image dataset for the evaluation of computer vision based precision agriculture tasks. In *Computer Vision - ECCV 2014 Workshops*, pages 105–116. Springer.

[53] Hitrec, T., Luppi, M., Bastianini, S., Squarcio, F., Berteotti, C., Lo Martire, V., Martelli, D., Occhinegro, A., Tupone, D., Zoccoli, G., et al. (2019). Neural control of fasting-induced torpor in mice. *Scientific reports*, 9(1):15462.

[54] Hong, J., Fulton, M., and Sattar, J. (2020). Trashcan: A semantically-segmented dataset towards visual detection of marine debris. *arXiv preprint arXiv:2007.08097*.

[55] Hoover, A., Kouznetsova, V., and Goldbaum, M. (2000). Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical imaging*, 19(3):203–210.

[56] Hosseinpoor, S. (2019). Semantic Terrain Segmentation with an Original RGB Data Set, Targeting Elevation Differences. Master's thesis.

[57] Humans In The Loop (2020). Semantic segmentation of aerial imagery. Kaggle. https://www.kaggle.com/datasets/humansintheloop/semantic-segmentation-of-aerial-imagery.

[58] IDEA-Research (2023). Grounded-SAM. https://github.com/IDEA-Research/Grounded-Segment-Anything.

[59] Iglovikov, V., Mushinskiy, S., and Osin, V. (2017). Satellite imagery feature detection using deep convolutional neural network: A kaggle competition. *arXiv preprint arXiv:1706.06169*.

[60] Institute of Computer Graphics and Vision (2019). Semantic Drone Dataset. http://dronedataset.icg.tugraz.at/.

[61] Islam, M. J., Edge, C., Xiao, Y., Luo, P., Mehtaz, M., Morse, C., Enan, S. S., and Sattar, J. (2020). Semantic segmentation of underwater imagery: Dataset and benchmark. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1769–1776.

[62] Jha, D., Ali, S., Emanuelsen, K., Hicks, S. A., Thambawita, V., Garcia-Ceja, E., Riegler, M. A., de Lange, T., Schmidt, P. T., Johansen, H. D., et al. (2021). Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. *MultiMedia Modeling: 27th International Conference, MMM 2021*, pages 218–229.

[63] Jha, D., Smedsrud, P. H., Riegler, M. A., Halvorsen, P., de Lange, T., Johansen, D., and Johansen, H. D. (2020). Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*, pages 451–462. Springer.

[64] Jiang, P., Osteen, P., Wigness, M., and Saripalli, S. (2021). Rellis-3d dataset: Data, benchmarks and analysis. In *IEEE international conference on robotics and automation (ICRA)*, pages 1110–1116. IEEE.

[65] Kamran, S. A., Hossain, K. F., Tavakkoli, A., Zuckerbrod, S. L., Sanders, K. M., and Baker, S. A. (2021). Rv-gan: Segmenting retinal vascular structure in fundus photographs using a novel multi-scale generative adversarial network. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pages 34–44. Springer.

[66] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. *arXiv preprint arXiv:2304.02643*.

[67] Lai, C.-Y., Kingslake, J., Wearing, M. G., Chen, P.-H. C., Gentine, P., Li, H., Spergel, J. J., and van Wessem, J. M. (2020). Vulnerability of Antarctica's ice shelves to meltwater-driven fracture. *Nature*, 584(7822):574–578.

[68] Laves, M.-H., Bicker, J., Kahrs, L. A., and Ortmaier, T. (2019a). A dataset of laryngeal endoscopic images with comparative study on convolution neural network-based semantic segmentation. *International journal of computer assisted radiology and surgery*, 14:483–492.

[69] Laves, M.-H., Bicker, J., Kahrs, L. A., and Ortmaier, T. (2019b). A dataset of laryngeal endoscopic images with comparative study on convolution neural network-based semantic segmentation. *International journal of computer assisted radiology and surgery*, 14:483–492.

[70] Le, T.-N., Nguyen, T. V., Nie, Z., Tran, M.-T., and Sugimoto, A. (2019). Anabranch network for camouflaged object segmentation. *Computer vision and image understanding*, 184:45–56.

[71] Lehmann, T. M., Schubert, H., Keysers, D., Kohnen, M., and Wein, B. B. (2003). The IRMA code for unique classification of medical images. *Medical Imaging 2003: PACS and Integrated Medical Information Systems: Design and Evaluation*, 5033:440 – 451.

[72] Li, J., Zhao, J., Wei, Y., Lang, C., Li, Y., Sim, T., Yan, S., and Feng, J. (2017). Multiple-human parsing in the wild. *arXiv preprint arXiv:1705.07206*.

[73] Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., et al. (2022a). Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975.

[74] Li, X., Wei, T., Chen, Y. P., Tai, Y.-W., and Tang, C.-K. (2020). Fss-1000: A 1000-class dataset for few-shot segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

[75] Li, Z., Chen, P., Shuai, L., Wang, M., Zhang, L., Wang, Y., and Mu, J. (2022b). A Copy Paste and Semantic Segmentation-Based Approach for the Classification and Assessment of Significant Rice Diseases. *Plants*, 11(22):3174.

[76] Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., and Marculescu, D. (2023). Open-vocabulary semantic segmentation with mask-adapted CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070.

[77] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. pages 740–755. Springer.

[78] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al. (2023). Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

[79] Liu, Y., Yao, J., Lu, X., Xie, R., and Li, L. (2019). Deepcrack: A deep hierarchical feature learning architecture for crack segmentation. *Neurocomputing*, 338:139–153.

[80] Liu, Z., Jin, L., Chen, J., Fang, Q., Ablameyko, S., Yin, Z., and Xu, Y. (2021). A survey on applications of deep learning in microscopy image analysis. *Computers in Biology and Medicine*, 134:104523.

[81] Lu, H., Li, Y., Zhang, Y., Chen, M., Serikawa, S., and Kim, H. (2017). Underwater optical image processing: a comprehensive review. *Mobile networks and applications*, 22:1204–1211.

[82] Lyu, Y., Vosselman, G., Xia, G.-S., Yilmaz, A., and Yang, M. Y. (2020). UAVid: A semantic segmentation dataset for UAV imagery. *ISPRS journal of photogrammetry and remote sensing*, 165:108–119.

[83] Mahbod, A., Schaefer, G., Bancher, B., Löw, C., Dorffner, G., Ecker, R., and Ellinger, I. (2021). CryoNuSeg: A dataset for nuclei instance segmentation of cryosectioned H&E-stained histological images. *Computers in biology and medicine*, 132:104349.

[84] Mateo-Garcia, G., Veitch-Michaelis, J., Smith, L., Oprea, S. V., Schumann, G., Gal, Y., Baydin, A. G., and Backes, D. (2021). Towards global flood mapping onboard low cost satellites with machine learning. *Scientific reports*, 11(1):1–12.

[85] Mayer, K., Rausch, B., Arlt, M.-L., Gust, G., Wang, Z., Neumann, D., and Rajagopal, R. (2022). 3D-PV-Locator: Large-scale detection of rooftop-mounted photovoltaic systems in 3D. *Applied Energy*, 310:118469.

[86] Mazurowski, M. A., Dong, H., Gu, H., Yang, J., Konz, N., and Zhang, Y. (2023). Segment anything model for medical image analysis: An experimental study. *Medical Image Analysis*, 89:102918.

[87] Meena, S. R., Nava, L., Bhuyan, K., Puliero, S., Soares, L. P., Dias, H. C., Floris, M., and Catani, F. (2022). HR-GLDD: A globally distributed dataset using generalized DL for rapid landslide mapping on HR satellite imagery. *Earth System Science Data Discussions*, pages 1–21.

[88] Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al. (2014). The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024.

[89] Miao, S., Du, Y., Lu, X., and Guan, H. (2023). Semantic segmentation of vehicle vision based on two-branch enet network. In *2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA)*, pages 477–481. IEEE.

[90] Minervini, M., Fischbach, A., Scharr, H., and Tsaftaris, S. A. (2016). Finely-grained annotated datasets for image-based plant phenotyping. *Pattern recognition letters*, 81:80–89.

[91] Mortimer, P. and W¨unsche, H.-J. (2022). TAS-NIR: A VIS+ NIR Dataset for Fine-grained Semantic Segmentation in Unstructured Outdoor Environments. *arXiv preprint arXiv:2212.09368*.

[92] Mottaghi, R., Chen, X., Liu, X., Cho, N.-G., Lee, S.-W., Fidler, S., Urtasun, R., and Yuille, A. (2014). The role of context for object detection and semantic segmentation in the wild. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 891–898.

[93] Neuhold, G., Ollmann, T., Rota Bulo, S., and Kontschieder, P. (2017). The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999.

[94] Nickerson, R. C., Varshney, U., Muntermann, J., and Nickerson, R. C. (2013). A method for taxonomy development and its application in information systems. *European Journal of Information Systems*.

[95] Nir, G., Hor, S., Karimi, D., Fazli, L., Skinnider, B. F., Tavassoli, P., Turbin, D., Villamil, C. F., Wang, G., Wilson, R. S., et al. (2018). Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts. *Medical image analysis*, 50:167–180.

[96] Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. (2012). Cats and dogs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3498–3505. IEEE.

[97] Pradham, P., Younan, N. H., and King, R. L. (2008). 16 - concepts of image fusion in remote sensing applications. In Stathaki, T., editor, *Image Fusion*, pages 393–428. Academic Press, Oxford.

[98] Proença, P. F. and Simoes, P. (2020). Taco: Trash annotations in context for litter detection. *arXiv preprint arXiv:2003.06975*.

[99] Quesada, J., Sathidevi, L., Liu, R., Ahad, N., Jackson, J., Azabou, M., Xiao, J., Liding, C., Jin, M., Urzay, C., et al. (2022). Mtneuro: A benchmark for evaluating representations of brain structure across multiple levels of abstraction. *Advances in Neural Information Processing Systems*, 35:5299–5314.

[100] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*.

[101] Rahnemoonfar, M., Chowdhury, T., Sarkar, A., Varshney, D., Yari, M., and Murphy, R. R. (2021). Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9:89644–89654.

[102] Ros, G., Sellart, L., Materzynska, J., Vazquez, D., and Lopez, A. M. (2016). The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243.

[103] Ross, T., Reinke, A., Full, P. M., Wagner, M., Kenngott, H., Apitz, M., Hempe, H., Filimon, D. M., Scholz, P., Tran, T. N., et al. (2020). Robust medical instrument segmentation challenge 2019. *arXiv preprint arXiv:2003.10299*.

[104] Roy, S. (2021). Thermal Dog Dataset. Kaggle. https://www.kaggle.com/datasets/sagnik1511/thermal-dog-dataset-instance-segmentation.

[105] Sakaridis, C., Dai, D., and Gool, L. V. (2019). Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7374–7383.

[106] Saleh, A., Laradji, I. H., Konovalov, D. A., Bradley, M., Vazquez, D., and Sheaves, M. (2020). A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Scientific Reports*, 10(1):14671.

[107] Santos, T., de Souza, L., dos Santos, A., and Sandra, A. (2019). Embrapa Wine Grape Instance Segmentation Dataset–Embrapa WGISD. *Zenodo*.

[108] Seibold, C., Reiß, S., Sarfraz, S., Fink, M. A., Mayer, V., Sellner, J., Kim, M. S., Maier-Hein, K. H., Kleesiek, J., and Stiefelhagen, R. (2022). Detailed Annotations of Chest X-Rays via CT Projection for Report Understanding. *Proceedings of the 33th British Machine Vision Conference*.

[109] Severstal (2019). Severstal: Steel Defect Detection. Kaggle. https://www.kaggle.com/competitions/severstal-steel-defect-detection/overview.

[110] Shermeyer, J., Hogan, D., Brown, J., Van Etten, A., Weir, N., Pacifici, F., Hansch, R., Bastidas, A., Soenen, S., Bacastow, T., et al. (2020). SpaceNet 6: Multi-sensor all weather mapping dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 196–197.

[111] Shilkrot13, R., Narasimhaswamy, S., Vazir, S., and Hoai12, M. (2019). WorkingHands: A hand-tool assembly dataset for image segmentation and activity mining.

[112] Shivakumar, S. S., Rodrigues, N., Zhou, A., Miller, I. D., Kumar, V., and Taylor, C. J. (2020). PST900: RGB-thermal calibration, dataset and segmentation network. *IEEE international conference on robotics and automation (ICRA)*, pages 9441–9447.

[113] Sirinukunwattana, K., Pluim, J. P., Chen, H., Qi, X., Heng, P.-A., Guo, Y. B., Wang, L. Y., Matuszewski, B. J., Bruni, E., Sanchez, U., et al. (2017). Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35:489–502.

[114] Steininger, D., Trondl, A., Croonen, G., Simon, J., and Widhalm, V. (2023). The CropAndWeed Dataset: A Multi-Modal Learning Approach for Efficient Crop and Weed Manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3729–3738.

[115] Steininger, D., Widhalm, V., Simon, J., Kriegler, A., and Sulzbachner, C. (2021). The aircraft context dataset: understanding and optimizing data variability in aerial domains. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3823–3832.

[116] Swan, R. M., Atha, D., Leopold, H. A., Gildner, M., Oij, S., Chiu, C., and Ono, M. (2021). Ai4mars: A dataset for terrain-aware autonomous driving on mars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1982–1991.

[117] Sylvester, A. (2021). Neural Net Mapping of Hudson Bay Sea Ice. Github repository. https://github.com/asylve/Sea-Ice.

[118] Teikari, P., Santos, M., Poon, C., and Hynynen, K. (2016). Deep learning convolutional networks for multiphoton microscopy vasculature segmentation. *arXiv preprint arXiv:1606.02382*.

[119] Tlebaldinova, A., Karmenova, M., Ponkina, E., and Bondarovich, A. (2022). Cnn-based approaches for weed detection. In *2022 10th International Scientific Conference on Computer Science (COMSCI)*, pages 1–4. IEEE.

[120] Tong, X.-Y., Xia, G.-S., Lu, Q., Shen, H., Li, S., You, S., and Zhang, L. (2020). Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment*, 237:111322.

[121] Trotter, C., Atkinson, G., Sharpe, M., Richardson, K., McGough, A. S., Wright, N., Burville, B., and Berggren, P. (2020). NDD20: A large-scale few-shot dolphin dataset for coarse and fine-grained categorisation. *arXiv preprint arXiv:2005.13359*.

[122] Twinanda, A. P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., and Padoy, N. (2016). Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97.

[123] Upchurch, P. and Niu, R. (2022). A Dense Material Segmentation Dataset for Indoor and Outdoor Scene Parsing. In *European Conference on Computer Vision (ECCV)*, pages 450–466. Springer.

[124] Urooj, A. and Borji, A. (2018). Analysis of hand segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4710–4719.

[125] Vitale, S., Orlando, J. I., Iarussi, E., and Larrabide, I. (2020). Improving realism in patient-specific abdominal ultrasound simulation using cyclegans. *International journal of computer assisted radiology and surgery*, 15(2):183–192.

[126] Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. (2011). Caltech-UCSD Birds 200. *California Institute of Technology*.

[127] Wang, B., Zhang, L., Wen, L., Liu, X., and Wu, Y. (2021a). Towards real-world prohibited item detection: A large-scale x-ray benchmark. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5412–5421.

[128] Wang, D., Zhang, J., Du, B., Xia, G.-S., and Tao, D. (2022a). An empirical study of remote sensing pretraining. *IEEE Transactions on Geoscience and Remote Sensing*.

[129] Wang, J., Zheng, Z., Ma, A., Lu, X., and Zhong, Y. (2021b). LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*.

[130] Wang, L., Li, R., Duan, C., Zhang, C., Meng, X., and Fang, S. (2022b). A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5.

[131] Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., and Atkinson, P. M. (2022c). Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:196–214.

[132] Wang, Q., Paynabar, K., and Pacella, M. (2022d). Online automatic anomaly detection for photovoltaic systems using thermography imaging and low rank matrix decomposition. *Journal of Quality Technology*, 54(5):503–516.

[133] Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., et al. (2023). Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14408–14419.

[134] Waqas Zamir, S., Arora, A., Gupta, A., Khan, S., Sun, G., Shahbaz Khan, F., Zhu, F., Shao, L., Xia, G.-S., and Bai, X. (2019). iSAID: A large-scale dataset for instance segmentation in aerial images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–37.

[135] Wu, X., Fu, X., Liu, Y., Lim, E.-P., Hoi, S. C., and Sun, Q. (2021). A large-scale benchmark for food image segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 506–515.

[136] Xu, M., Zhang, Z., Wei, F., Hu, H., and Bai, X. (2023). Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954.

[137] Xu, M., Zhang, Z., Wei, F., Lin, Y., Cao, Y., Hu, H., and Bai, X. (2022). A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, pages 736–753. Springer.

[138] Yildirim, B. and Cole, J. M. (2021). Bayesian particle instance segmentation for electron microscopy image quantification. *Journal of Chemical Information and Modeling*, 61(3):1136–1149.

[139] Yogamani, S., Hughes, C., Horgan, J., Sistu, G., Varley, P., O'Dea, D., Uricár, M., Milz, S., Simon, M., Amende, K., et al. (2019). Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9308–9318.

[140] Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., and Darrell, T. (2020). BDD100K: A diverse driving dataset for heterogeneous multitask learning. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645.

[141] Zendel, O., Murschitz, M., Zeilinger, M., Steininger, D., Abbasi, S., and Beleznai, C. (2019). Railsem19: A dataset for semantic rail scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0.

[142] Zhang, H., Li, F., Zou, X., Liu, S., Li, C., Yang, J., and Zhang, L. (2023). A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1020–1031.

[143] Zhang, L., Zhou, S., Stent, S., and Shi, J. (2022a). Fine-Grained Egocentric Hand-Object Segmentation: Dataset, Model, and Applications. In *European Conference on Computer Vision (ECCV)*, pages 127–145. Springer.

[144] Zhang, X., Zhao, W., Zhang, W., Peng, J., and Fan, J. (2022b). Guided filter network for semantic image segmentation. *IEEE Transactions on Image Processing*, 31:2695–2709.

[145] Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H., et al. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890.

[146] Zheng, S., Yang, F., Kiapour, M. H., and Piramuthu, R. (2018). Modanet: A large-scale street fashion dataset with polygon annotations. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1670–1678.

[147] Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., and Torralba, A. (2019). Semantic understanding of scenes through the ADE20K dataset. *International Journal of Computer Vision*, 127:302–321.

[148] Zou, C., Yu, Q., Du, R., Mo, H., Song, Y.-Z., Xiang, T., Gao, C., Chen, B., and Zhang, H. (2018). Sketchyscene: Richly-annotated scene sketches. In *European Conference on Computer Vision (ECCV)*, pages 421–436.

[149] Zou, X., Dou, Z.-Y., Yang, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L., et al. (2023). Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127.