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

**Benedikt Blumenstiel**[*]
Karlsruhe Institute of Technology
IBM Research Europe
benedikt.blumenstiel@kit.edu

**Johannes Jakubik**[*]
Karlsruhe Institute of Technology
IBM Research Europe
johannes.jakubik@kit.edu

**Hilde Kühne**
University of Bonn
MIT-IBM Watson AI Lab
hildegard.kuehne@ibm.com

**Michael Vössing**
Karlsruhe Institute of Technology
IBM Germany
michael.voessing@kit.edu

## Abstract

While semantic segmentation has seen tremendous improvements in the past, there are still significant labeling efforts necessary and the problem of limited generalization to classes that have not been present during training. To address this problem, zero-shot semantic segmentation makes use of large self-supervised vision-language models, allowing zero-shot transfer to unseen classes. In this work, we build a benchmark for Multi-domain Evaluation of Semantic Segmentation (MESS), which allows a holistic analysis of performance across a wide range of domain-specific datasets such as medicine, engineering, earth monitoring, biology, and agriculture. To do this, we reviewed 120 datasets, developed a taxonomy, and classified the datasets according to the developed taxonomy. We select a representative subset consisting of 22 datasets and propose it as the MESS benchmark. We evaluate eight recently published models on the proposed MESS benchmark and analyze characteristics for the performance of zero-shot transfer models. The toolkit is available at https://github.com/blumenstiel/MESS.

## 1   Introduction

Zero-shot semantic segmentation utilizes self-supervised models such as CLIP to minimize labeling requirements during training and to improve model generalization. Recent models are already able to include classes during inference that were not present during training. For this reason, zero-shot semantic segmentation is becoming increasingly relevant for real-world scenarios. In particular, the performance on domain-specific datasets such as earth monitoring datasets, as visualized in Figure 1, becomes more and more relevant. Current standard benchmarks tend to focus on in-domain tasks but do not capture performance comparisons across domains. This is problematic because it limits insight into the applicability of zero-shot semantic segmentation to new domains. It also makes it difficult to assess whether architectures might be suitable for datasets that pose additional challenges (e.g., different sensor types or specialized vocabulary). To better understand the behavior of zero-shot semantic segmentation models on a wider range of more complex, domain-specific datasets, we propose a holistic Multi-domain Evaluation of Semantic Segmentation (MESS). To this end, we have examined 120 datasets and classified them within a developed taxonomy. We leverage our benchmark to evaluate eight recently published models for zero-shot semantic segmentation including
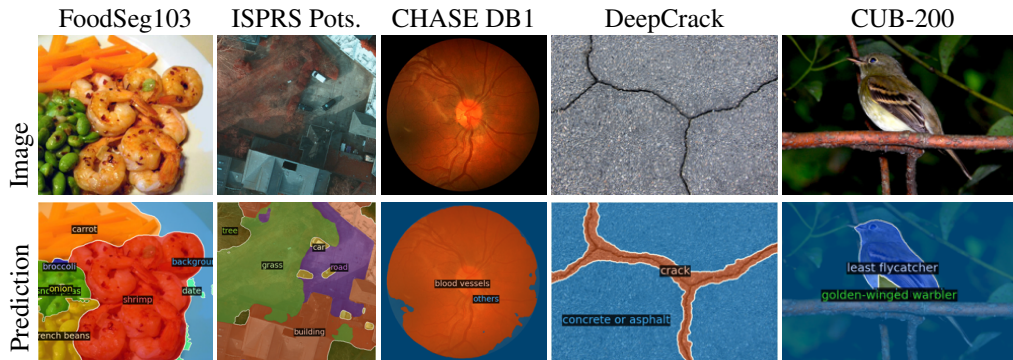
---

[*]Equal contributions

Figure 1: CAT-Seg-L [7] predictions for a range of domain-specific datasets. The model achieves promising predictions on everyday and satellite images, while it faces difficulties in segmenting small segments such as blood vessels, and distinguishing similar classes such as bird species.

the state-of-the-art models on 22 datasets from the fields of medical sciences, earth monitoring, agriculture and biology, engineering as well as a general domain including datasets on, e.g., driving scenes, maritime scenes, paintings, and body parts. Our evaluation focuses on zero-shot text-to-mask models—also known as open-vocabulary semantic segmentation (OVSS)—and later also compares their performance with zero-shot point-to-mask and box-to-mask approaches of SAM [20]. Using the proposed benchmark, we identify and analyze several characteristics that influence the performance of OVSS models, i.a., showing that the semantic, textual similarity of classes as well as the underlying sensor type, significantly affect the performance of current models.

Our experiments reveal various challenges for the application of zero-shot semantic segmentation on domain-specific datasets, e.g., we found that the selection of class labels can significantly affect the quality of predictions. We also observe that the models are sensitive to the semantics of the textual prompts, e.g., general terminology leads to better performance than domain-specific terminology. Overall, we hope that our benchmark will support accelerating zero-shot semantic segmentation and improve the real-world applicability of semantic segmentation in general.

We summarize the contributions of this work as follows: (1) We develop a taxonomy based on a quantitative and qualitative analysis of a broad variety of semantic segmentation datasets. (2) We propose a new benchmark for multi-domain semantic segmentation. (3) We evaluate eight zero-shot models on the MESS benchmark with an in-depth analysis of the task characteristics.

## 2 Related work

### 2.1 Zero-shot semantic segmentation

Large-scale self-supervised pre-training has revolutionized the field of computer vision over the last couple of years. One stream of work focuses on vision-language pre-training such as in recent foundation model architectures like CLIP [37], ALIGN [19], and Florence [54]. These models are trained on image-text pairs and encode both visual and text semantics in a shared embedding space. This approach particularly enables so-called open-vocabulary image classification by computing the similarity between the embeddings of the image and the embeddings of natural language describing the classes in the image. The text describing the images can be any arbitrary textual sequence and might describe classes on the images that have been unseen during training. This is in contrast to recent segmentation models, like Segment-Anything (SAM, see [20]), which are trained only on image data and therefore do not include a text encoder to encode semantic concepts. Hence, segmentation models like SAM do not facilitate open-vocabulary out of the box and need to be adapted to support the processing of textual information (e.g., by using additional models that generate text embeddings or models that provide bounding boxes as input such as Grounding DINO [27]).

Early approaches in OVSS have been built upon standard zero-shot semantic segmentation, such as ZS3Net [5], using simple word2vec text encoders. Subsequent two-stage approaches made use of mask proposals based on MaskFormer [6] in stage 1 followed by predictions of each mask by

CLIP [9, 10, 25, 51]. Recently, one-stage frameworks like SAN [50] generate masks in a side adapter network during the CLIP inference. Therefore, CLIP does not classify many mask proposals but only the image ones, resulting in a faster inference. Other mask-based models like GroupViT [49] and ViewCo [40] are grouping pixels into larger segments which are then classified.

Decoder-focused approaches such as DenseCLIP [39] and LSeg [23] encode the image with CLIP and obtain the pixel-level patch embeddings. Because the pre-training is focused on the class embedding, the approaches append additional decoders to refine the patch embeddings. For this refinement, CAT-Seg [7] utilizes multiple stages of cost aggregation to generate the final segmentation mask. PACL [34] aligns patch embeddings and class embeddings during training and, as a result, does not require segmentation-specific training data or additional modules. Zero-shot semantic segmentation models have been combined with other tasks as well. OpenSeeD [55] implements open-vocabulary for object detection and segmentation. SEEM [58] processes text prompts and additional inputs like visual prompts similar to SAM. Apart from differences in the architecture, the models vary in the training process—particularly in fine-tuning CLIP's vision encoder.

## 2.2 Evaluation and benchmarking of zero-shot semantic segmentation

Zero-shot semantic segmentation models are typically evaluated on datasets consisting exclusively of everyday images, such as ADE20K [56], Pascal Context [33], and Pascal VOC [12]. These dataset are the *de facto* standard for evaluating these models (see [7, 14, 25, 50, 51, 57]). Few studies have considered additional datasets. Notably, Zou et al. [57] proposed a Segmentation in the Wild (SegInW) benchmark with 25 datasets. However, the majority of the datasets in SegInW still consist of everyday images with only two exceptions: brain tumor segmentation and a bird's eye view in stables. To the best of our knowledge, zero-shot semantic segmentation and OVSS have not been evaluated on other datasets. Outside of zero-shot semantic segmentation and OVSS, semantic segmentation is usually evaluated based on collections of datasets, like MSeg [22]. These datasets generally only include everyday images, indoor scenes, and driving datasets and lack domain-specific datasets. SAM has been evaluated on 23 instance segmentation datasets in a point-to-mask setting [20]. This collection of datasets is the most extensive for segmentation tasks but still misses domains, such as engineering and earth monitoring. Other works evaluate specifically domain dataset collections such as medical tasks [32] or satellite data [21].

Table 1: Multi-domain benchmark for zero-shot semantic segmentation models consisting of 22 downstream tasks, a total of 448 classes, and 25,079 images.

| Dataset | Domain | Sensor type | Segment size | Number of classes | Class similarity | Vocabulary | Number of images | Task |
|---|---|---|---|---|---|---|---|---|
| BDD100K [53] | | Visible spectrum | Medium | 19 (Medium) | Low | Generic | 1,000 | Driving |
| Dark Zurich [41] | | Visible spectrum | Medium | 20 (Medium) | Low | Generic | 50 | Driving |
| MHP v1 [24] | General | Visible spectrum | Small | 19 (Medium) | High | Task-spec. | 980 | Body parts |
| FoodSeg103 [47] | | Visible spectrum | Medium | 104 (Many) | High | Generic | 2,135 | Ingredients |
| ATLANTIS [11] | | Visible spectrum | Small | 56 (Many) | Low | Generic | 1,295 | Maritime |
| DRAM [8] | | Visible spectrum | Medium | 12 (Medium) | Low | Generic | 718 | Paintings |
| iSAID [46] | | Visible spectrum | Small | 16 (Medium) | Low | Generic | 4,055 | Objects |
| ISPRS Potsdam [4] | Earth | Multispectral | Small | 6 (Few) | Low | Generic | 504 | Land use |
| WorldFloods [31] | Monitoring | Multispectral | Medium | 3 (Binary) | Low | Generic | 160 | Floods |
| FloodNet [38] | | Visible spectrum | Medium | 10 (Few) | Low | Task-spec. | 5,571 | Floods |
| UAVid [29] | | Visible spectrum | Small | 8 (Few) | High | Task-spec. | 840 | Objects |
| Kvasir-Inst. [18] | | Visible spectrum | Medium | 2 (Binary) | Low | Generic | 118 | Endoscopy |
| CHASE DB1 [13] | Medical | Microscopic | Small | 2 (Binary) | Low | Domain-spec. | 20 | Retina scan |
| CryoNuSeg [30] | Sciences | Microscopic | Small | 2 (Binary) | Low | Domain-spec. | 30 | WSI |
| PAXRay-4 [42] | | Electromagnetic | Large | 4x2 (Binary) | Low | Domain-spec. | 180 | X-Ray |
| Corrosion CS [3] | | Visible spectrum | Medium | 4 (Few) | High | Task-spec. | 44 | Corrosion |
| DeepCrack [28] | Engineering | Visible spectrum | Small | 2 (Binary) | Low | Generic | 237 | Cracks |
| ZeroWaste-f [2] | | Visible spectrum | Medium | 5 (Few) | High | Generic | 929 | Conveyor |
| PST900 [43] | | Electromagnetic | Small | 5 (Few) | Low | Generic | 288 | Thermal |
| SUIM [17] | Agriculture | Visible spectrum | Medium | 8 (Few) | Low | Generic | 110 | Underwater |
| CUB-200 [45] | and Biology | Visible spectrum | Medium | 201 (Many) | High | Domain-spec. | 5,794 | Bird species |
| CWFID [15] | | Visible spectrum | Small | 3 (Few) | High | Generic | 21 | Crops |

3

# 3   MESS benchmark

Following the HELM benchmark [26] proposed for the evaluation of large language models, we develop a taxonomy with task characteristics for semantic segmentation and retrieve a set of more than 500 datasets that we review as part of the benchmark creation. For the development of the taxonomy, we use a method proposed by Nickerson et al. [35]. We start the development of the taxonomy by specifying the so-called meta-characteristic of the taxonomy (i.e., our goal): *identify visual and language characteristics of downstream tasks influencing the performance of zero-shot semantic segmentation models*. We then initialize the taxonomy in a conceptual-to-empirical cycle based on a review of other benchmarks and literature. Next, we refine the taxonomy in multiple empirical-to-conceptual iterations. We reviewed semantic segmentation datasets on Papers with Code, Kaggle, and additional test datasets used by recent segmentation models. We repeatedly reduced the dimensions of the taxonomy to the most meaningful ones for the meta-characteristic. We then conducted a statistical analysis of potential taxonomy dimensions to identify and remove similar or overlapping dimensions (see supplementary material). We identified multiple dimensions that highly correlate with each other like color map and sensor type, segment size and segments per image, as well as viewpoint and domain. Based on this analysis, we discarded color map, resolution, segments per image, and viewpoint. The final taxonomy matches all ending conditions [35]. While the proposed taxonomy identifies the most important dimensions and characteristics validated based on 120 classified datasets, there may be additional dimensions that influence the performance of zero-shot semantic segmentation models in specific cases. Overall, we observe that certain characteristics are more likely to co-occur. E.g, binary datasets typically imply a low class similarity, whereas task-specific vocabulary is often associated with a high similarity between the task-specific classes. We account for this imbalance in the distribution of the characteristics and reflect it in our benchmark.

Following the taxonomy development, we selected a representative set of datasets so that the MESS benchmark is informative, reproducible, and manageable. Specifically, we filtered the 120 classified datasets based on four exclusion criteria: each dataset has an official and annotated validation or test set, high annotation quality, moderate disk usage, and sufficient image size. Next, we selected a subset that consists of complementing use cases to avoid duplication and covers all characteristics of the taxonomy. We present the 22 selected datasets and their characteristics within the taxonomy's dimensions in Table 1 . These datasets cover a variety of applications, resulting in a holistic evaluation of domain-specific applications. We publish this new MESS benchmark at `https://blumenstiel.github.io/mess-benchmark` and invite others to suggest additional datasets and refine classes for future versions.

During dataset selection, we have not identified any ethical issues with these datasets based on the information provided by the data sources. Our use follows the terms and conditions set by the data providers, and we list the corresponding licenses in the supplementary material. However, we acknowledge the importance of considering the societal impact of our work. FMs, such as CLIP, are pre-trained on vast corpora of data that may contain biases. We refer to Agarwal et al. [1] for a detailed analysis of biases in CLIP. While the majority of MESS datasets are less prone to such biases, some may include data specific to gender or geographic regions. We believe that assessing model performance across a range of datasets can help to identify and mitigate the impact of biases.

# 4   Experimental setup

In this section, we provide a brief definition of the zero-shot semantic segmentation task, describe the metrics, and outline implementation details.

## 4.1   Task

Let $I$ denote an image with a set of candidate classes $\mathcal{C} = \{C_1, C_2, ..., C_N\}$, where each candidate class $C_i$ is described in natural language. Zero-shot semantic segmentation models then assign a class $C_i$ to each pixel of $I$. The number of candidate classes $N$ can vary during inference (e.g., different downstream tasks) and, additionally, the model may not have seen the candidate classes during training. This is in contrast to traditional semantic segmentation, where the set of classes is fixed during training and inference [7]. Each dataset represents a set of images with the same label set, and in our evaluations, none of the models is trained on the datasets from the benchmark

or the same set of candidate classes. However, it is reasonable to assume the evaluated classes have been present in the pre-training of the underlying vision-language models (like CLIP). All evaluated models have been trained on images with three channels (i.e., RGB). To account for datasets with varying numbers of input channels we mapped them to RGB (i.e., inputs with a single channel are mapped to RGB, for multispectral inputs we selected a subset of three channels).

## 4.2 Implementation

Following common practice, we evaluate all models using the mean of class-wise intersection over union (mIoU) [6, 7, 25, 50, 51]. We split very large images from the earth monitoring datasets into smaller patches of $1024 \times 1024$ pixels. Further, we use an IRRG color map for multispectral datasets (ISPRS Potsdam and WorldFloods) and select the thermal data in PST900. All other datasets include images with one or three channels. Across our implementation, we use PyTorch [36] and Detectron2 [48] for implementing the data loaders. For the convenience of users and contributors to our benchmark, we additionally provide wrappers for torchvision and MMSeg to process datasets in the Detectron2 dataset catalog. We did not train any models but used the publicly available weights and model configurations. The evaluation was conducted on an NVIDIA V100S.

## 4.3 Models

We utilize our MESS benchmark to evaluate a range of recent models for zero-shot semantic segmentation including the state-of-the-art, selecting models based on the reported performance and the availability of official code and weights. OVSeg [25], SAN [50], and CAT-Seg [7] represent the state-of-the-art across different approaches in the architecture for zero-shot semantic segmentation (i.e., two-stage mask-based, one-stage mask-based, and pixel-based). We additionally consider ZSSeg [51] and ZegFormer [9] which are frequently consulted as baseline models e.g. by [7, 25, 50]. The previously listed models are based on CLIP and use COCO Stuff to train the additional segmentation modules. Additionally, OVSeg uses COCO Captions for fine-tuning. X-Decoder [57] and OpenSeeD [55] are part of our evaluation since these approaches do not make use of CLIP but are based on UniCL [52] (i.e., their public versions). X-Decoder and OpenSeeD are trained on multiple datasets which we detail in the supplementary material.

To account for recent developments in the field, we additionally include SAM [20] in our evaluations. Standard SAM can only process visual prompts and does not facilitate text-to-mask settings. Therefore, we validated other ways to make use of SAM. We implement Grounded-SAM [16] using the predicted bounding boxes from Grounding DINO [27] as input for SAM and thereby enabling an open-vocabulary setting (i.e., text-to-mask). This serves as a baseline to better understand the potential of SAM-based text-to-mask models. The overall evaluation time per model on the MESS benchmark varies in our experiments between 1 hour for SAN-B and 14.5 hours for OVSeg-L.

# 5 Experiments

In the following, we provide a holistic overview of the performance of multiple zero-shot semantic segmentation models based on our MESS benchmark. We conduct a range of in-detail analyses of model performances across the dimensions of our taxonomy including sensor types, the class similarity, and the vocabulary—additional experiments are included in the supplementary material.

## 5.1 Multi-domain zero-shot semantic segmentation

We provide a quantitative comparison across models and all datasets summarized by their domain in Table 2 and per dataset results in Fig. 2 and 3. We add a random prediction as a lower bound by calculating the expected mIoU value with uniformly distributed predictions over all classes. In addition, we report fully supervised results based on the current SOTA from supervised semantic segmentation (see supplementary material). Overall, CAT-Seg-L achieves a strong performance across domains with an average mIoU of 38.14%, followed by its base and huge version. CAT-Seg is followed by SAN-L with a performance of 30.06%. Notably, the performance of zero-shot CAT-Seg-L in the general domain is only 8.69pp (average mIoU) below the performance of supervised SOTA approaches. In comparison, CAT-Seg-L reaches on average 50.36% of the supervised performance in earth monitoring and 54.18% on medical sciences. The performance gap compared to supervised

Table 2: mIoU results averaged by the dataset domain. Best-performing models are highlighted in bold, and the second-best are underlined. Random represents the randomly expected mIoU with uniformly distributed predictions. The best supervised models are separately selected for each dataset (see supplementary material for the supervised models and results).

| Model | Parameters | Inference (s/iter) | General | Earth Monit. | Medical Sciences | Engineer. | Agri. and Biology | Mean |
|---|---|---|---|---|---|---|---|---|
| *Random (LB)* | | | *1.17* | *7.11* | *29.51* | *11.71* | *6.14* | *10.27* |
| *Best supervised (UB)* | | | *48.62* | *79.12* | *89.49* | *67.66* | *81.94* | *70.99* |
| ZSSeg-B [51] | 211M | 0.49 | 19.98 | 17.98 | <u>41.82</u> | 14.0 | 22.32 | 22.73 |
| ZegFormer-B [9] | 210M | 0.18 | 13.57 | 17.25 | 17.47 | 17.92 | <u>25.78</u> | 17.57 |
| X-Decoder-T [57] | 164M | 0.1 | 22.01 | 18.92 | 23.28 | 15.31 | 18.17 | 19.8 |
| SAN-B [50] | 158M | 0.04 | <u>29.35</u> | <u>30.64</u> | 29.85 | <u>23.58</u> | 15.07 | <u>26.74</u> |
| OpenSeeD-T [55] | 116M | 0.08 | 22.49 | 25.11 | **44.44** | 16.5 | 10.35 | 24.33 |
| CAT-Seg-B [7] | 181M | 0.17 | **34.96** | **34.57** | 41.65 | **26.26** | **29.32** | **33.74** |
| OVSeg-L [25] | 531M | 1.64 | 29.54 | 29.04 | 31.9 | 14.16 | 28.64 | 26.94 |
| SAN-L [50] | 437M | 0.14 | 36.18 | <u>38.83</u> | 30.27 | 16.95 | 20.41 | 30.06 |
| CAT-Seg-L [7] | 490M | 0.33 | **39.93** | **39.85** | **48.49** | <u>26.04</u> | <u>34.06</u> | **38.14** |
| CAT-Seg-H [7] | 1049M | 0.5 | <u>37.98</u> | 37.74 | <u>34.65</u> | **29.04** | **37.76** | <u>35.66</u> |

models is even larger for the two other domains. Looking at the dataset-specific performance in Fig. 2 and 3, we observe that the performance varies between datasets and models. While SAN-L is the best-performing model on CUB-200 and DRAM, it has significantly lower performance on CWFID or CHASE DB1 compared to CAT-Seg-L. The model achieves scores between 50% and over 100% of the performance of supervised state-of-the-art in the general domain. Within the other domains, CAT-Seg-L has a performance gap of more than 25pp for most of the datasets.

The inference time varies between the models and, in particular, between different model architectures with some models requiring more than ten times higher computational effort indicated by higher inference times. In general, we observe the highest inference times for two-stage mask-based approaches, such as ZSSeg and OVSeg, which are between five to twelve times higher than other mask-based approaches (X-Decoder, OpenSeeD, and SAN). The point-based CAT-Seg uses a sliding window approach which requires five passes and therefore results in higher inference times than SAN. Overall, SAN represents the fastest model in our experiments.
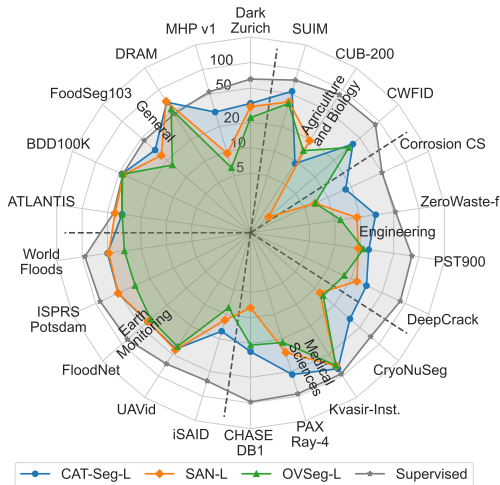


Figure 2: mIoU results for large models on a log scale. The datasets are grouped by their domain and sorted by supervised performance.
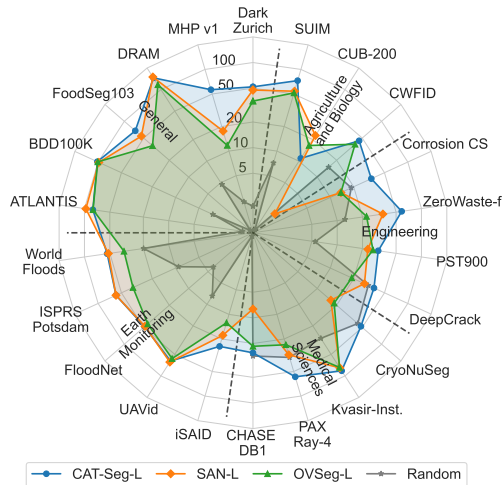


Figure 3: mIoU results of large models in relative comparison to the supervised mIoU on a log scale. 100 is equal to the supervised mIoU.

## 5.2 Sensor type evaluation

All considered models have been developed for the visual spectrum (i.e., RGB). In the following, we investigate the performance of three different sensor types: multispectral, electromagnetic, and microscopic. Three datasets from MESS allow for a direct comparison between different sensor types. For multispectral sensors, the MESS benchmark includes the IRRG color map for ISPRS Potsdam and WorldFloods. The models are able to process the different color maps and profit from the visual highlighting of vegetation through the infrared channel. This insight might be limited to commonly used color maps because other color maps might be less represented in the pre-training data of CLIP. On electromagnetic and thermal imagery, none of the evaluated models can regularly segment objects on the PST900 dataset. We compared this result to the aligned RGB images from PST900. All models perform significantly better on the RGB images. E.g., CAT-Seg-L reaches a mIoU of 65.55% on RGB images compared to only 25.26% for thermal data. We also tested a pseudo color map that maps the grayscale thermal data to a pseudo color scale, resulting in a similar low performance. Therefore, we conclude that zero-shot semantic segmentation models are currently not able to sufficiently segment objects in thermal images. Most models are also not able to correctly segment X-ray images in the PAXRay dataset, the second benchmark dataset with an electromagnetic sensor type. However, X-rays do include much more visual features compared to thermal images and CAT-Seg is able to segment some anatomical structures like the lungs. Further, the benchmark includes retina scans in CHASE DB1 and WSI images in CrypNuSeg to evaluate microscopic imagery. Similar to the PAXRay results, most models fail to segment the structures. But CAT-Seg and ZSSeg can locate the requested class. Thus, we assume that CLIP and zero-shot semantic segmentation can understand microscopic concepts but the correct segmentation is not achieved because of the small segments instead of the image type.

## 5.3 Multi-domain vs. in-domain evaluation

Most zero-shot semantic segmentation models are currently evaluated on five datasets: Pascal VOC, ADE20K-150, ADE20K-847, Pascal Context-59, and Pascal Context-459. Figure 4 compares the average results of the evaluated models on these common datasets (i.e., in-domain datasets) to a multi-domain setting with datasets of MESS benchmark. Note that the multi-domain datasets contain fewer classes on average, resulting in a much higher random mIoU. We provide the results for each dataset in the supplementary material. While SAN-L has comparable performance to the CAT-Seg models on common datasets, it has a significantly lower mIoU on domain datasets. Further, X-Decoder has a generally lower mIoU on domain datasets compared to other models. X-Decoder does not use CLIP which may explain the limited generalizability of the model. Overall, CAT-Seg is the only model architecture with a higher average mIoU on the domain datasets than common datasets.

## 5.4 Language characteristics

The differentiation between related classes is relevant in domain-specific use cases like biology. We analyze the influence of class similarity on class-wise IoU in Figure 5. Following Xu et al. [50], we calculated the class similarity as the maximum cosine similarity of the embedding to all other CLIP text embeddings in the label set. Overall, the class IoU does not correlate with the similarity. However, none of the classes with high similarity reaches a desirable IoU (e.g., the Corrosion CS dataset with three classes describing different corrosion stages). All models face difficulties in differentiating these classes. In additional experiments, the model performance significantly improved when considering similar classes as a single class. Also, specialized terms affect the model performance, specifically,

Table 3: Comparision of mIoU results for images with different sensor types. Pseudo refers to thermal data mapped to a pseudo color map.

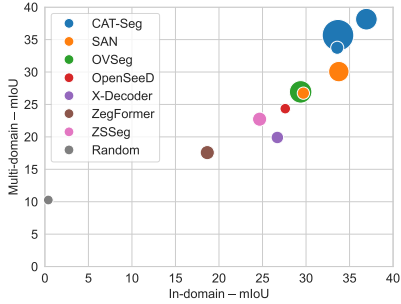| Model | ISPRS Potsdam | | WorldFloods | | PST900 | | |
| | IRRG | RGB | IRRG | RGB | Thermal | Pseudo | RGB |
|---|---|---|---|---|---|---|---|
| OVSeg-L [25] | 31.03 | 35.46 | 31.48 | 22.86 | _21.89_ | _21.63_ | 42.9 |
| SAN-L [50] | **51.45** | **52.06** | _48.24_ | **45.93** | 19.01 | 19.41 | _49.02_ |
| CAT-Seg-L [7] | _51.42_ | _51.29_ | **49.86** | _45.39_ | **25.26** | **25.43** | **65.55** |

7

Figure 4: Cross domain settings can be challenging: Average mIoU of commonly used evaluation datasets in comparison to the results on the MESS benchmark. The size represents the parameter count of the models.
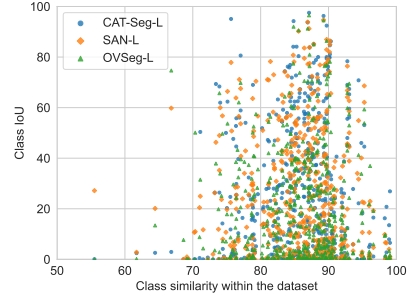


Figure 5: The class-wise IoU in comparison with the similarity to other labels within the dataset. The similarity is measured by the minimum cosine distance of the class label to all other CLIP text embeddings within the dataset.

domain-specific and task-specific labels. Our evaluation covers domain-specific words from medicine and biology, i.e., bird species and anatomical structures like the *mediastinum*. It shows that CLIP is able to understand domain-specific concepts to a limited extent. We observed higher performance for generic terminology. E.g., all models achieve higher performances on the Kvasir-Instrument dataset when using a generic vocabulary like *tool*. Utilizing a more precise term like *surgical instrument* reduces the mIoU. We refer to classes with specified conditions as task-specific classes. In our evaluations, CAT-Seg achieves the best results on task-specific classes. However, CAT-Seg still confuses classes and, e.g., predicts the right shoe and right leg significantly more often than the left side in MHP v1. CAT-Seg models are further biased towards the *parked car* class in UAVid images, while SAN and OVSeg mostly assign masks to the label *moving car*. Overall, domain-specific and task-specific vocabulary limits the performance of zero-shot semantic segmentation models.

Table 4: Domain-averaged mIoU results for Grounded-SAM and SAM with oracle inputs in a point-to-mask and box-to-mask setting. Random, supervised and CAT-Seg-L are provided for reference.

| Model | Input prompt | General | Earth Monitoring | Medical Sciences | Engineering | Agri. and Biology | Mean |
|---|---|---|---|---|---|---|---|
| *Random (LB)* | | *1.17* | *7.11* | *29.51* | *11.71* | *6.14* | *10.27* |
| *Best supervised (UB)* | | *48.62* | *79.12* | *89.49* | *67.66* | *81.94* | *70.99* |
| CAT-Seg-L [7] | | 39.93 | 39.85 | 48.49 | 26.04 | 34.06 | 38.14 |
| Gr.-SAM-B [16] | | 29.51 | 25.97 | 37.38 | **29.51** | 17.66 | 28.52 |
| Gr.-SAM-L [16] | Grounding DINO [27] | **30.32** | **26.44** | **38.69** | <u>29.25</u> | **17.73** | **29.05** |
| Gr.-SAM-H [16] | | <u>30.27</u> | **26.44** | <u>38.45</u> | 28.16 | <u>17.67</u> | <u>28.78</u> |
| SAM-B [20] | | **50.41** | <u>38.72</u> | 43.7 | 45.16 | <u>57.84</u> | <u>46.59</u> |
| SAM-L [20] | Oracle points [44] | <u>45.99</u> | **44.03** | <u>55.74</u> | **50.0** | **58.23** | **49.99** |
| SAM-H [20] | | 36.05 | 34.82 | **59.58** | <u>47.35</u> | 39.91 | 43.0 |
| SAM-B [20] | Oracle bounding boxes | **78.5** | **73.56** | **68.14** | **73.29** | <u>86.0</u> | **75.67** |
| SAM-L [20] | | <u>78.0</u> | <u>73.27</u> | 64.98 | <u>73.09</u> | **86.99** | <u>74.97</u> |
| SAM-H [20] | | 65.23 | 59.61 | <u>66.58</u> | 66.4 | 78.63 | 66.55 |

## 5.5 Comparison to SAM

For a better understanding of current text-to-mask zero-shot semantic segmentation approaches, we compare them with grounded and oracle versions of SAM. SAM cannot directly process textual inputs, instead, it uses visual prompt inputs, i.e., bounding boxes or points. For the comparison, we implemented three versions of SAM. First, we made use of existing available demos combining Grounding DINO and SAM and extended them by a comprehensive quantitative evaluation. Second, oracle point-to-mask SAM refers to a model that provides a single point for every connected segment in the ground truth mask to simulate the visual input. We use the point sampling approach from

RITM [44]. Third, oracle box-to-mask SAM utilizes a single box for every segment in the ground truth mask to simulate the visual input. We consider up to 100 input prompts per image to avoid a large number of very small segments. We later combine all predicted masks by taking the maximum logit value for each pixel. Pixels with only negative logit values are assigned to the background class or marked *unlabeled* in datasets without a background class. Note that inputting text data as in the models before is fundamentally different from utilizing visual inputs as in our two oracle SAM implementations and our analyses are not intended for a direct comparison but to better understand the potentials of SAM for zero-shot text-to-mask models.

In Table 4, we observe that the non-oracle implementation of SAM utilizing Grounding DINO generally exhibits limited performance compared to CAT-Seg text-to-mask models. Oracle versions of SAM receive significantly improved information on the location of the object and, therefore, show a strong performance. Given the perfect information on the location of objects in the image with oracle bounding boxes, the oracle box-to-mask SAM implementation even outperforms supervised semantic segmentation models. Overall, we observe that SAM models achieve a strong performance based on oracle information on the location of the objects. However text-to-mask zero-shot semantic segmentation models like CAT-Seg outperform the combination of Grounding DINO and SAM. Similar to X-Decoder and OpenSeeD, Grounding DINO does not use CLIP, which results in limited multi-domain performance. The results with oracle bounding boxes suggest that future combinations of SAM with open-vocabulary object detection models based on FMs like CLIP may outperform the current state-of-the-art in zero-shot semantic segmentation.

## 5.6 Qualitative analyses

In the following, we quantitatively compare the predictions of the three promising text-to-mask zero-shot semantic segmentation models with the ground truth and the grounding version of SAM on four different datasets (autonomous driving, satellite imagery, medical science, and engineering). We visually observe the following characteristics: First, CAT-Seg also visually surpasses the predictions of the other models. Second, across different domains, the predictions of CAT-Seg are largely in line with the ground truth and the segmentation is comparatively fine-grained. Third, we observe that Grounding DINO does not locate most segments and, therefore, Grounded-SAM tends to predict the background class. These qualitative observations are largely in line with our quantitative experiments.



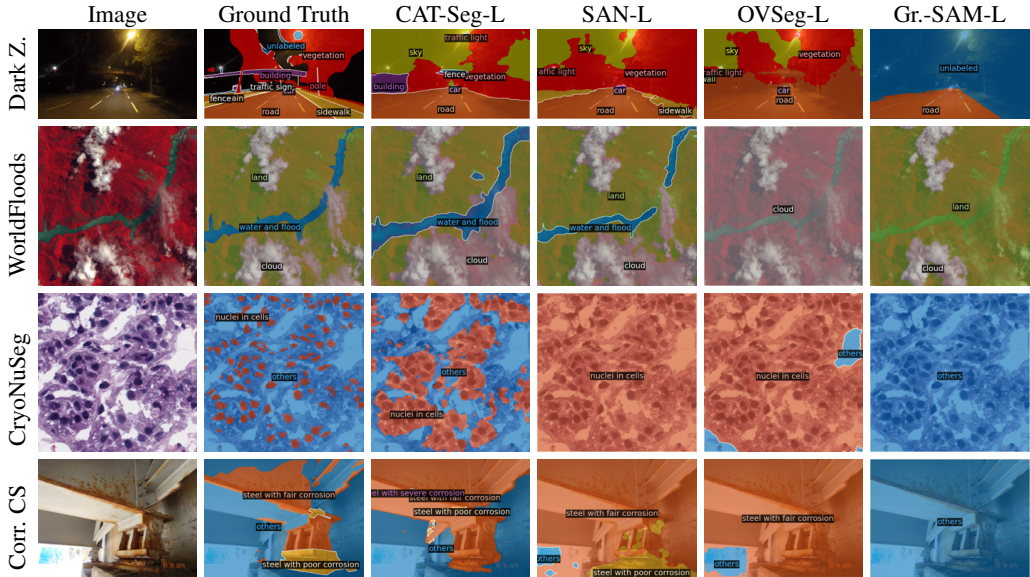Figure 6: Predictions from selected datasets based on CAT-Seg-L [7], SAN-L [50], OVSeg-L [25], and Grounded-SAM [16].

Zero-shot semantic segmentation achieves a remarkable performance on in-domain datasets [7, 50]. Based on the MESS benchmark, we observe that these models can solve some tasks from other domains, however, are limited in their applicability to domains like medical science, engineering,

9

and agriculture. We identified a range of challenges: First, we observe that domain-specific and task-specific vocabulary are difficult to handle. Models tend to be confused by labels with a high class similarity as in Corrosion CS. Therefore, we recommend to utilize a generic vocabulary with common class names, which led to improved performances in our experiments (e.g., *tool* instead of *medical instrument* in Kvasir-Instrument). Second, differences in the type of the sensor influence the performance of these models which are generally trained on the visual spectrum—for example, thermal data is hard to process. Third, we observe that state-of-the-art text-to-mask approaches outperform Grounded-SAM across multiple domains.

## 6   Conclusion

Zero-shot semantic segmentation has the potential to make segmentation models more accurate, cheap, flexible, and interactive. However, the current evaluation is limited to in-domain datasets, and previous analyses focused on model properties rather than task characteristics. With the MESS benchmark, we enable a holistic evaluation and invite others to utilize this benchmark to accelerate the field of semantic segmentation across domains to improve its real-world applicability.

## 7   Acknowledgements

## References

[1]  Agarwal, S., Krueger, G., Clark, J., Radford, A., Kim, J. W., and Brundage, M. (2021). Evaluating clip: Towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*.

[2]  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.

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

[4]  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.

[5]  Bucher, M., Vu, T.-H., Cord, M., and Pérez, P. (2019). Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32.

[6]  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.

[7]  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*.

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

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

[10] Ding, Z., Wang, J., and Tu, Z. (2022b). Open-vocabulary panoptic segmentation with maskclip. *arXiv preprint arXiv:2208.08984*.

[11] 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.

[12] 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.

[13] 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.

[14] Ghiasi, G., Gu, X., Cui, Y., and Lin, T.-Y. (2022). Scaling open-vocabulary image segmentation with image-level labels. pages 540–557. Springer.

[15] 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.

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

[17] 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.

[18] 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.

[19] Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

[20] 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*.

[21] Lacoste, A., Sherwin, E. D., Kerner, H., Alemohammad, H., Lütjens, B., Irvin, J., Dao, D., Chang, A., Gunturkun, M., Drouin, A., et al. (2021). Toward foundation models for earth monitoring: Proposal for a climate change benchmark. *arXiv preprint arXiv:2112.00570*.

[22] Lambert, J., Liu, Z., Sener, O., Hays, J., and Koltun, V. (2020). MSeg: A composite dataset for multi-domain semantic segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2879–2888.

[23] Li, B., Weinberger, K. Q., Belongie, S., Koltun, V., and Ranftl, R. (2022). Language-driven semantic segmentation. In *International Conference on Learning Representations*.

[24] 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*.

[25] 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.

[26] Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. (2022). Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

[27] 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*.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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. *arXiv preprint arXiv:2304.10517v1*.

[33] 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.

[34] Mukhoti, J., Lin, T.-Y., Poursaeed, O., Wang, R., Shah, A., Torr, P. H., and Lim, S.-N. (2022). Open vocabulary semantic segmentation with patch aligned contrastive learning. *arXiv preprint arXiv:2212.04994*.

[35] 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*.

[36] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

[37] 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*.

[38] 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.

[39] Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., and Lu, J. (2022). DenseCLIP: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091.

[40] Ren, P., Li, C., Xu, H., Zhu, Y., Wang, G., Liu, J., Chang, X., and Liang, X. (2023). Viewco: Discovering text-supervised segmentation masks via multi-view semantic consistency. *arXiv preprint arXiv:2302.10307*.

[41] 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.

[42] 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 (BMVC)*.

[43] 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.

[44] Sofiiuk, K., Petrov, I. A., and Konushin, A. (2022). Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3141–3145. IEEE.

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

[46] 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.

[47] 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.

[48] Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., and Girshick, R. (2019). Detectron2. https://github.com/facebookresearch/detectron2.

[49] Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., and Wang, X. (2022). Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144.

[50] 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.

[51] Xu, M., Zhang, Z., Wei, F., Lin, Y., Cao, Y., Hu, H., and Bai, X. (2021). A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. *arXiv preprint arXiv:2112.14757*.

[52] Yang, J., Li, C., Zhang, P., Xiao, B., Liu, C., Yuan, L., and Gao, J. (2022). Unified contrastive learning in image-text-label space. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19163–19173.

[53] 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.

[54] Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., Liu, C., Liu, M., Liu, Z., Lu, Y., Shi, Y., Wang, L., Wang, J., Xiao, B., Xiao, Z., Yang, J., Zeng, M., Zhou, L., and Zhang, P. (2021). Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432v1*.

[55] Zhang, H., Li, F., Zou, X., Liu, S., Li, C., Gao, J., Yang, J., and Zhang, L. (2023). A simple framework for open-vocabulary segmentation and detection. *arXiv preprint arXiv:2303.08131*.

[56] 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.

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

[58] Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Gao, J., and Lee, Y. J. (2023b). Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*.

# 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

|  | Class similarity | Color map | Domain | Number of classes | Resolution | Segment size | Segments per image | Sensor type | Viewpoint | Vocabulary | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Class similarity | 1.00 | 0.10 | 0.12 | 0.36 | 0.03 | 0.00 | 0.17 | 0.07 | 0.14 | 0.44 | 0.24 |
| Color map | 0.03 | 1.00 | 0.35 | 0.14 | 0.17 | 0.06 | 0.13 | 0.58 | 0.31 | 0.13 | 0.29 |
| Domain | 0.06 | 0.34 | 1.00 | 0.28 | 0.36 | 0.12 | 0.17 | 0.44 | 0.47 | 0.28 | 0.35 |
| Number of classes | 0.36 | 0.29 | 0.64 | 1.00 | 0.04 | 0.07 | 0.15 | 0.36 | 0.50 | 0.09 | 0.35 |
| Resolution | 0.03 | 0.48 | 0.88 | 0.04 | 1.00 | 0.10 | 0.18 | 0.63 | 0.82 | 0.33 | 0.45 |
| Segment size | 0.00 | 0.15 | 0.21 | 0.07 | 0.10 | 1.00 | 0.62 | 0.17 | 0.21 | 0.10 | 0.26 |
| Segments per image | 0.17 | 0.22 | 0.37 | 0.15 | 0.18 | 0.62 | 1.00 | 0.18 | 0.22 | 0.23 | 0.33 |
| Sensor type | 0.04 | 0.65 | 0.49 | 0.23 | 0.31 | 0.11 | 0.11 | 1.00 | 0.44 | 0.25 | 0.36 |
| Viewpoint | 0.08 | 0.33 | 0.59 | 0.29 | 0.41 | 0.12 | 0.12 | 0.44 | 1.00 | 0.21 | 0.36 |
| Vocabulary | 0.34 | 0.17 | 0.41 | 0.07 | 0.23 | 0.08 | 0.13 | 0.35 | 0.27 | 1.00 | 0.30 |

(a) Interim dimensions

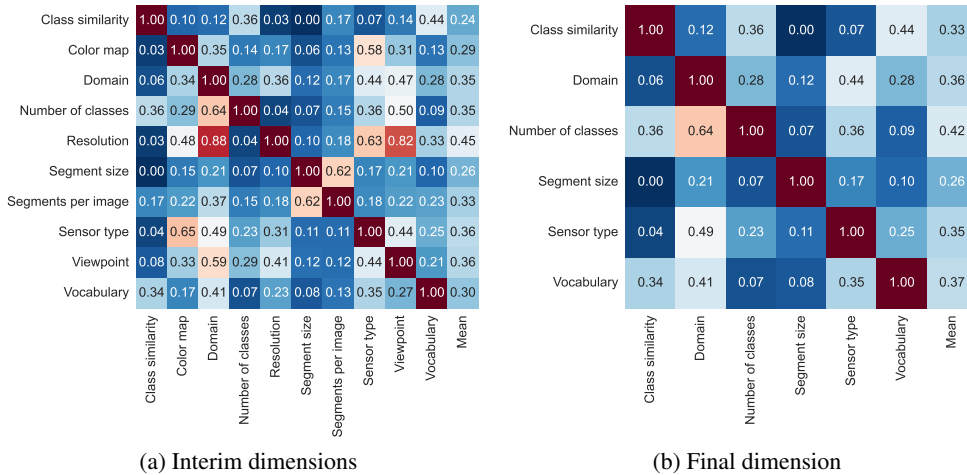|  | Class similarity | Domain | Number of classes | Segment size | Sensor type | Vocabulary | Mean |
|---|---|---|---|---|---|---|---|
| Class similarity | 1.00 | 0.12 | 0.36 | 0.00 | 0.07 | 0.44 | 0.33 |
| Domain | 0.06 | 1.00 | 0.28 | 0.12 | 0.44 | 0.28 | 0.36 |
| Number of classes | 0.36 | 0.64 | 1.00 | 0.07 | 0.36 | 0.09 | 0.42 |
| Segment size | 0.00 | 0.21 | 0.07 | 1.00 | 0.17 | 0.10 | 0.26 |
| Sensor type | 0.04 | 0.49 | 0.23 | 0.11 | 1.00 | 0.25 | 0.35 |
| Vocabulary | 0.34 | 0.41 | 0.07 | 0.08 | 0.35 | 1.00 | 0.37 |

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

| Dataset | Task | Domain | Sensor type | Segment size | Number of classes | Class similarity | Vocabulary |
|---|---|---|---|---|---|---|---|
| GLaS [113] | WSI | | 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 | Medical | Microscopic | Small | 3 (Few) | High | Domain-spec. |
| Neuronal Cells [53] | Brain cells | Sciences | 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 | | 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 | Engineering | 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 | | 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 | Agriculture | Visible spectrum | Medium | 2 (Medium) | Low | Domain-spec. |
| COD [38] | Animals | and Biology | 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.
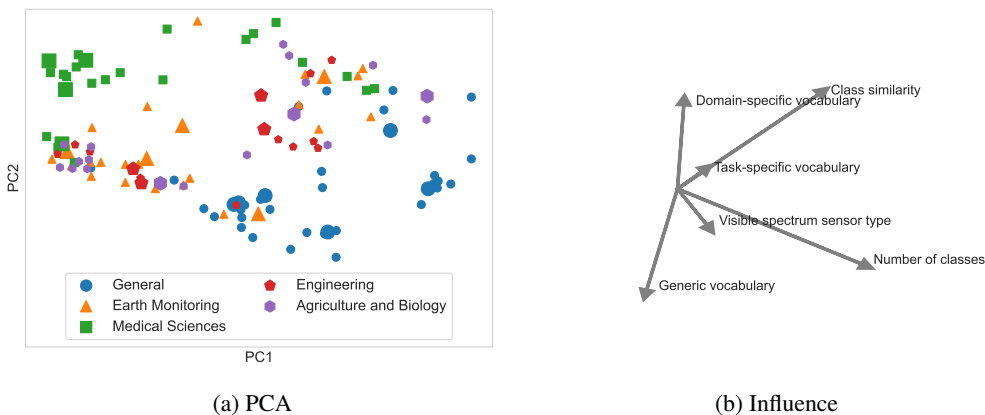


(a) PCA

(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.a., 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).

|  | BDD100K | Dark Zurich | MHP v1 | FoodSeg103 | ATLANTIS | DRAM | iSAID | ISPRS Potsdam | WorldFloods | FloodNet | UAVid | Kvasir-Instrument | CHASE DB1 | CryoNuSeg | PAXRay-4 | Corrosion CS | DeepCrack | PST900 | ZeroWaste-f | SUIM | CUB-200 | CWFID |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Minimum train similarity | 83 | 83 | 79 | 65 | 76 | 90 | 79 | 95 | 90 | 84 | 95 | 89 | 77 | 76 | 77 | 84 | 84 | 83 | 87 | 75 | 34 | 82 |
| Mean train similarity | 96 | 96 | 87 | 85 | 91 | 99 | 90 | 99 | 93 | 92 | 97 | 92 | 86 | 85 | 89 | 88 | 88 | 91 | 96 | 80 | 66 | 91 |
| Minimum intra similarity | 82 | 82 | 76 | 69 | 74 | 76 | 76 | 84 | 83 | 84 | 84 | 89 | 77 | 76 | 76 | 75 | 78 | 75 | 84 | 73 | 56 | 79 |
| Mean intra similarity | 88 | 88 | 89 | 86 | 87 | 87 | 87 | 87 | 84 | 89 | 89 | 89 | 77 | 76 | 82 | 92 | 78 | 82 | 89 | 80 | 84 | 83 |
| Maximum intra similarity | 93 | 93 | 99 | 96 | 95 | 90 | 91 | 88 | 85 | 95 | 92 | 89 | 77 | 76 | 88 | 98 | 78 | 87 | 94 | 85 | 97 | 85 |

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 $CoI \subseteq C$, and $IoU_i$ being the intersection over union for class $i$, we calculate the metric $CoI\text{-}mIoU = \frac{\sum_{i \in CoI} IoU_i}{|CoI|}$. In binary segmentation tasks, this is similar to the $IoU_{pos}$ of the positive class [86].

Figure 4 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



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.

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.
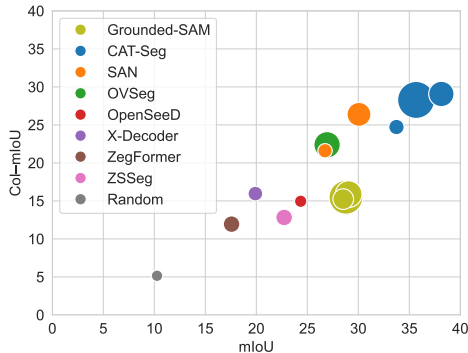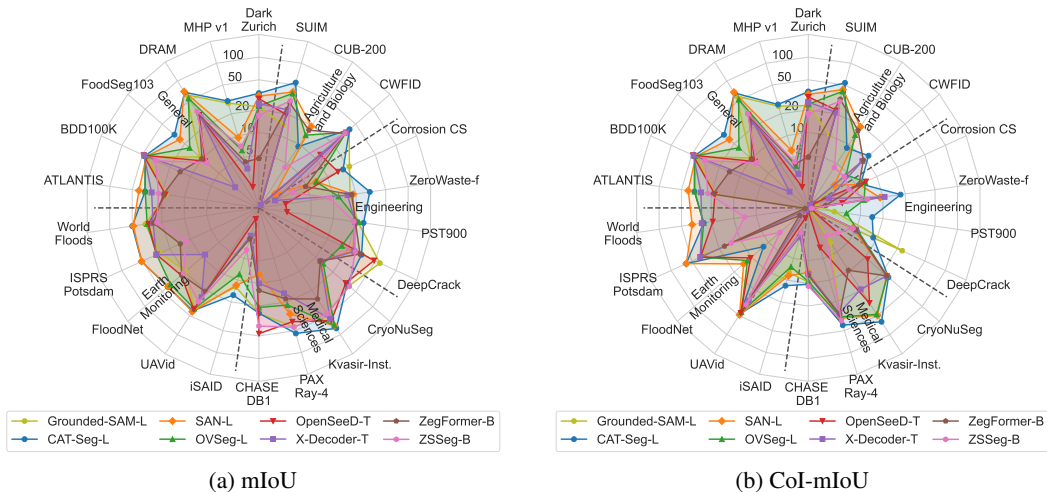
(a) mIoU            (b) CoI-mIoU

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].
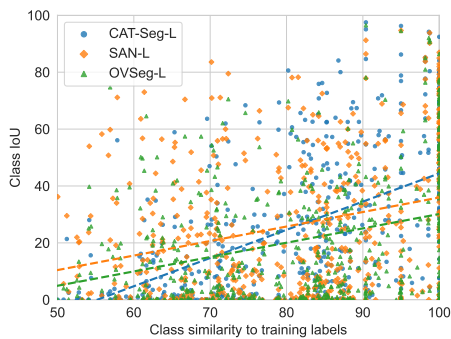


Figure 6: Class IoU in comparison to the class similarity with the labels in COCO-Stuff, represented by the maximum cosine similarity.
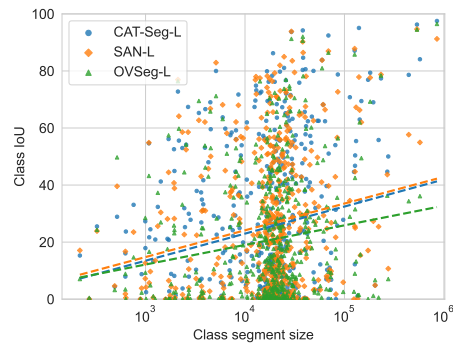


Figure 7: Class IoU in comparison to the class segment size on a log scale. The segment size is the class-wise average pixel count of a segment.

## 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

9

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 | 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.

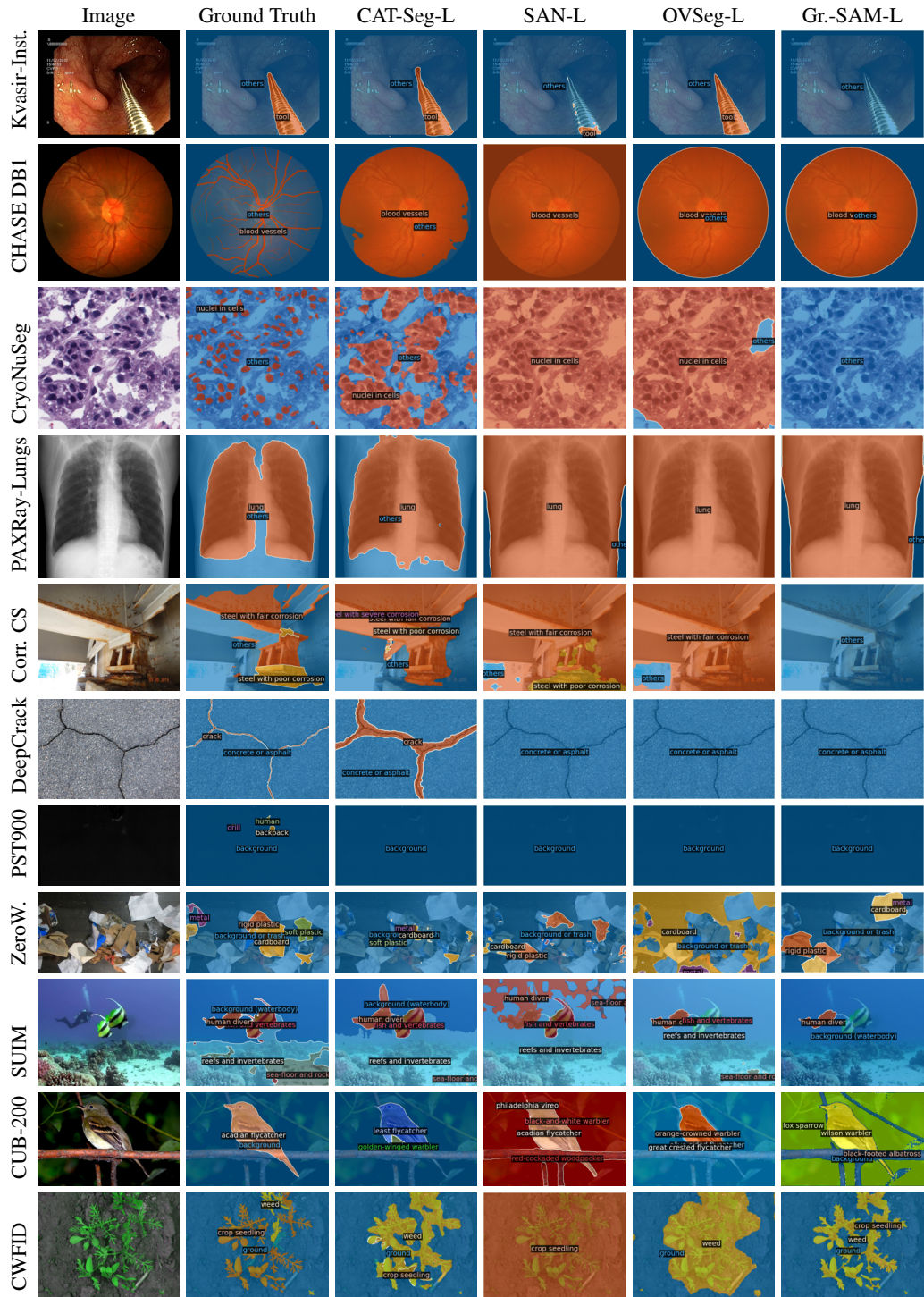Figure 8: Predictions for datasets of the domains general and earth monitoring.

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. 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., Kühl, 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 endo-scopic 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 endo-scopic 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. *arXiv preprint arXiv:2304.10517v1*.

[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 Wünsche, 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 (BMVC)*.

[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. (2021). A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. *arXiv preprint arXiv:2112.14757*.

[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., Gao, J., Yang, J., and Zhang, L. (2023). A simple framework for open-vocabulary segmentation and detection. *arXiv preprint arXiv:2303.08131*.

[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.