Segment anything, from space?

Simiao Ren[1]   Francesco Luzi*[1]   Saad Lahrichi*[2]   Kaleb Kassaw*[1]

Leslie M. Collins[1]   Kyle Bradbury[1,3]   Jordan M. Malof[4]

[1] Electrical and Computer Engineering, Duke University
[2] Division of Natural and Applied Sciences, Duke Kunshan University
[3] Nicholas Institute for Energy, Environment & Sustainability, Duke University
[4] Computer Science, University of Montana

{simiao.ren, francesco.luzi, saad.lahrichi, kaleb.kassaw}@duke.edu,
{leslie.collins, kyle.bradbury}@duke.edu, jordan.malof@umontana.edu

Abstract

Recently, the first foundation model developed specifically for image segmentation tasks was developed, termed the "Segment Anything Model" (SAM). SAM can segment objects in input imagery based on cheap input prompts, such as one (or more) points, a bounding box, or a mask. The authors examined the zero-shot image segmentation accuracy of SAM on a large number of vision benchmark tasks and found that SAM usually achieved recognition accuracy similar to, or sometimes exceeding, vision models that had been trained on the target tasks. The impressive generalization of SAM for segmentation has major implications for vision researchers working on natural imagery. In this work, we examine whether SAM's performance extends to overhead imagery problems and help guide the community's response to its development. We examine SAM's performance on a set of diverse and widely studied benchmark tasks. We find that SAM does often generalize well to overhead imagery, although it fails in some cases due to the unique characteristics of overhead imagery and its common target objects. We report on these unique systematic failure cases for remote sensing imagery that may comprise useful future research for the community.

1. Introduction

Foundation models are large deep learning models (e.g., in terms of free parameters) that have been trained on massive datasets, giving them the ability to generalize well to novel down-stream tasks (e.g., novel datasets, or prediction targets) with little or no additional training (i.e., so-called few-shot or zero-shot generalization). Recent foundation models have been focused on natural language process-
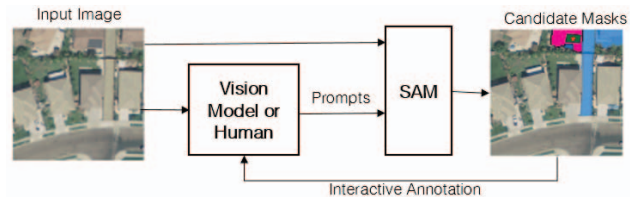

Figure 1. Illustration of using SAM for segmenting solar arrays in satellite imagery.

ing (NLP) tasks (e.g., BERT [13], GPT-3 [7]), while some work also focused on text and imagery (e.g., CLIP [43] and ALIGN [20]). Recently the first foundation model primarily designed for Segmentation tasks was developed, termed the "Segment Anything Model" (SAM) [25]. SAM is designed to output a segmentation mask for a given input image based upon one (or several) of the following input prompts: one (or more) points, a bounding box, or a segmentation mask (e.g., one that is coarse).

The goal of SAM is to segment *any* object in *any* image only based upon these cheap input prompts, and without the need for additional task-specific or dataset-specific adaptation (e.g., training) [25]. SAM's flexible input prompting [25]) and zero-shot segmentation scheme make it a highly flexible model that could be employed as a key component in a variety of vision systems. To illustrate SAM's flexibility and potential impact, the authors in [25] demonstrated its effectiveness on several potential application scenarios. Two important scenarios - which we aim to replicate in overhead imagery in this work - were *interactive annotation* and so-called *model composition*, which are broadly illustrated in Fig. 1 (see Sec. 4 and Sec. 5 for more detail).

Interactive Annotation. In this scenario SAM is em-

ployed to reduce the cost of manual (human) annotation of target object instances. Rather than drawing a mask by hand, the annotator provides cheap input prompts such as bounding boxes or point prompts to SAM. The annotator can optionally provide additional prompts as feedback to SAM, to iteratively refine its segmentation, or *mask*, predictions. SAM was shown to generally outperform existing interactive annotation methods - often by a large margin - on natural imagery tasks and to be capable of producing highly accurate annotations with relatively little guidance [25].

Model Composition. In this scenario SAM is employed to predict a segmentation mask based on prompts from another vision model: e.g., a bounding box or another (potentially low-quality) mask. For example, given limited data, a detection model could be trained to produce a (potentially noisy) bounding box, which could then be used to prompt SAM. In this scenario, and on natural imagery, SAM often achieved segmentation accuracy that was comparable to existing models that had been trained specifically for the target task (i.e., with full target label supervision) [25].

1.1. Contributions of this Work

The authors of SAM reported comprehensive empirical evidence demonstrating the effectiveness of SAM for natural imagery, however, they only investigated a single overhead imagery dataset. Vision problems for overhead imagery exhibit unique properties and challenges compared to natural imagery. For example, the types of target objects commonly sought in overhead imagery (e.g., buildings, roads, land-cover) are usually much larger-scale than those often sought in natural imagery (e.g., people, road signs, vegetation), and collected at a very different spatial resolution and perspective; these may lead to distinct statistical properties between the two tasks. Another example is that target objects can usually be encountered at any orientation in overhead imagery, so that models may benefit by being rotationally equivariant [29, 34, 35]. Due to these distinctions between overhead and natural imagery - among others - it is unclear whether, or to what extent, the effectiveness of SAM transfers to tasks involving overhead imagery. In other words, we ask whether SAM can segment anything *from space*. For example, Fig. 2 presents some out-of-the-box results with SAM on varying tasks, illustrating a variety of performance levels. To answer this question, we replicate the experiments from [25] that correspond to two important potential applications in overhead imagery: interactive annotation and model composition.

We adapt the aforementioned experiments of [25] to seven existing public benchmark datasets of overhead imagery (see Table 1), encompassing 5 million square kilometers of surface area. The benchmarks include a variety of widely-studied object classes (e.g., buildings, roads, cloud cover, farming crops), image resolutions (e.g., 0.3m - 30m),

and geographic locations (e.g., including Africa, Europe, Asia, and North America). To our knowledge, this is the first systematic and comprehensive investigation of SAM on overhead imagery, including experiments relevant to both model composition and interactive annotation, thereby providing valuable guidance to the research community. From our experiments, we also identify SAM failure cases that are unique to overhead imagery and suggest potential future areas of research for the community.

2. Related Work

Investigations of Segment Anything Model (SAM). In the short time since its publication, SAM has been used in numerous applications including image dehazing [22], image tagging [58], shadow segmentation [49], sonar segmentation [48], foreground segmentation [50], electron microscopy segmentation [10,26], and many medical segmentation applications [2,14–16,24,27,47,55]. There have also been several applications of SAM in remote sensing, including segmentation of geological features and landforms in planetary structures [23], segmentation of glaciers [45] and sea ice [53], road segmentation [19], and building segmentation [19,56]. Our work represents the first comprehensive evaluation of SAM on overhead imagery (e.g., including a variety of target classes, resolutions, and geographic locations), and the first to include both model composition and interactive annotation scenarios. Collectively, therefore, our results are the first that assess how well SAM transfers to overhead imagery applications.

Recognition in Overhead Imagery. This is a widely-studied application area in the computer vision community (e.g., [11, 39]), and especially since the emergence of deep learning (e.g., [5, 28, 31, 61]). Although not an exhaustive list, some notable major applications areas are the recognition (e.g., segmentation, classification, or detection) of buildings (e.g., [12, 17, 32], roads (e.g., [1, 21, 51], solar arrays (e.g., [6, 33, 54], cloud mapping [18, 36, 38], and land cover mapping [41, 42, 52, 57]. For these reasons we include the aforementioned tasks in our benchmark experiments to examine SAM under a diverse and representative set of tasks, and image resolutions. Although the sensing modality used in these tasks may vary, to maintain a tractable scope for our study, we focus on overhead color imagery.

3. Benchmark Datasets

For our experiments, we utilized eight benchmark datasets of overhead imagery that were selected for inclusion based on several criteria. We first had two strict criteria for inclusion: (i) the availability of class-level or instance-level segmentation labels; and (ii) the datasets were publicly available, to enable further study by the community. Among

the datasets that satisfied these criteria, we also applied several softer criteria for inclusion: (i) we prioritized datasets that were of greater interest to the overhead imagery community, as evidenced by their inclusion in many prior publications; (ii) benchmarks that involved widely-studied target objects (e.g., buildings, roads, land use); and (iii) datasets that collectively resulted in a representative set of key properties (e.g., geographic location, image resolution, or target classes). The resulting set of eight datasets that we selected is shown in Table 1, along with references and key details. Further details about our benchmark datasets can be found in the Supplement.

Additional Data Preparation Details. The datasets we ultimately used for experimentation were constructed based upon the benchmarks in Table 1. In many cases, we used the datasets without modification: Solar, 38-Cloud, DeepGlobe Roads, SpaceNet 2. However, we made some modifications to other datasets. Following a recent large-scale comparison of building segmentation models [30], we combine the Inria and the DeepGlobe Building datasets into a single dataset, *DG+Inria Building*, allowing us to compare our results with SAM to the strong supervised models from [30]. The Deep-Globe Land dataset comprises seven classes, and all pixels are assigned to one of seven classes, resulting in a large number of potential object instances; in our experiments, we treat each connected component within each class as an instance (see Sec. 4). To make the problem less computationally intensive, we chose a subset of three classes of increasing visual complexity (water, agriculture, and urban) and in each case, we set their labels to one, and all other classes to zero, resulting in three separate problems: *DG-land-water*, *DG-land-urban*, *DG-land-agri*. Lastly, the Parcel Delineation dataset is natively an edge detection dataset; however, we designated each isolated component of non-edge pixels as an object instance.

4. Model Composition Experiments

In the Composition scenario, SAM is utilized to enhance the predictions made by other vision models. Specifically, following [25], we assume that SAM is prompted with either a single point, p, or a bounding box, b, that has been produced by some other vision model. The goal of SAM is to produce an accurate instance mask based upon these simpler, and potentially imperfect, prompts. To emulate the desired model-based prompts, we train U-Net models to generate the bounding box prompts, and we use the ground truth labels from each benchmark to generate the point-based prompts. Our experimental design is illustrated in Fig 3.

U-Net Models. We train a U-Net segmentation model for each of our benchmark datasets, denoted *U-Net (ours)*. Note that most of our benchmarks have class-level segmentation labels (as opposed to instance-level) and therefore

Dataset	# of Unique Locations	Classes	Size (km²)	Resolution (m)	Task Type
Solar [6]	4	Solar Panel	1,352.25	0.30	Seg
Inria [32]	5	Building	405	0.30	Seg
DeepGlobe [12]	4	Building	398.28	0.31	Seg
38-Cloud [37]	Unknown	Cloud	2,188,800	30	Seg
DeepGlobe Roads [12]	3	Roads	2,220	0.50	Seg
Parcel Delineation [3]	1	Crop Boundaries	4,403.84	10	Edge
DeepGlobe Land [12]	3	Land Use	1,716.9	0.50	Seg
SpaceNet 2 [46]	4	Building	3011	0.30	Seg

Table 1. A summary of the quantity, resolution, and type of data we evaluate. The resolution is in units of meters per pixel. For task type, we have "Seg" for segmentation and "Edge" for edge detection.

we train our models to predict class-level masks. For each benchmark we create disjoint training and testing data partitions for model training and evaluation. Where possible we use pre-established partitions for each benchmark (e.g., Inria, DeepGlobe, SpaceNet). For some of our benchmarks - where they are available - we also report the IoU of a recent State-of-the-Art (SOTA) U-net model, denoted *U-Net (SOTA)* [4, 30, 40, 59]. Full details of the our models and the SOTA models can be found in the Supplementary Materials (e.g., architecture, training procedures, source publications). These models reflect the current performance that can be achieved using fully supervised techniques, providing a useful comparison to SAM.

Bounding Box Prompts. Based upon the output of the class-level masks output by the *U-Net (ours)* model, we extract instance-level bounding boxes that can then be used to prompt SAM. This is achieved by treating each connected component in the U-Net output as an instance-level mask. Then for each of these instance-level masks, we compute the smallest bounding box that encloses it and use the resulting box to prompt SAM. From these prompts, SAM produces instance-level mask predictions, denoted $\hat{m}_i$, where i refers to the mask generated by the ith box prompt. To evaluate the accuracy of SAM's masks, we use them to create a single class-level mask $\hat{m} = \cup_i \hat{m}_i$, which can be scored against the class-level ground truth masks available with our benchmark datasets.

Point Prompts. Our point-based prompts are generated using the ground truth labels that are available with each benchmark dataset. Except for SpaceNet, our benchmark datasets all provide class-level labels. We treat connected components in the ground truth masks as instance-level masks. Within each mask, we generate two point prompts: one by selecting a random point within the mask, and one

by selecting the center point in the mask (where "center" refers to the point most distant from the mask boundary). We score SAM's performance once using only the random points, and once only the center points. When SAM is prompted with a single point it produces three candidate masks, denoted $\hat{m}^k$, and a score for each mask indicating the model's estimate of the IoU of that mask with the true underlying object, denoted $\hat{c}^k$. SAM returns as a prediction the mask with the highest IoU prediction, which we denote $\hat{m}^*$. Similar to bounding boxes, we produce a single class-level mask by taking the union of all instance-level mask predictions.

4.1. Results and Discussion

The results of our composition experiments are presented in Fig. 4. The results indicate that the task-specific supervised models almost always achieve the highest IoU. This is unsurprising since these models were trained using a large quantity of task-specific segmentation masks. The *U-Net (SOTA)* model always outperforms the *U-Net (ours)*, reflecting the additional effort invested in its specialized design for each benchmark. Therefore, the IoU of the supervised models represents the performance that can be achieved using contemporary vision models out-of-the-box - in the case of *U-Net (ours)* - and contemporary vision models with significant design investment - in the case of *U-Net (SOTA)* model. We also see that SAM consistently performs significantly worse with point prompts compared to more informative (but more expensive) box prompts, and that the position of point prompts (center versus random) has little effect on results. These findings for overhead imagery are largely consistent with those reported for SAM for natural imagery in [25] (e.g., on COCO and LVIS), where SAM's performance was also always inferior to a supervised model, and less-informative prompts performed worse. However, we highlight a few key differences in our findings here.

One key difference in our findings is that the gap in performance between SAM and supervised models (the U-Net models in our case) is significantly larger on most of our benchmarks. This is especially noticeable on the largest and most widely studied benchmark target objects, such as buildings and roads. We hypothesize that this reflects SAM's bias towards objects in natural imagery. For example, the building class is highly complex (e.g., buildings are often large, and comprise many object-like subcomponents), and therefore they likely cannot be well-characterized using a generic "objectness" concept inferred based upon natural imagery. However, SAM also performs relatively poorly (compared to supervised models) on solar arrays, despite their apparent visual simplicity, suggesting that it is not guaranteed to generalize to overhead imagery even when the target objects are relatively simple. As we

discuss in **Sec. 6**, SAM's performance can depend strongly upon the scale of the objects (i.e., we find simply artificially up-sampling the imagery can be beneficial), potentially impacting its performance on solar arrays or other small object instances. Lastly, we see that SAM achieves substantially lower performance, and low performance overall, on the road class. This highlights a major challenge with some target classes in overhead imagery where the notion of an object instance is not well-defined. We discuss this issue further in **Sec. 6**.

Overall our results suggest that SAM can sometimes generalize well to overhead imagery tasks, providing competitive zero-shot results compared with supervised models in some cases. However, in most cases, it performs relatively poorly compared to supervised models, especially heavily-engineered models (e.g., on buildings and clouds), and sometimes it completely fails (e.g., roads). Many of these challenges may be overcome by fine-tuning SAM on overhead imagery classes (e.g., buildings, clouds, land use), or by producing class-level segmentation rather than instance-level segmentations in cases where an instance concept is poorly defined (e.g., roads).

5. Interactive Annotation Experiments

In the Interactive Annotation scenario, SAM is employed to reduce the time and effort required for a human to draw instance-level segmentation masks. Rather than drawing a full mask by hand, the annotator can provide one (or more) cheap input prompts, from which SAM can provide richer instance-level masks. Specifically, following [25], we assume that SAM is prompted with either a series of points, denoted p_i for the ith point; or a bounding box, denoted b. In contrast to the composition scenario in Sec. 4, here we assume that the prompts are provided by a human annotator and this changes some of the experimental design. Our experimental design, which follows [25], is illustrated in Fig 5 and described below.

Bounding Box Prompts. Bounding box prompts are created in the same fashion as described in Sec. 4 except that we use ground truth class-level masks available with our benchmark datasets instead of those generated by a supervised model. Ground truth masks are generally more accurate than those generated by a model, reflecting the superior accuracy of a human annotator. Recapitulating this procedure, we extract instance-level prompts from the ground truth of each dataset, which are then used to prompt SAM. From these prompts, SAM produces instance-level mask predictions, denoted $\hat{m}_i$, where i refers to the mask generated by the ith box prompt. To evaluate the accuracy of SAM's masks, we again use them to create a single class-level mask $\hat{m} = \cup_i \hat{m}_i$, which can be scored against the class-level ground truth masks available with our benchmark datasets.
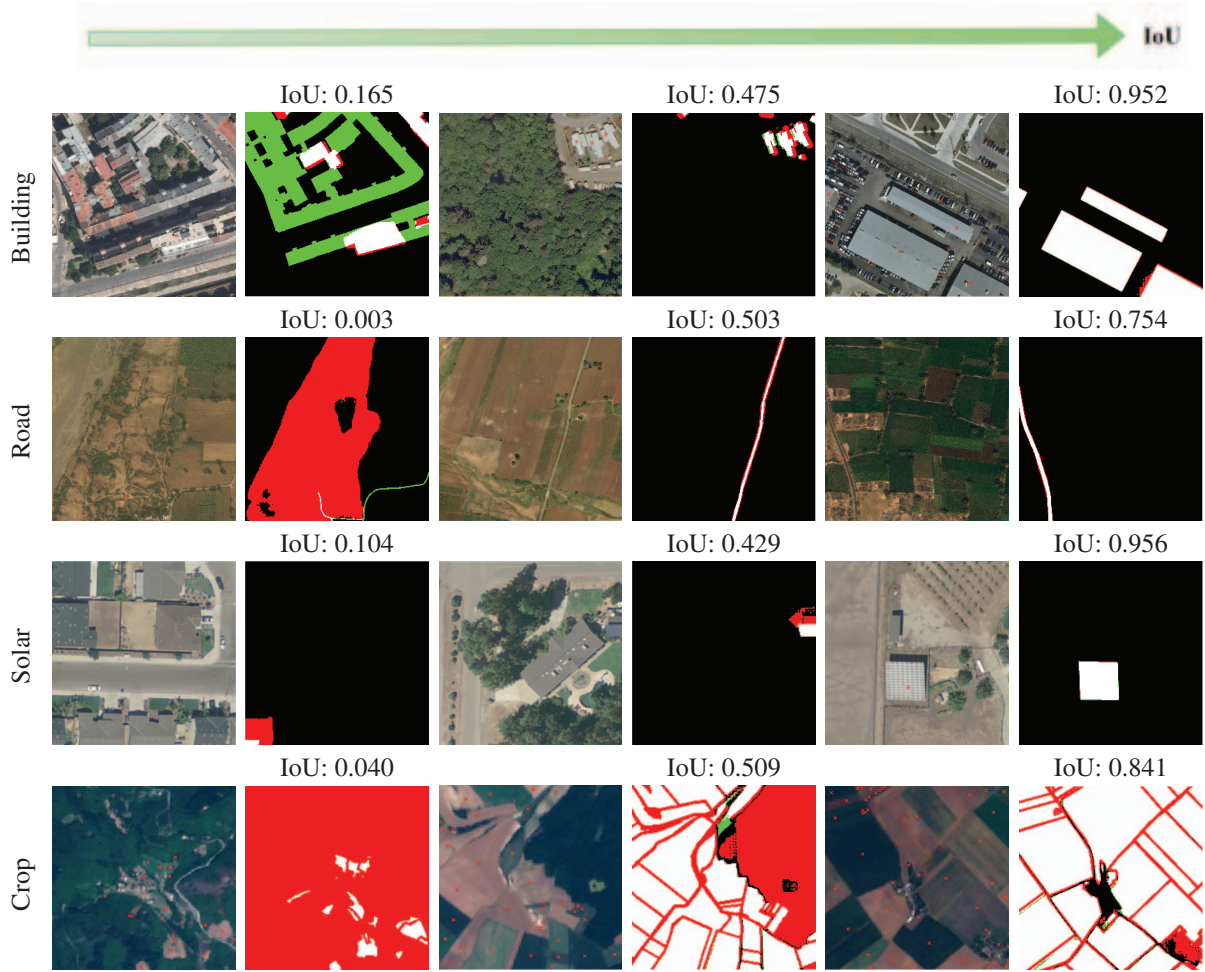
Figure 2. The input image, prompts, and segmentation output for SAM is shown for the building, road, solar panel, and crop segmentation tasks. Each row displays a specific segmentation task and the input output image pairs are ordered from worst performing to best performing. Red X's in the input images represent the prompts given to SAM. The output is shown in three colors where white represents the correctly predicted pixels, red shows the false positive pixels, and green shows the missed pixels. The image-wise IoU is shown above each output image.

Single Point Prompts. In these experiments we evaluate SAM's performance when prompted with a single point, similar to Sec. 4. For this purpose we use exactly the same procedure for generating the point prompts as was described in Sec. 4. In summary, we extract instance-level masks from the ground truth masks available for each of our benchmark datasets and then prompt SAM either with (i) a random point, or (ii) a centered point, in each instance-level mask. When prompted with a point, SAM produces three candidate masks and corresponding estimates of their IoU, $(\hat{m}_t^k, \hat{c}_t^k)$, where k indexes over the three masks, t indexes the iteration of interactive segmentation ($t = 1$ in the single point case). In the composition experiments in Sec. 4 we selected the mask with the largest c value as the final mask, however, a human annotator can evaluate the *true*

IoU of each candidate mask, denoted c_t^k. Therefore, in this case, we select as the output mask, the one that has the highest true IoU, denoted $\hat{m}^*$. Once we have $\hat{m}^*$, we aggregate SAM's instance-level mask predictions in the same way as Sec. 4 and report the resulting IoU for each benchmark. Also following [25] we compare to the Reviving Iterative Training with Mask Guidance (RITM) interactive annotation model [44], which is evaluated using the same random and centered point prompts, respectively, that were used to prompt SAM. To our knowledge there is no interactive annotation model specifically for overhead imagery.

Interactive Point Prompts. The goal of this experiment is to emulate an interactive point-prompting scenario, in which a human annotator provides a series of point prompts, and where each point is used to improve upon the masks
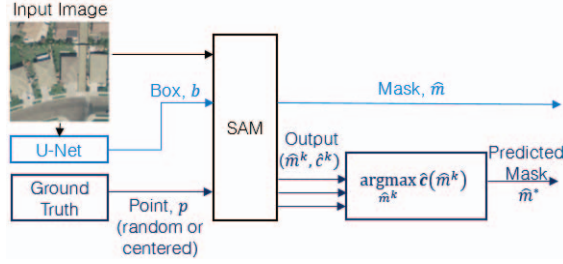
Figure 3. Illustration of the Model Composition experimental design, adapted from [25] to our Overhead Imagery application. We consider two general interactive annotation scenarios: a single bounding box prompt (light blue), and multi-point prompts (dark blue). In the first scenario (light blue) SAM is provided with a single bounding box prompt produced by our fine-tuned U-Net model, denoted b and it returns a single estimated mask, denoted $\hat{m}$. In the second scenario (dark blue), SAM is initially provided with a single point prompt, denoted p_1, and returns as output three mask-score pairs, $(\hat{m}_i^k, \hat{c}_i^k)$ each with an estimate of the IOU obtained.

generated using all the previous prompts. Specifically, and following [25], we begin by prompting SAM with an initial point, p_1, which is obtained using the center point prompt from the Single Point Prompt above. The final mask chosen for the next iteration of annotation is the one with the highest true IoU, denoted $\hat{m}_t^*$. Based upon $\hat{m}_t^*$, another point prompt, p_{t+1} is generated with the following process: we identify the largest contiguous error regions (either false positive or false negative) and then produce a point prompt in the center of that region. The point prompt also assigned a number, which is input to SAM, indicating whether the point corresponds to a false negative or false positive region. Then we input p_{t+1} and $\hat{m}_t^*$ to SAM to generate another set of candidate masks. We evaluate the performance of SAM and RITM as a function of T, the total number of iterations (and total number of point prompts) provided by SAM.

5.1. Results and Discussion

The results of our single-point and single-box prompt experiments are reported in Fig. 6. The results indicate that the IoU of SAM tends to decrease as the informativeness of the prompts decreases, where bounding boxes are most informative, and random center points are least informative. SAM seems relatively insensitive to whether the prompt points are centered or random. We also find that SAM almost always outperforms RITM, and often by a substantial margin. For example, on the Solar and Inria+DG Buildings datasets, SAM outperforms RITM by 50 and 45 IoU points, respectively. These results are largely consistent with those reported for natural imagery in [25], although the advantage of SAM over RITM seems even larger on overhead imagery than on natural imagery. Also, notably, and similar to our

model composition experiments in Sec. 4, all approaches perform very poorly on Inria+DG Roads, which we discuss in Sec 6.

The results of our multi-point prompt experiments are reported in Fig. 7. The results indicate that the IOU of both SAM and RITM consistently improves as the number of prompt points increases. Notably SAM outperforms RITM in nearly every combination of benchmark or number of points, and SAM typically has the greatest performance advantages when there is a small number of points. Once $T = 10$ both approaches typically have similar IoU, and often achieve high overall IoUs that we expect would often be satisfying for extraction of ground truth masks (e.g., SAM achieves an IOU of 0.75, or larger, on 6 out of our 10 benchmarks when T=10). These results are largely consistent with those reported for SAM on natural imagery [25], except that the overall IoUs are lower. Collectively these results suggest that SAM offers state-of-the-art interactive annotation accuracy, and also often (though not always) achieves sufficiently high IoU to be a useful tool for human-interactive instance annotation. Similar to our findings in the model annotation case, we hypothesize that substantial improvements in IoU could be obtained through fine-tuning SAM for overhead imagery, or for class-level segmentation.

6. Additional Analysis

In this section we discuss additional findings based on the collective experimental results in Sec. 4 and Sec. 5.

Sensitivity of SAM to Image Resolution. One finding of SAM was that it suffers from some sensitivity to image resolution. It is well-known that using higher-resolution overhead imagery will enable higher accuracy for image recognition models, which is often attributed to the greater information content in the imagery. However, we found that simply up-sampling our overhead imagery (i.e., adding no new information) can significantly improve SAM's IoU. Fig. 8 shows the IoU of SAM as a function of up-sampling factor for two benchmark problems, where in each case we artificially up-sampled the imagery. The results indicate that the impact of up-sampling varies, but that SAM is generally sensitive to the number of pixels per unit of ground area, even if the information content in the image has not changed. We hypothesize that SAM may have a bias in its expected size of object instance. In particular, SAM was trained on very high-resolution natural imagery, which in many cases biases it towards more pixels per object.

Instance Segmentation is Sometimes Ill-Posed. SAM is designed to perform object instance segmentation, however, some classes in overhead imagery may not be well-conceptualized in this manner. Fundamentally this problem can arise if there is no clear definition of an instance of a target class, or the definition may vary significantly across geography or application areas. One example is roadways,
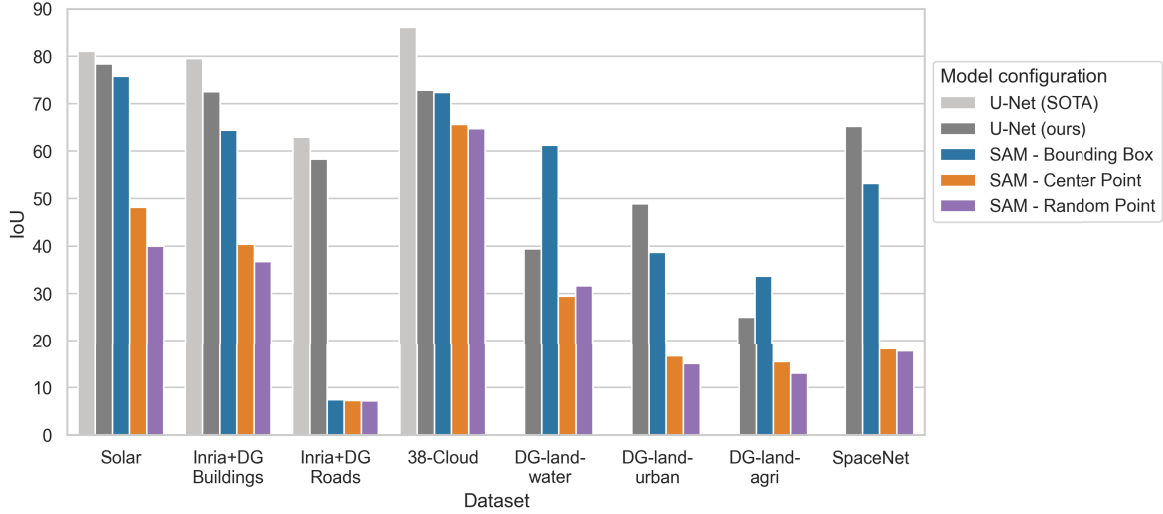
8345

Figure 4. The intersection-over-union for all models included in our Model Composition experiments. Performance of the SOTA models shown from left to right are taken from the following papers [40], [30], [59], and [4]. For the DeepGlobe Land and SpaceNet dataset we performed binary segmentation for a single class and as such there was no SOTA U-Net model to compare with.
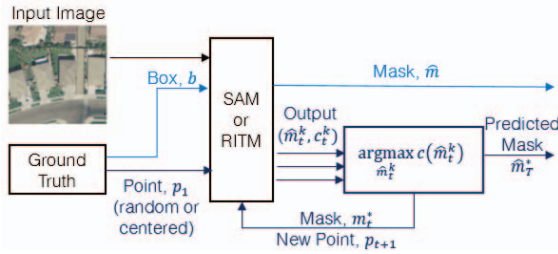

Figure 5. Illustration of the experimental designs adopted from [25] to evaluate SAM for human-interactive annotation. We consider two general interactive annotation scenarios: a single bounding box prompt (light blue), and multi-point prompts (dark blue). In the first scenario (light blue) SAM is provided with a single bounding box prompt, denoted b and it returns a single estimated mask, denoted $\hat{m}$. In the second scenario (dark blue), SAM is initially provided with a single point prompt, denoted p_1, and returns as output three mask-score pairs, $(\hat{m}_i^k, \hat{c}_i^k)$ each with an estimate of the IOU obtained.

a widely-studied problem in overhead imagery [8,9,59,60], which is challenging because roads are large objects that are spatially connected over very large geographic regions. Therefore it is unclear how to define when one road instance ends, and another begins. Consequently, SAM consistently performs very poorly (IoU<10) on the road segmentation task (i.e., Inria+DG Roads) in all of our experiments compared to models that have been trained for class-level segmentation, which does not suffer from these problems. This problem may be addressed by fine-tuning SAMs decoder to produce class-level segmentation masks for certain classes

where this is more appropriate.

7. Conclusions

In this work, we investigate whether the recently-proposed Segment Anything Model (SAM) [25] generalizes well to tasks involving overhead imagery. We evaluated SAM on a large and diverse set of eight overhead imagery benchmark datasets. For each benchmark, we investigated SAM's performance on two important potential application scenarios, in a similar fashion to [25]: model composition (COMP), and interactive segmentation (INTER). We summarize our conclusions, labeled by the application area.

- **(INTER)** SAM nearly always outperforms RITM (a baseline state-of-the-art annotation approach), and often by a large margin.

- **(INTER)** SAM achieves somewhat lower IOUs on overhead imagery compared to comparable experimental settings with natural imagery reported in [25].

- **(INTER)** However, SAM can often achieve IoUs that would likely be satisfying for many practical annotation scenarios (e.g., SAM achieves >0.75 IoU on 6 of 10 benchmarks, with 10 point-based prompts).

- **(COMP)** SAM can often achieve performance that is comparable to out-of-the-box, but state-of-the-art, supervised models that have been trained on each task. However, its behavior is highly variable: in most cases it performs somewhat worse, but in some cases it performs superior to supervised models, while in others it performs much worse.
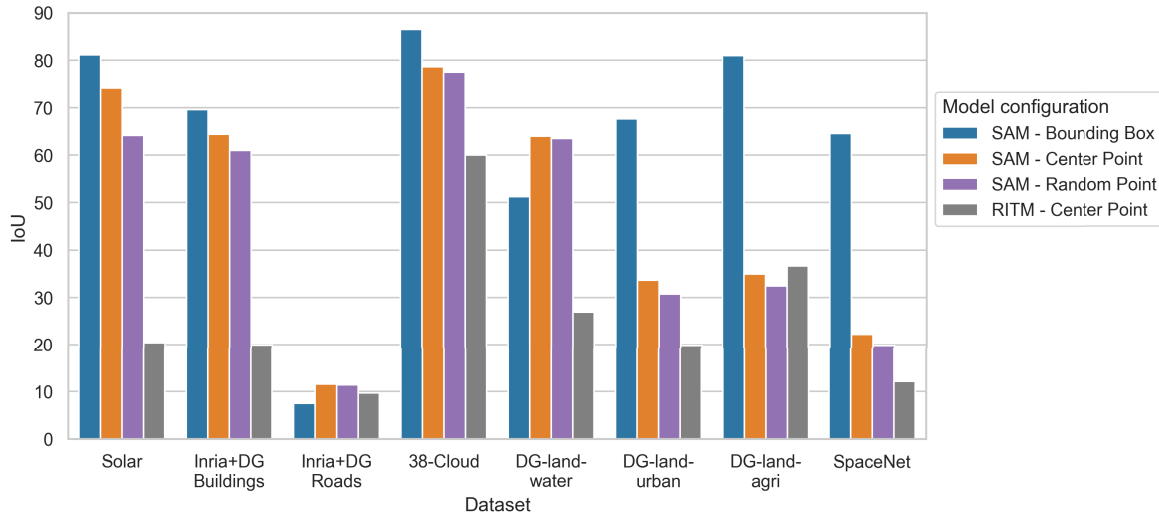
Figure 6. Results of (pixel) intersection-over-union scores for each interactive annotation scenario.
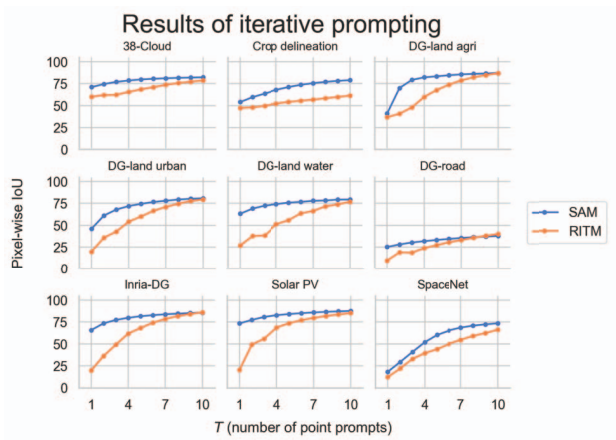

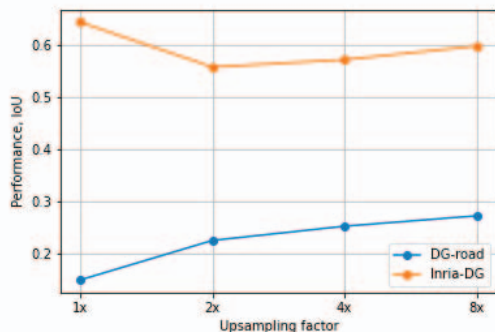Figure 7. Results of iterative prompts.

Figure 8. Results of iterative prompts.

- **(COMP)** The aforementioned (COMP) finding pertain to bounding-box prompts provided by other vision models. SAM generally performs poorly when prompted with point prompts obtained from other vision models.

- We also find that SAM is sensitive to the resolution of the imagery; e.g., artificially up-sampling the imagery can significantly improve or lower SAM's performance.

- SAM is designed to produce instance segmentations, however, this may not be well-suited to some object classes, such as roads, where we observe very poor performance.

Recommendations. The performance of SAM can vary substantially on overhead imagery tasks. This is typical of models being applied to novel data domains (i.e., overhead imagery instead of natural imagery), but users should be aware of this variability. We suspect that fine-tuning SAM's decoder alone (as opposed to its larger encoder) may yield substantial improvements in its performance and reliability on overhead imagery. In particular, retraining the decoder to better recognize common target class features (e.g., buildings, land-use) may be greatly beneficial. Also, many tasks in overhead imagery are better-suited for class-level segmentation (e.g., roadways), and the decoder could be re-trained for class-level segmentation to substantially improve performance on road segmentation, or similar classes.

References

[1] Abolfazl Abdollahi, Biswajeet Pradhan, Nagesh Shukla, Subrata Chakraborty, and Abdullah Alamri. Deep learning approaches applied to remote sensing datasets for road extraction: A state-of-the-art review. *Remote Sensing*, 12(9):1444, 2020. 2

[2] Mohsen Ahmadi, Masoumeh Farhadi Nia, Sara Asgarian, Kasra Danesh, Elyas Irankhah, Ahmad Gholizadeh Lonbar, and Abbas Sharifi. Comparative Analysis of Segment Anything Model and U-Net for Breast Tumor Detection in Ultrasound and Mammography Images. *arXiv preprint arXiv:2306.12510*, 2023. 2

[3] Han Lin Aung, Burak Uzkent, Marshall Burke, David Lobell, and Stefano Ermon. Farm Parcel Delineation Using Spatio-temporal Convolutional Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 340–349, Seattle, WA, USA, June 2020. IEEE. 3

[4] Gaétan Bahl, Lionel Daniel, Matthieu Moretti, and Florent Lafarge. Low-power neural networks for semantic segmentation of satellite images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3, 7

[5] John E Ball, Derek T Anderson, and Chee Seng Chan. Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community. *Journal of applied remote sensing*, 11(4):042609–042609, 2017. 2

[6] Kyle Bradbury, Raghav Saboo, Timothy L Johnson, Jordan M Malof, Arjun Devarajan, Wuming Zhang, Leslie M Collins, and Richard G Newell. Distributed solar photovoltaic array location and extent dataset for remote sensing object identification. *Scientific data*, 3(1):1–9, 2016. 2, 3

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[8] Alexander Buslaev, Selim Seferbekov, Vladimir Iglovikov, and Alexey Shvets. Fully Convolutional Network for Automatic Road Extraction from Satellite Imagery. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 207–210, 2018. 7

[9] Ziyi Chen, Liai Deng, Yuhua Luo, Dilong Li, José Marcato Junior, Wesley Nunes Gonçalves, Abdul Awal Md Nurunnabi, Jonathan Li, Cheng Wang, and Deren Li. Road extraction in remote sensing data: A survey. *International Journal of Applied Earth Observation and Geoinformation*, 112:102833, 2022. 7

[10] Ao Cheng, Guoqiang Zhao, Lirong Wang, and Ruobing Zhang. AxonCallosumEM Dataset: Axon Semantic Segmentation of Whole Corpus Callosum cross section from EM Images. *arXiv preprint arXiv:2307.02464*, 2023. 2

[11] Gong Cheng and Junwei Han. A survey on object detection in optical remote sensing images. *ISPRS journal of photogrammetry and remote sensing*, 117:11–28, 2016. 2

[12] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. DeepGlobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 172–181, 2018. 2, 3

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[14] Sheng He, Rina Bao, Jingpeng Li, P Ellen Grant, and Yangming Ou. Accuracy of Segment-Anything Model (SAM) in medical image segmentation tasks. *arXiv preprint arXiv:2304.09324*, 2023. 2

[15] Fabian Hörst, Moritz Rempe, Lukas Heine, Constantin Seibold, Julius Keyl, Giulia Baldini, Selma Ugurel, Jens Siveke, Barbara Grünwald, Jan Egger, et al. CellViT: Vision Transformers for Precise Cell Segmentation and Classification. *arXiv preprint arXiv:2306.15350*, 2023. 2

[16] Xinrong Hu, Xiaowei Xu, and Yiyu Shi. How to Efficiently Adapt Large Segmentation Model (SAM) to Medical Images. *arXiv preprint arXiv:2306.13731*, 2023. 2

[17] Bohao Huang, Kangkang Lu, Nicolas Audeberr, Andrew Khalel, Yuliya Tarabalka, Jordan Malof, Alexandre Boulch, Bertr Le Saux, Leslie Collins, Kyle Bradbury, et al. Large-scale semantic classification: outcome of the first year of inria aerial image labeling benchmark. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 6947–6950. IEEE, 2018. 2

[18] Jacob Høxbroe Jeppesen, Rune Hylsberg Jacobsen, Fadil Inceoglu, and Thomas Skjødeberg Toftegaard. A cloud detection algorithm for satellite imagery based on deep learning. *Remote sensing of environment*, 229:247–259, 2019. 2

[19] Wei Ji, Jingjing Li, Qi Bi, Wenbo Li, and Li Cheng. Segment Anything Is Not Always Perfect: An Investigation of SAM on Different Real-world Applications. *arXiv preprint arXiv:2304.05750*, 2023. 2

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1

[21] Xiaoying Jin and Curt H Davis. An integrated system for automatic road mapping from high-resolution multi-spectral satellite imagery by information fusion. *Information Fusion*, 6(4):257–273, 2005. 2

[22] Zheyan Jin, Shiqi Chen, Yueting Chen, Zhihai Xu, and Huajun Feng. Let Segment Anything Help Image Dehaze. *arXiv preprint arXiv:2306.15870*, 2023. 2

[23] Sahib Julka and Michael Granitzer. Knowledge distillation with Segment Anything (SAM) model for Planetary Geological Mapping. *arXiv preprint arXiv:2305.07586*, 2023. 2

[24] Heejong Kim, Victor Ion Butoi, Adrian V Dalca, and Mert R Sabuncu. Empirical Analysis of a Segmentation Foundation Model in Prostate Imaging. *arXiv preprint arXiv:2307.03266*, 2023. 2

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer White-

head, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1, 2, 3, 4, 5, 6, 7

[26] Rasmus Larsen, Torben L Villadsen, Jette K Mathiesen, Kirsten MØ Jensen, and Espen D Boejesen. NP-SAM: Implementing the Segment Anything Model for Easy Nanoparticle Segmentation in Electron Microscopy Images. 2023. 2

[27] Haipeng Li, Dingrui Liu, Yu Zeng, Shuaicheng Liu, Tao Gan, Nini Rao, Jinlin Yang, and Bing Zeng. Single-Image-Based Deep Learning for Segmentation of Early Esophageal Cancer Lesions. *arXiv preprint arXiv:2306.05912*, 2023. 2

[28] Ying Li, Haokui Zhang, Xizhe Xue, Yenan Jiang, and Qiang Shen. Deep learning for remote sensing image classification: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(6):e1264, 2018. 2

[29] Kangcheng Lin, Bohao Huang, Leslie M Collins, Kyle Bradbury, and Jordan M Malof. A simple rotational equivariance loss for generic convolutional segmentation networks: preliminary results. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 3876–3879. IEEE, 2019. 2

[30] Francesco Luzi, Aneesh Gupta, Leslie Collins, Kyle Bradbury, and Jordan Malof. Transformers For Recognition In Overhead Imagery: A Reality Check. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3778–3787, 2023. 3, 7

[31] Lei Ma, Yu Liu, Xueliang Zhang, Yuanxin Ye, Gaofei Yin, and Brian Alan Johnson. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS journal of photogrammetry and remote sensing*, 152:166–177, 2019. 2

[32] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can Semantic Labeling Methods Generalize to Any City? The Inria Aerial Image Labeling Benchmark. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2017. 2, 3

[33] Jordan M Malof, Kyle Bradbury, Leslie M Collins, and Richard G Newell. Automatic detection of solar photovoltaic arrays in high resolution aerial imagery. *Applied energy*, 183:229–240, 2016. 2

[34] Diego Marcos, Michele Volpi, Benjamin Kellenberger, and Devis Tuia. Land cover mapping at very high resolution with rotation equivariant cnns: Towards small yet accurate models. *ISPRS journal of photogrammetry and remote sensing*, 145:96–107, 2018. 2

[35] Diego Marcos, Michele Volpi, Nikos Komodakis, and Devis Tuia. Rotation equivariant vector field networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5048–5057, 2017. 2

[36] Andrea Meraner, Patrick Ebel, Xiao Xiang Zhu, and Michael Schmitt. Cloud removal in sentinel-2 imagery using a deep residual neural network and sar-optical data fusion. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:333–346, 2020. 2

[37] S. Mohajerani, T. A. Krammer, and P. Saeedi. "A Cloud Detection Algorithm for Remote Sensing Images Using Fully Convolutional Neural Networks". In *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5, Aug 2018. 3

[38] Sorour Mohajerani and Parvaneh Saeedi. Cloud-net: An end-to-end cloud detection algorithm for landsat 8 imagery. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 1029–1032. IEEE, 2019. 2

[39] Giorgos Mountrakis, Jungho Im, and Caesar Ogole. Support vector machines in remote sensing: A review. *ISPRS journal of photogrammetry and remote sensing*, 66(3):247–259, 2011. 2

[40] Spencer Paul, Ethan Hellman, and Rodri Neito. SolarX: Solar Panel Segmentation and Classification. 3, 7

[41] Charlotte Pelletier, Silvia Valero, Jordi Inglada, Nicolas Champion, and Gérard Dedieu. Assessing the robustness of random forests to map land cover with high resolution satellite image time series over large areas. *Remote Sensing of Environment*, 187:156–168, 2016. 2

[42] Darius Phiri, Matamyo Simwanda, Serajis Salekin, Vincent R Nyirenda, Yuji Murayama, and Manjula Ranagalage. Sentinel-2 data for land cover/use mapping: A review. *Remote Sensing*, 12(14):2291, 2020. 2

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[44] Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. Reviving Iterative Training with Mask Guidance for Interactive Segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3141–3145. IEEE, 2022. 5

[45] Leigh Stearns, Cornelis van der Veen, and Siddharth Shankar. Segment Anything in Glaciology: An initial study implementing the Segment Anything Model (SAM). 2023. 2

[46] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. SpaceNet: A Remote Sensing Dataset and Challenge Series. *arXiv preprint arXiv:1807.01232*, 2018. 3

[47] An Wang, Mobarakol Islam, Mengya Xu, Yang Zhang, and Hongliang Ren. SAM Meets Robotic Surgery: An Empirical Study on Generalization, Robustness and Adaptation. *arXiv preprint arXiv:2308.07156*, 2023. 2

[48] Lin Wang, Xiufen Ye, Liqiang Zhu, Weijie Wu, Jianguo Zhang, Huiming Xing, and Chao Hu. When SAM Meets Sonar Images. *arXiv preprint arXiv:2306.14109*, 2023. 2

[49] Yonghui Wang, Wengang Zhou, Yunyao Mao, and Houqiang Li. Detect Any Shadow: Segment Anything for Video Shadow Detection. *arXiv preprint arXiv:2305.16698*, 2023. 2

[50] Muduo Xu, Jianhao Su, and Yutao Liu. AquaSAM: Underwater Image Foreground Segmentation. *arXiv preprint arXiv:2308.04218*, 2023. 2

[51] Yongyang Xu, Zhong Xie, Yaxing Feng, and Zhanlong Chen. Road extraction from high-resolution remote sensing imagery using deep learning. *Remote Sensing*, 10(9):1461, 2018. 2

[52] Christopher Yeh, Chenlin Meng, Sherrie Wang, Anne Driscoll, Erik Rozi, Patrick Liu, Jihyeon Lee, Marshall Burke, David B Lobell, and Stefano Ermon. Sustainbench: Benchmarks for monitoring the sustainable development goals with machine learning. *arXiv preprint arXiv:2111.04724*, 2021. 2

[53] Anzhu Yu, Wenjun Huang, Qing Xu, Qun Sun, Wenyue Guo, Song Ji, Bowei Wen, and Chunping Qiu. Sea Ice Extraction via Remote Sensed Imagery: Algorithms, Datasets, Applications and Challenges. *arXiv preprint arXiv:2306.00303*, 2023. 2

[54] Jiafan Yu, Zhecheng Wang, Arun Majumdar, and Ram Rajagopal. Deepsolar: A machine learning framework to efficiently construct a solar deployment database in the united states. *Joule*, 2(12):2605–2617, 2018. 2

[55] Wenxi Yue, Jing Zhang, Kun Hu, Yong Xia, Jiebo Luo, and Zhiyong Wang. SurgicalSAM: Efficient Class Promptable Surgical Instrument Segmentation. *arXiv preprint arXiv:2308.08746*, 2023. 2

[56] Jielu Zhang, Zhongliang Zhou, Gengchen Mai, Lan Mu, Mengxuan Hu, and Sheng Li. Text2Seg: Remote Sensing Image Semantic Segmentation via Text-Guided Visual Foundation Models. *arXiv preprint arXiv:2304.10597*, 2023. 2

[57] Pengbin Zhang, Yinghai Ke, Zhenxin Zhang, Mingli Wang, Peng Li, and Shuangyue Zhang. Urban land use and land cover classification using novel deep learning models based on high spatial resolution satellite imagery. *Sensors*, 18(11):3717, 2018. 2

[58] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize Anything: A Strong Image Tagging Model. *arXiv preprint arXiv:2306.03514*, 2023. 2

[59] Lichen Zhou, Chuang Zhang, and Ming Wu. D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 182–186, 2018. 3, 7

[60] Qiqi Zhu, Yanan Zhang, Lizeng Wang, Yanfei Zhong, Qingfeng Guan, Xiaoyan Lu, Liangpei Zhang, and Deren Li. A global context-aware and batch-independent network for road extraction from VHR satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 175:353–365, 2021. 7

[61] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE geoscience and remote sensing magazine*, 5(4):8–36, 2017. 2