Segment anything, from space?
Supplemental Materials

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

1. Experimental Design: additional details

Difficulties with SAM Mask input. The mask prompt input space is the 'low-resolution logit' space of SAM output, and it is much harder to be incorporated into the current remote sensing segmentation pipeline (a segmentation map output from a trained model cannot be readily consumed by SAM).

Edge extraction process for Parcel Delineation Dataset Using images from this dataset as input, we prompt SAM using points on a 16-by-16-pixel grid. (For the 224-by-224-pixel field delineation images, this corresponds to the center-points of 14-by-14-pixel patches.) The resulting masks are then filtered using non-maximum suppression, and masks covering more than 95 percent of the image area are removed. We then take the argmax of remaining masks at each pixel and pass the resulting output through a Sobel filter, leaving only extracted edges between masks. To smooth and thicken lines, we apply edge NMS (i.e., a Canny filter [4]) to the resulting image and perform a dilated convolution with a 3x3 kernel.

2. Benchmark Datasets: Additional Details

We evaluate a total eight datasets over several binary segmentation tasks. Of these eight, three are building segmentation datasets. We emphasizes the task of building segmentation due to its frequency in the literature and ease of finding diverse datasets.

Solar. The Solar PV dataset [3] is a collection of images from four major cities in California where solar PVs are commonly used. The dataset contains over 19,000 examples of solar PVs across 601 5000×5000 images and over 1,352 km^2 of area. The dataset is labeled with polygons containing the solar PVs as well as a segmentation mask for pixel-wise binary segmentation. This dataset contains an author-defined train-test split.

Inria. The Inria Aerial Image dataset [10] contains images taken at 0.30 meters per pixel resolution over ten cities (only five with labels). The cities were chosen such that there would be a mix of American and European cities in the dataset covering both sparse and densely populated areas. Equal coverage was given to each city and binary segmentation masks were labeled with building / no building. Following the convention in [9], we use the first six tiles in the dataset as our test set, as the official test set is withheld from the public.

DeepGlobe Buildings. The DeepGlobe Building Detection dataset [7] contains building segmentation labels for four major cities in four different countries. The dataset consists of 650×650 non-overlapping images taken at 0.31 meters per pixel resolution. Images were taken from the Spacenet dataset [16] using a WorldView-3 sensor. We randomly select one-sixth of the dataset to be the test set, per the convention in [9].

DeepGlobe Roads. The DeepGlobe Road Extraction dataset [7] is composed of images taken from the DigitalGlobe Vivid+ Images dataset and filtered to include interesting or useful areas for the road extraction task. Images were taken from three countries at 0.50 meters per pixel resolution and segmentation labels were added denoting roads as the only class. We randomly select one-sixth of the dataset to be the test set, per the convention in [9].

DeepGlobe Land. The DeepGlobe Land Cover Classification Dataset [7] contains 1,146 satellite images collected from the DigitalGlobe Vivid+ dataset, spanning 1,717 square kilometers of predominantly rural locations.

Model	Prompt Source	Prompt Method	Mask Selection	Solar	Inria + DG (Building)	DeepGlobe (Roads)	38-Cloud	Parcel Delineation
Unet [14]	NA	NA	NA	81.05 [12]	79.53 [9]	62.94 [17]	86.08 [2]	36.99 [1]
SwinUnet [5]	NA	NA	NA	-	78.48 [9]	-	-	-
TransUnet [6]	NA	NA	NA	-	79.96 [9]	-	-	-
D-LinkNet [17]	NA	NA	NA	-	-	64.12 [17]	-	-
C-UNet [2]	NA	NA	NA	-	-	-	86.50 [2]	-
Spatio-temporal UNet [1]	NA	NA	NA	-	-	-	-	43.88 [1]
Unet†	NA	NA	NA	78.39	72.50	58.35	72.88	-
SAM [8]	GT	Center Point	Max confidence	48.18	40.40	7.24	65.64	-
	GT	Center Point	Oracle	74.17	64.41	11.53	78.63	-
	GT	Random Point	Max confidence	39.96	36.76	7.17	64.70	-
	GT	Random Point	Oracle	64.19	61.01	11.41	77.46	-
	GT	Bounding Box	Single output	81.12	69.61	7.47	86.48	-
	Unet	Bounding Box	Single output	75.79	64.42	7.41	72.33	-
	-	Grid points	Max confidence	-	-	4.5	-	10.1

Table 1. Table of experimental results with baseline model comparison. Performance is reported in pixel-wise IoU.

Model	Inria					DeepGlobe				City
	Austin	Chicago	Kitsap County	West Tyrol	Vienna	Las Vegas	Paris	Shanghai	Khartoum	Average
Unet [14]	81.92	71.58	69.00	80.44	82.53	85.50	72.26	77.13	73.64	77.11
SwinUnet [5]	80.21	69.62	68.70	80.33	81.85	84.87	70.57	76.25	72.36	76.08
TransUnet [6]	81.94	73.21	69.19	81.46	82.94	85.45	72.92	77.31	73.79	77.58
SAM oracle center point [8]	91.73	83.93	67.81	65.10	76.13	80.79	58.58	87.00	56.63	74.19
SAM max conf center point [8]	60.68	83.36	58.07	46.48	67.69	69.56	57.89	87.00	59.31	65.56

Table 2. Comparison of Inria + DeepGlobe building segmentation broken down by city, measured by the intersection-over-union (IoU). The Unet, TransUnet, and SwinUnet results are taken from [9].

Classes identified in this dataset are urban, agriculture, rangeland, forest, water, barren, or unknown. We randomly select one-sixth of the dataset to be the test set, per the convention in [9].

38-Cloud. The 38-Cloud dataset [11] leverages Landsat 8 [15] to collect 38 images, taken from the QA band, of cloud cover at 30 meters per pixel resolution. The 38 images are split into 9700 patches for easier processing and filtered to remove snow / ice coverage. The dataset provides segmentation labels by human annotators specifying the pixels that contain cloud coverage. This dataset contains an author-defined train-test split.

Farm Parcel Delineation The Farm Parcel Delineation dataset [1] is comprised of 224×224 images at 10 meter per pixel resolution taken over three different time periods. The dataset uses imagery from Sentinel-2 which contains at least one plot of farmland. The images were queried from several different regions around the world but the vast majority of them were taken from France's countryside. The Farm Parcel Delineation dataset is labeled pixel-wise as boundary or non-boundary between parcels of farmland with 2-pixel width for the boundary lines. This dataset contains an author-defined train-test split.

SpaceNet The SpaceNet 2 dataset [16] contains aerial imagery with segmented buildings from four cities: Las Vegas, Paris, Shanghai, and Khartoum. In total, approximately 300 thousand buildings are segmented, nearly all of which are in urban and suburban regions. This dataset contains an author-defined train-test split.

3. Model Composition and Training: Additional Details

Ours. We first train a U-Net model (code) with a ResNet-50 encoder initialized with ImageNet weights (code), and the decoder architecture of the original U-Net [14] initialized from scratch. Both the encoder and decoder are then trained for 100 epochs using a batch size of 16, a learning rate of 10e-3 with the Adam optimizer and the soft intersection over union (IoU) loss [13]. We normalize the images using ImageNet pixel mean value. During inference we binarize the output mask using a threshold of 0.5. We use these same training and inference procedures for all of the datasets in our study.

SOTA. For the SOTA Unet results we used performance values reported in other published works. All references are shown in Table 1 next to the corresponding performance metric, with the individual Inria and DeepGlobe Building city results being shown in Table 2. For implementation and training details of the reported SOTA Unet models, see the corresponding references: Solar [12], Inria [9], DeepGlobe Building [9], DeepGlobe Roads [17], 38-Cloud [2], and Parcel Delineation [1].

References

[1] 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. 2, 3

[2] 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. 2, 3

[3] 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. 1

[4] John Canny. A Computational Approach To Edge Detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-8:679–698, Dec. 1986. 1

[5] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. *arXiv preprint arXiv:2105.05537*, 2021. 2

[6] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 2

[7] 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. 1

[8] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2

[9] 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. 1, 2, 3

[10] 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. 1

[11] 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. 2

[12] Spencer Paul, Ethan Hellman, and Rodri Neito. SolarX: Solar Panel Segmentation and Classification. 2, 3

[13] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 2

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2

[15] David P Roy, Michael A Wulder, Thomas R Loveland, Curtis E Woodcock, Richard G Allen, Martha C Anderson, Dennis Helder, James R Irons, David M Johnson, Robert Kennedy, et al. Landsat-8: Science and product vision for terrestrial global change research. *Remote sensing of Environment*, 145:154–172, 2014. 2

[16] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. SpaceNet: A Remote Sensing Dataset and Challenge Series. *arXiv preprint arXiv:1807.01232*, 2018. 1, 2

[17] 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. 2, 3