Segment Anything Model

Minwoo Jung
mai-lab.net, *Medical Artificial Intelligence Laboratory*
Electrical and Electronic Engineering
Yonsei University

Medical Artificial
Intelligence Laboratory
At Yonsei University

Segmentation

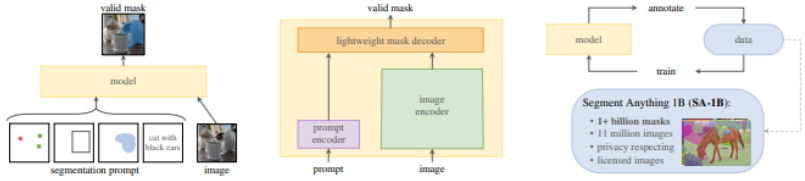
SAM

ICCV

2023

Segment Anything

Alexander Kirillov[1,2,4] Eric Mintun[2] Nikhila Ravi[1,2] Hanzi Mao[2] Chloe Rolland[3] Laura Gustafson[3]
Tete Xiao[3] Spencer Whitehead Alexander C. Berg Wan-Yen Lo Piotr Dollár[4] Ross Girshick[4]
[1]project lead [2]joint first author [3]equal contribution [4]directional lead
Meta AI Research, FAIR



(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

Figure 1: We aim to build a foundation model for segmentation by introducing three interconnected components: a promptable segmentation *task*, a segmentation *model* (SAM) that powers data annotation and enables zero-shot transfer to a range of tasks via prompt engineering, and a *data* engine for collecting SA-1B, our dataset of over 1 billion masks.

Abstract

We introduce the Segment Anything (SA) project: a new task, model, and dataset for image segmentation. Using our efficient model in a data collection loop, we built the largest segmentation dataset to date (by far), with over 1 ***billion*** *masks on 11M licensed and privacy respecting images. The model is designed and trained to be promptable, so it can transfer zero-shot to new image distributions and tasks. We evaluate its capabilities on numerous tasks and find that its zero-shot performance is impressive – often competitive with or even superior to prior fully supervised results. We are releasing the Segment Anything Model (SAM) and corresponding dataset (SA-1B) of 1B masks and 11M images at segment-anything.com to foster research into foundation models for computer vision.* **We recommend reading the full paper at: arxiv.org/abs/2304.02643.**

matching in some cases) fine-tuned models [10, 20]. Empirical trends show this behavior improving with model scale, dataset size, and total training compute [54, 10, 20, 49].

Foundation models have also been explored in computer vision, albeit to a lesser extent. Perhaps the most prominent illustration aligns paired text and images from the web. For example, CLIP [80] and ALIGN [53] use contrastive learning to train text and image encoders that align the two modalities. Once trained, engineered text prompts enable zero-shot generalization to novel visual concepts and data distributions. Such encoders also compose effectively with other modules to enable downstream tasks, such as image generation (*e.g.*, DALL·E [81]). While much progress has been made on vision and language encoders, computer vision includes a wide range of problems beyond this scope, and for many of these, abundant training data does not exist.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4015-4026).

Segmentation

SAM

Background & Goal

Background

- Foundation model, a model pre-trained with a huge dataset, shows tremendous generalizability for the task
- CV includes a wide range of problems, and for many of these, abundant training data does not exist

Goal

- To build a Foundation model for image segmentation
- To develop a prompt-able model and pre-train it on a broad dataset using a task that enables generalization

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4015-4026).

Segmentation

SAM

Task

Promptable Segmentation task

- To return a valid segmentation mask given any segmentation prompt
- The requirement of a valid output mask means that even when a prompt is ambiguous and could refer to multiple objects, the output should be a reasonable mask for at least one of those objects

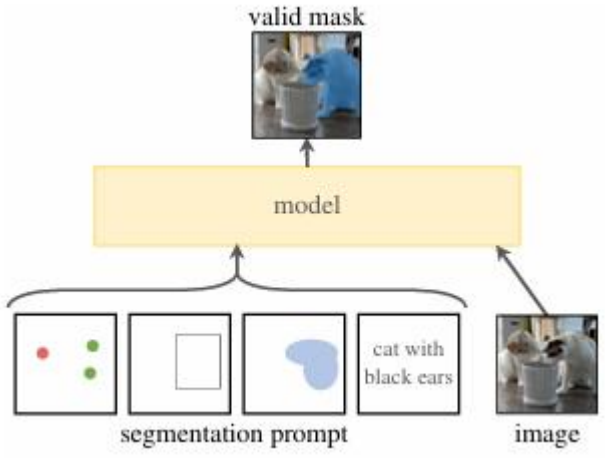


Fig 1. A promptable segmentation task

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4015-4026).

Segmentation

SAM

Task

Promptable Segmentation task

- To return a valid segmentation mask given any segmentation prompt
- To always predict a valid mask for any prompt even when the prompt is ambiguous

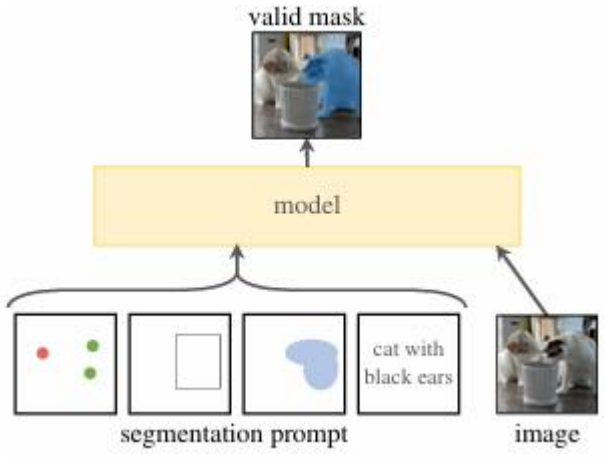


Fig 1. A promptable segmentation task

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4015-4026).

Segmentation

SAM

Segment Anything Model

Image Encoder

- Use an MAE pre-trained Vision Transformer

Prompt Encoder

- Sparse set(points, boxes, text): Represent sparse sets by positional encodings summed with learned embeddings for each prompt type and free-form text with a CLIP
- Dense set(masks): Embedded using Convolutions and Summed element-wise with the image embedding

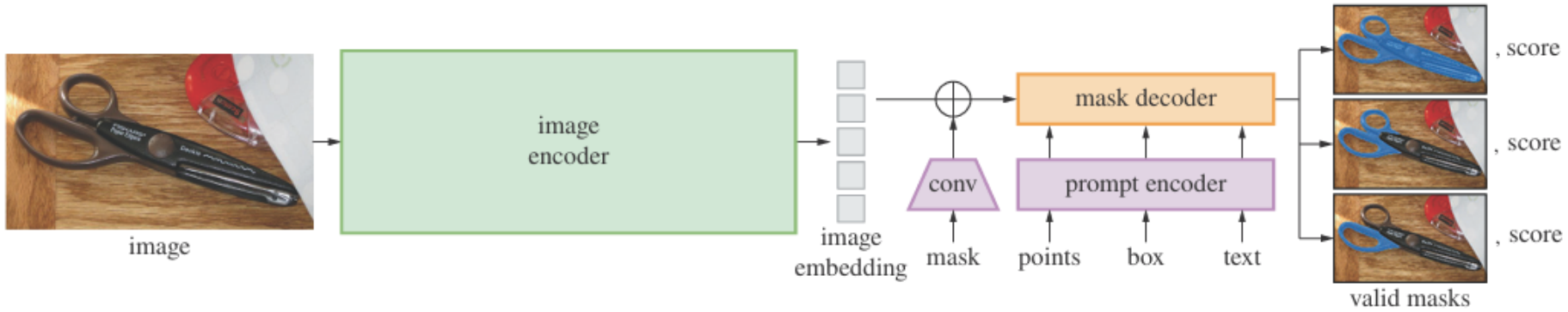


Fig 2. Segment Anything Model Overview

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4015-4026).

Segmentation

SAM

Segment Anything Model

Mask Decoder

- Efficiently maps the image embedding, prompt embeddings, and an output token to a masks
- Use prompt self-attention and cross-attention in two directions to update all embeddings

Resolving Ambiguity

- Modify the model to predict three output masks for a single prompt
- Backprop only the minimum loss over masks
- To rank masks, the model predicts a confidence score (i.e., estimated IoU) for each mask

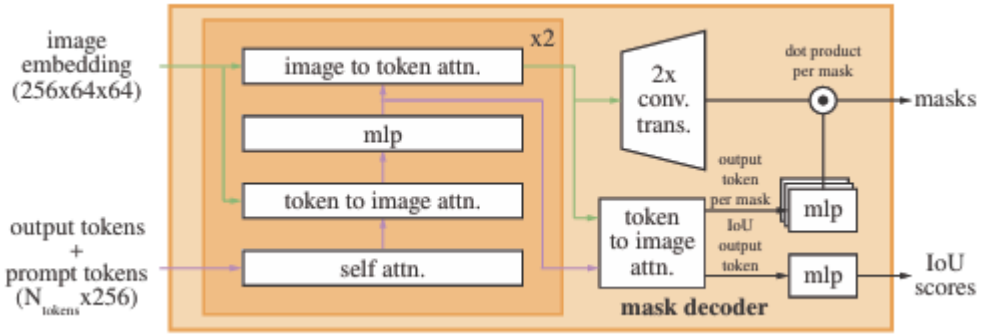


Fig 3. Details of the lightweight mask Decoder

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4015-4026).

Segmentation

SAM

Segment Anything Data Engine

Assisted-manual stage

- A team of professional annotators labeled masks by clicking foreground / background object points using a browser-based interactive segmentation tool powered by SAM → collect 4.3M masks from 120k images

Semi-automatic stage

- Automatically detect masks through the model and provide images to annotators, asking them to annotate additional unannotated objects → collect an additional 5.9M masks from 180k images

Fully automatic stage

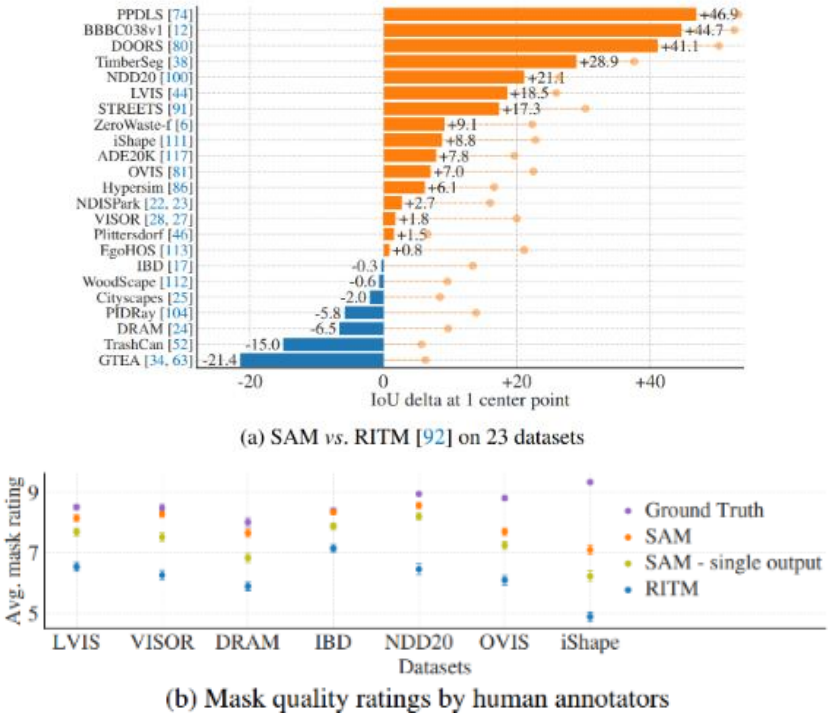
- Annotation was fully automatic → collect 11M images in dataset, producing a total of 1.1B high-quality masks

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4015-4026).

Segmentation

SAM

Experiments



(a) SAM *vs.* RITM [92] on 23 datasets

(b) Mask quality ratings by human annotators

Fig 4. Zero-shot single point valid mask evaluation

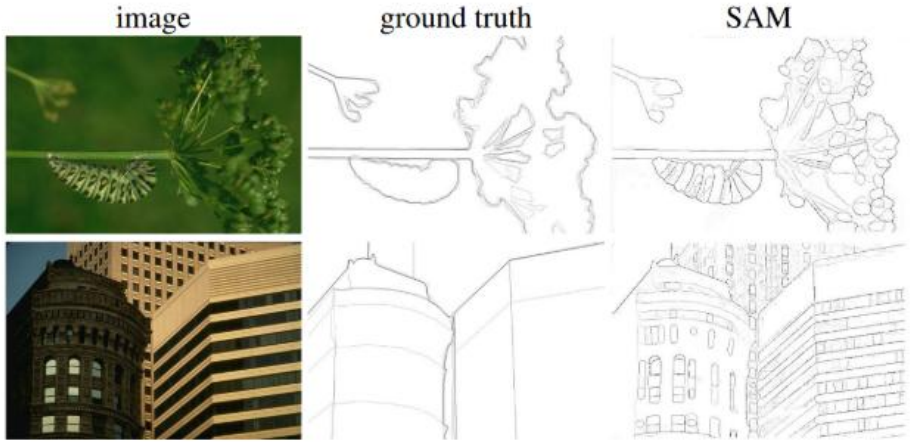


Fig 5. Results of Zero-shot Edge Detection

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4015-4026).

Segmentation

SAM

Experiments

method	mask AR@1000 all	small	med.	large	freq.	com.	rare
ViTDet-H [62]	63.0	51.7	80.8	87.0	63.1	63.3	58.3
zero-shot transfer methods:							
SAM – single out.	54.9	42.8	76.7	74.4	54.7	59.8	62.0
SAM	59.3	45.5	81.6	86.9	59.1	63.9	65.8

Table 1. Results of Zero-shot Object Proposals

method	COCO [66] AP	AP^S	AP^M	AP^L	LVIS v1 [44] AP	AP^S	AP^M	AP^L
ViTDet-H [62]	51.0	32.0	54.3	68.9	46.6	35.0	58.0	66.3
zero-shot transfer methods (segmentation module only):								
SAM	46.5	30.8	51.0	61.7	44.7	32.5	57.6	65.5

Table 2. Results of Zero-shot Instance Segmentation

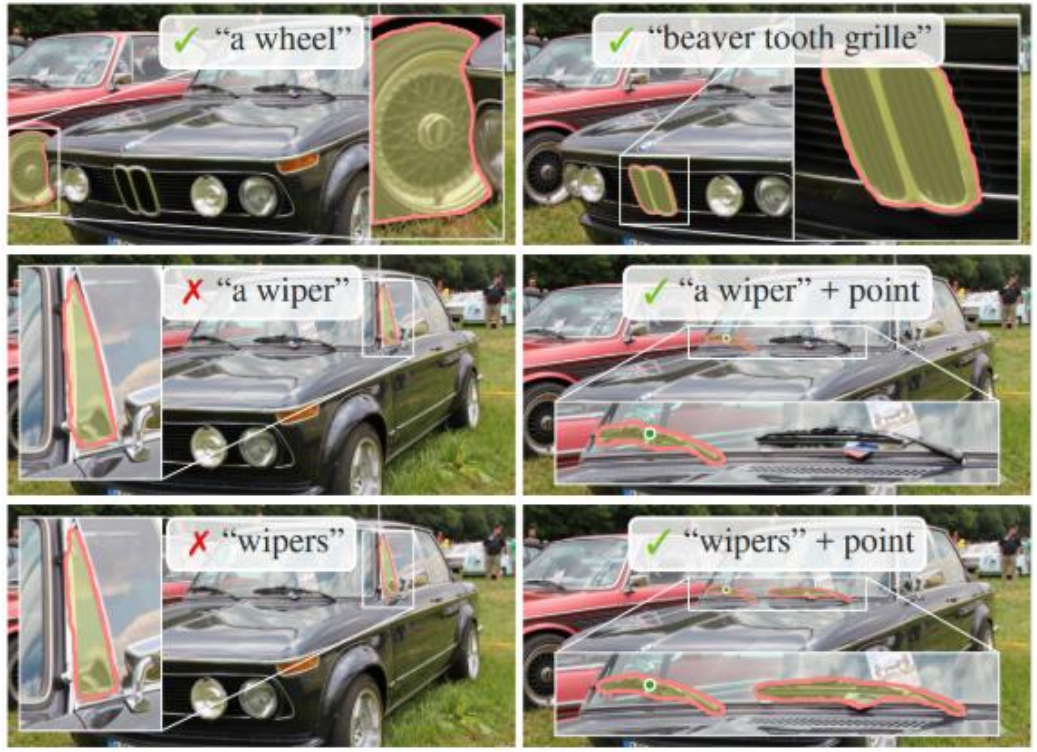


Fig 6. Results of Zero-shot Text-to-Mask

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4015-4026).

Segmentation

SAM

Implications & Limitations

Implications

- Attempt to lift image segmentation into the era of foundation models
- Principal contributions are a new task(promptable segmentation), model(SAM), and dataset(SA-1B)

Limitations

- SAM can miss fine structures, hallucinates small disconnected components, and does not produce boundaries as crispy as computationally intensive methods

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4015-4026).