A review of **Segment Anything (SAM)**, 2023

Zachary Parent

MAI

April 25, 2025

Outline

1. Introduction
2. The Task: Promptable Segmentation
3. The Model: SAM Architecture
4. The Dataset: SA-1B
5. Evaluation Results
6. Strengths
7. Limitations Future Work
8. Conclusion

The Vision: A Foundation Model for Segmentation

Goal: Build a foundation model analogous to LLMs for vision, specifically for segmentation.

- Like GPT, but for pixels.
- Enables zero-shot generalization via prompt engineering.
- Three core components: Task, Model, Dataset.

Released model (SAM) and dataset (SA-1B) to foster research.

Key Research Questions

1. What task will enable zero-shot generalization?
2. What is the corresponding model architecture?
3. What data can power this task and model?

Defining the Task: Promptable Segmentation

Input: An image and a prompt (point, box, mask, text).
Output: A valid segmentation mask.

- "Valid" means outputting a reasonable mask even if the prompt is ambiguous.
- Inspired by interactive segmentation, but goal is always predicting a valid mask immediately.

TODO: Add graphic illustrating prompt types (point, box, text, mask) leading to segmentation. (Fig 1 or similar)

Zero-Shot Transfer via Prompting

The core idea: Train a model on the general promptable task, then solve specific downstream tasks by designing the right prompts.

- Example: Use bounding box outputs from a cat detector as prompts to get cat instance segmentation.
- SAM acts as a composable component in a larger system.
- Can adapt to many (but not all) existing and new tasks.

Model Requirements & Design

Requirements:

- Flexible Prompts: Handle various input types.
- Real-time: Amortized real-time inference ($\sim$50ms) for interactivity.
- Ambiguity-aware: Output multiple valid masks for ambiguous prompts.

Design: Decoupled architecture for efficiency.

1. Heavyweight Image Encoder (ViT-H MAE pre-trained) - run once per image.
2. Lightweight Prompt Encoder (Positional embeddings, CLIP for text).
3. Lightweight Mask Decoder (Transformer) - combines image and prompt embeddings.

TODO: Add graphic illustrating the SAM model architecture (Fig 2).

Encoder/Decoder Details

Image Encoder: MAE pre-trained ViT. **Prompt Encoder:**

- Sparse (Points, Boxes, Text): Positional encodings + learned embeddings per prompt type. Text uses CLIP.
- Dense (Masks): CNN downscales mask, combines with image embedding.

Mask Decoder:

- Transformer decoder updates prompt tokens via self/cross-attention.
- Upscales image embedding.
- Predicts mask logits via MLP + dynamic linear classifier using output tokens.

Loss: Linear combination of Focal Loss and Dice Loss.

The Need for Data & The Data Engine

Problem: No web-scale segmentation data exists. Solution: Build a ”Data Engine” - iterate between model-assisted annotation and model improvement. **Three Stages:**

1. **Assisted-Manual:** Annotators label interactively with SAM's help. Model retrained iteratively.
2. **Semi-Automatic:** SAM proposes masks for likely objects; annotators focus on the rest.
3. **Fully Automatic:** Prompt SAM with a regular grid (32x32) of points → 100 masks/image.

TODO: Add graphic illustrating the data engine loop (Fig 4).

SA-1B Dataset

Result of the fully automatic stage.

- **1.1 Billion** high-quality masks.
- **11 Million** diverse, high-resolution images (licensed from provider).
- **400x** more masks than previous largest datasets (COCO, LVIS).
- Masks automatically generated and filtered for quality/stability (IoU, stability threshold).
- Geographically and economically diverse image sources.

Zero-Shot Evaluation

Extensive evaluation on 23 diverse segmentation datasets.

- Single point prompt performance often near manually annotated GT.
- Strong results on zero-shot downstream tasks via prompt engineering:
  - Edge Detection
  - Object Proposal Generation
  - Instance Segmentation
  - Text-to-Mask (preliminary)

TODO: Add a key result figure/table (e.g., comparing point prompt performance, Fig 6/7).

Strengths

- **Foundation Model Vision:** Pioneering promptable, general segmentation.
- **Unifying Task:** Promptable segmentation enables flexibility and zero-shot transfer.
- **Massive Dataset (SA-1B):** Unprecedented scale and diversity.
- **Strong Zero-Shot Performance:** Impressive generalization across many tasks/datasets.
- **Efficient Architecture:** Decoupled design allows real-time interaction.

Limitations & Future Work

- **Dataset Quality Assurance:** Transparency needed on automated mask filtering effectiveness.
- **Downstream Implementation:** Lack of practical integration examples.
- **Prompt Engineering:** Methodology underexplored.
- **Shallow Semantics:** Text understanding needs improvement.
- **Bias Analysis:** Surface-level; deeper sociotechnical audit needed.
- **Generalist vs Specialist:** May not beat specialized models in niche domains.

Conclusion

Segment Anything introduces a powerful new paradigm for image segmentation.

- Successfully implements a promptable task, model, and large-scale dataset.
- Demonstrates strong zero-shot capabilities.
- Opens up many avenues for future research in foundation models for vision, prompt engineering, and dataset quality.

A significant step towards general-purpose, interactive image understanding.

Thank You / Q&A

Questions?

Model and dataset available at:
`https://segment-anything.com`