SA

Segment Anything

2025.03.14

Seung min chung

Contents

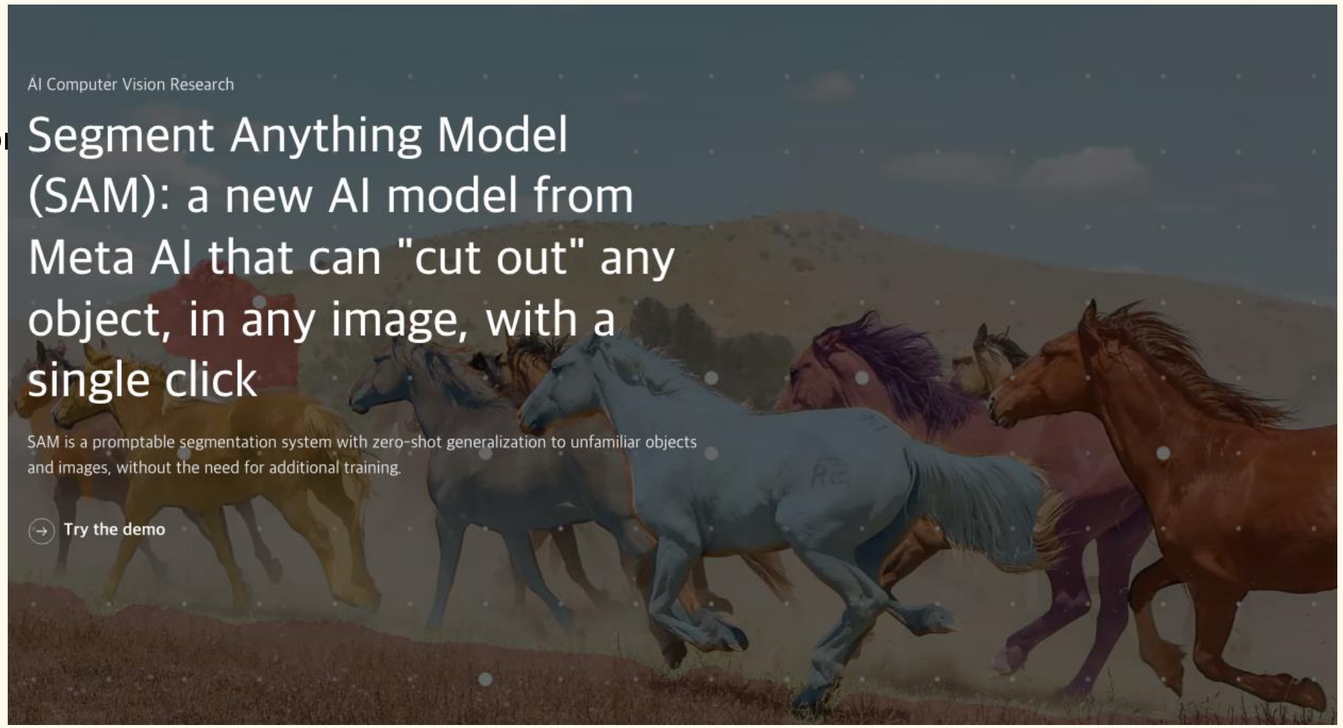
01 Abstract

SA project

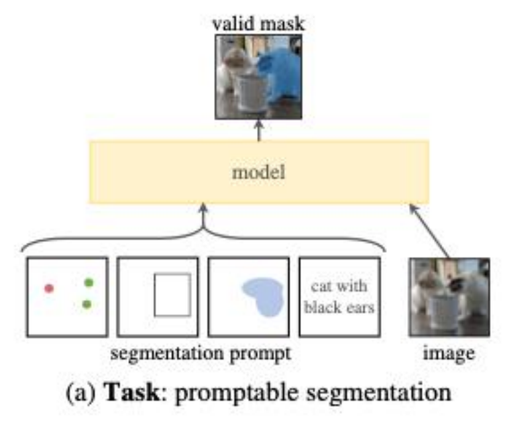
In this project, propose “foundation model” for segmentatio
model.

핵심

- promptable Model
- Task, Model, Dataset



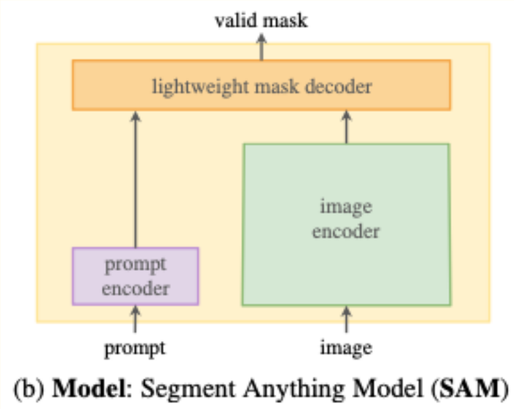
02 Introduction



(a) **Task**: promptable segmentation

In segmentation, various prompt are needed
Such as dots, text, boxes, or masks.
Thus, the foundation model is trained using prompt

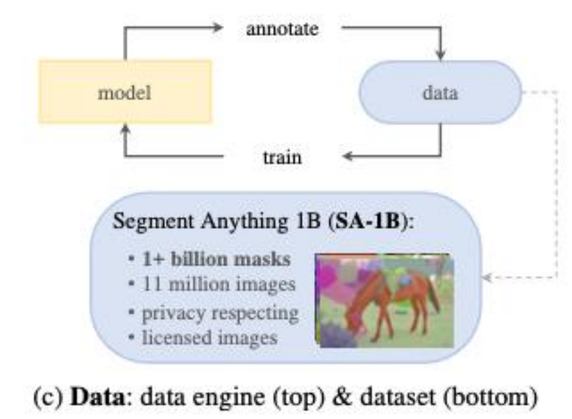
Task



(b) **Model**: Segment Anything Model (**SAM**)

SAM can treat flexible prompts and must be ambiguity-aware.

Model



(c) **Data**: data engine (top) & dataset (bottom)

To achieve string generalization to new data distributions., we found it necessary to train SAM on large and diverse set of masks.

Data

02 Introduction

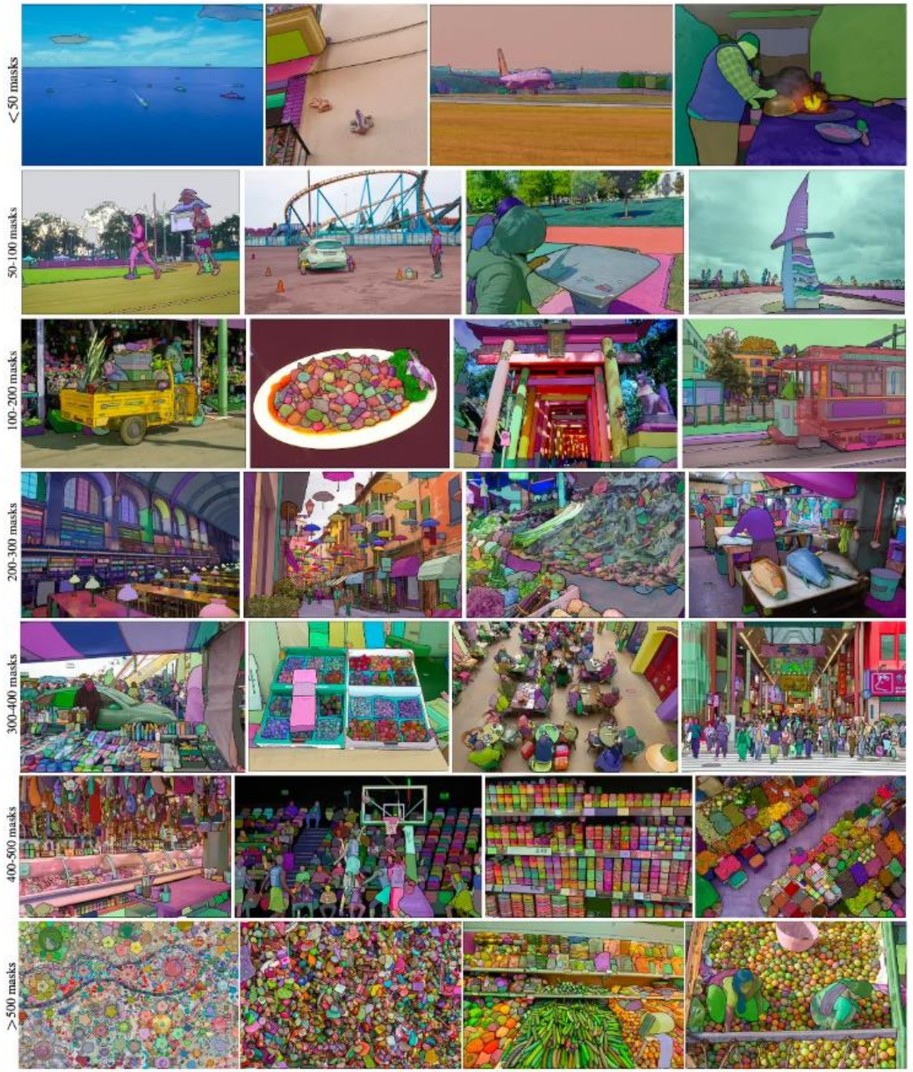


Figure 2: Example images with overlaid masks from our newly introduced dataset, **SA-1B**. SA-1B contains 11M diverse, high-resolution, licensed, and privacy protecting images and 1.1B high-quality segmentation masks. These masks were annotated *fully automatically* by SAM, and as we verify by human ratings and numerous experiments, are of high quality and diversity. We group images by number of masks per image for visualization (there are ~100 masks per image on average).

Promptable Task

This project is inspired by LLM models that predict the next token using prompts. SA’s goal is to segment anything, even when ambiguous prompts are provided, so that it can solve any sub-task

Pre-training

Various positive answer masks exist, and we automatically generate various prompts from each answer mask. The goal is to build a model that can accurately predict our answers when given these prompts.

04 segment anything model

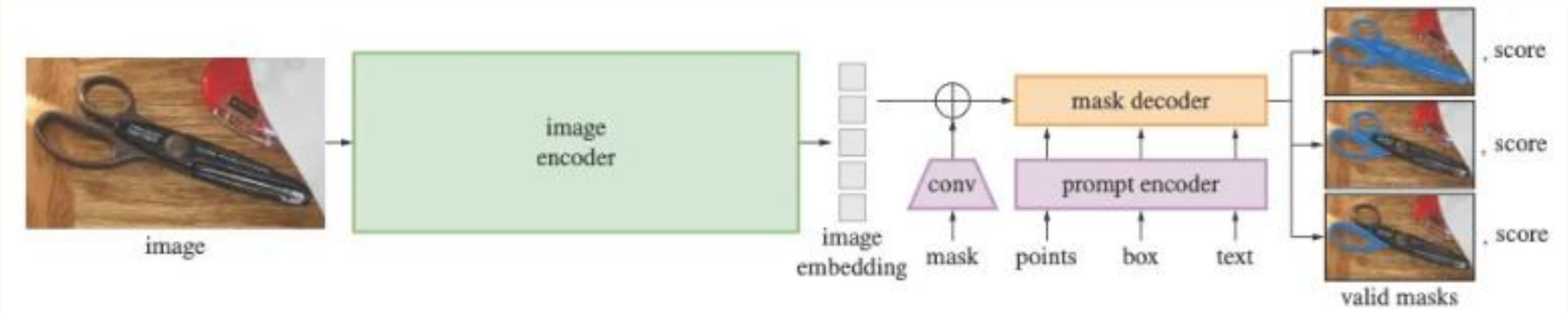


Figure 4: Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks at amortized real-time speed. For ambiguous prompts corresponding to more than one object, SAM can output multiple valid masks and associated confidence scores.

Image Encoder

Built on a pre-trained ViT using the MAE method, with slight modifications to effectively handle high-resolution images.

Prompt Encoder

Uses a flexible encoder that applies different embedding methods based on prompt type.

• Divides prompts into Sparse (points, boxes, text via CLIP encoder) and Dense (mask processed with convolution and pixel-level fusion).

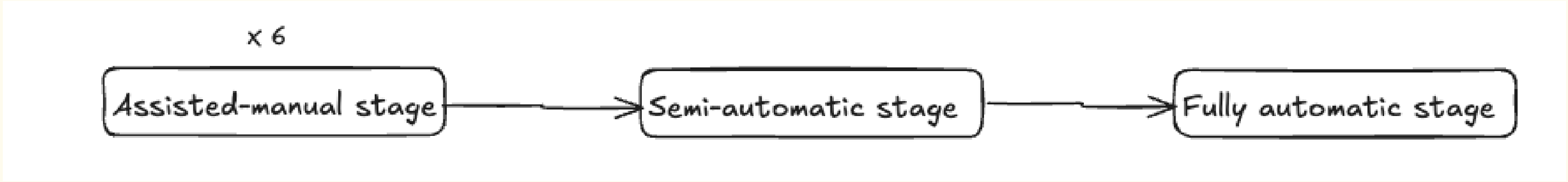
Mask Decoder

A modified Transformer decoder enables bidirectional attention between prompt and image embeddings, updating both.
• Upsampled image embeddings are mapped via an MLP-based dynamic linear classifier to predict per-pixel object probabilities, forming the final segmentation mask.

Losses and Training

The model is trained with a loss function that linearly combines Focal Loss and Dice Score.
• For each mask, various prompt types are input, with 11 random prompt simulations integrated into the training pipeline.

05 segment anything task data engine



Assisted-Manual Stage

A pre-trained ViT model extracts image features from public datasets, and human annotators label key regions manually.
• Six cycles of training and annotation improved efficiency, resulting in 4.3M masks from 120k images.

Semi-Automatic Stage

Confident masks are automatically generated, while annotators only refine areas with less salient objects.
• This approach supplements automatically detected masks to enhance overall data diversity and quality.

Fully Automatic Stage

Using a 32×32 grid-based prediction, the model generates multiple masks per point and automatically selects the correct ones based on confidence.
• This stage allows for large-scale, high-quality mask generation without any human intervention.

06 segment anything task dataset

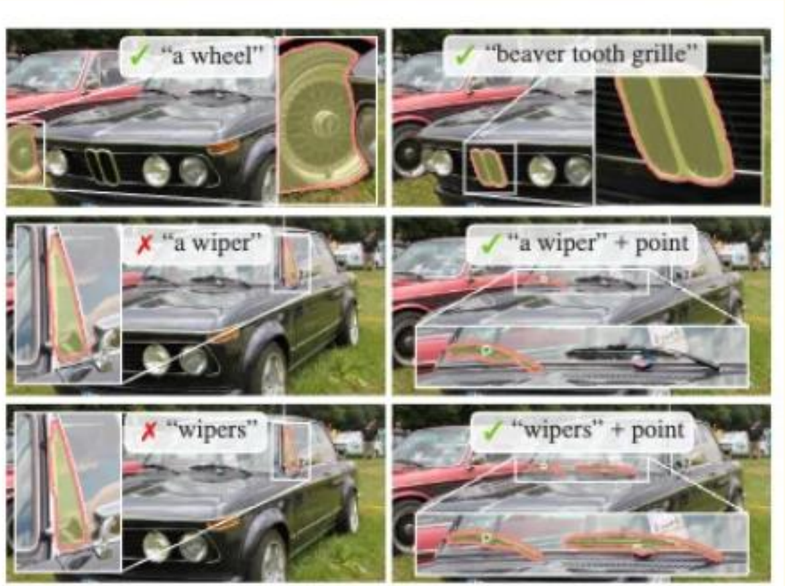


Figure 12: Zero-shot text-to-mask. SAM can work with simple and nuanced text prompts. When SAM fails to make a correct prediction, an additional point prompt can help.

Figure 10: Zero-shot edge prediction on BSDS500. SAM was not trained to predict edge maps nor did it have access to BSDS images or annotations during training.

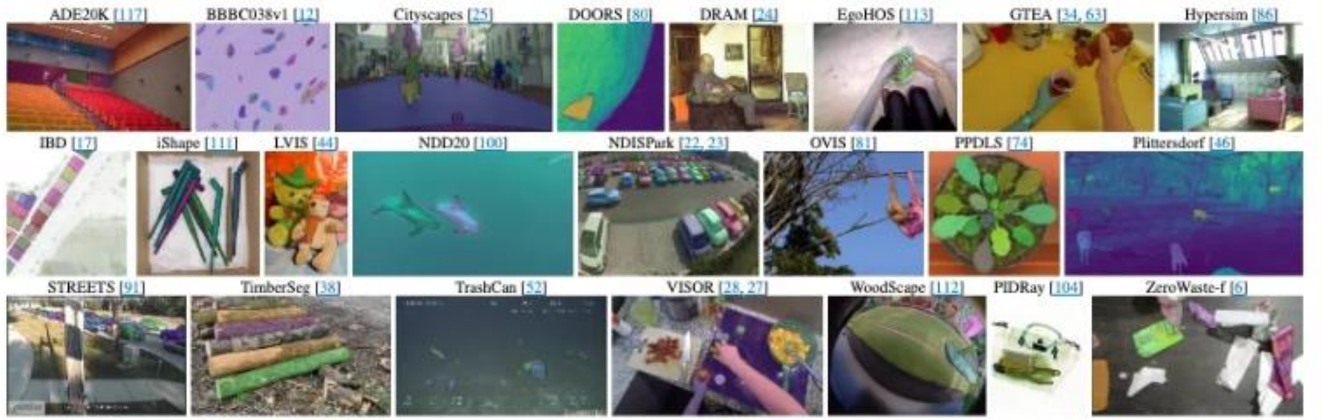


Figure 8: Samples from the 23 diverse segmentation datasets used to evaluate SAM's zero-shot transfer capabilities.

- The dataset boasts higher resolution than other datasets and includes a variety of licenses.
- Images are provided on the web, downsampled to a minimum side length of 1500 pixels.
- Additionally, 99.1% of the masks are automatically generated, with 94% of them achieving an IOU greater than 90, indicating high mask reliability.

07 summary

- The concept of foundation models is highlighted along with the significance of the SAM model.

- SAM is a universal model designed for various downstream segmentation tasks through large-scale supervised learning and prompt-based segmentation, effectively segmenting objects within images and integrating easily with other systems.

- Its limitations include challenges in handling fine structures or complex boundaries perfectly and lower performance in some domains compared to specialized tools.

- Ultimately, by releasing over 1B masks and the SA-1B dataset, the project aims to usher in a new era of foundation models in image segmentation.

Thanks