Segment Anything

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, Ross Girshick

Presented to: Rutleia
Presented by: Rogers Aloo
Date: May 2023

Source: SAM Original Paper

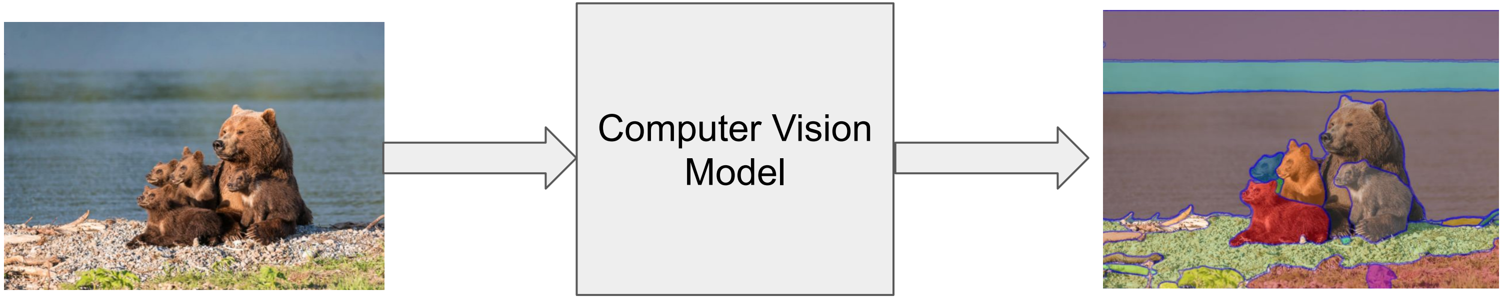
Outline

1. Explain SAM model and theory
    1.1. Introduction, Prompting & Segmentation
    1.2. SAM model Architecture
    1.3. Training
    1.4. SA-1B dataset
    1.5. Zero-shot transfer experiments & results
    1.6. Results
2. Explain SAM code
3. Fine tune methods for SAM
4. Applications of SAM on MVTEC dataset

Introduction

- Foundation models in NLP common with zero-shot learning via prompting ie chat gpt.
- NLP is successful due to scale of data and no labels.
- FM models for computer vision is a problem due to annotations & masks.

prompting for zero-shot a problem in vision

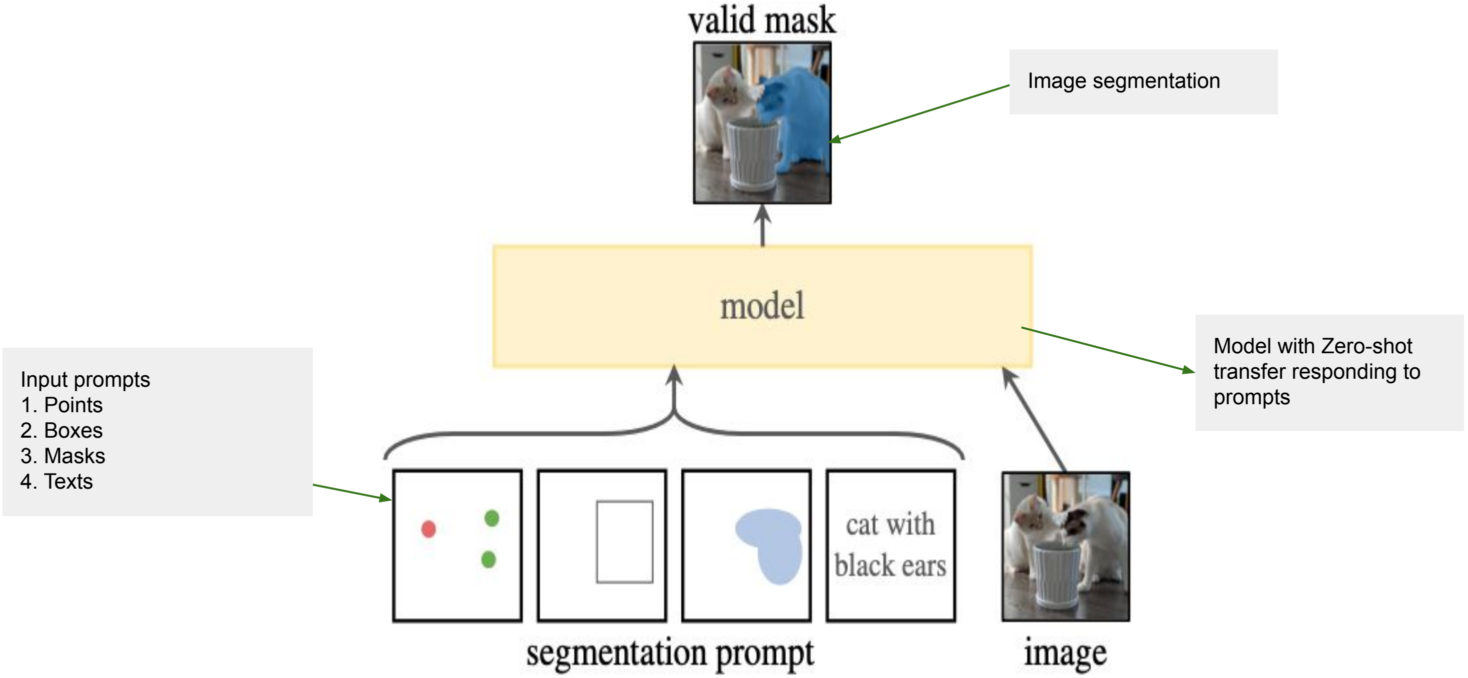


- SAM “Foundation model” trained on diverse dataset for segmentation problems.
- Goal: Seek to develop a promptable model and pre-train it on a broad dataset using a task that enables powerful generalization.

Sourcee: SAM official [website demo]

Prompting and segmentation

- In computer vision prompting of points, masks, texts, boxes can be used for image segmentation.



Source: Task Image, Segment Anything pg 16. Taken from the [original paper.]

SAM Model

- Has three components image encoder, flexible prompt encoder & fast mask decoder.

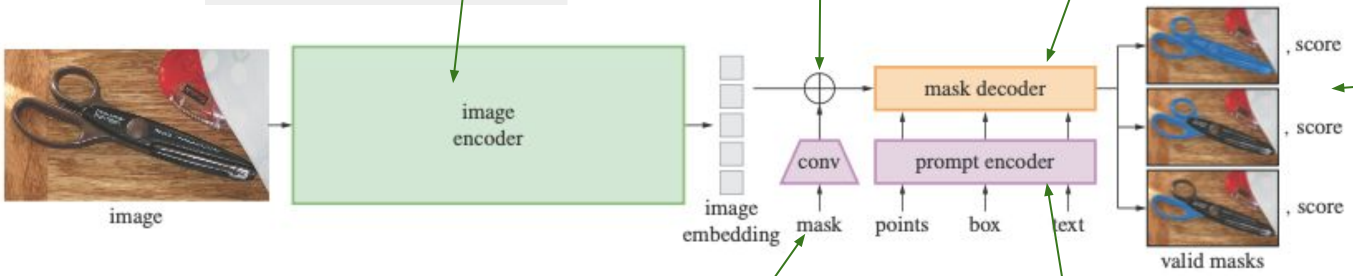
1. Input image to embeddings
Use Masked autoencoder(MAE) Vision Transformer(ViT) handling HR images

4. Element Wise summation of embeddings and prompts

5. Decoder up-samples embedding to original image size
Modified ViT block

6. Trained on Focal + Dice Loss (20:1)
Output has 3 masks to remove ambiguity

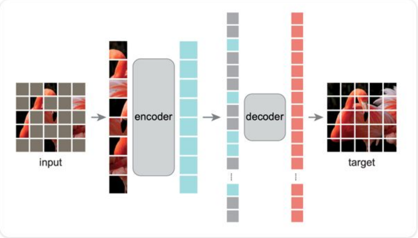
Model trains on three level of granularity (deep, part and

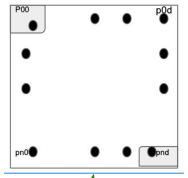


2. Dense prompts mask use convolutions

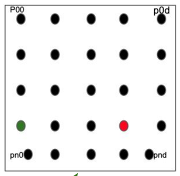
3. Sparse Prompts
Points & box use positional ecodings
Texts use clip encodings

Masked autoencoder(MAE) Vision Transformer(ViT) h
Encoder masked patches
decoder unmasked patches + embeddings





BOX: Sum of positional encoding + learned embedding indicating foreground or background



POINTS: Sum of positional encoding + learned embedding indicating foreground or background
Sample points using NMS

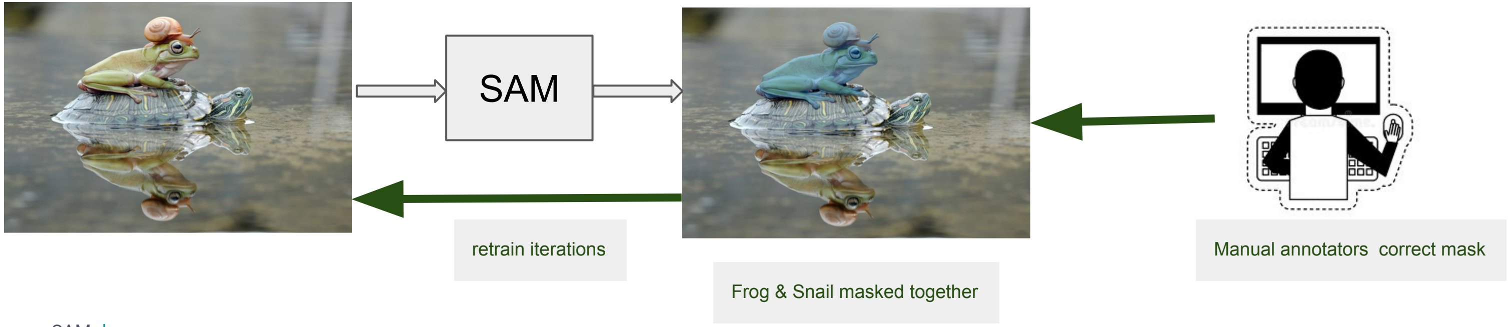
Source: Image 2, MAE architecture original paper. | Box & point description self interpretation

SAM Training - (1/3)

- SAM training different from convention since its a foundation model. Lack of public dataset hence built data engine(with three stages).

1. **Assisted Manual stage**

- Initially trained on commonly available datasets. Manual annotators correct masksmasks via web interaction tools.
- Retrain the model with obtained data in **6 iterations** (encoder increase from ViT_b to VitH).
- Output 120k images, 4.3M masks, approx 44 masks per image.

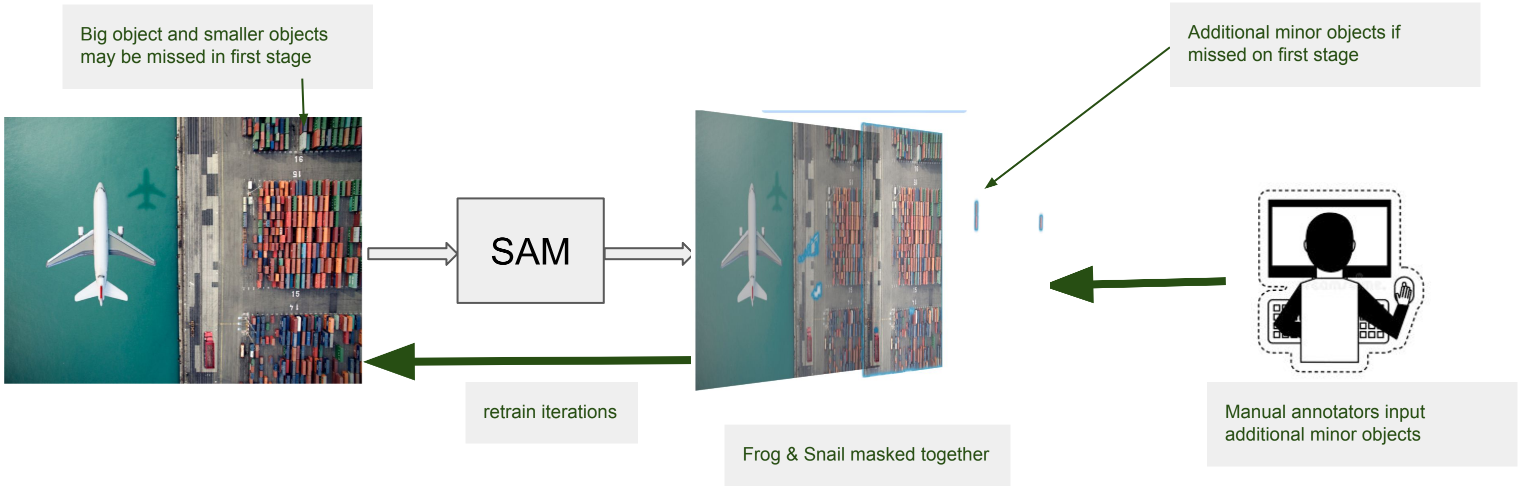


Source: Images SAM demo page

SAM Training - (2/3)

2. **Semi automatic stage**

- Improve diversity of Masks. Annotators label additional unlabelled objects.
- Retrain the model with added label in **5 iterations**.
- Output 180k images, 5.9M masks, approx 72 masks per image.

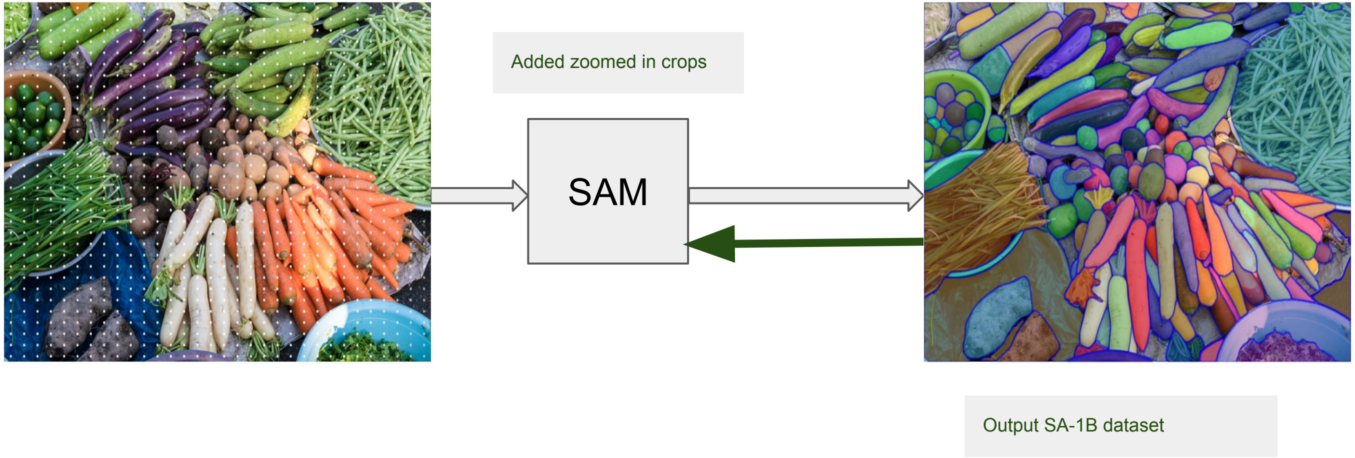


Source: Images SAM demo page

SAM Training - (3/3)

3. **Fully automatic stage**

- Introduce prompting with 32x32 grids.
- Added zoomed in crops on images to improve quality.
- Output 11M images, 1.1B masks.



Source: Images SAM demo page

SA-1B dataset - (1/2)

- SAM produces foundation model and dataset (SA-1B dataset).
- Size 11M images, 1.1B masks 99.1% auto-generate. High resolution images(3300 x 4950).
- Evenly distributed masks on the images. (photographer bias).
- Random sample 500 to mask quality. Based on IoU threshold of pair between masks vs professional annotaters.
    - 94% of pair have greater than 90% IoU.
    - 97% of pair have greater than 75% IoU.

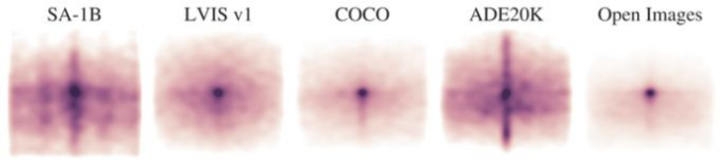
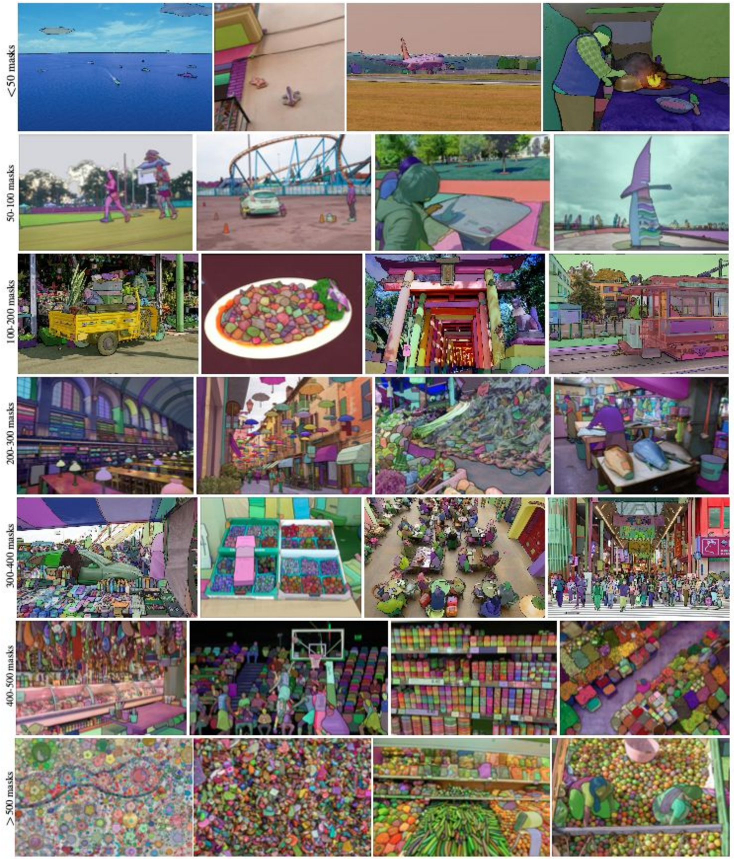


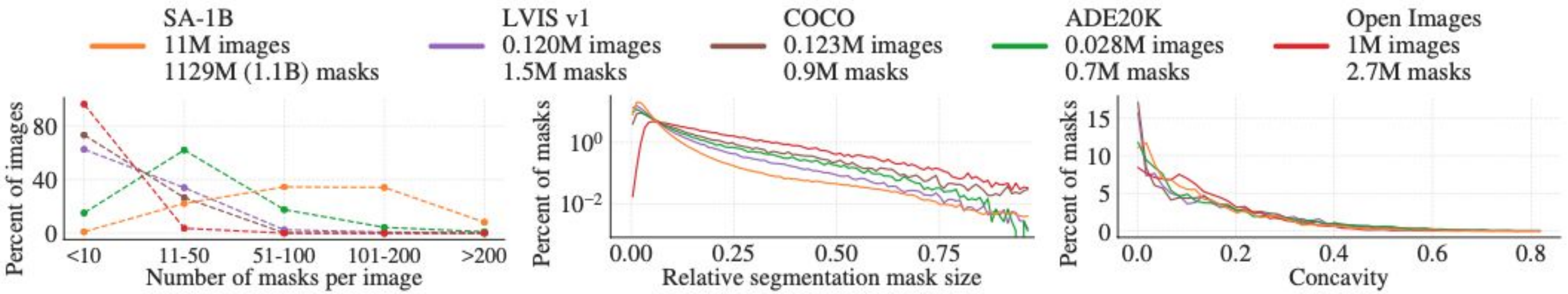
Figure 5: Image-size normalized mask center distributions.



Source: Images [SAM paper]

SA-1B dataset - (2/2)

- SA-1B has 11× more images and 400× more masks than the second largest, Open Images.
- More masks per image, it also tends to include a greater percentage of small and medium relative-size masks.
- SA-1B includes all regions 10× of previous datasets. Average masks/image fairly consistent on all regions.



Source: Images [SAM paper]

Zero-shot transfer Results - (1/5)

1. Zero-Shot Single Point Valid Mask Evaluation

- Single input point and mask outputs. Evaluation and SAM is superior in 16/23 datasets compared to RITM.

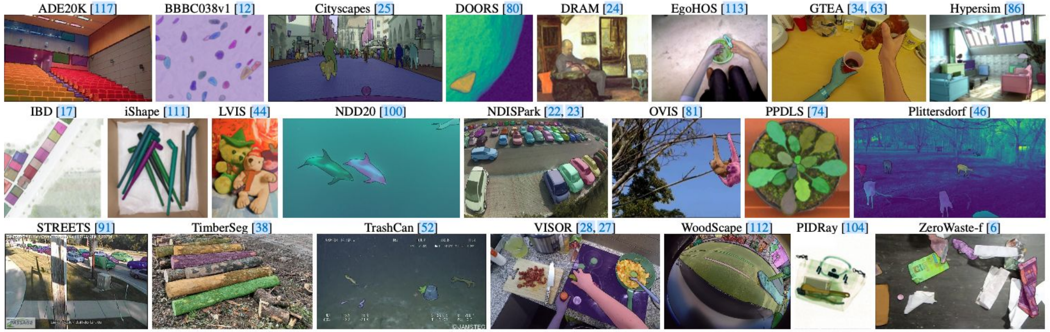
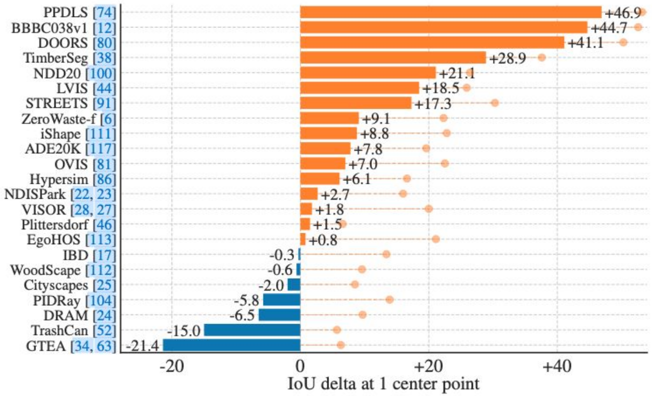
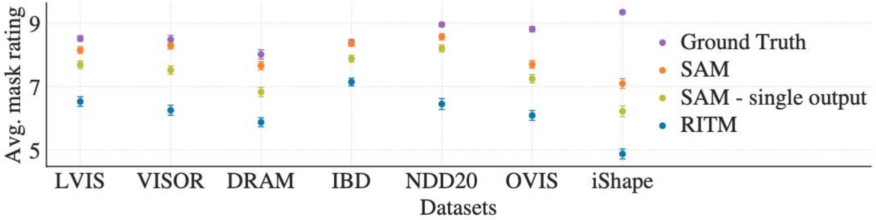


Figure 8: Samples from the 23 diverse segmentation datasets used to evaluate SAM's zero-shot transfer capabilities.



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



(b) Mask quality ratings by human annotators

Source: Images SAM paper

Zero-shot transfer Results - (2/5)

2. Zero-Shot Edge Detection

- Identify edges in the image. Evaluate on BSDS500 dataset with 16X16 grid prompts resulting in 768 predicted masks(3 per point).
- SAM doesn't understand edges to suppress because its a FM and doesn't learn dataset bias.
- High recall R50 is high to precision and trail state-of-the-art methods for bias learning of BSDS500.

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.

method	year	ODS	OIS	AP	R50
HED [108]	2015	.788	.808	.840	.923
EDETR [79]	2022	.840	.858	.896	.930
zero-shot transfer methods:					
Sobel filter	1968	.539	-	-	-
Canny [13]	1986	.600	.640	.580	-
Felz-Hutt [35]	2004	.610	.640	.560	-
SAM	2023	.768	.786	.794	.928

Table 3: Zero-shot transfer to edge detection on BSDS500.

Source: Images SAM paper

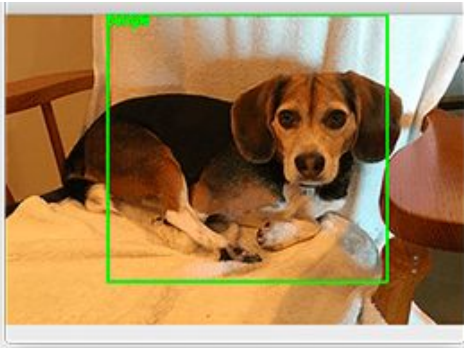
Zero-shot transfer Results - (3/5)

3. Zero-Shot Object Proposal

- SAM modified to convert output masks as proposal bounding boxes.
- Use LVIS dataset to evaluate and compare to ViTDet-H detector on recall metric.
- Outperforms ViTDet-H detector on medium and large tasks only.
- “Ambiguity unaware”-SAM performs worse than SAM.

	mask AR@1000						
method	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 4: Object proposal generation on LVIS v1. SAM is applied zero-shot, *i.e.* it was not trained for object proposal generation nor did it access LVIS images or annotations.



Source: Images SAM paper | Object proposal Image

Zero-shot transfer Results - (4/5)

4. Zero-Shot Instance Segmentation

- Instant segmentation same objects of a particular class within an image.
- Prompt SAM with output from object proposal ie ViTDet-H and evaluate with AP metric.
- SAM performs behind VitDet-H though have more crispier mask boundaries(Evaluate claim with human rating)

	COCO [66]				LVIS v1 [44]			
method	AP	AP^S	AP^M	AP^L	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 5: Instance segmentation results. SAM is prompted with ViTDet boxes to do zero-shot segmentation. The fully-supervised ViTDet outperforms SAM, but the gap shrinks on the higher-quality LVIS masks. Interestingly, SAM outperforms ViTDet according to human ratings (see Fig. 11).

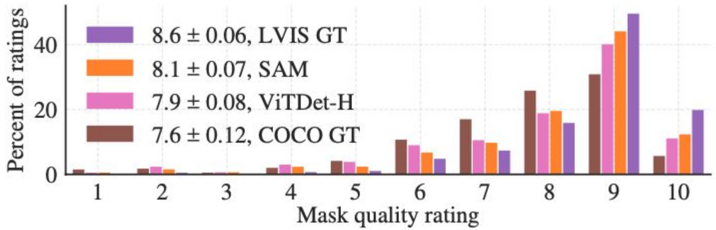


Figure 11: Mask quality rating distribution from our human study for ViTDet and SAM, both applied to LVIS ground truth boxes. We also report LVIS and COCO ground truth quality. The legend shows rating means and 95% confidence intervals. Despite its lower AP (Table 5), SAM has higher ratings than ViTDet, suggesting that ViTDet exploits biases in the COCO and LVIS training data.

Source: Images SAM paper

Zero-shot transfer Results (5/5)

5. Zero-Shot Text-to-Mask

- PoC task to test SAM ability in utilizing text prompts.
- Utilize CLIPs image embedding to align to text embeddings.
- SAM segments points using the text but in failure cases an additional point fixes prediction.

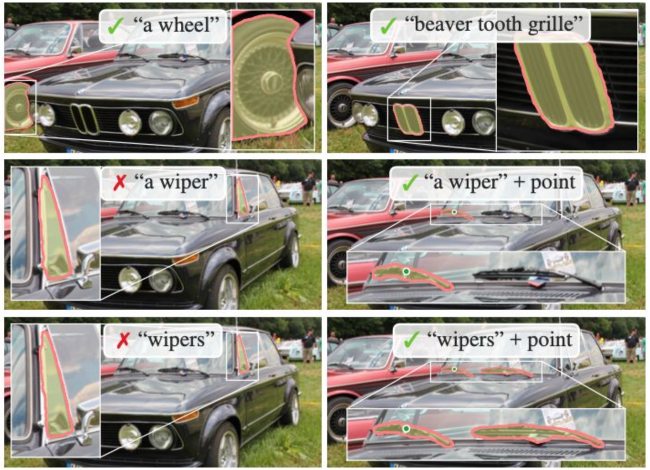
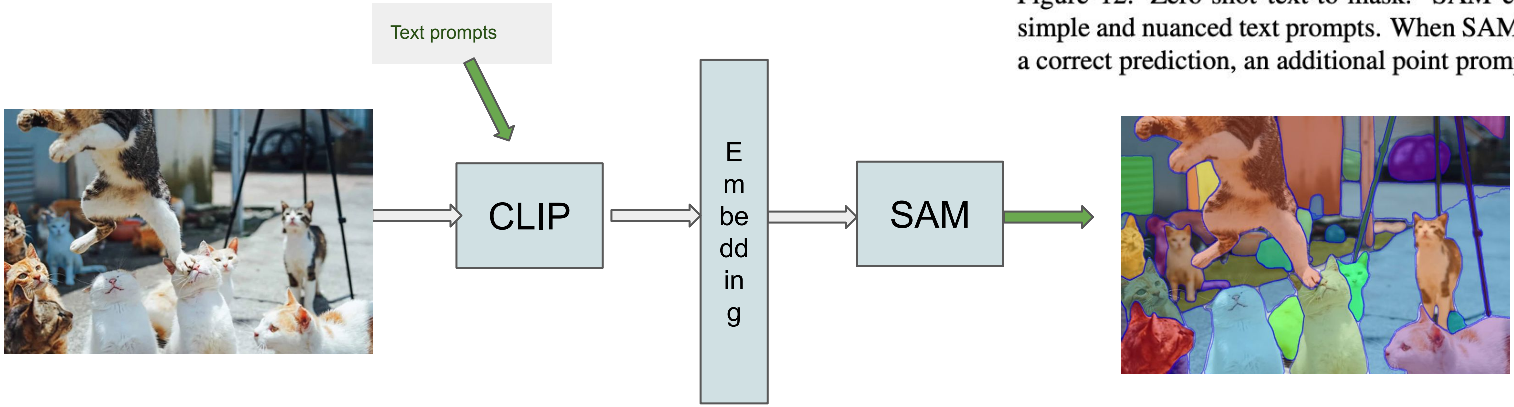


Figure 12: Zero-shot text-to-mask. SAM can work with simple and nuanced text prompts. When SAM fails to make a correct prediction, an additional point prompt can help.



Source: Images [SAM paper]

SAM code explanation - (1/2)

- Segment-anything.modelling
    1. common.py - *MLPBlock*, *LayerNorm2d* for image decoder
    2. Image_encoder.py - *ImageEncoderViT*, *Block*, *Attention*, *PatchEmbed*
    3. Prompt_encoder.py - *PromptEncoder*, *PositionEmbeddingRandom*
    4. mask_decorder.py - *MaskDecoder*, *MLP*
    5. sam.py - *Sam* end-end model, predictor with pre & post processing
    6. Transformer.py - *TwoWayTransformer*, *TwoWayAttentionBlock*, *Attention*
- Scripts
    1. amg - auto-generate masks for the data engine in the 3rd step
    2. export_onnx_model - convert SAM (pytorch) to onnx form

Source: Adapted from SAM github repo

SAM code explanation (2/2)

- segment-anything.modeling
  1. automatic_mask_generator..py - *SamAutomaticMaskGenerator*
  2. build_sam.py - model versions, concat (image encoder, prompt encoder, image decoder)
  3. predictor.py - *SamPredictor*, use SAM to calculate image embeddings > mask prediction
- Notebooks
  1. Automatic_mask_generator_example.ipynb- amg of segmentation masks
  2. onnx_model_example.ipynb - ONNX format generate masks from prompts > web runtime
  3. predictor_example.ipynb - *SamPredictor*, mask predictor given object prompts

Source: Adapted from SAM github repo

Fine Tuning SAM

- Foundation models untrained on all datasets. SAM trained on 23 diverse datasets.
- Fine-tuning allow adaptation to specified tasks.
- Don't unfreeze encoder on similar dataset override the learning rate and vice versa.

Train procedure

1. Wrap image encoder with no gradient flow.
2. Generate prompt embeds with no grad flow.
3. Generate the masks -(use case output mask 1).

Hyper-params

1. Learning rate to 1e-06 (SAM - 8e-4 > 8 e-3 > 8e-2).
2. Change image decoder loss type to MSE (SAM - Focal : dice).
3. Adam optimizer on image decoder (SAM - AdamW).

*Tuning of amg masks *SamAutomaticMaskGenerator* algorithms possible to generate more masks.

Application on MVTEC AD dataset

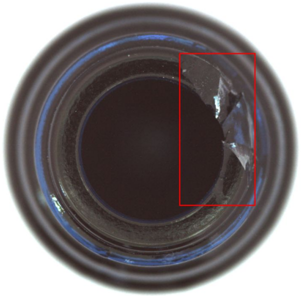
- Apply & Fine tune SAM on MVTEC AD dataset specifically identifying broken bottles.
- Tuning improves masking of broken bottles (1e-06 (SAM - 8e-4 > 8 e-3 > 8e-2) >MSE (SAM - Focal : dice)> Adam optimizer on image decoder (SAM - AdamW)..

Ground Truth broken bottle

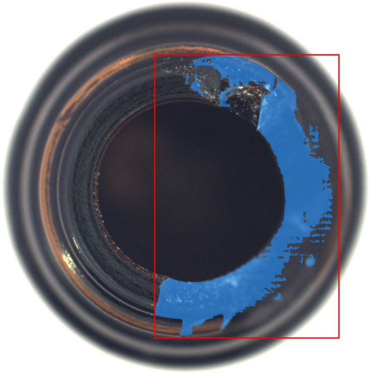
Normal Bottle

Ground Truth Mask

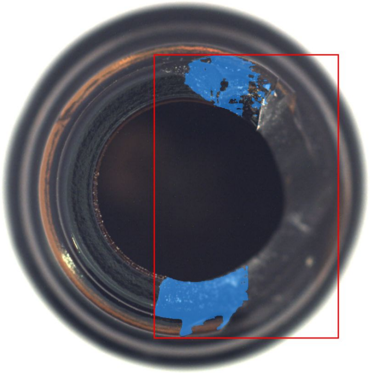
Ground Truth bounding box



Mask with Tuned Model



Mask with Untuned Model



Open in Colab

Source: Colab Notebook |Dataset MVTec AD

Reference

- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. arXiv preprint [arXiv:2304.02643]
- He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 16000-16009) arXiv:2111.06377
- Pal, A., & Balasubramanian, V. N. (2019). Zero-shot task transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2189-2198) arXiv:1903.01092
- Fine Tuning github repo
- Segment Anything web_ui

ありがとうございます