SEGMENT ANYTHING

Meta AI

Group 3

2019313007 SEUNGKWON YANG
2021310397 EUNSUR KIM
2021313014 HAEMIN CHOI

CONTENTS

INTRODUCTION

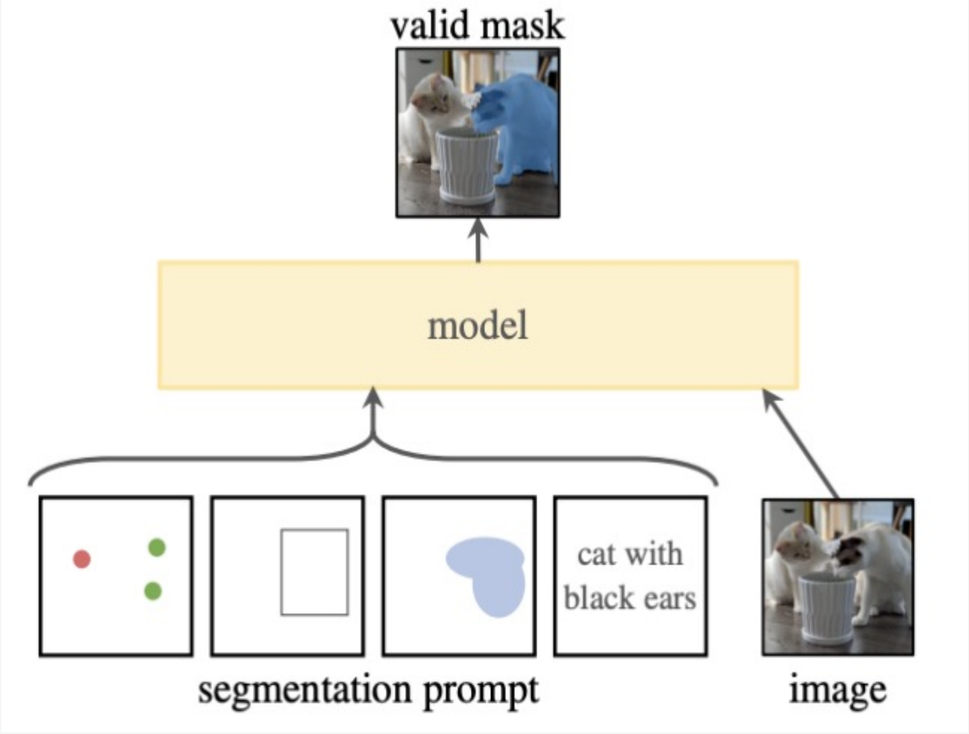
✧ GOAL : To build a foundation model for image segmentation
→ promptable model, enable zero-shot learning

What is a foundation model?

- A model that can powerfully generalize to tasks and data distributions beyond those seen during training

- Zero-shot learning: Making models to be able to perform tasks that has not learned in the course of learning

INTRODUCTION



What is a prompt?
Designating what to segment within the image
→ An information specifying the segmentation target

Sparse prompt	Dense prompt
Points, Boxes, Text	Masks

INTRODUCTION

How to success the plan? (Authors threw three questions)

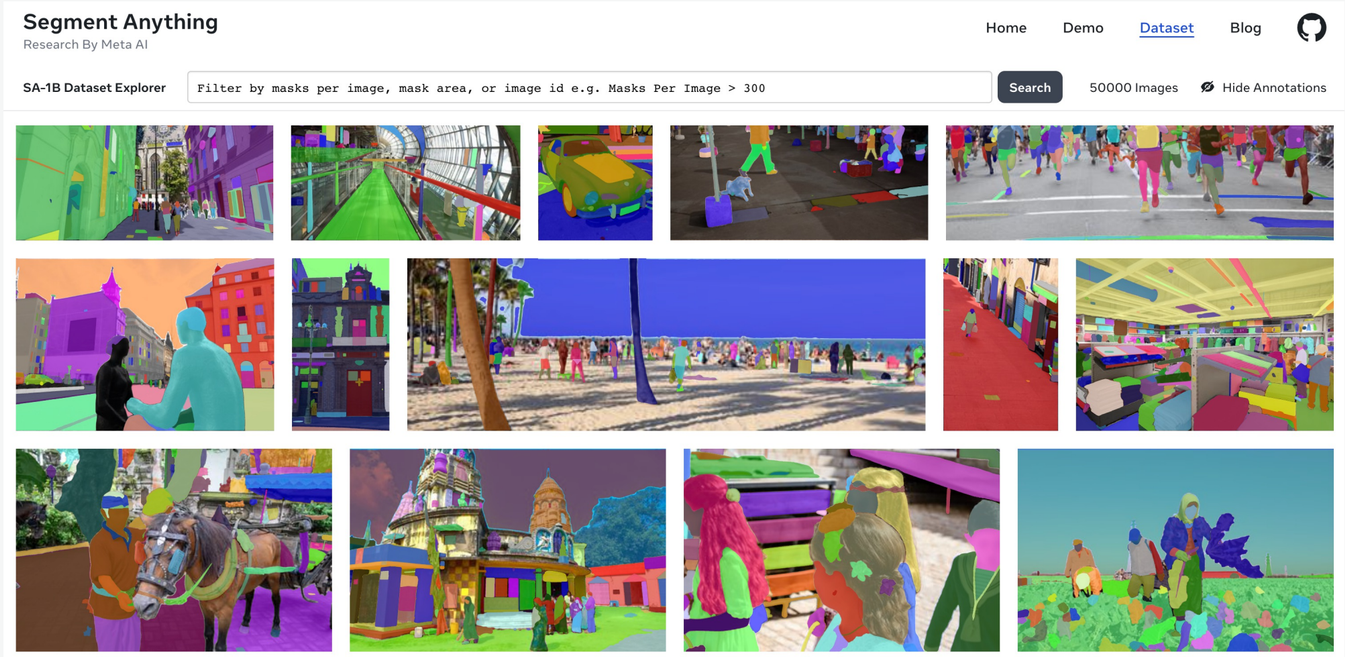
- ✦ What task will enable zero-shot generalization?
- ✦ What model is suitable to use?
- ✦ What data will fit to this task and model?

METHOD1 - SA-1B DATASET

SA-1B DATASET

- **11M images & 1.1B segmentation masks** in high quality
- 99.1% of the masks were generated automatically
- 94% of pairs(automatically predicted & professionally corrected) have greater than 90% IoU

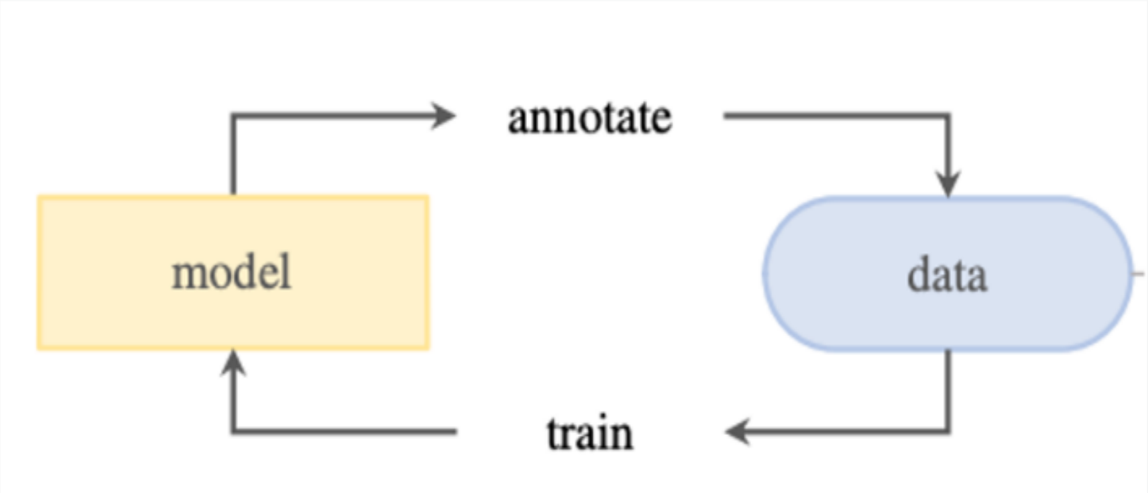
IoU (Intersection over Union) : A metric that evaluates overlap of the ground truth and prediction region



METHOD1 - SA-1B DATASET

✦ Why SA-1B?

Segmentation masks are not abundant on the internet
→ Built a **data engine** to collect mask dataset



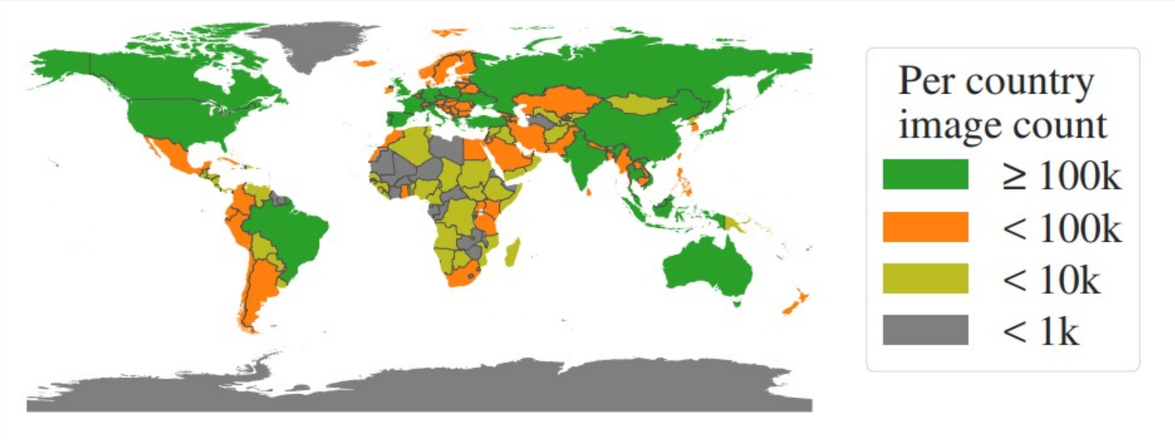
METHOD1 - SA-1B DATASET

Segment Anything Responsible AI Analysis

To be equitable for real-world (reduce biases)

1. Geographic and income distribution of countries

→ SA-1B has higher representation in different parts of the world

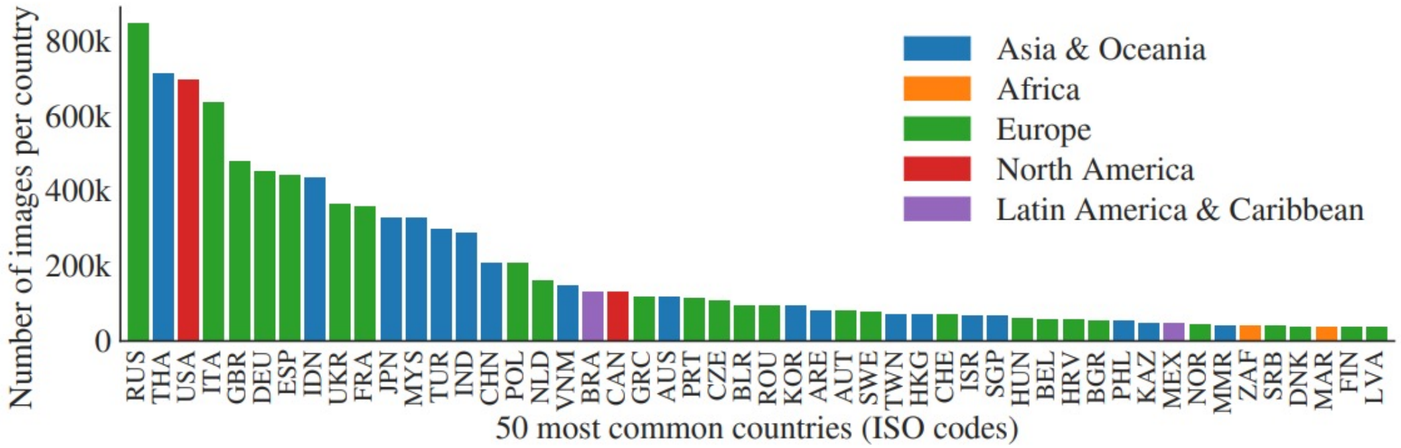


METHOD1 - SA-1B DATASET

Segment Anything Responsible AI Analysis

To be equitable for real-world (reduce biases)

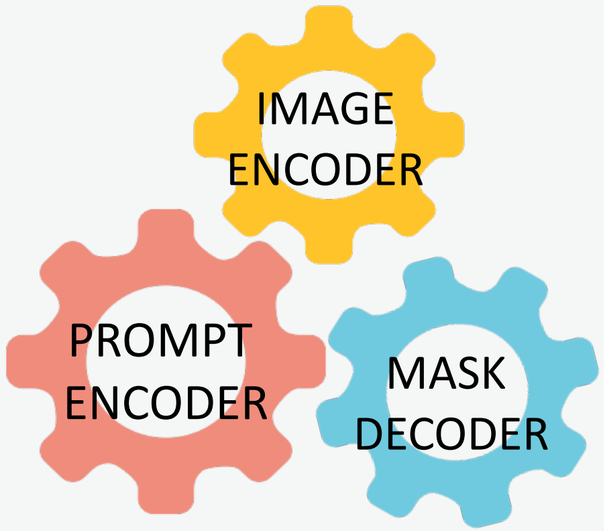
2. Fairness of SAM across different groups of people (gender, age group, skin tone)
   - Gender groups -> Similar performance
   - Age groups -> Performs best on elders
   - Skin tone -> No significant difference across groups found

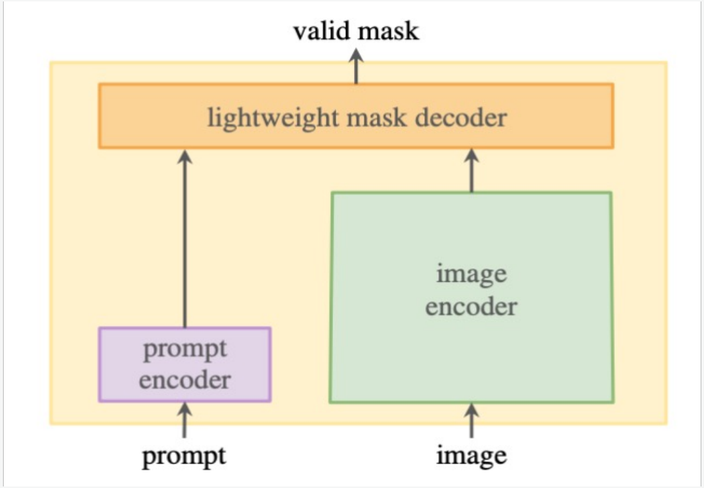


METHOD2 - SEGMANT ANYTHING MODEL

✦ Segment Anything Model? SAM!

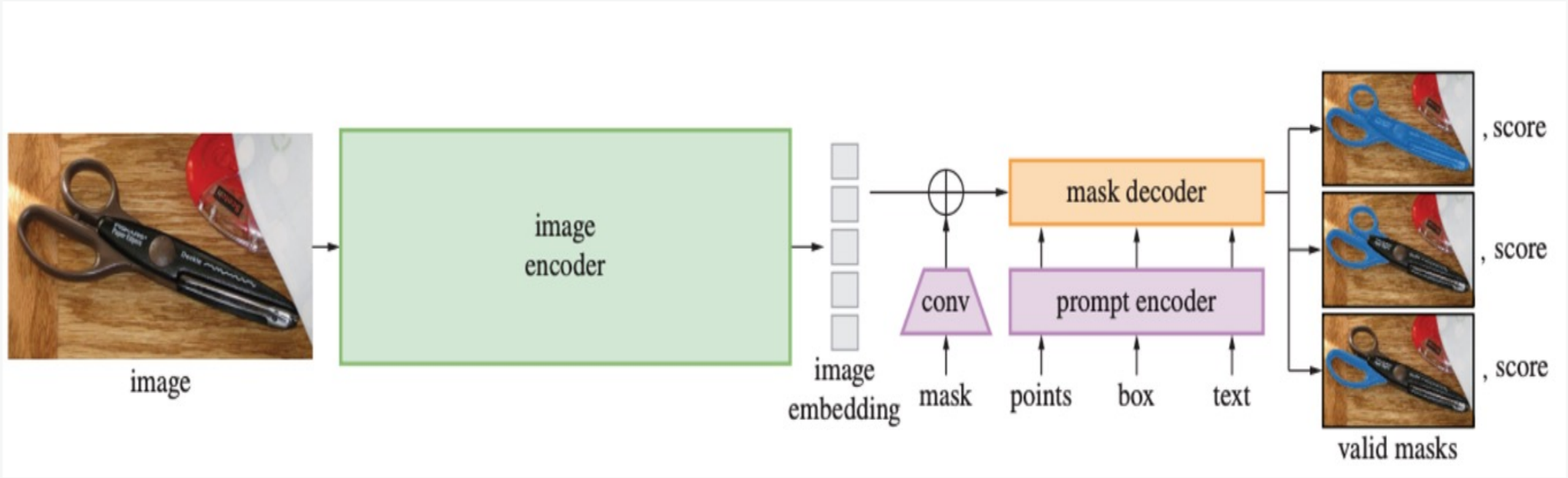
SAM uses Masked Auto Encoding pre-trained ViT as a baseline model.





- Support flexible prompts + Compute masks in real-time + Ambiguity-aware
- Zero-shot performance

METHOD2 - SEGMANT ANYTHING MODEL



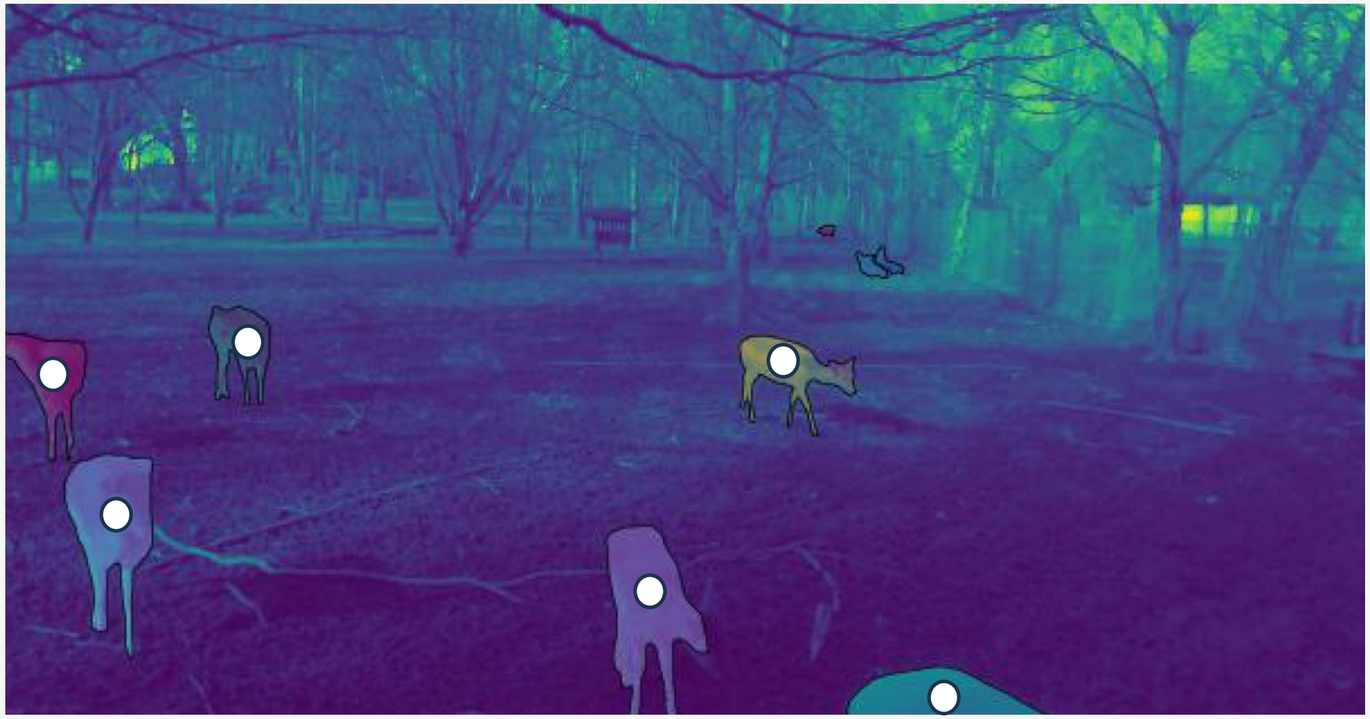
	input	output
image encoder	image	image embedding
promt encoder	prompt	prompt embedding
mask decoder	image embedding & prompt embedding	Segmentation mask

EVALUATION - ZERO-SHOT EXPERIMENTS

ZERO-SHOT SINGLE POINT VALID MASK EVALUATION

Approach

Extract foreground segmentation mask with single foreground point



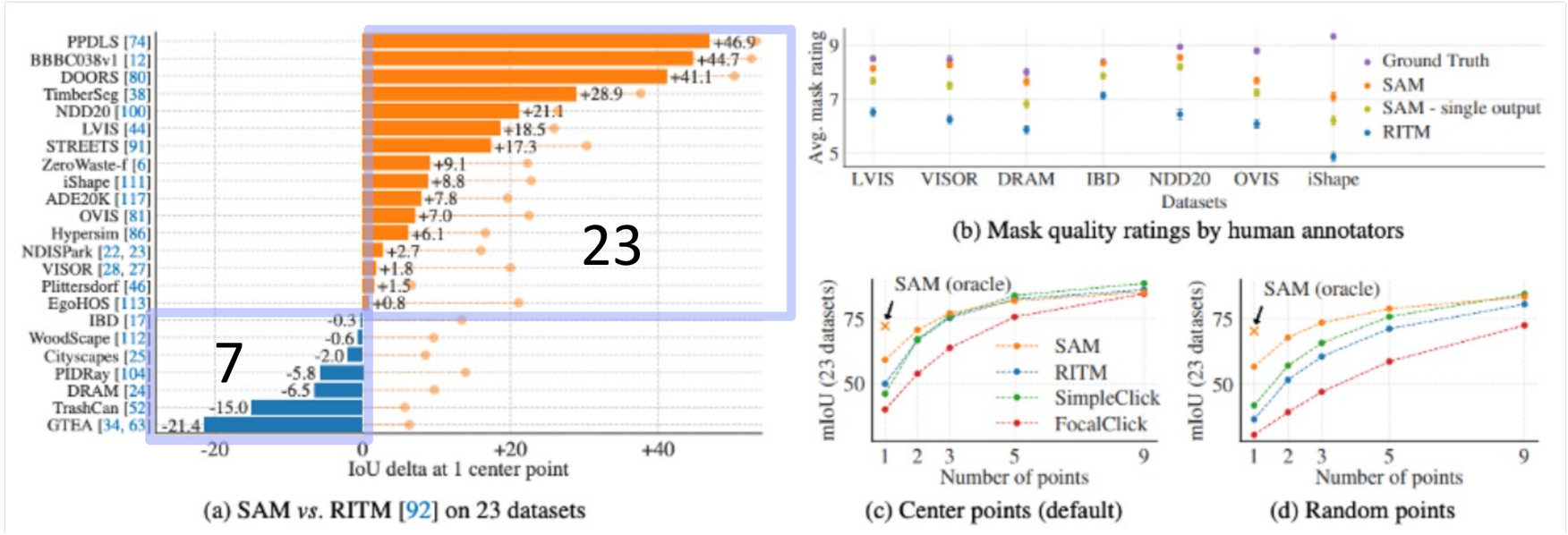
EVALUATION - ZERO-SHOT EXPERIMENTS

ZERO-SHOT SINGLE POINT VALID MASK EVALUATION

Results

- 16 of 23 datasets yields higher IoU than RITM among 23
- SAM outperforms other segmenters with varying number of points
  → SAM learned to segment masks from a single point

RITM : a strong interactive segmenter that performs best on the benchmark compared to other strong baselines

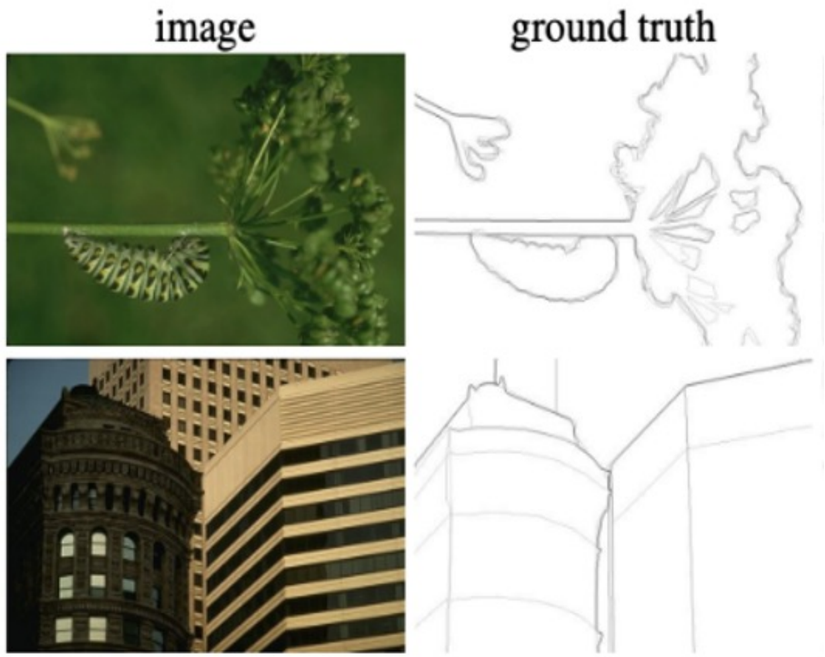


EVALUATION - ZERO-SHOT EXPERIMENTS

ZERO-SHOT EDGE DETECTION

Approach

- In order to extract ONLY the edge of an image
- Used BSDS500 dataset

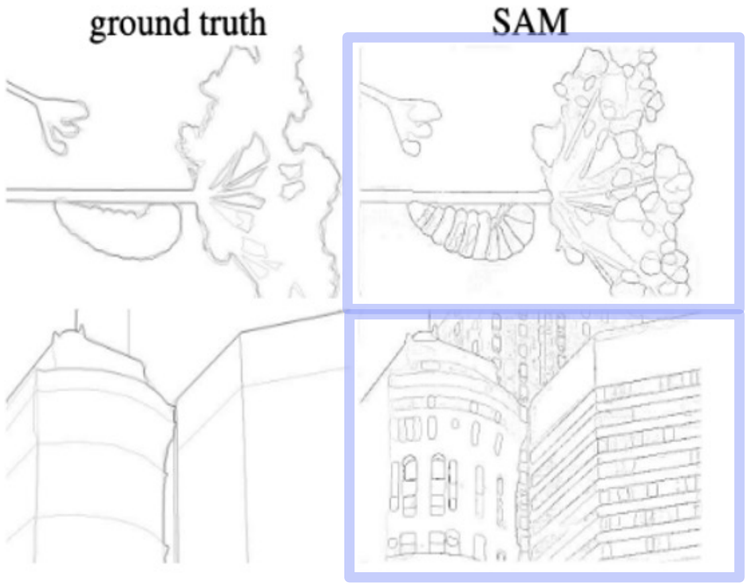


EVALUATION - ZERO-SHOT EXPERIMENTS

ZERO-SHOT EDGE DETECTION

Results

- Reasonable edge maps though never trained of edge detection
- Even included sensible ones that were not in the original dataset



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

TRAINING METHOD - MODEL EVALUATION

Total Loss = Mask Loss + IoU Prediction Loss

1. Mask Loss : 20*Focal Loss + Dice Loss

2. IoU Prediction Loss : MSE Loss between the IoU prediction and the predicted mask's IoU with the ground truth mask

Losses. We supervise mask prediction with a linear combination of focal loss [65] and dice loss [73] in a 20:1 ratio of focal loss to dice loss, following [20, 14]. Unlike [20, 14], we observe that auxiliary deep supervision after each decoder layer is unhelpful. The IoU prediction head is trained with mean-square-error loss between the IoU prediction and the predicted mask's IoU with the ground truth mask. It is added to the mask loss with a constant scaling factor of 1.0.

TRAINING METHOD - TRAINING OUTLINE

1. Training Mask Decoder of SAM from Scratch

- Used the smallest model, ViT-b
- Used smaller dataset, 10 images of SA-1B with only 10 masks
- Initialized only mask decoder

2. Fine-Tuning SAM for NDISPark Dataset

- Used the smallest model, ViT-b
- Used smaller custom dataset, Night and Day Instance Segmentation Park (NDISPark)
- Fine-tuned only mask decoder

TRAINING METHOD - TRAINING SETTING

1. Training Mask Decoder of SAM from Scratch

	Paper	Our Group
Dataset	SA-1B	10 images from SA-1B with 10 masks
Num_epochs	90k iterations (~2 SA-1B epochs)	3 epochs
Batch Size	128	1
Hyperparameters	Learning Rate(lr) = 8e-4 Learning Rate Decay(ld) = 0.6 (stepwise & layerwise ld applied) Weight Decay(wd) = 0.1 Drop Path(dp) = 0.6	Learning Rate(lr) = 8e-4 Learning Rate Decay(ld) = 0.6 (no module or code for layerwise ld → only stepwise ld applied) Weight Decay(wd) = 0.1 DropPath(dp) = X (no module or code for dp)
Training Time	3~5 days on 256 A100 GPUs	20s for 3 epochs on T4 GPU

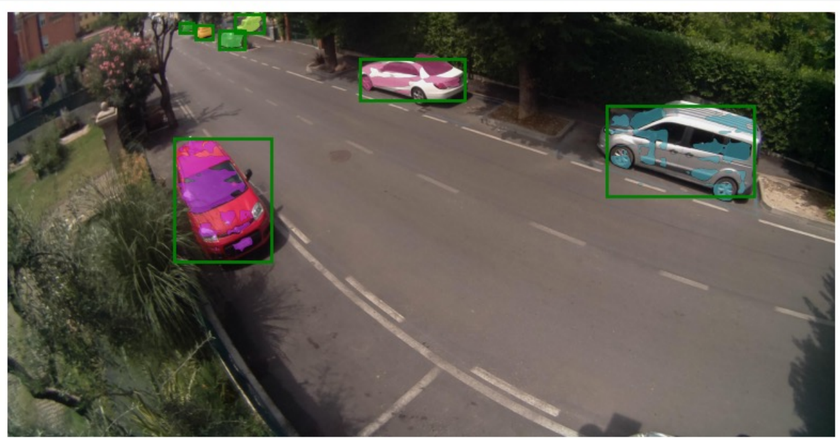
TRAINING METHOD - TRAINING SETTING

2. Fine-Tuning SAM for NDISPark Dataset

- A collection of images of parking lots for vehicle detection, segmentation
- Each image is labeled with pixel-wise masks and bounding boxes localizing vehicles.
- NDISPark Dataset Link

TRAINING METHOD - TRAINING SETTING

2. Fine-Tuning SAM for NDISPark Dataset

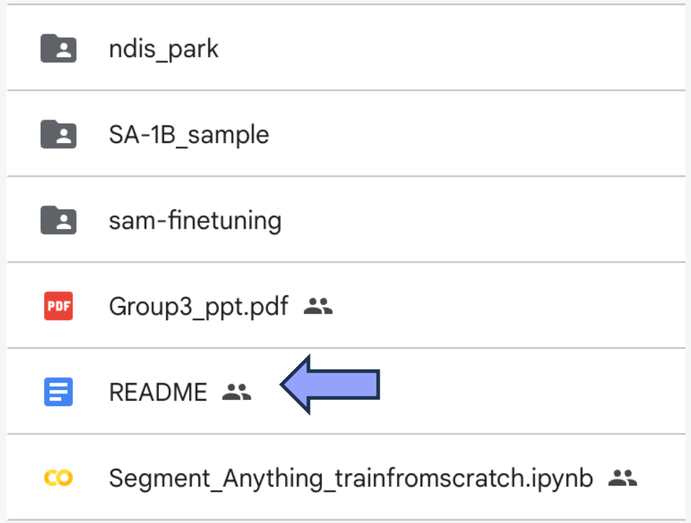


	Our Group
Dataset	NDISPark (111 training images, 30 validation images)
Num_epochs	2 epochs
Batch Size	1
Hyperparameters	Learning Rate(lr) = 8e-4 Weight Decay(wd) = 1e-4 DropPath(dp) = X
Training Time	2min for 1 epoch on T4 GPU

CODE DEMO

Follow this link below,
and please read the **README file** before running the codes!

https://docs.google.com/document/d/1tNYN58cqSFkA-1KVR6XZHf_px64-DmZ4lK93e2F-1qw/edit



CONCLUSION & LIMITATION

CONCLUSION

- Segment Anything project is an attempt to lift image segmentation into the era of foundation models.

- Improved image segmentation scope contributing to promptable segmentation task, building a new model, and new dataset designed for training general-purpose object segmentation

LIMITATION

- Overall performance of SAM is not real-time when using a heavy image encoder even if it can process prompts in real-time.

- Domain-specific tools are expected to outperform SAM, since it is designed for generality.

THANK YOU !

Q&A