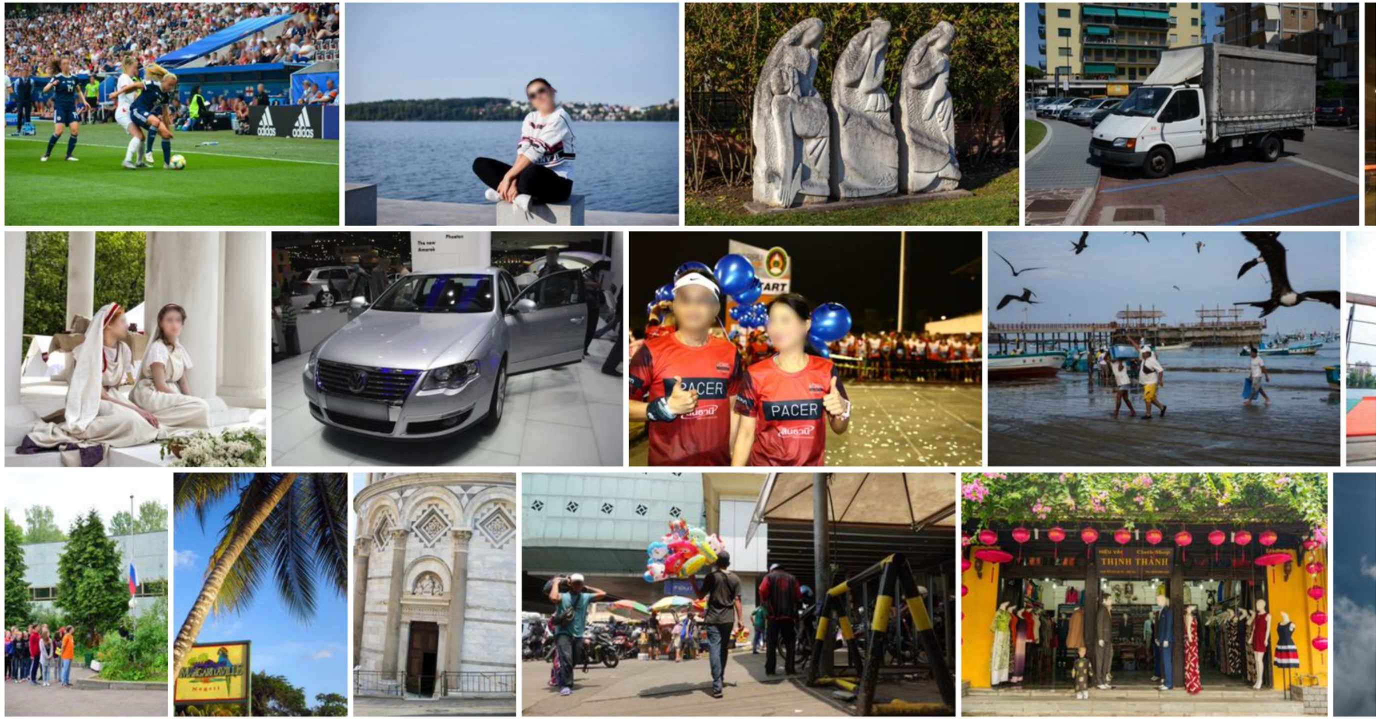
Segment Anything

Presenter: Dori Litvak

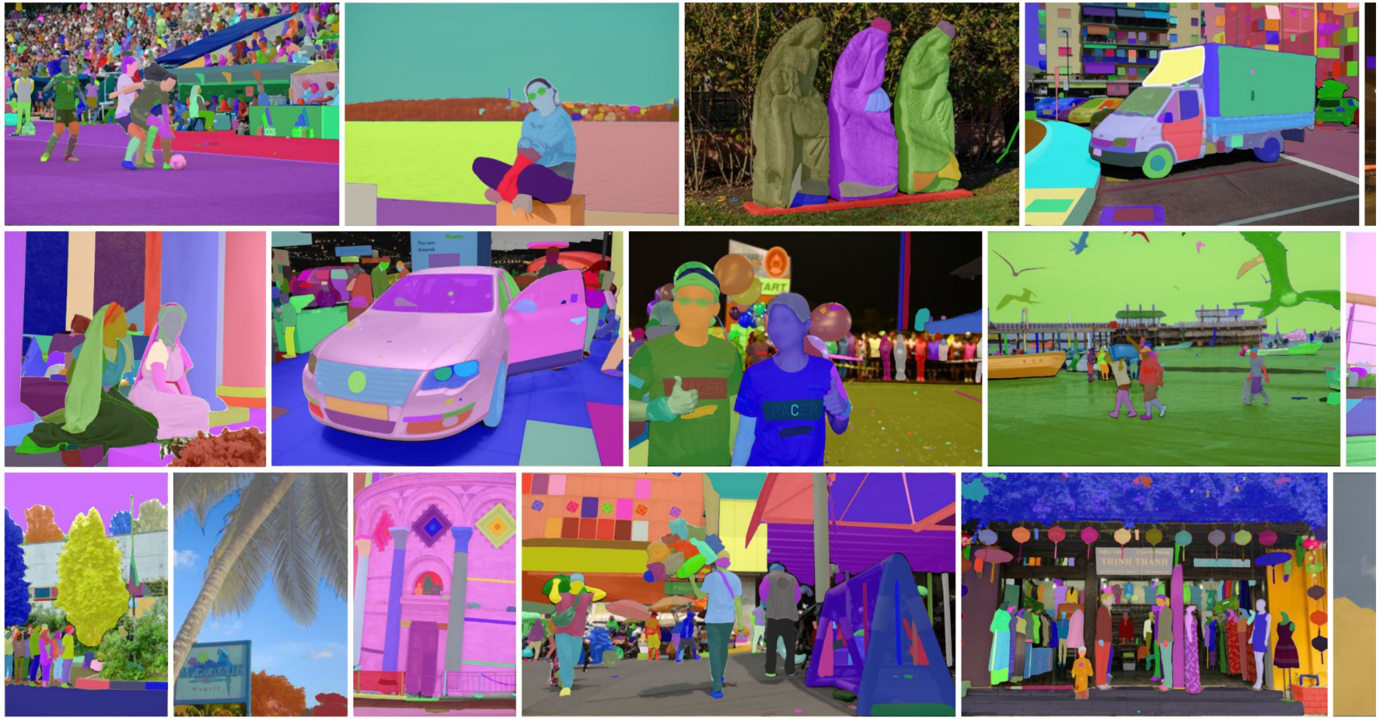
08/29/2023

The Segmentation Problem



https://segment-anything.com/dataset/index.html

The Segmentation Problem



https://segment-anything.com/dataset/index.html

Why is segmentation important?

- Simplification
- Focus
- Fast Analysis

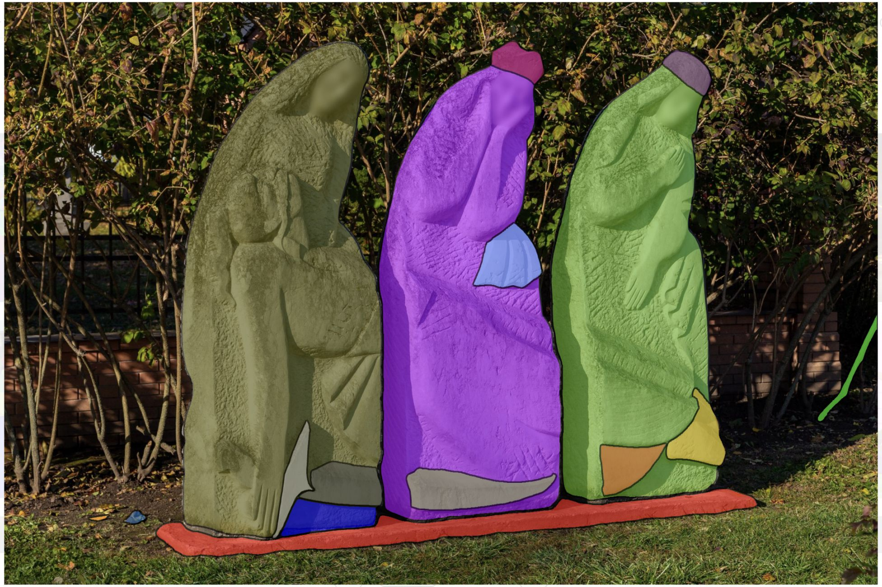
Why is Segmentation Important?

- Simplification

https://segment-anything.com/dataset/index.html ID: sa_10001804

Why is Segmentation Important?

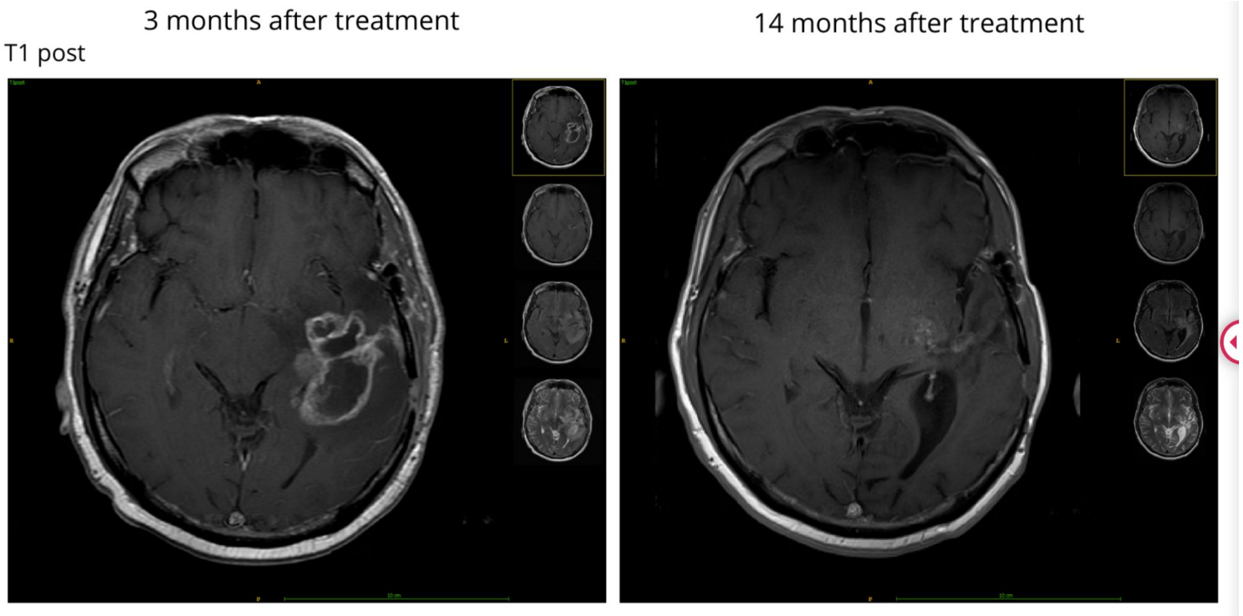
- Simplification



https://segment-anything.com/dataset/index.html ID: sa_10001804

Why is Segmentation Important?

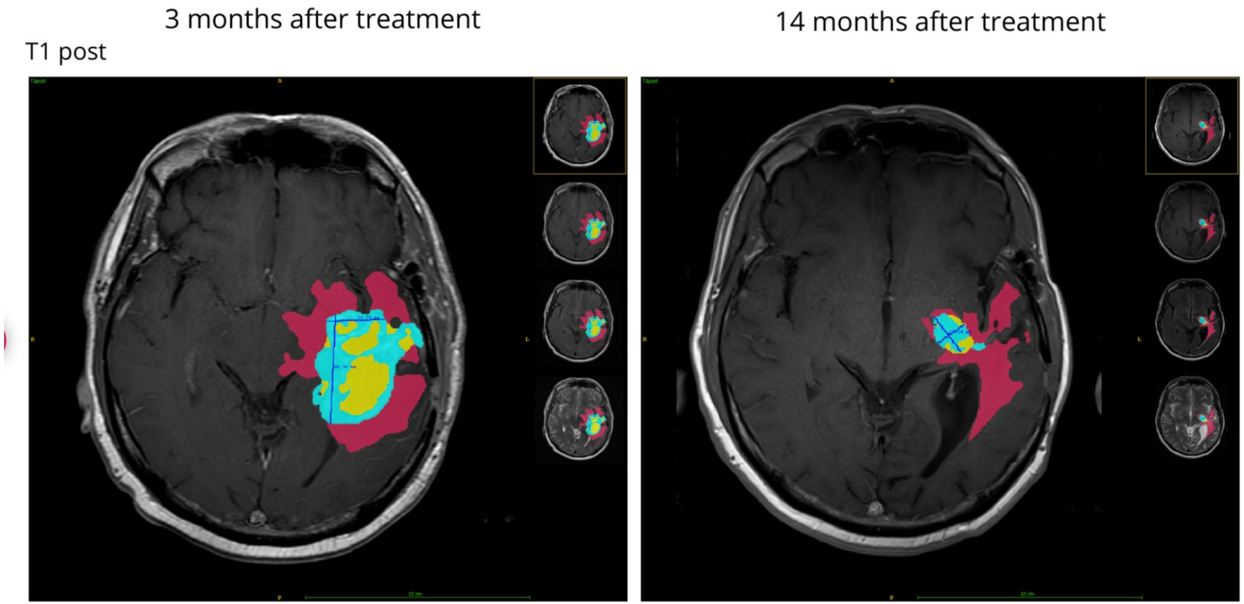
- Focus



https://graylight-imaging.com/project/automatic-brain-tumor-segmentation-with-subregions/

Why is Segmentation Important?

- Focus



https://graylight-imaging.com/project/automatic-brain-tumor-segmentation-with-subregions/

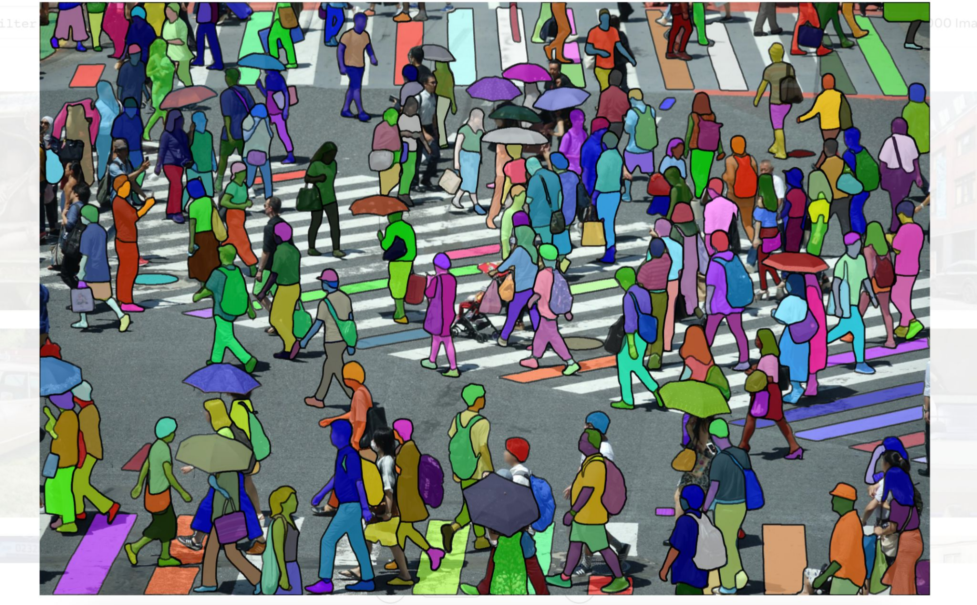
Why is Segmentation Important?

- Fast Analysis

https://segment-anything.com/dataset/index.html ID: sa_1017021

Why is Segmentation Important?

- Fast Analysis



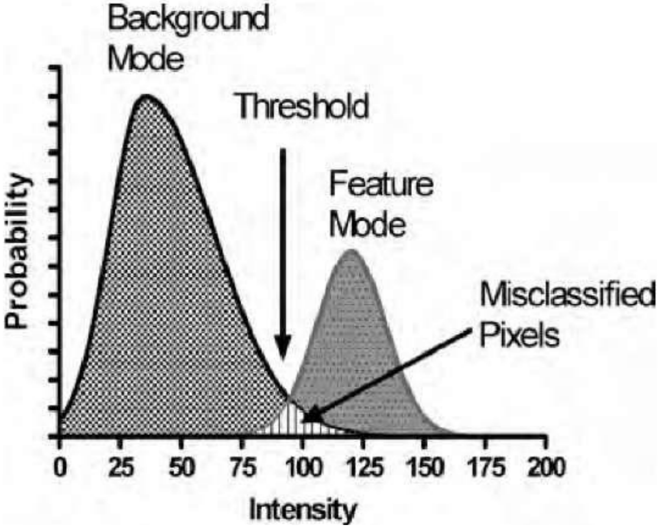
https://segment-anything.com/dataset/index.html ID: sa_1017021

Segmentation challenges

- Objects are complicated shapes
- Difference in Texture, Color, and Lightning.
- Not enough labeled data of segments we can use.

Previous Approaches

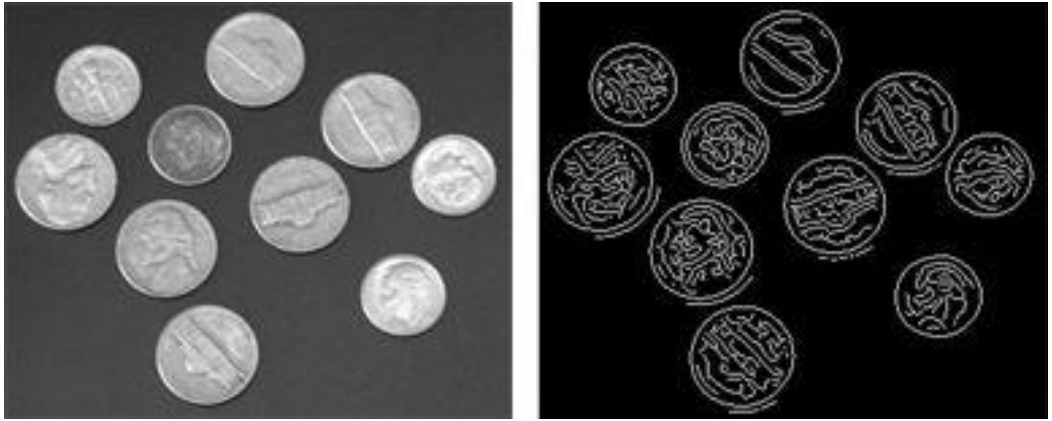
- Thresholding
- Edge detection
- Region growing
- Clustering
- Fully Connected Neural Network



http://what-when-how.com/biomedical-image-analysis/intensity-based-segmentation-thresholding-biomedical-image-analysis/

Previous Approaches

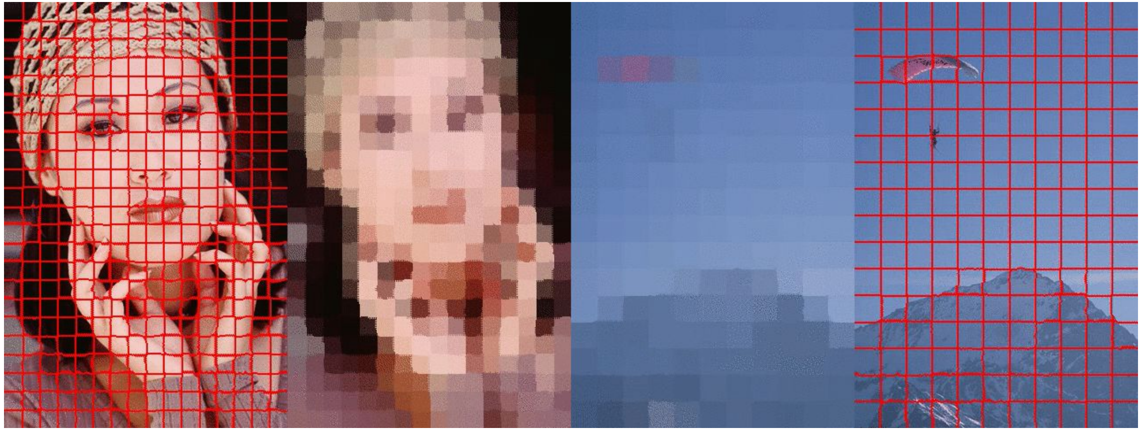
- Thresholding
- Edge detection
- Region growing
- Clustering
- Fully Connected Neural Network



https://www.mathworks.com/discovery/edge-detection.html

Previous Approaches

- Thresholding
- Edge detection
- Region growing
- Clustering
- Fully Connected Neural Network



Bayesian Adaptive Superpixel Segmentation, ICCV 2019, Uziel, Ronen and Freifeld

Previous Approaches

- Thresholding
- Edge detection
- Region growing
- Clustering
- Fully Connected Neural Network

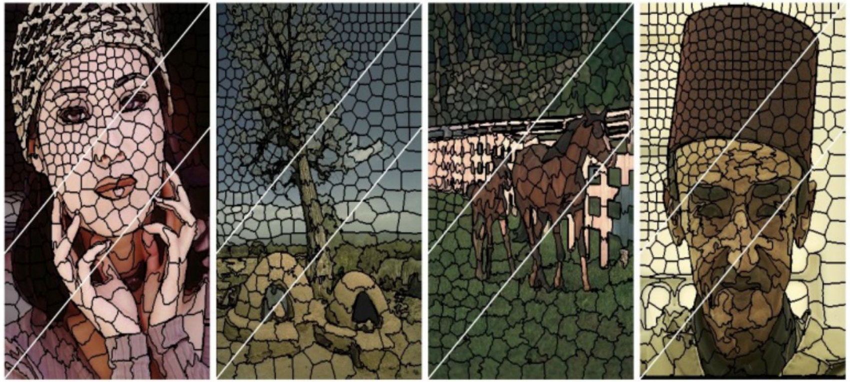


Figure 2: Images segmented into 1000/500/200 super pixels using the proposed LSC algorithm.

Image Segmentation by Using Linear Spectral Clustering

Previous Approaches

- Thresholding
- Edge detection
- Region growing
- Clustering
- Convolutional Neural Networks

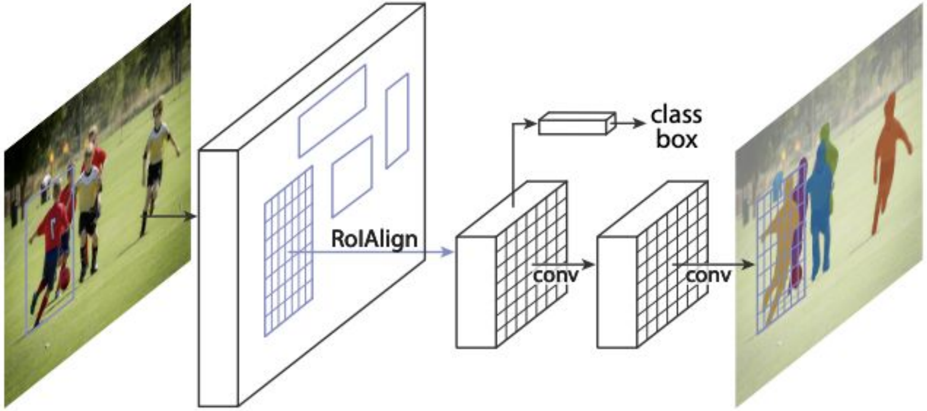


Figure 1. The **Mask R-CNN** framework for instance segmentation.

Image Segmentation by Using Linear Spectral Clustering

“ We introduce the Segment Anything (SA) project:
a new task,
model,
and dataset
for image segmentation. ”

- Meta AI Research, FAIR

Motivation

1. What task will enable zero-shot generalization?

2. What is the corresponding model architecture?

3. What data can power this task and model?

Segment Anything, Kirillov Et al.

Background

Definition: Foundation Models

Foundation models are large Artificial Intelligence (ML) models trained on broad data that can:

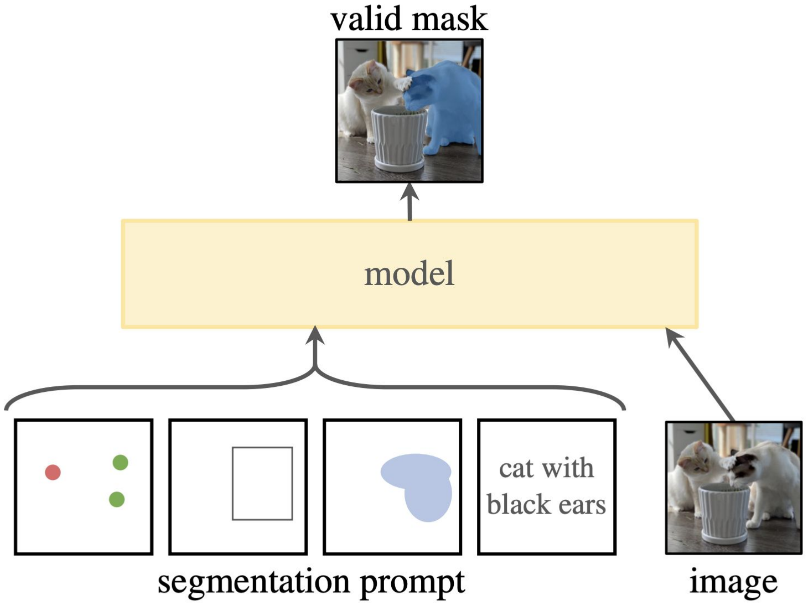
- *produce/generate wide variety of outputs.*
- *adapt to a wide range of downstream tasks.*
- *generalize beyond training data distributions.*

For more info: On the Opportunities and Risks of Foundation Models Center for Research on Foundation Models (CRFM), Stanford University

The Task

Definition: Promptable Segmentation Task

Given any segmentation prompt *specifies what to segment in an image*, *the goal is to return a valid segmentation mask.*



(a) **Task**: promptable segmentation

Segment Anything, Kirillov Et al. Figure 1 (a)

The Model - Segment Anything Model (SAM)

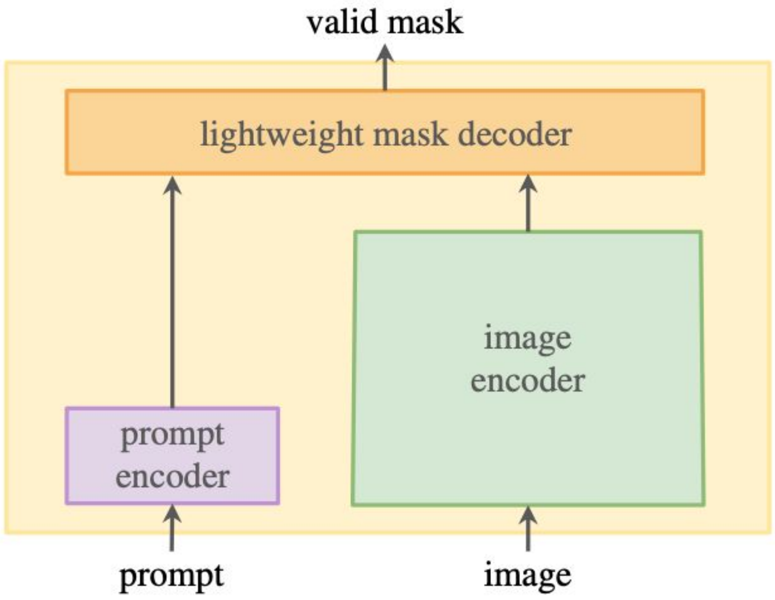
- *Support flexible prompts*
- *Real-time to allow interactive use*
- *Ambiguity-aware*

Segment Anything, Kirillov Et al.

The Model

*“a powerful **image encoder** computes an image embedding, a **prompt encoder** embeds prompts, and then the two information sources are combined in a **lightweight mask decoder** that predicts segmentation masks.*

We refer to this model as the Segment Anything Model, or SAM”



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

Segment Anything, Kirillov Et al. Figure 1 (b)

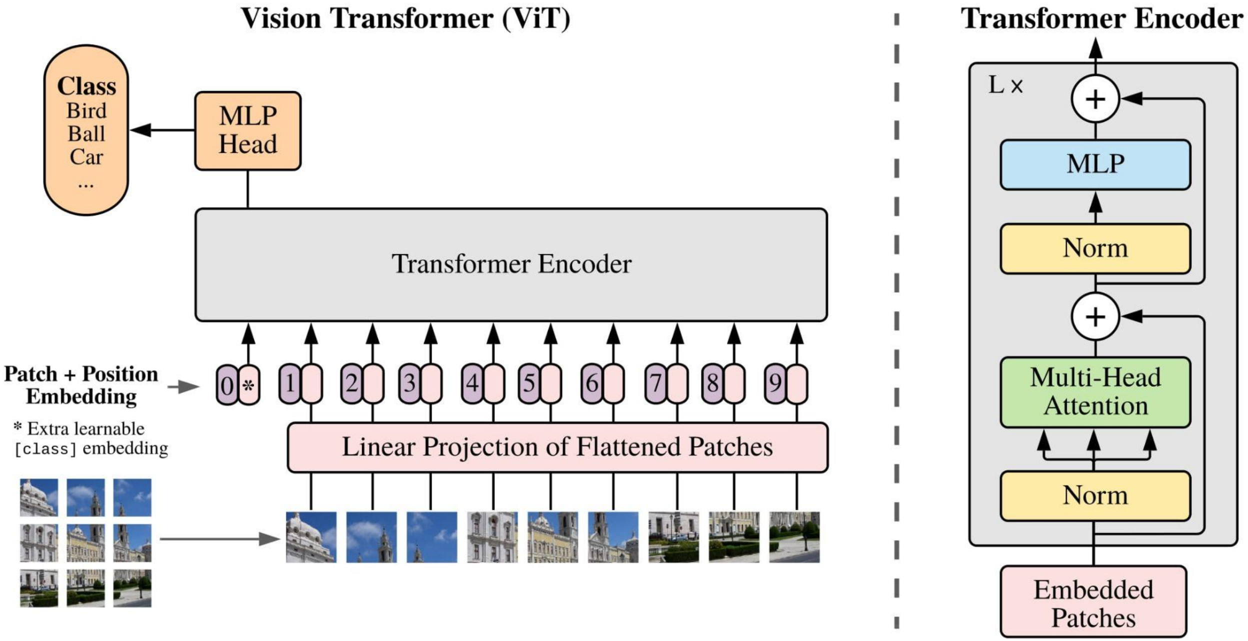
The Model

*A powerful **image encoder** computes an image embedding = **Vision Transformer (ViT)***

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale Dosovitskiy Et al.

The Model

*A powerful **image encoder** computes an image embedding = **Vision Transformer (ViT)***

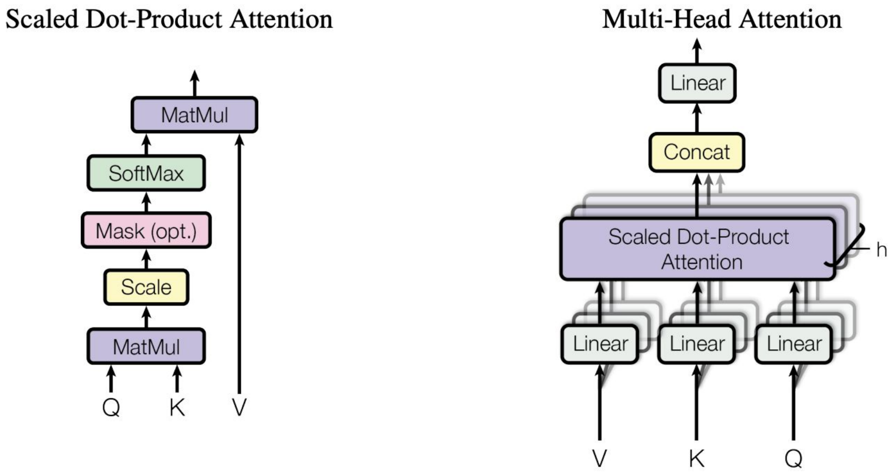


An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale Dosovitskiy Et al For more information: Attention Is All You Need Vaswani Et al.

For a nice tutorial: https://medium.com/mlearning-ai/vision-transformers-from-scratch-pytorch-a-step-by-step-guide-96c3313c2e0c Illustrations: https://www.youtube.com/watch?v=4Bdc55j80l8

The Model

*A powerful **image encoder** computes an image embedding = **Vision Transformer (ViT)***

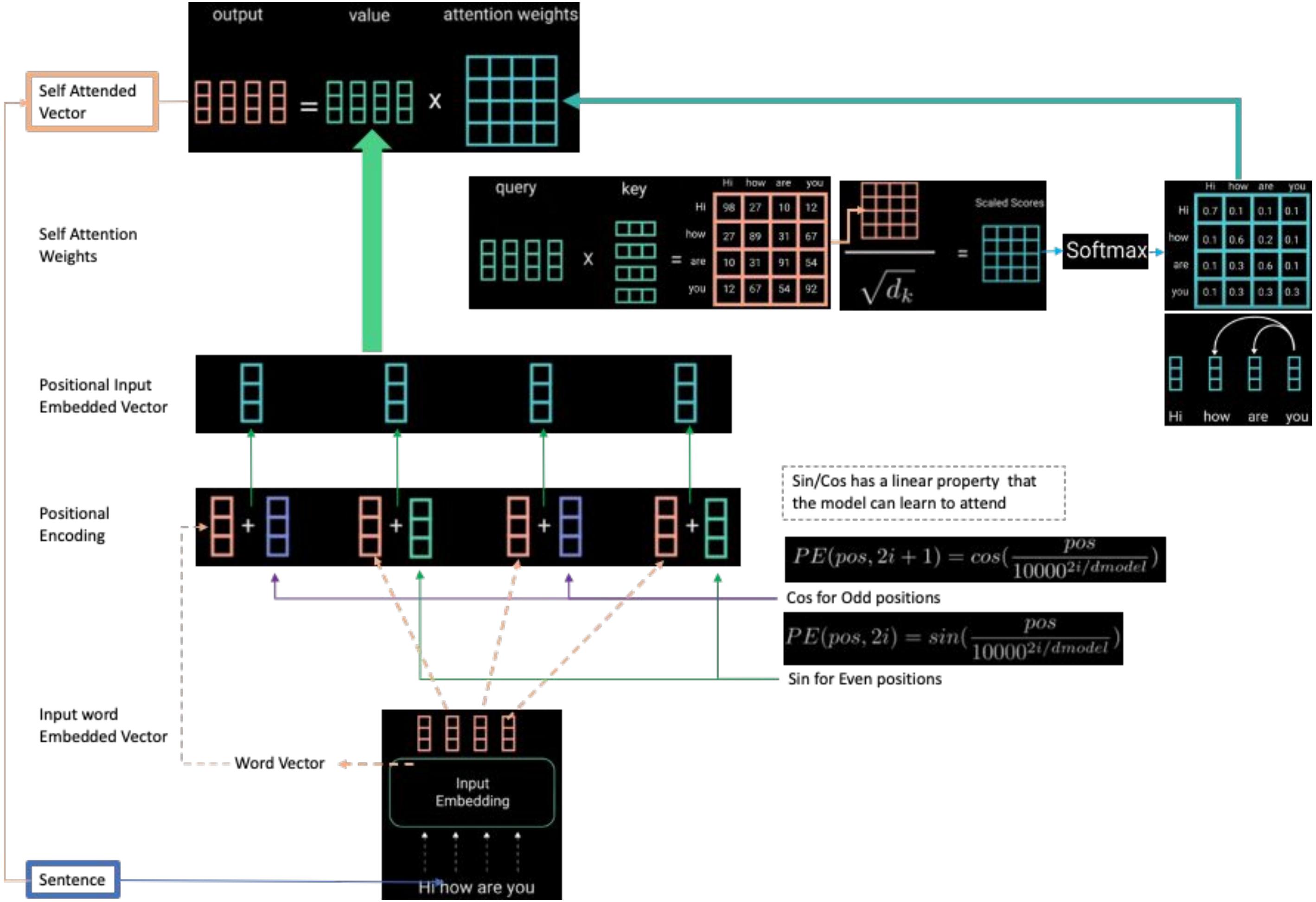


Scaled Dot-Product Attention

Multi-Head Attention

Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

Attention Is All You Need Vaswani Et al. For a nice tutorial: https://medium.com/mlearning-ai/vision-transformers-from-scratch-pytorch-a-step-by-step-guide-96c3313c2e0c



	Hi	how	are	you
Hi	98	27	10	12
how	27	89	31	67
are	10	31	91	54
you	12	67	54	92

$$\sqrt{d_k}$$

	Hi	how	are	you
Hi	0.7	0.1	0.1	0.1
how	0.1	0.6	0.2	0.1
are	0.1	0.3	0.6	0.1
you	0.1	0.3	0.3	0.3

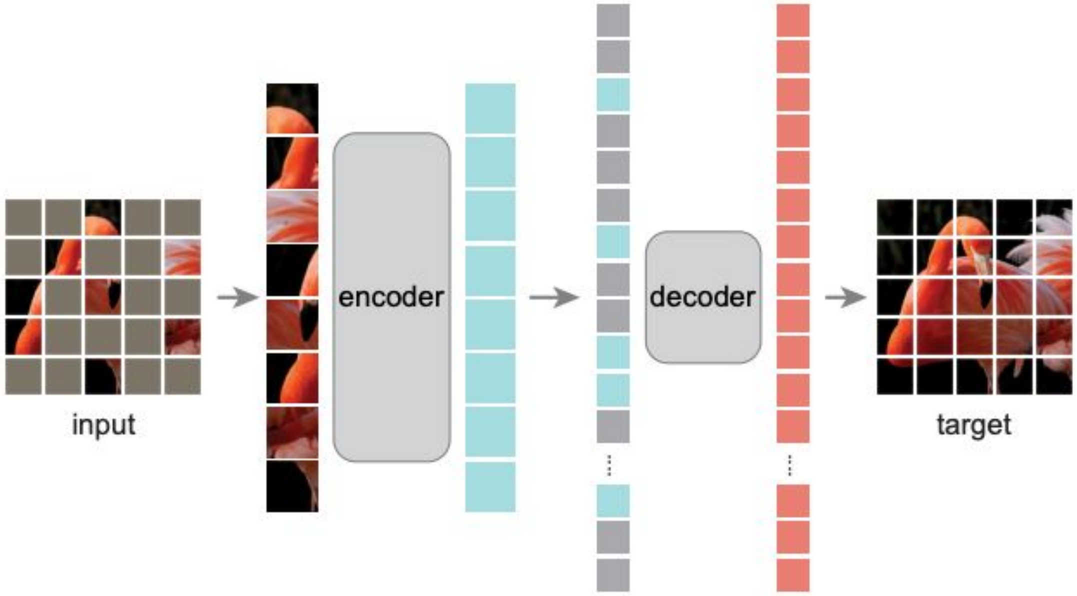
$$PE(pos, 2i+1) = cos(\frac{pos}{10000^{2i/dmodel}})$$

$$PE(pos, 2i) = sin(\frac{pos}{10000^{2i/dmodel}})$$

https://i.stack.imgur.com/xALqg.png

The Model

*A powerful **image encoder** computes an image embedding = **MAE**.*



Masked Autoencoders Are Scalable Vision Learners, He Et al. Figure 1

The Model

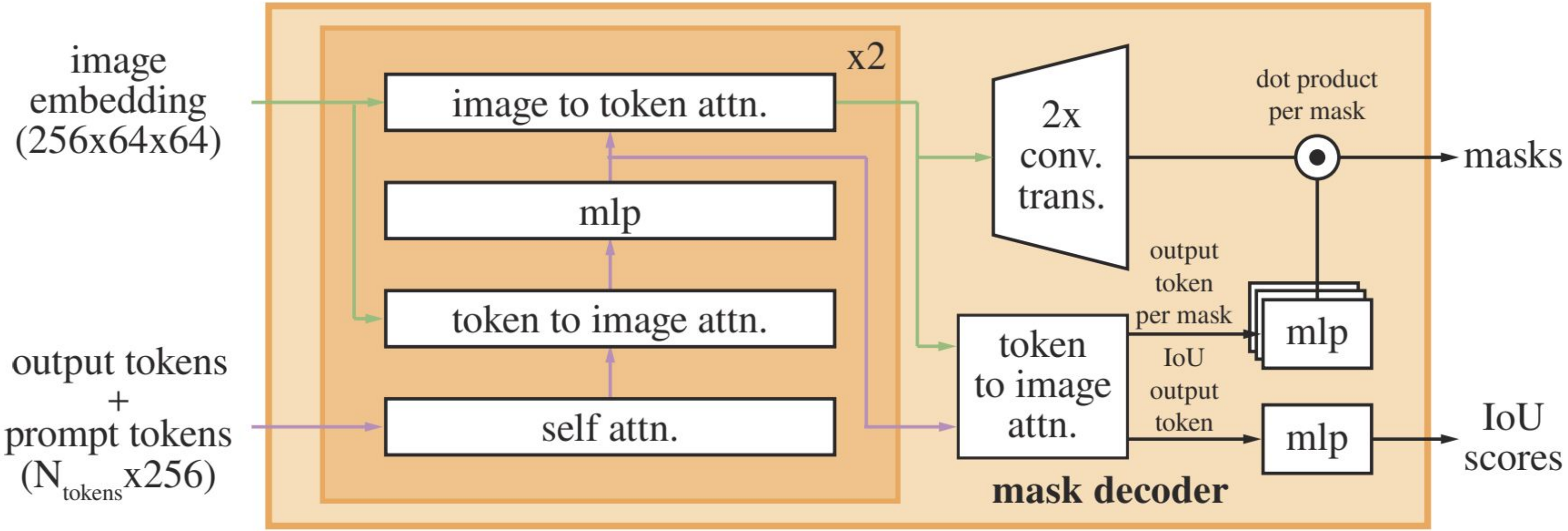
*A **prompt encoder** embeds prompts.*

- *Points and boxes - positional encodings [1] + embeddings for each prompt type.*
- *Free-form text - text encoder from CLIP [2].*
- *Masks - convolutions and summed element-wise with the image embedding.*

[1] Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains, [2] CLIP

The Model

*Combined in a **lightweight mask decoder** that predicts segmentation masks.*



Segment Anything, Kirillov Et al. Figure 14

Training Schedule

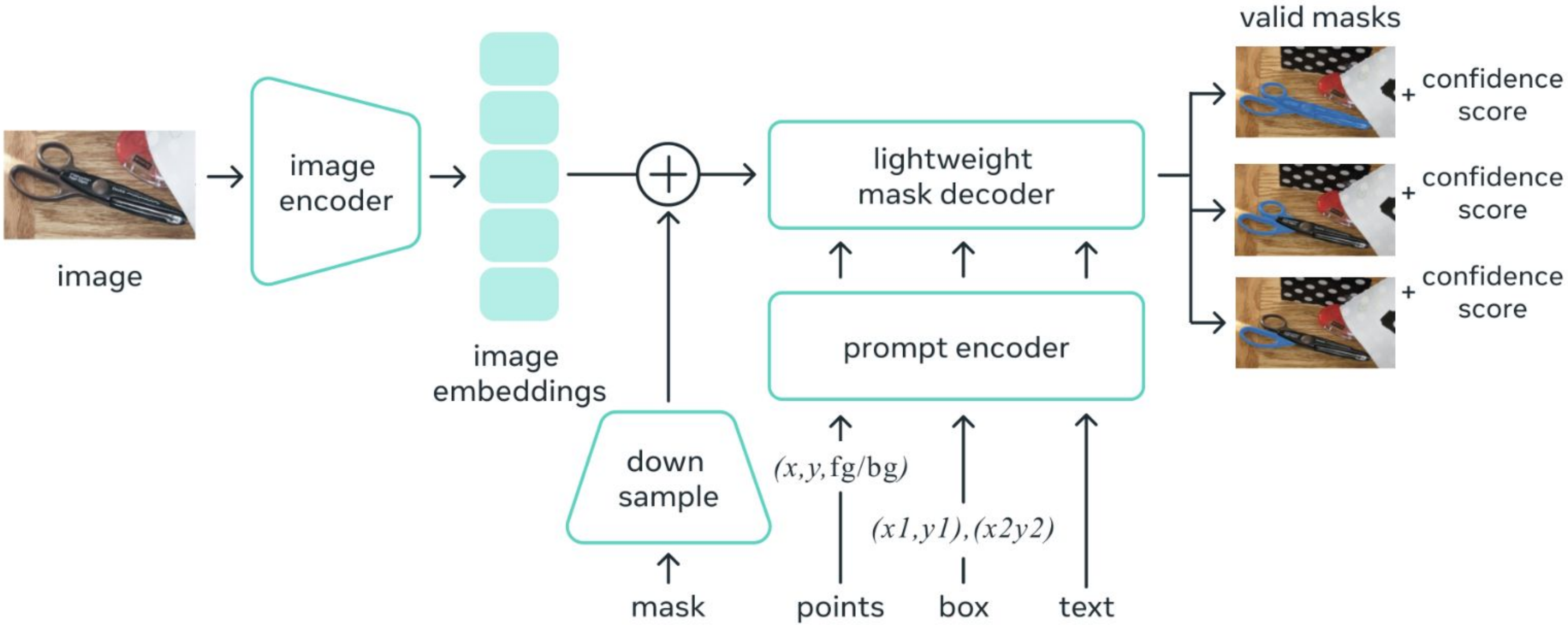


Figure: https://ai.meta.com/blog/segment-anything-foundation-model-image-segmentation/

The Model

To make SAM ambiguity-aware, it designed to predict multiple masks for a single prompt allowing SAM to naturally handle ambiguity, such as the shirt vs. person example.

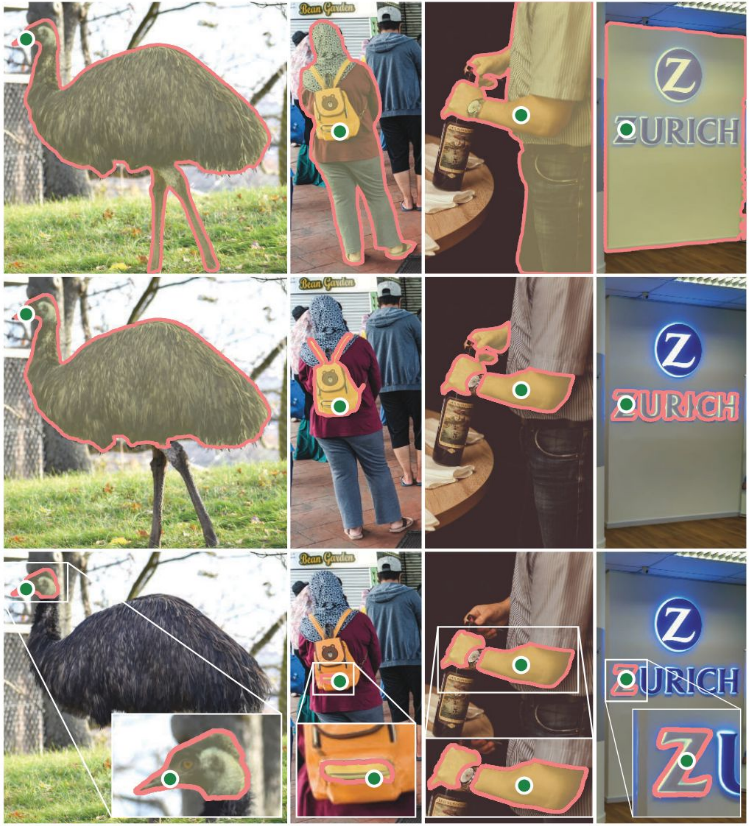
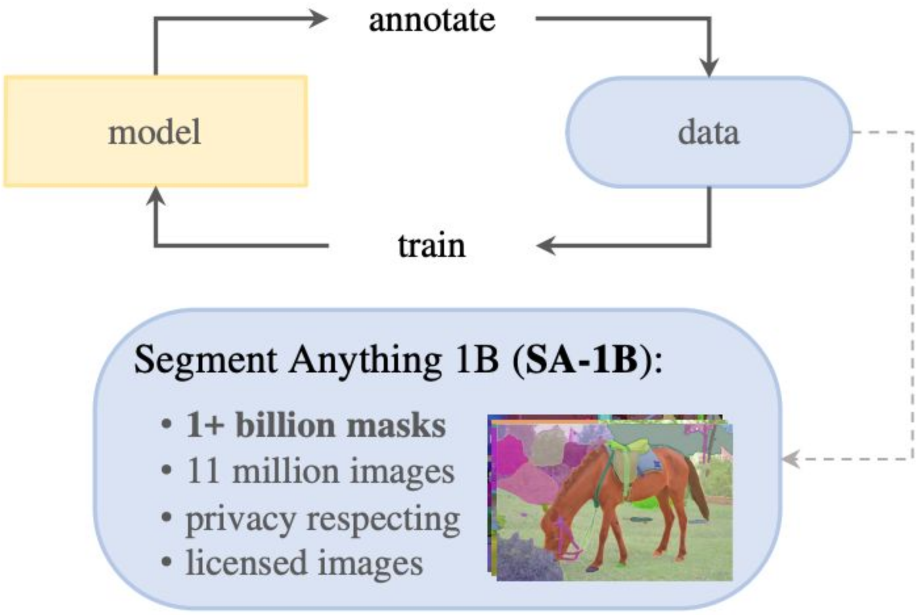


Figure 3: Each column shows 3 valid masks generated by SAM from a single ambiguous point prompt (green circle).

Segment Anything, Kirillov Et al. Figure 3

Data engine

To achieve strong generalization to new data distributions, it is necessary to train SAM on a large and diverse set of masks, beyond any segmentation dataset that was already exists.

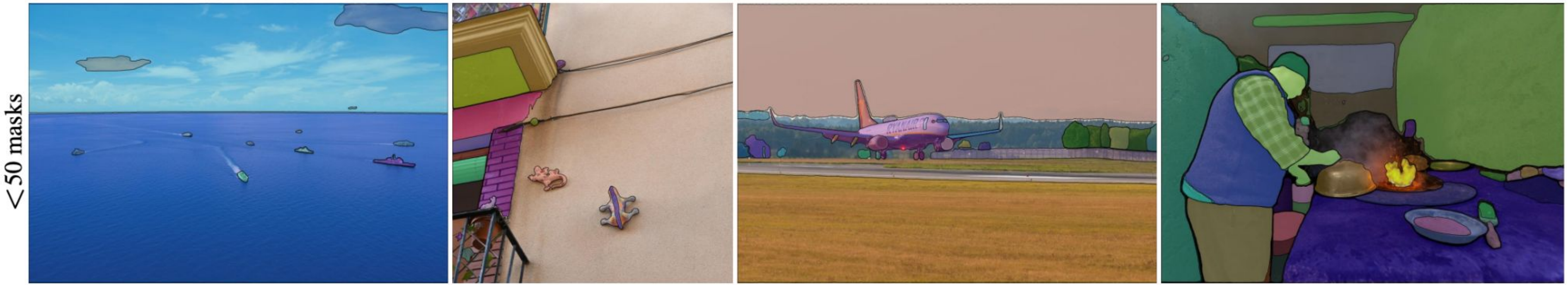


(c) **Data**: data engine (top) & dataset (bottom)

Segment Anything, Kirillov Et al. Figure 1 (c)

The Dataset

SA-1B dataset, includes more than 1B masks from 11M licensed and privacy-preserving images. High resolution (3300×4950 pixels on average), all masks was generated fully automatically.



Segment Anything, Kirillov Et al. Figure 2

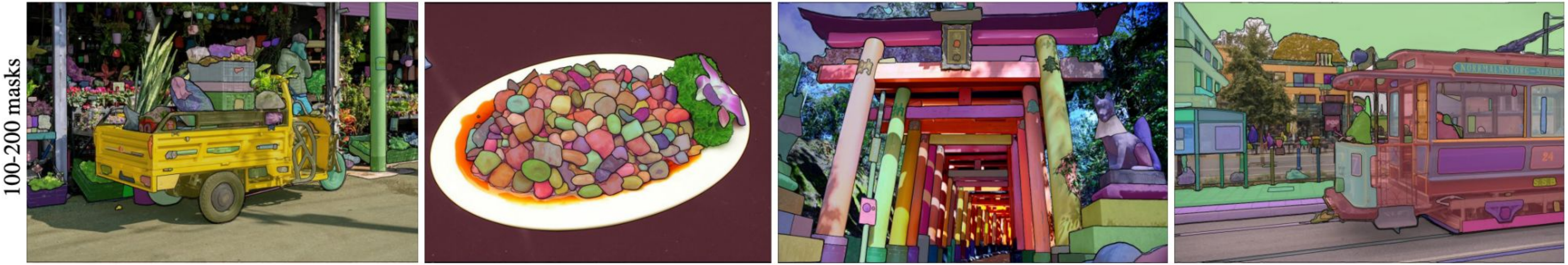
The Dataset

SA-1B dataset, includes more than 1B masks from 11M licensed and privacy-preserving images.

Segment Anything, Kirillov Et al. Figure 2

The Dataset

SA-1B dataset, includes more than 1B masks from 11M licensed and privacy-preserving images.



Segment Anything, Kirillov Et al. Figure 2

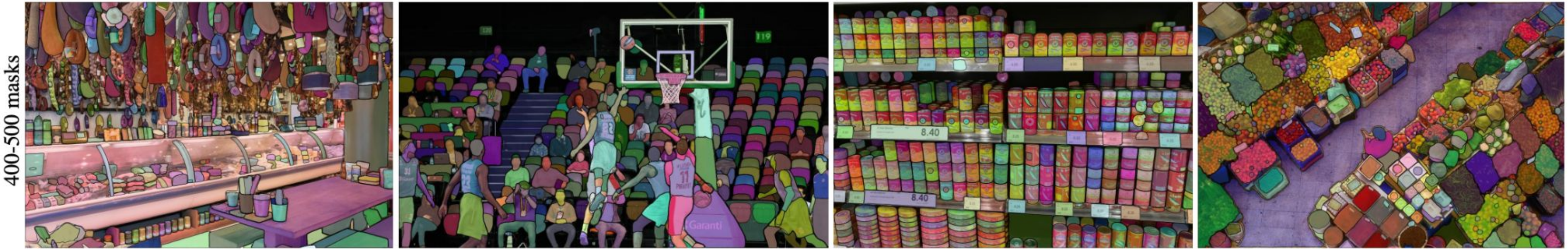
The Dataset

SA-1B dataset, includes more than 1B masks from 11M licensed and privacy-preserving images.

Segment Anything, Kirillov Et al. Figure 2

The Dataset

SA-1B dataset, includes more than 1B masks from 11M licensed and privacy-preserving images.



Segment Anything, Kirillov Et al. Figure 2

The Dataset

SA-1B dataset, includes more than 1B masks from 11M licensed and privacy-preserving images.

Segment Anything, Kirillov Et al. Figure 2

Zero-Shot Instance Segmentation

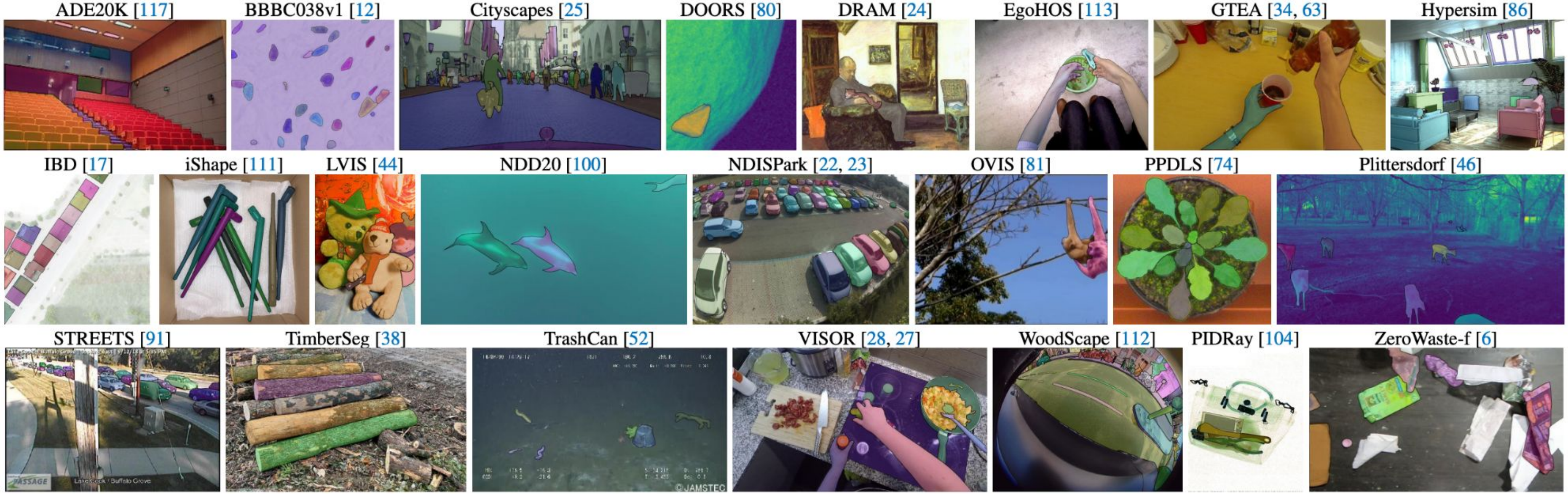


Figure 8: Samples from the 23 diverse segmentation datasets used to evaluate SAM's zero-shot transfer capabilities.

Segment Anything, Kirillov Et al. Figure 8

Zero-Shot text-to-mask

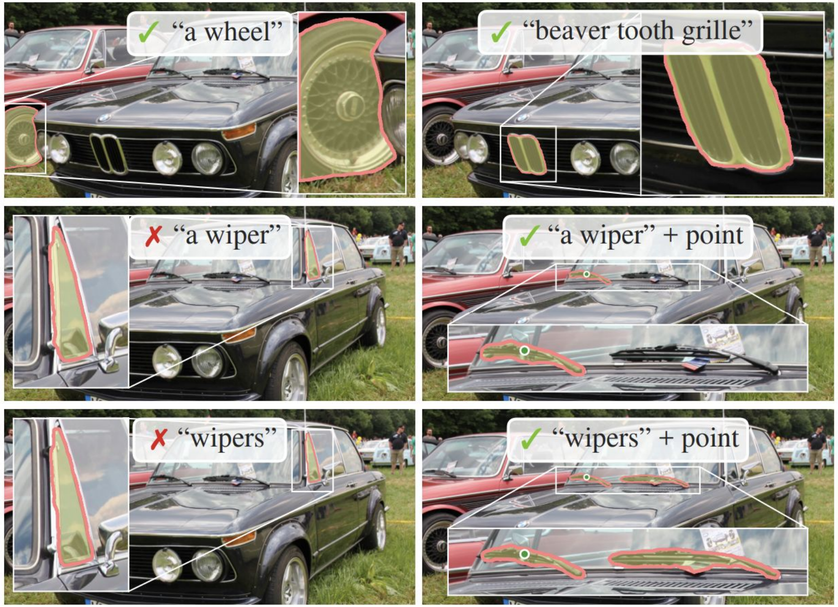


Figure 12: Zero-shot text-to-mask. SAM can work with simple and nuanced text prompts. When SAM fails to make a correct prediction, an additional point prompt can help.

Segment Anything, Kirillov Et al. Figure 12

Qualitative Results

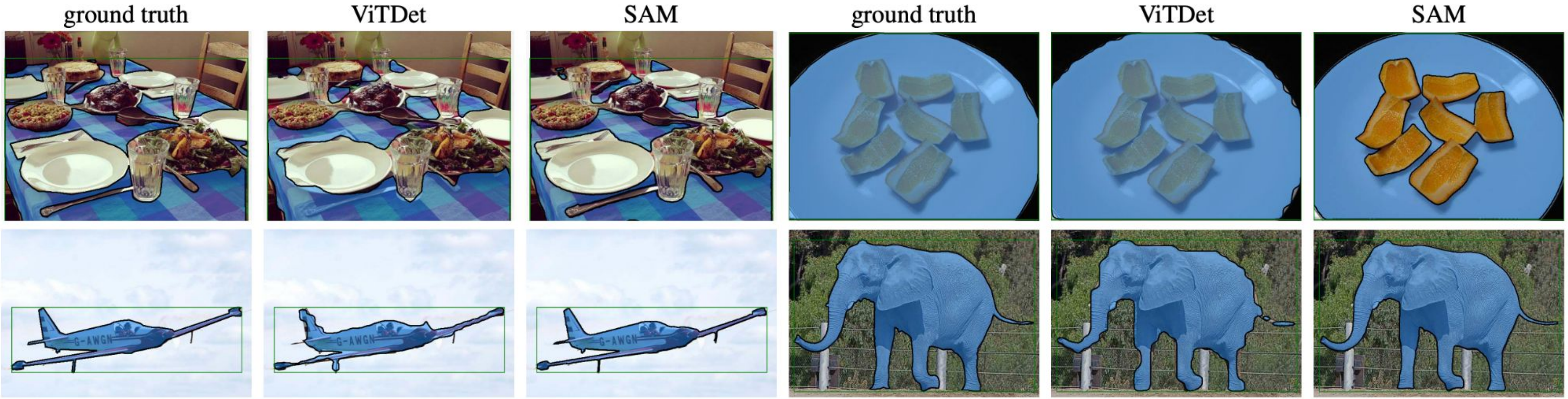
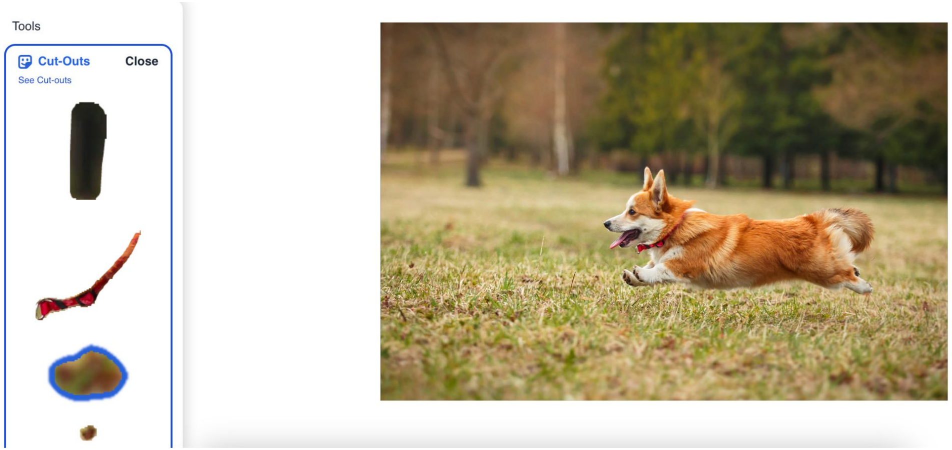


Figure 16: Zero-shot instance segmentation on LVIS v1. SAM produces higher quality masks than ViTDet. As a zero-shot model, SAM does not have the opportunity to learn specific training data biases; see top-right as an example where SAM makes a modal prediction, whereas the ground truth in LVIS is amodal given that mask annotations in LVIS have no holes.

Segment Anything, Kirillov Et al. Figure 16

Limitations

While SAM performs well in general, it is not perfect.
It can miss delicate structures, hallucinate small disconnected components at times,
and does not produce boundaries as well as more computationally intensive methods.
Dedicated interactive segmentation methods generally outperform SAM when many points are provided. SAM is designed
for generality rather than high
IoU interactive segmentation.



Conclusion

The Segment Anything project is lifting image segmentation into the era of foundation models.

The principal contributions are:

1. New task - promptable segmentation task.
2. New model (SAM) that allow generalize zero-shot segmentation task.
3. Generalize to other new tasks.
4. Present a new large dataset (SA-1B).

Try The Demo: https://segment-anything.com/

Conclusion

The Segment Anything project is lifting image segmentation into the era of foundation models.

The principal contributions are:

1. **New task - promptable segmentation task.**
2. **New model (SAM) that allow generalize zero-shot segmentation task.**
3. **Generalize to other new tasks.**
4. **Present a new large dataset (SA-1B).**

Try The Demo: https://segment-anything.com/