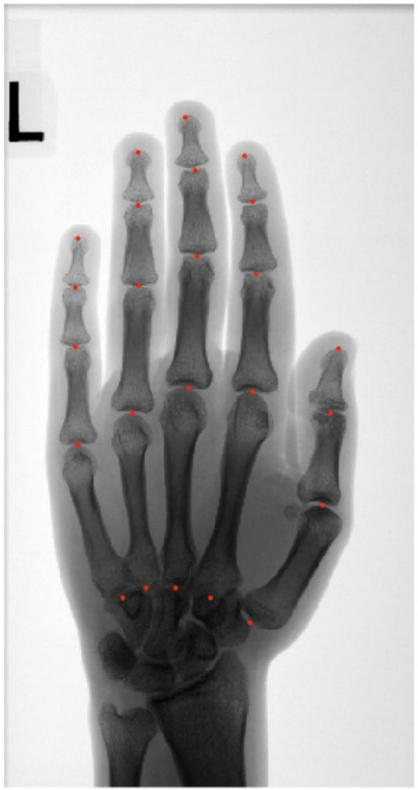
contact@mva.ai

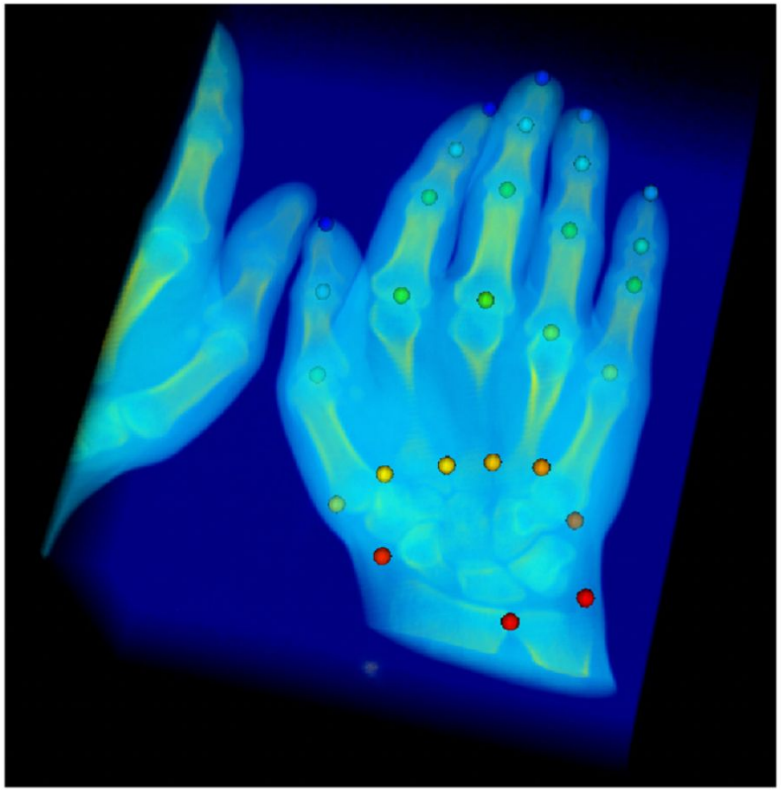
Segment Anything and the Rise of Foundation Models

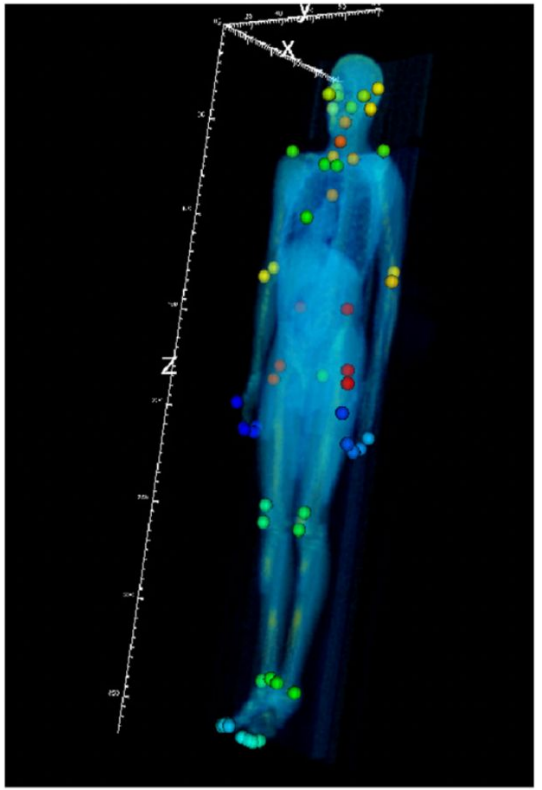
MEDICAL VOLUME ANNOTATOR

René Donner

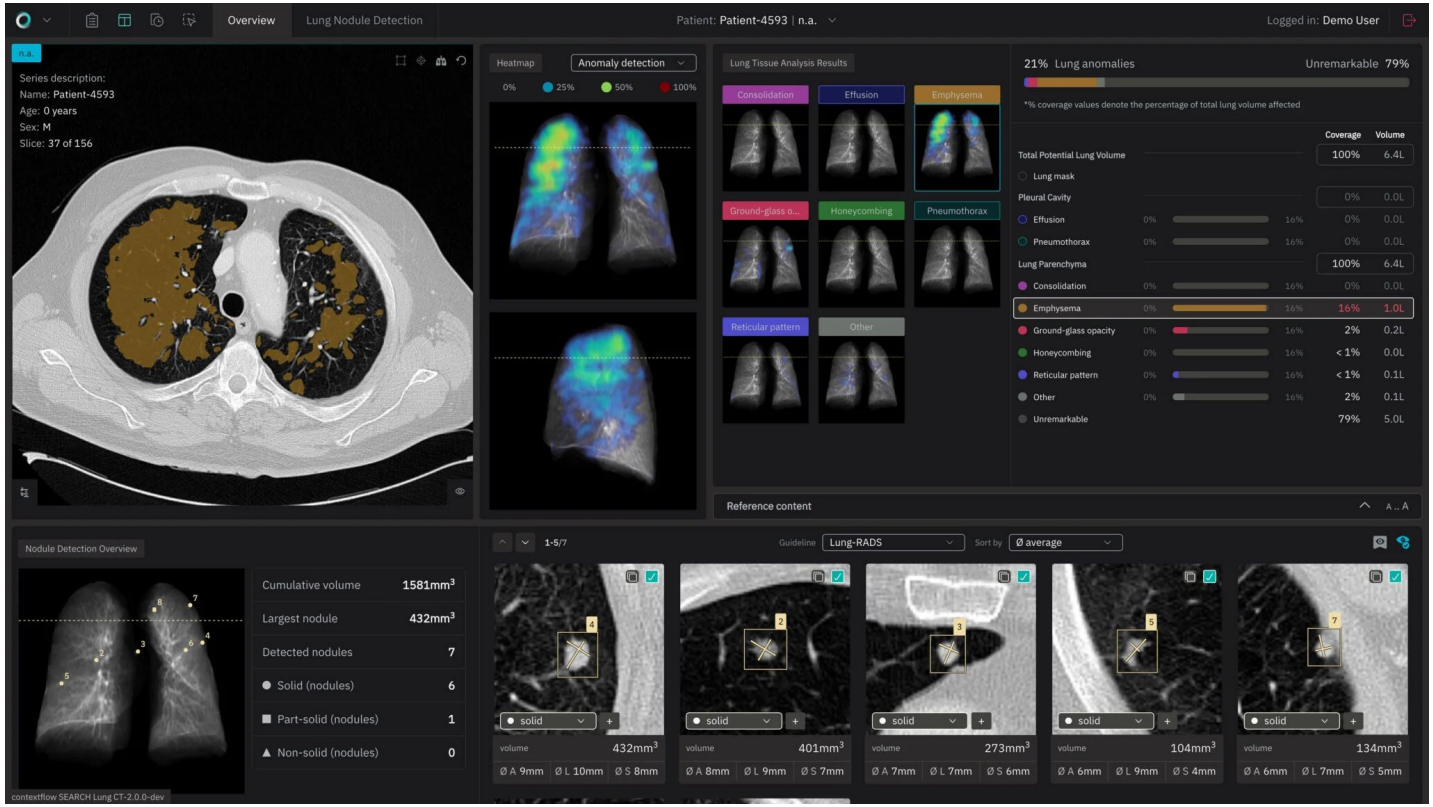
MUW / Radiologie / CIR





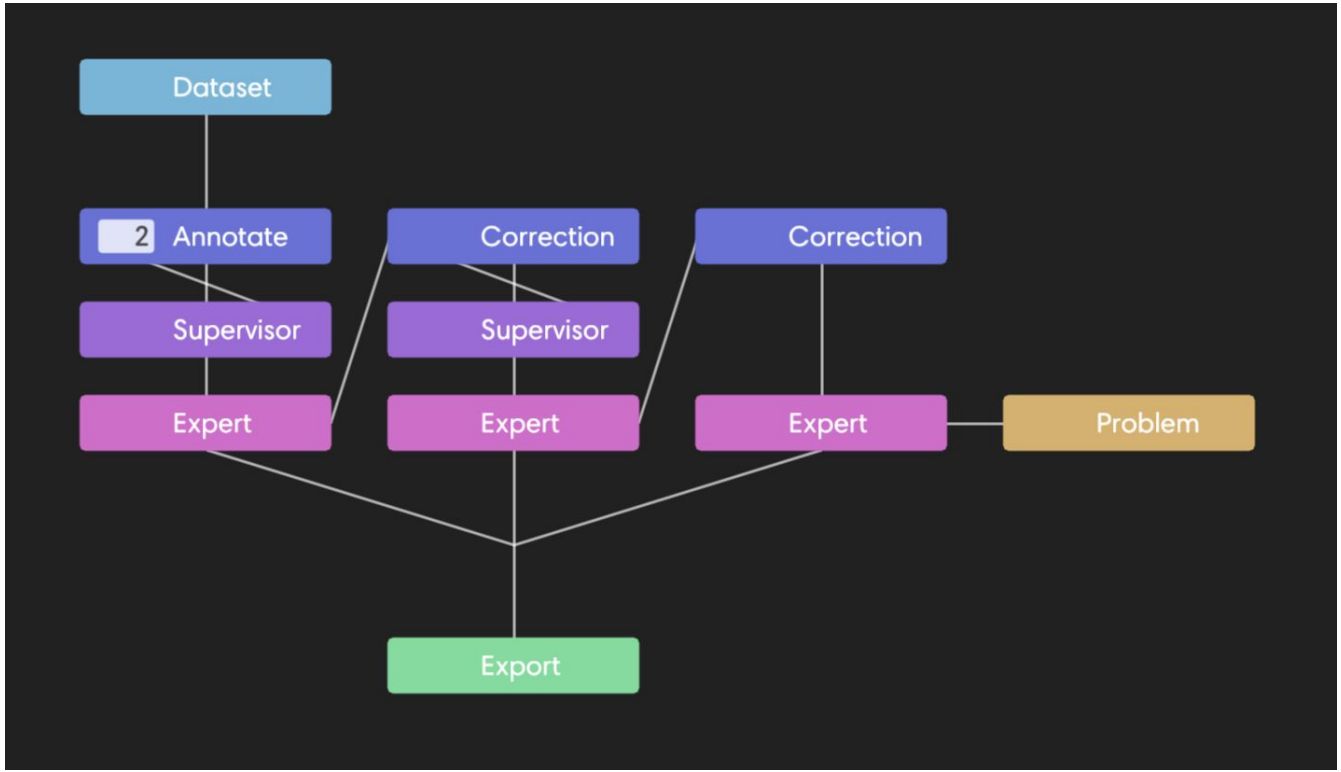


contextflow



Medical Volume Annotator – On-Site Annotation Platform

Medical Volume Annotator – Workflow Management, Statistics …



Medical Volume Annotator – Efficient Annotation

Smart Brushes – online learning of labeling models

Smart Propagation – distribute annotations spatially

Smart Regions – Super-pixels, over-segmentation, …

Medical Volume Annotator – Efficient Annotation

Smart Brushes – online learning of labeling models

Smart Propagation – propagation annotations spatially

Smart Regions – Super-pixels, over-segmentation, …

“Segment Anything Model (SAM): a new AI model from Meta AI that can ‘cut out’ any object, in any image, with a single click”

Segment Anything Model (SAM)

Segment Anything Model

Dataset – 11 Mio Images, 1.1 Billion masks // **“Foundation Model”**

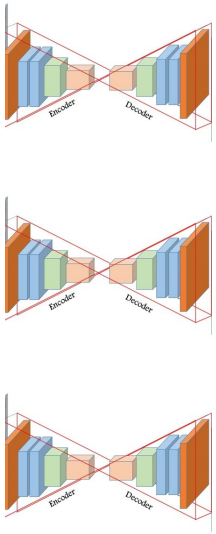
Foundation Models?

Large, pre-trained models

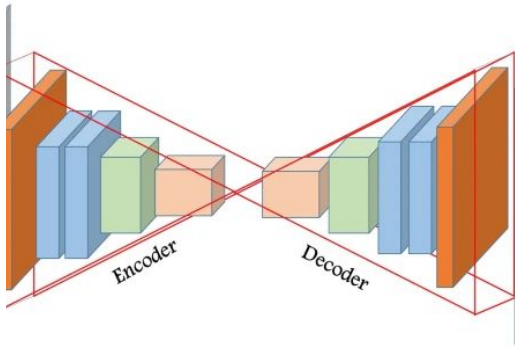
Strong generalization & zero-shot performance

Huge datasets, massive compute

Foundation Models?



From scratch



Segment Anything

+

Adapter

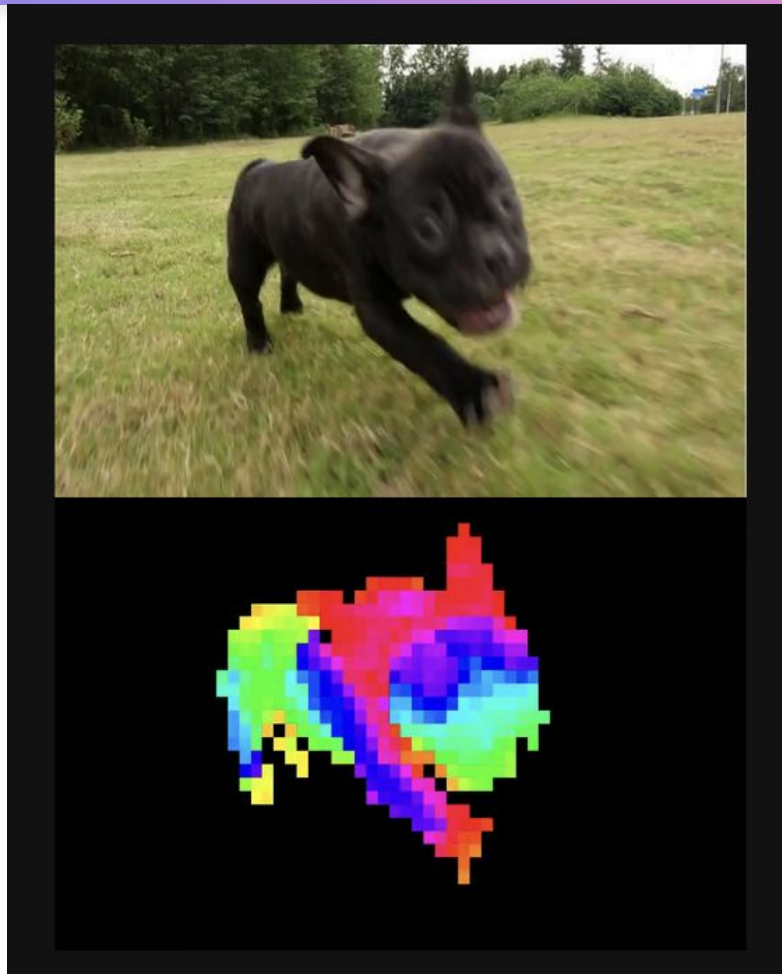
facebook DENOv2: Self-supervised Vision Transformer Model

Image-level visual tasks

- Image classification
- Instance retrieval
- Video understanding

Pixel-level visual tasks

- Depth estimation
- Semantic segmentation



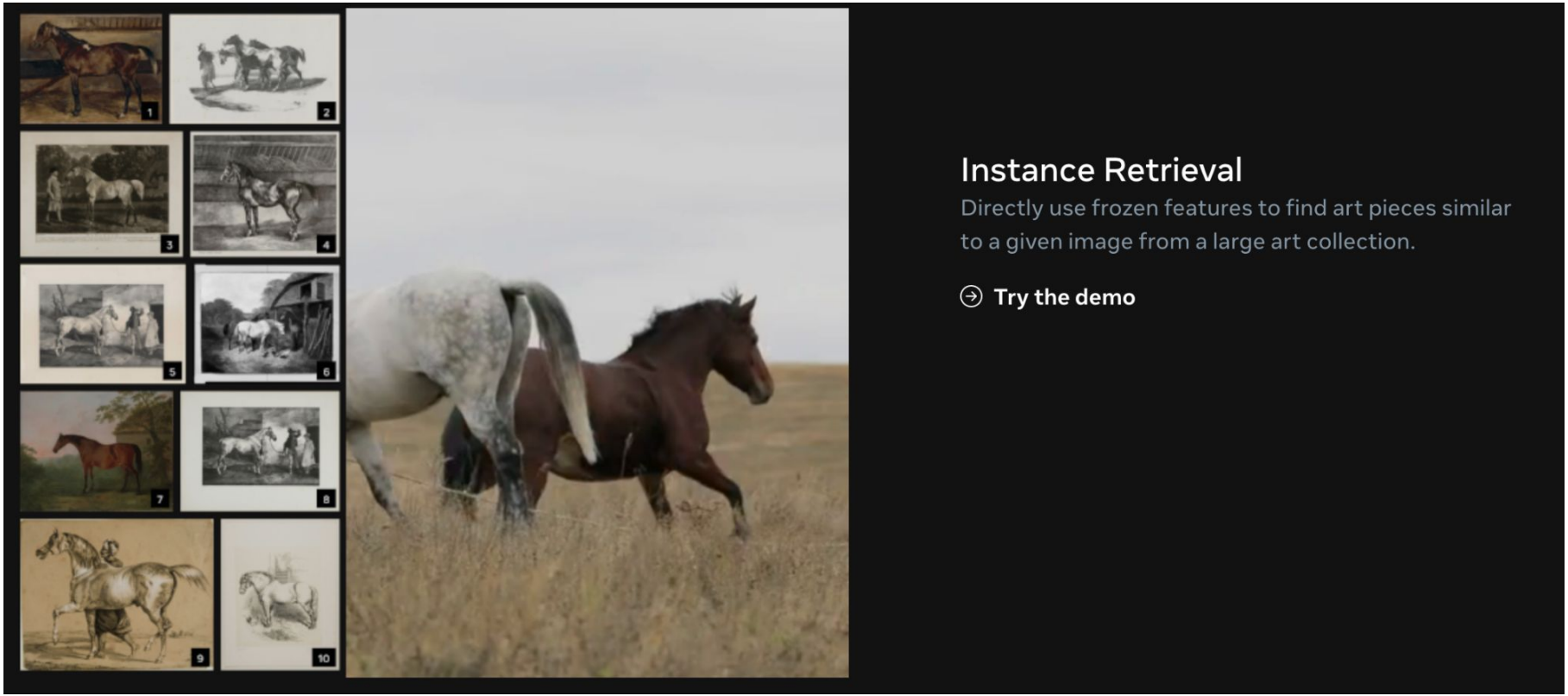
DENOv2: Self-supervised Vision Transformer Model

Depth Estimation

State-of-the-art results and strong generalization on estimating depth from a single image.

Try the demo

DENOv2: Self-supervised Vision Transformer Model



Microsoft's offerings

Language & Multilingual

> **UniLM**: unified pre-training for language understanding and generation

> **InfoXLM/XLM-E**: multilingual/cross-lingual pre-trained models for 100+ languages

> **DeltaLM/mT6**: encoder-decoder pre-training for language generation and translation for 100+ languages

> **MiniLM**: small and fast pre-trained models for language understanding and generation

> **AdaLM**: domain, language, and task adaptation of pre-trained models

> **EdgeLM**(`NEW`): small pre-trained models on edge/client devices

> **SimLM** (`NEW`): large-scale pre-training for similarity matching

> **E5** (`NEW`): text embeddings

Vision

> **BEiT/BEiT-2**: generative self-supervised pre-training for vision / BERT Pre-Training of Image Transformers

> **DiT**: self-supervised pre-training for Document Image Transformers

> **TextDiffuser** (`NEW`): Diffusion Models as Text Painters

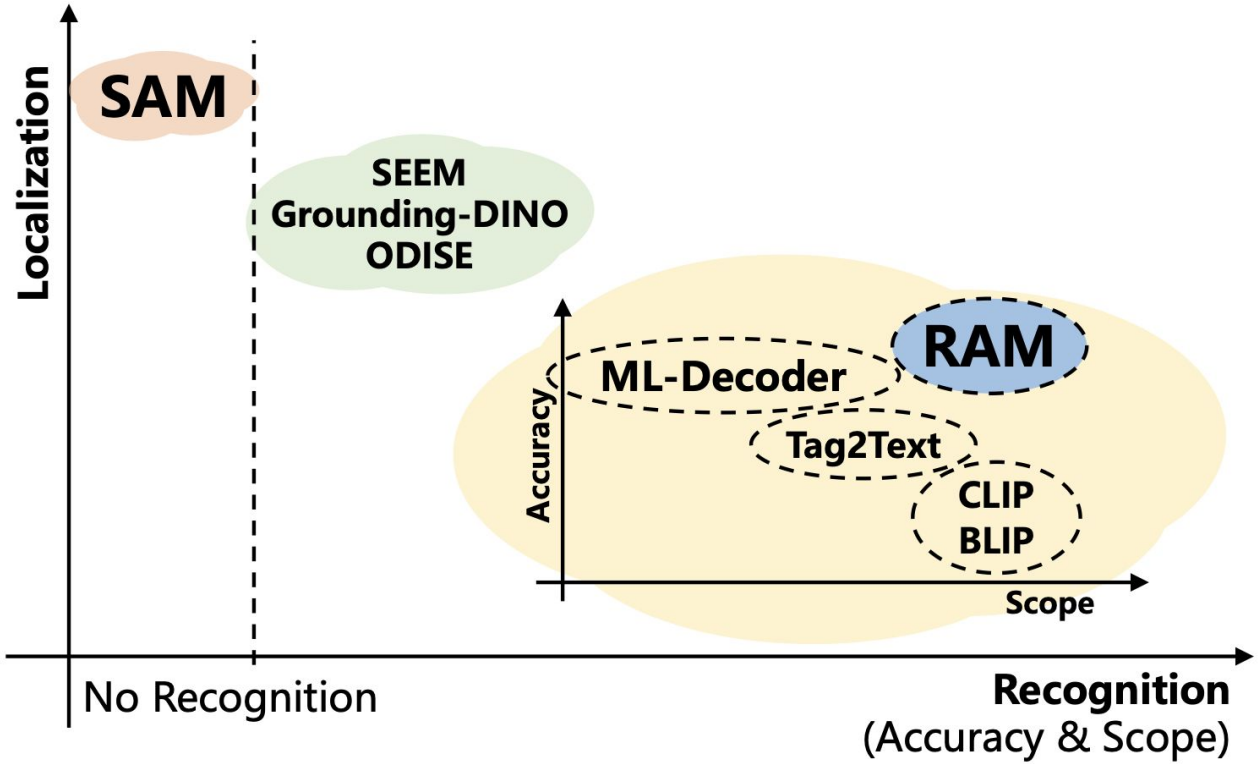
https://github.com/microsoft/unilm#foundation-models

“Anything Models” – Recognize Anything

RAM	living room, dog, blanket, carpet, couch, desk, furniture, pillow, plant, sit, wood floor, lamp	Christmas market, Christmas tree, stall, market square, snow, people, stroll, town, building
Tag2Text	living room, dog, sit on, blanket, couch, plant, modern Missing: lamp, carpet	Christmas market, Christmas tree, snow, town, people Missing: building
ML-Decoder	living room, lamp, houseplant, cushion, throw pillow, picture frame Bad: property, design, throw Missing: dog, couch, carpet, blanket	Christmas decoration, town square, market, snow, building Bad: human hair, human head, mixed-use

https://arxiv.org/abs/2306.03514

“Anything Models” – Recognize Anything



Segment Anything Model

What can SAM do?

"A new AI model from Meta AI that can 'cut out' any object, in any image, with a single click"

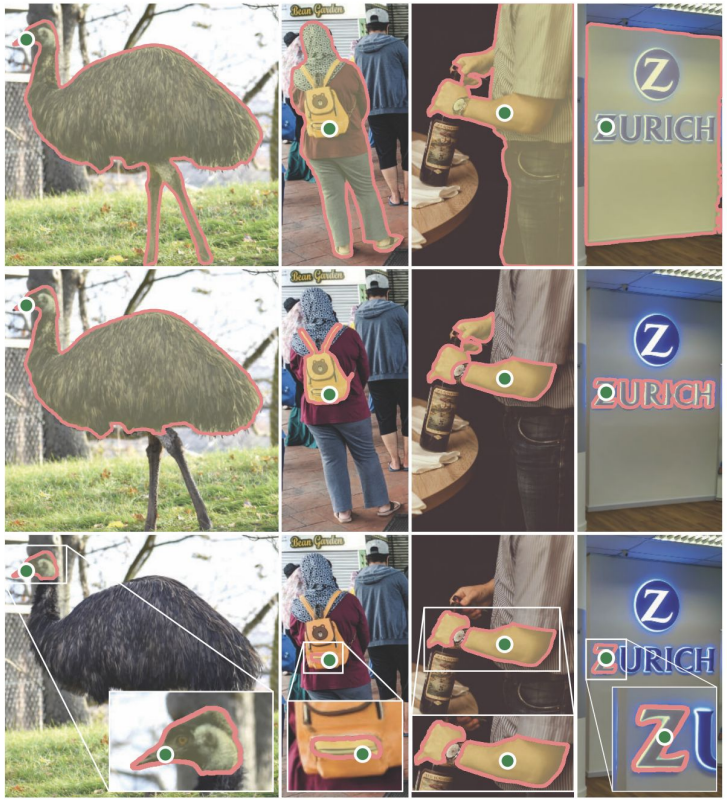
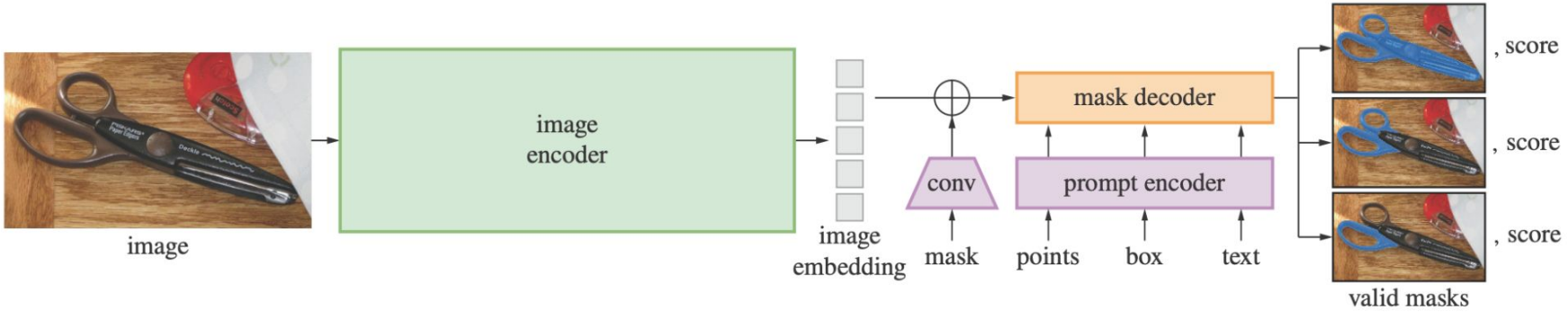
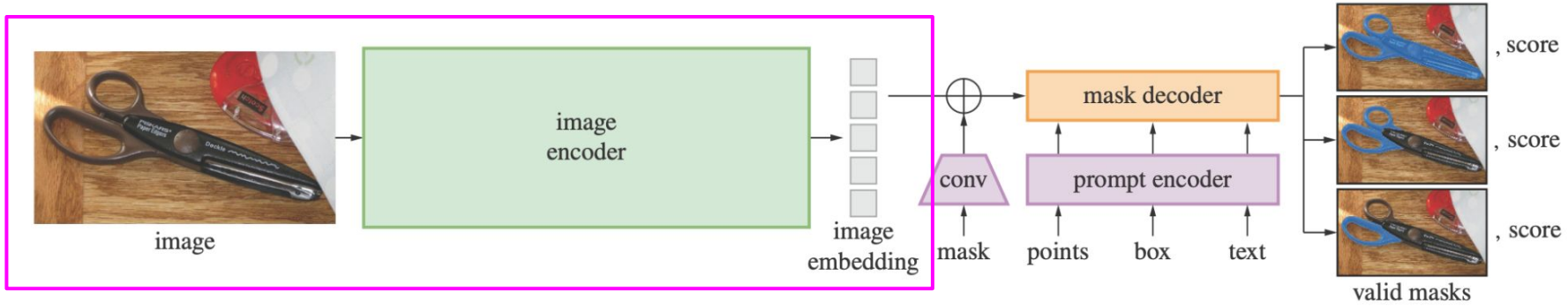


Figure 3: Each column shows 3 valid masks generated by SAM from a single ambiguous point prompt (green circle).

SAM Architecture



SAM Architecture



SAM Architecture

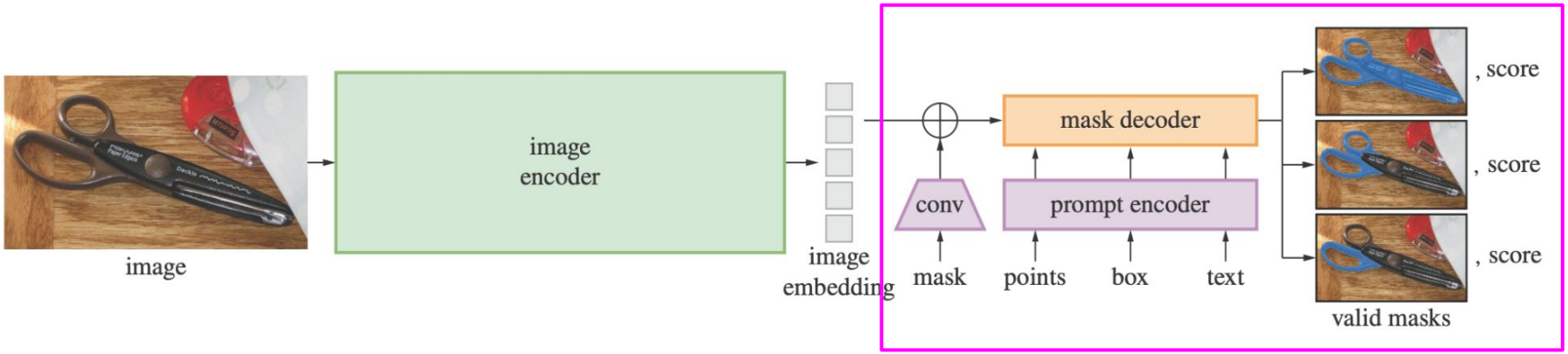
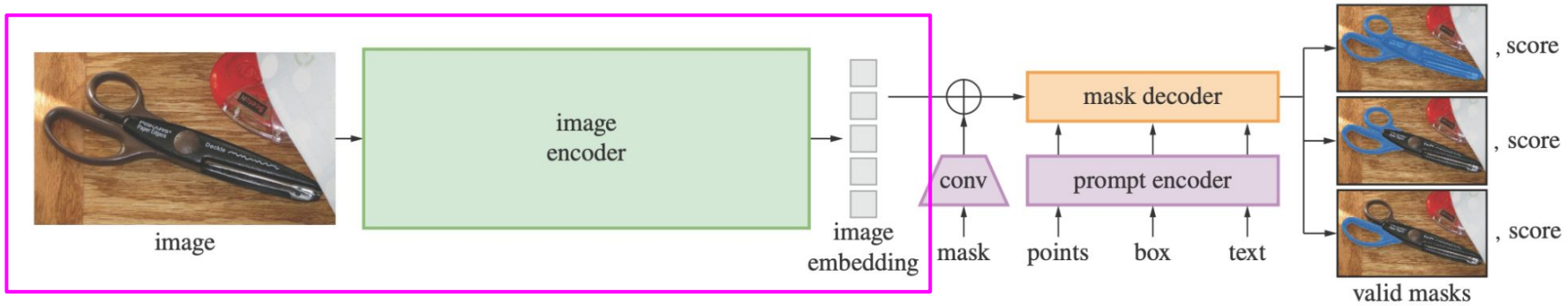


Image Encoder

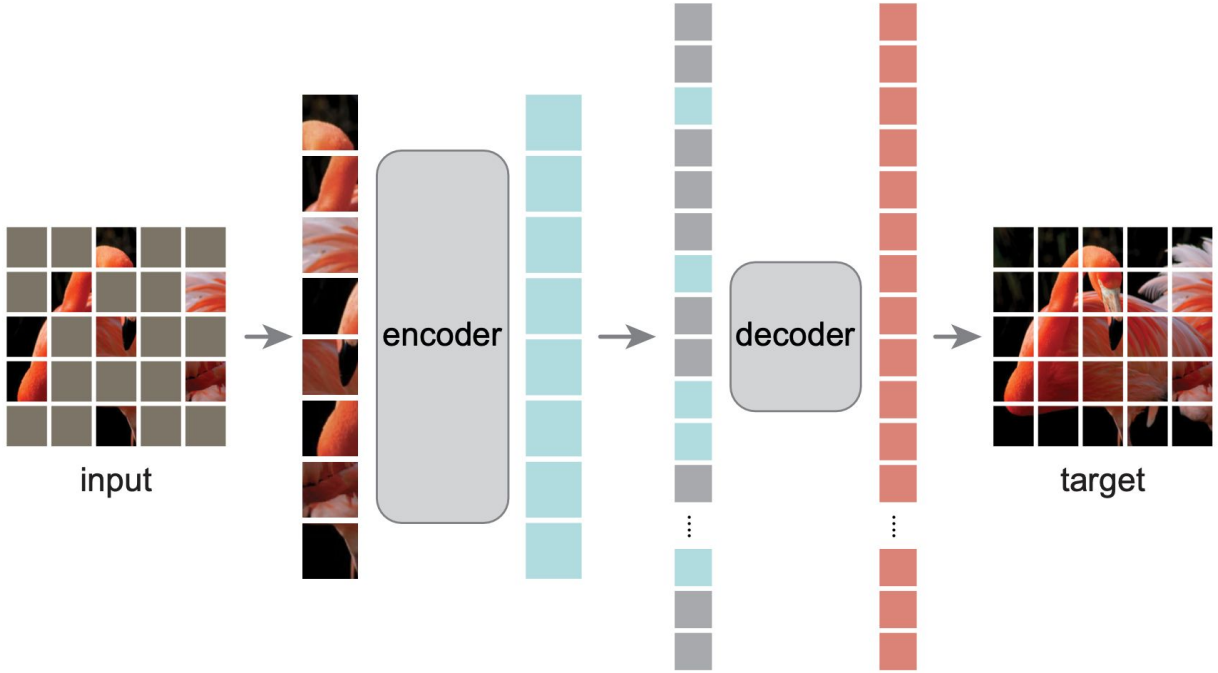


Encoding performed off-line, before user interaction

1024 x 1024 RGB image ⇒ 64 x 64 x 256 channels

Initialized with a pre-trained Masked Autoencoder, ViT-L backend

Masked Autoencoder



Masked Autoencoders Are Scalable Vision Learners, https://arxiv.org/abs/2111.06377

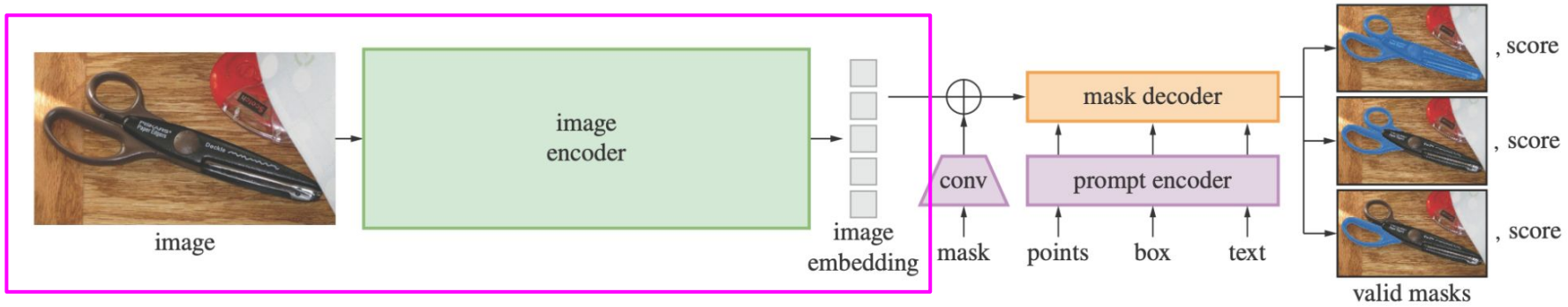
Masked Autoencoder

Masked Autoencoders Are Scalable Vision Learners, https://arxiv.org/abs/2111.06377

Masked Autoencoder

Masked Autoencoders Are Scalable Vision Learners, https://arxiv.org/abs/2111.06377

Image Encoder

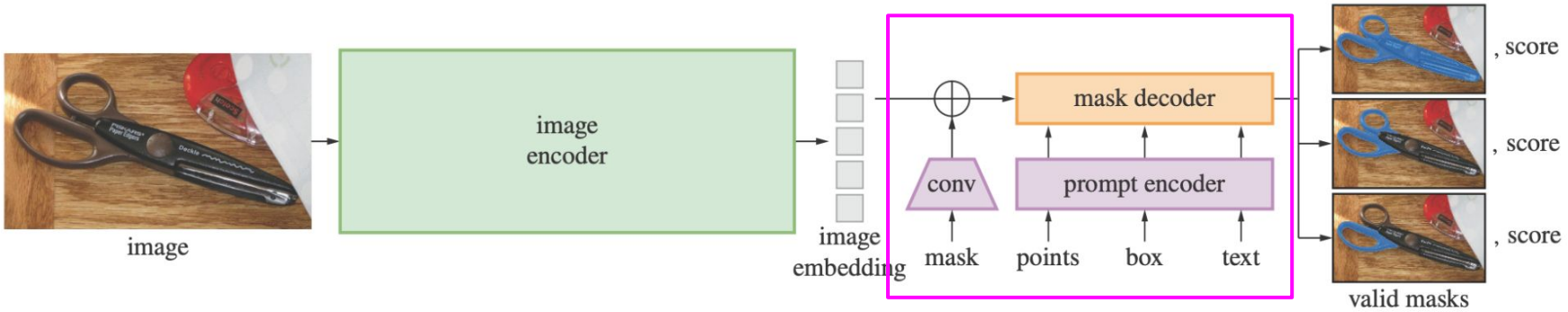


Encoding performed off-line, before user interaction

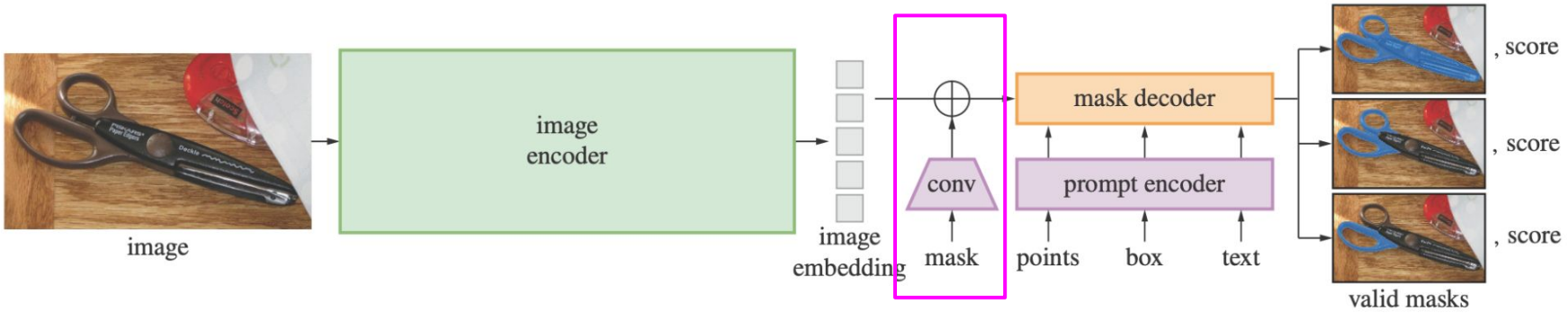
1024 x 1024 RGB image ⇒ 64 x 64 x 256 channels

Initialized with a pre-trained Masked Autoencoder, ViT-L backend

Prompt Encoder



Mask Encoder

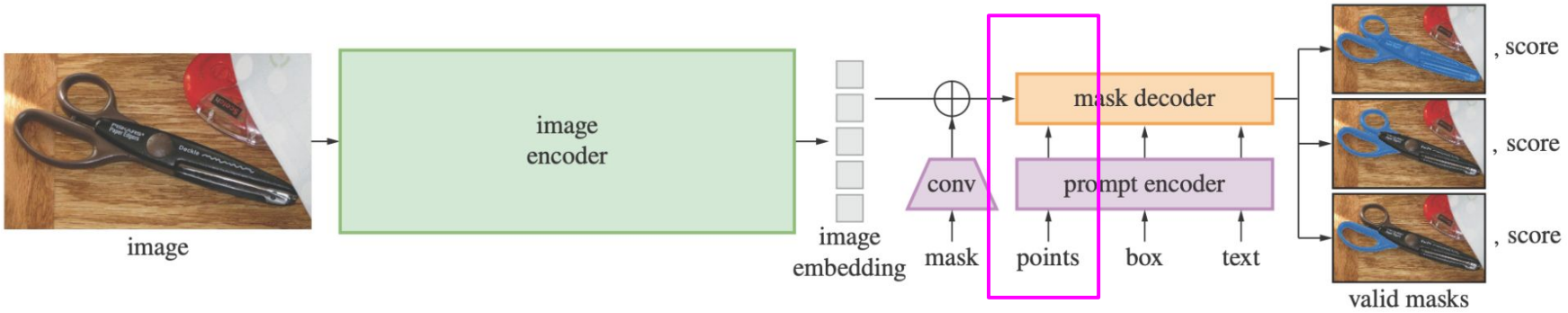


256 x 256 probability mask

Two 2 x 2, stride-2 convolutions, then 1 x 1 conv ⇒ 64 x 64 x 256

Element-wise addition
Learned (!) embedding for “no mask”

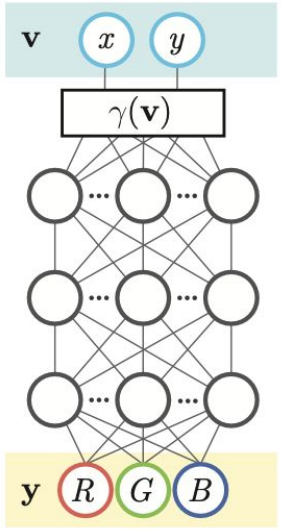
Prompt Encoder



Positional Encoding using Fourier Features

Summed with one of two learned embeddings for foreground/background

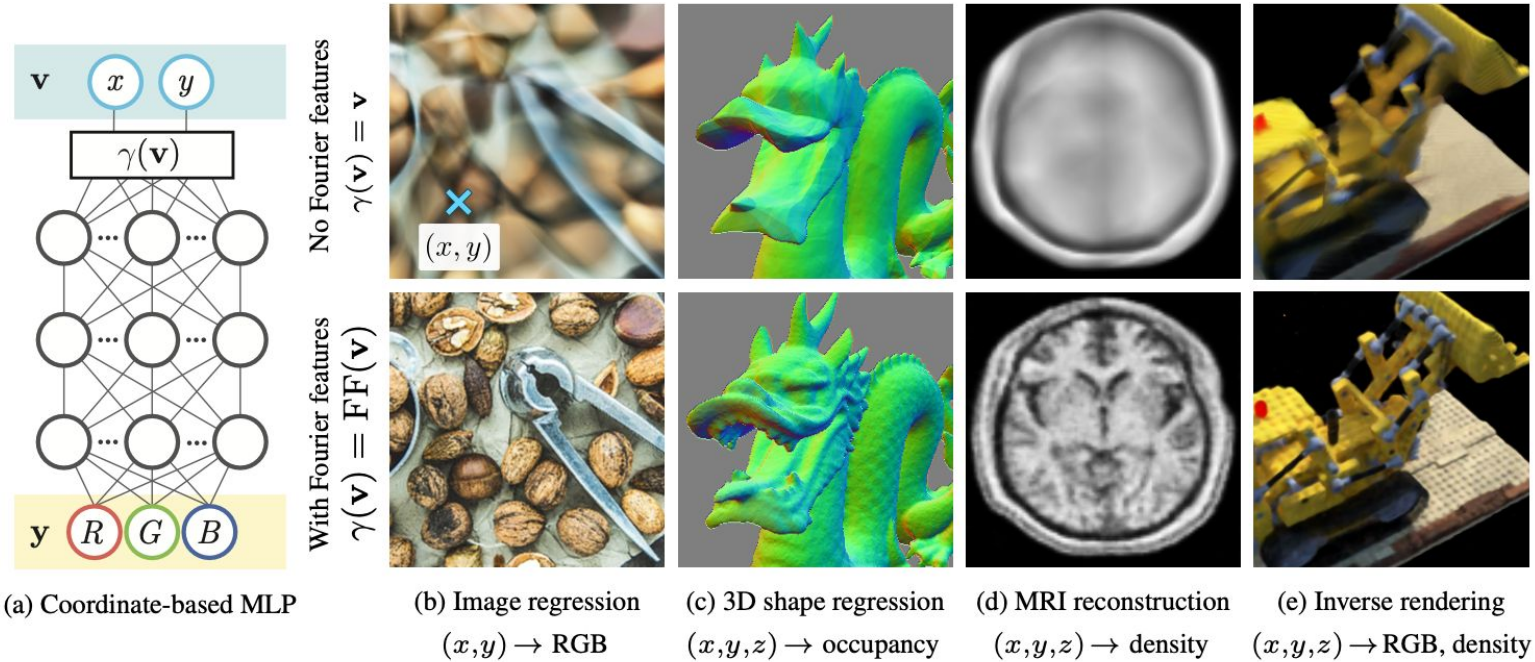
Positional Encoding – Fourier Features



(a) Coordinate-based MLP

Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains, https://arxiv.org/abs/2006.10739

Positional Encoding – Fourier Features



(a) Coordinate-based MLP

(b) Image regression
$(x,y) \rightarrow$ RGB

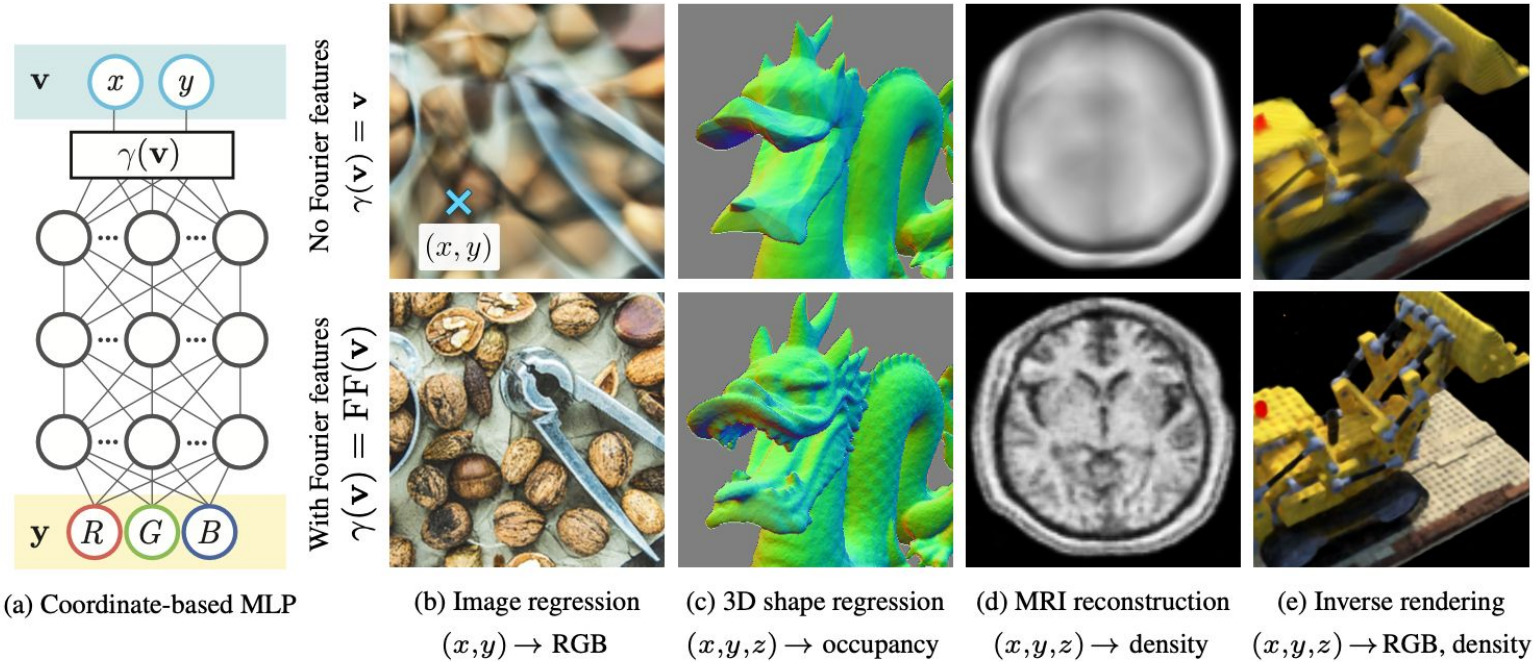
(c) 3D shape regression
$(x,y,z) \rightarrow$ occupancy

(d) MRI reconstruction
$(x,y,z) \rightarrow$ density

(e) Inverse rendering
$(x,y,z) \rightarrow$ RGB, density

Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains,
https://arxiv.org/abs/2006.10739

Positional Encoding – Fourier Features



(a) Coordinate-based MLP

(b) Image regression
$(x,y) \rightarrow$ RGB

(c) 3D shape regression
$(x,y,z) \rightarrow$ occupancy

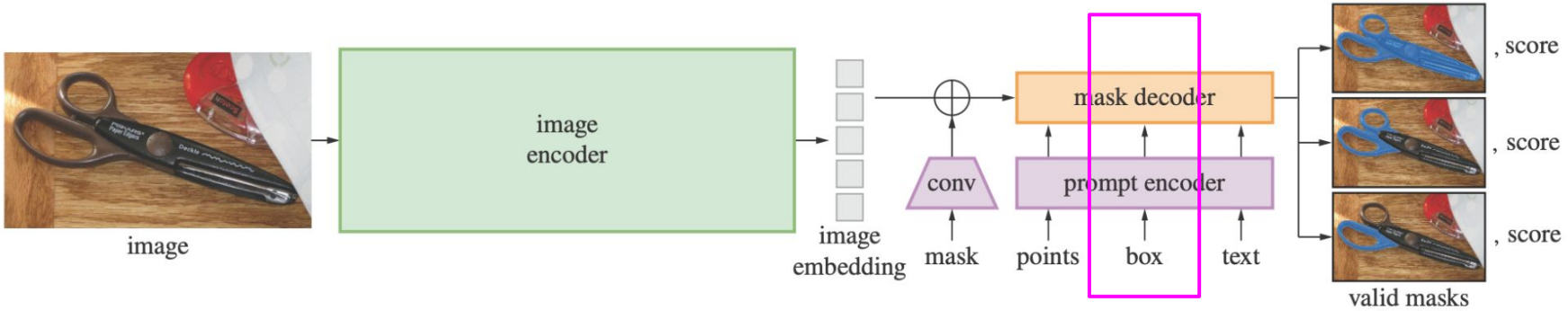
(d) MRI reconstruction
$(x,y,z) \rightarrow$ density

(e) Inverse rendering
$(x,y,z) \rightarrow$ RGB, density

$$\gamma(\mathbf{v}) = \left[a_1 \cos(2\pi \mathbf{b}_1^{\mathrm{T}} \mathbf{v}), a_1 \sin(2\pi \mathbf{b}_1^{\mathrm{T}} \mathbf{v}), \ldots, a_m \cos(2\pi \mathbf{b}_m^{\mathrm{T}} \mathbf{v}), a_m \sin(2\pi \mathbf{b}_m^{\mathrm{T}} \mathbf{v})\right]^{\mathrm{T}}$$

Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains,
https://arxiv.org/abs/2006.10739

Prompt Encoder

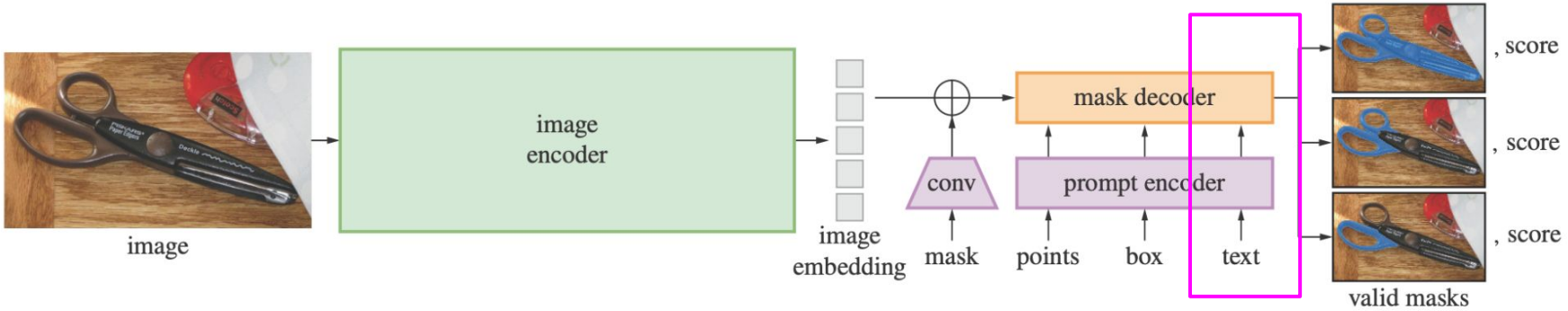


Embedding pair:

Positional encoding of top-left corner,
summed with a learned embedding “top-left corner”

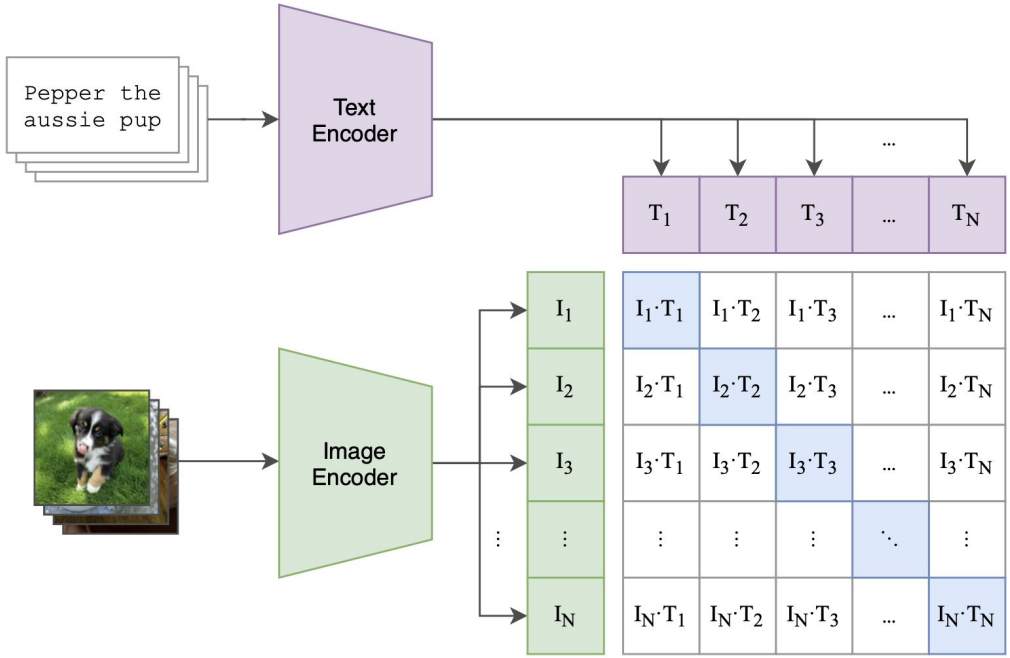
Same for “bottom- right corner”

Prompt Encoder



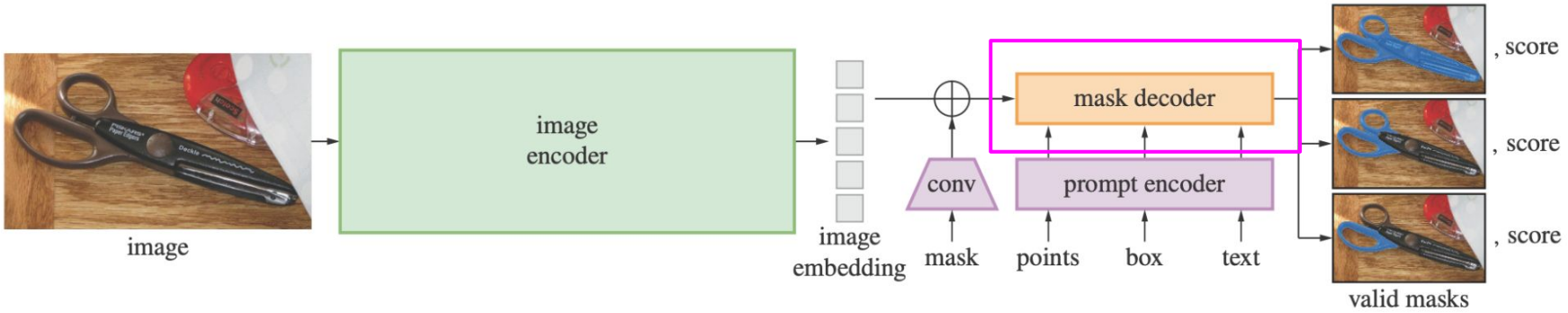
CLIP (Contrastive Language-Image Pre-training) text encoder

CLIP (Contrastive Language-Image Pre-training) text encoder



Learning Transferable Visual Models From Natural Language Supervision, https://arxiv.org/abs/2103.00020

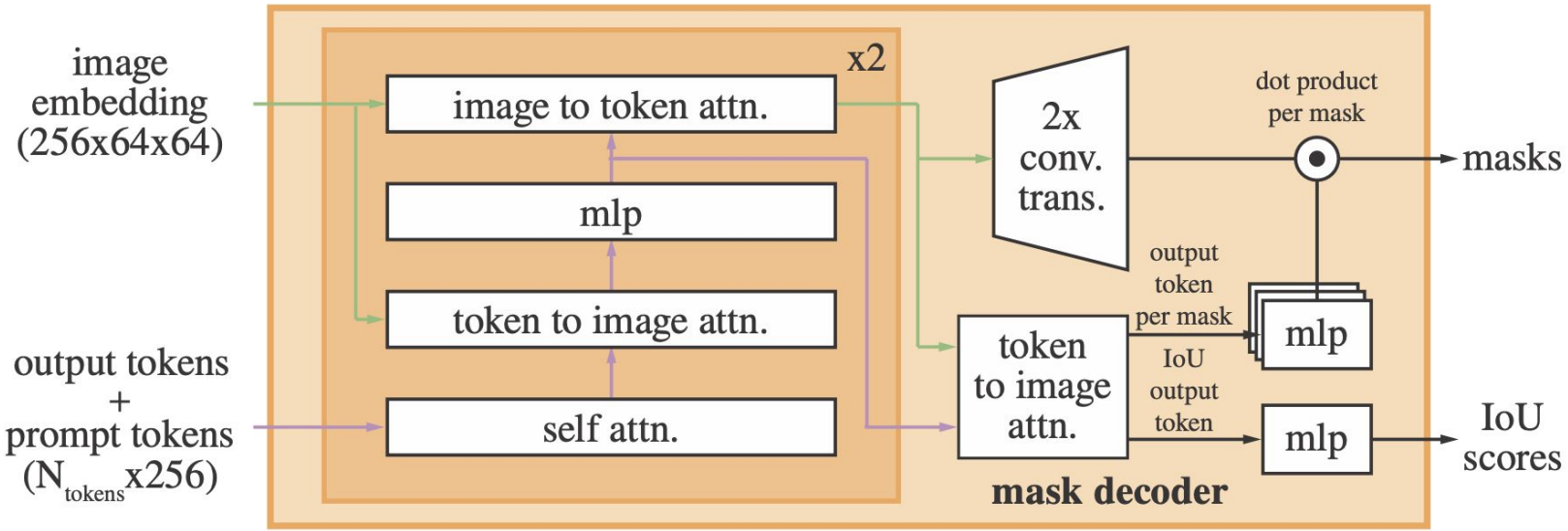
Mask Decoder



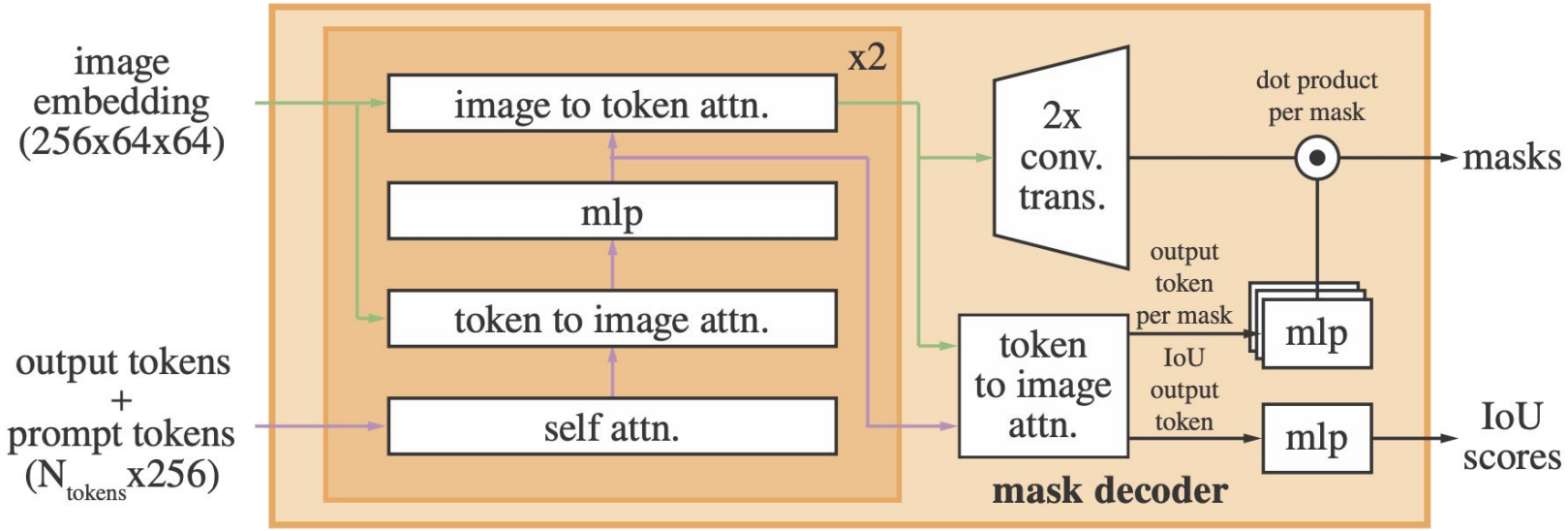
Based on Transformer segmentation models

Produces 3 masks and predicted IoU (Intersection over Union) scores

Mask Decoder



Mask Decoder



To ensure the decoder has access to critical geometric information the positional encodings are added to the image embedding whenever they participate in an attention layer.

Additionally, the entire original prompt tokens (including their positional encodings) are re-added to the updated to- kens whenever they participate in an attention layer.

Training Regime & Loss

Simulated User interactions

Three predicted masks: Whole, part, sub-part

Focal loss + dice loss + mean-square-error for IoU predictions

Compute 3 losses, but only backpropagate the smallest one

Training Regime & Loss

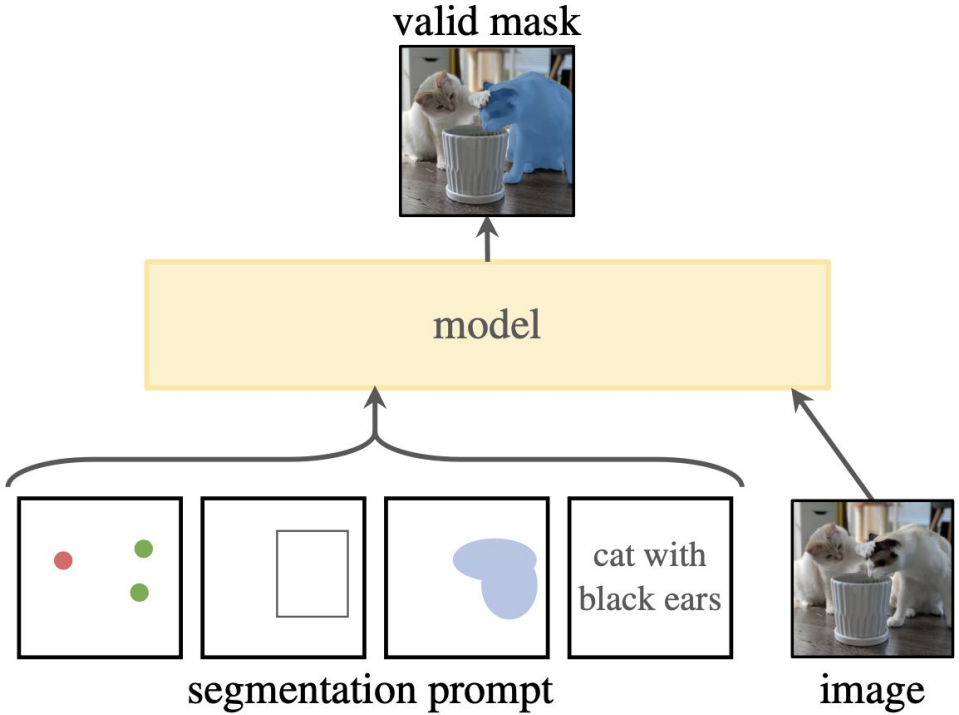
Simulated User interactions

Three predicted masks: Whole, part, sub-part

Focal loss + dice loss + mean-square-error for IoU predictions

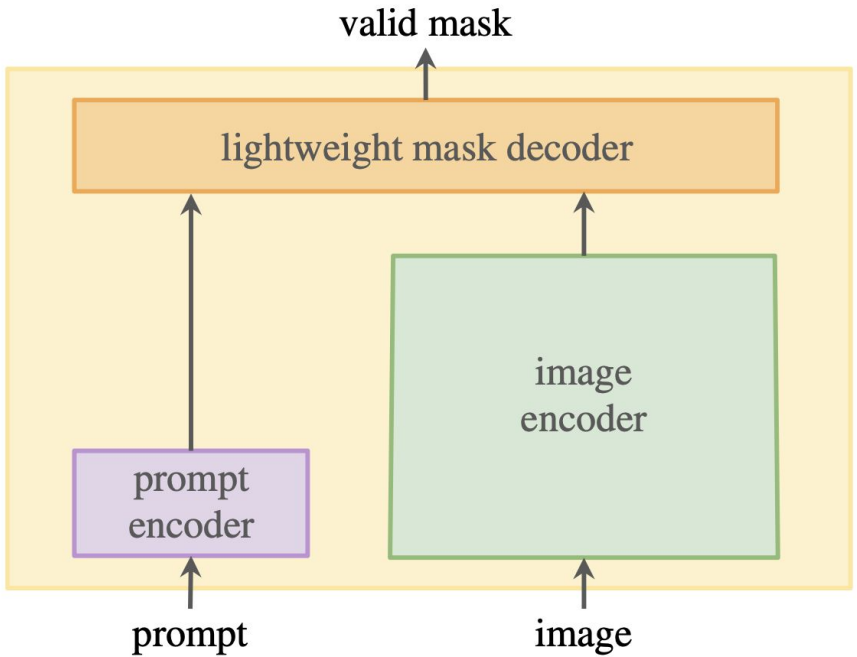
Compute 3 losses, but only backpropagate the smallest one

SAM Contributions 1 / 3



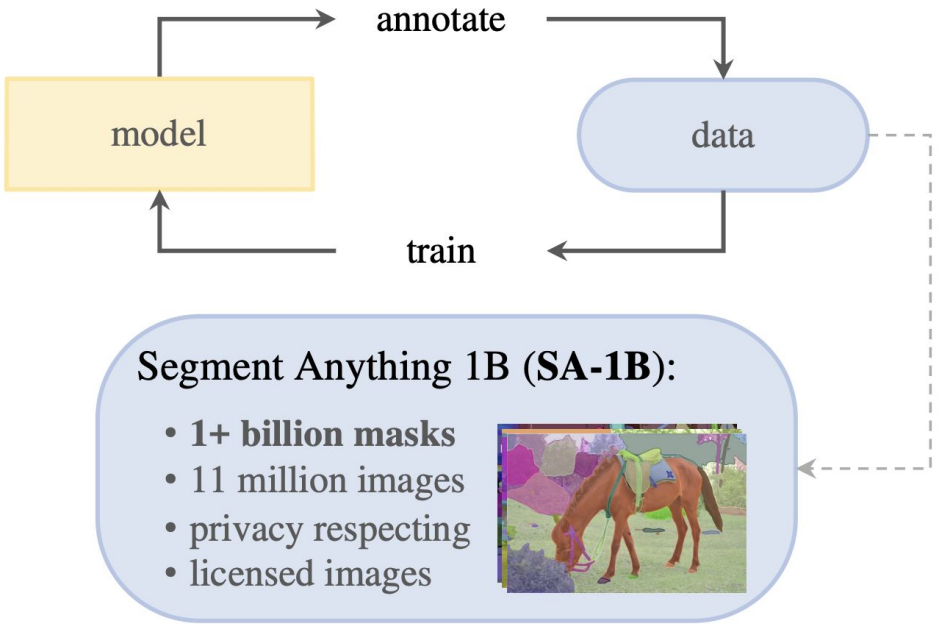
(a) **Task**: promptable segmentation

SAM Contributions 2 / 3



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

SAM Contributions 3 / 3



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

SAM Dataset

Manual

4.3M masks / 120k images

Semi-automatic

5.9M masks / 180k images

Fully automatic

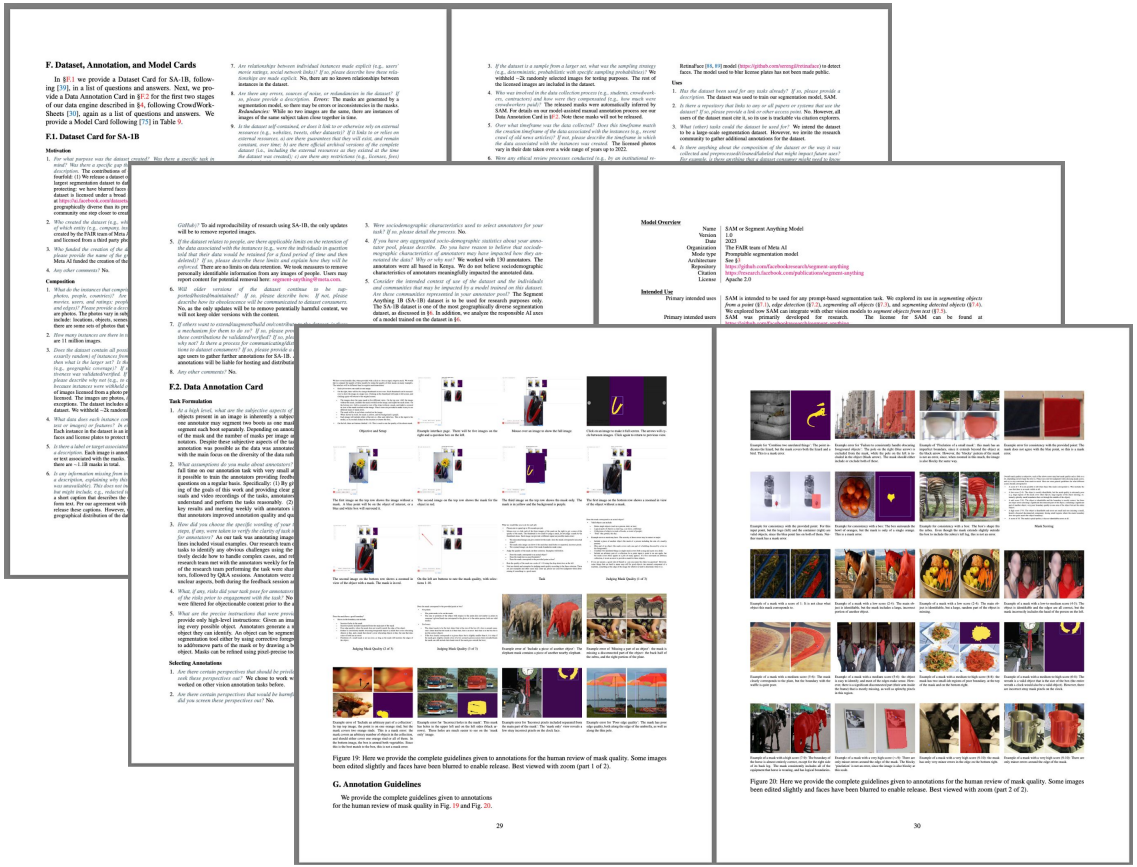
1.1B masks / 11M images

SAM Dataset – Responsible AI

Data set card

Annotation card

Model card



SAM Dataset – Responsible AI

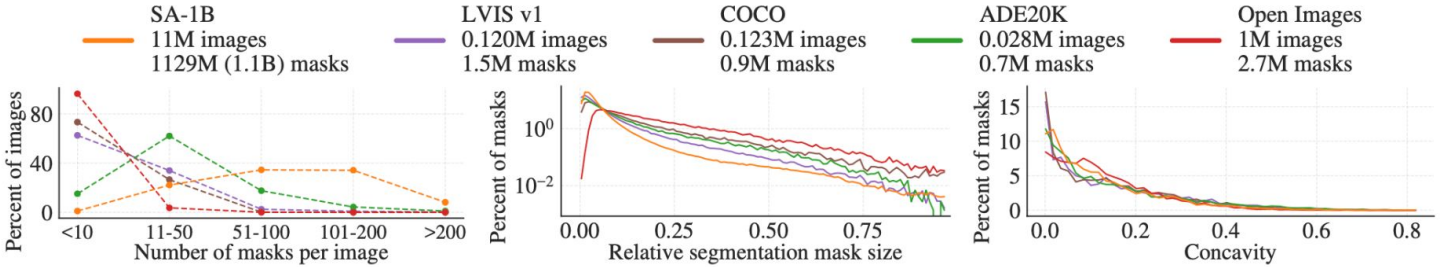


Figure 6: Dataset mask properties. The legend references the number of images and masks in each dataset. Note, that SA-1B has 11× more images and 400× more masks than the largest existing segmentation dataset Open Images [60].

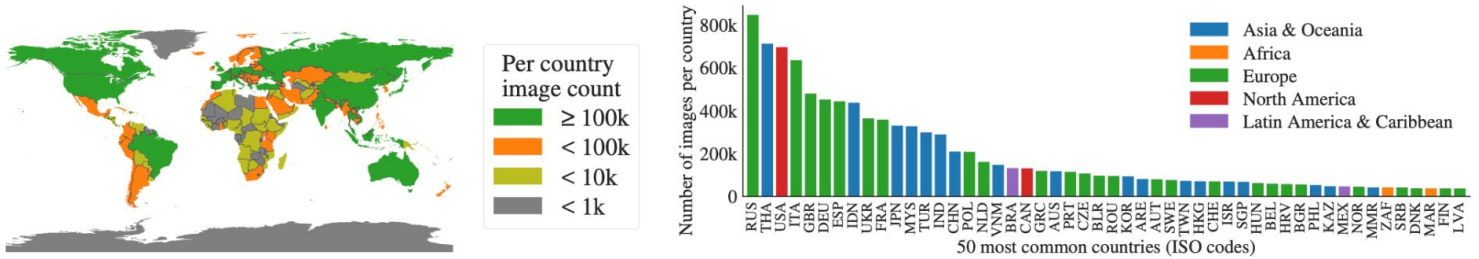
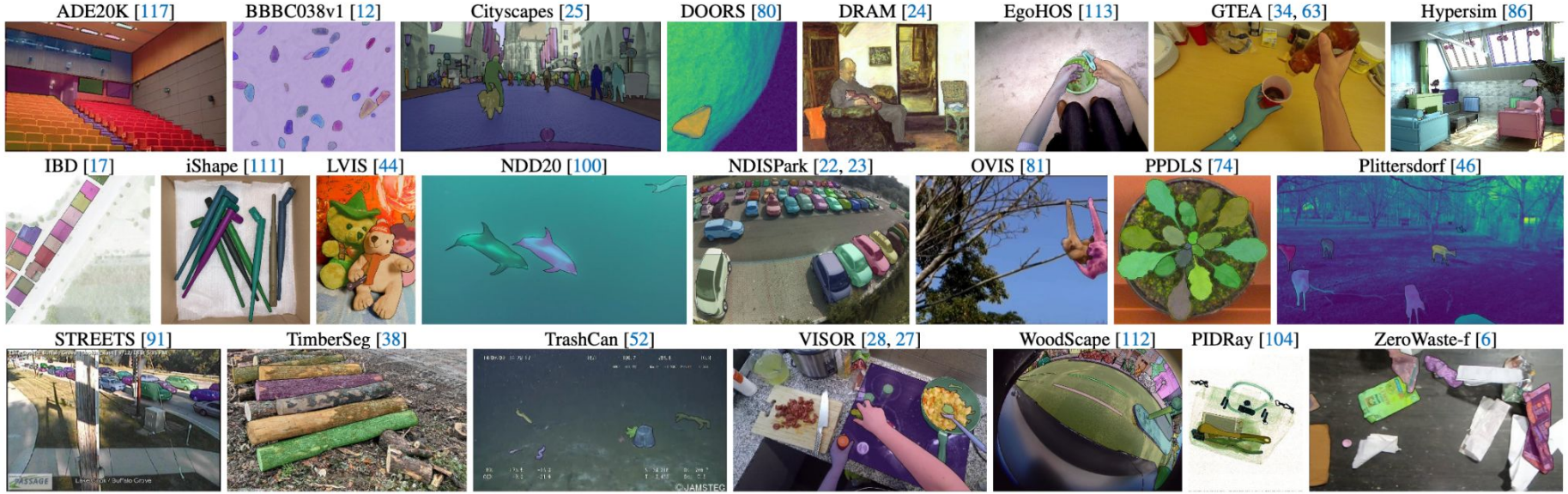


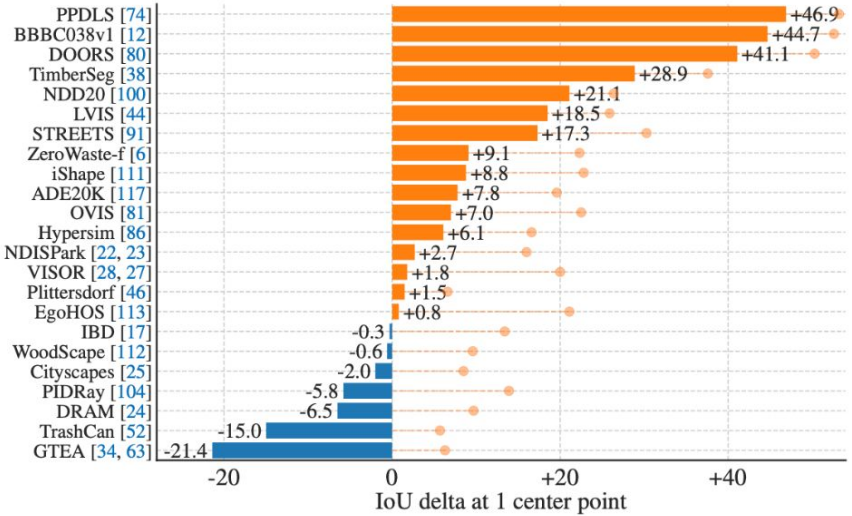
Figure 7: Estimated geographic distribution of SA-1B images. Most of the world's countries have more than 1000 images in SA-1B, and the three countries with the most images are from different parts of the world.

Zero-shot Generalization

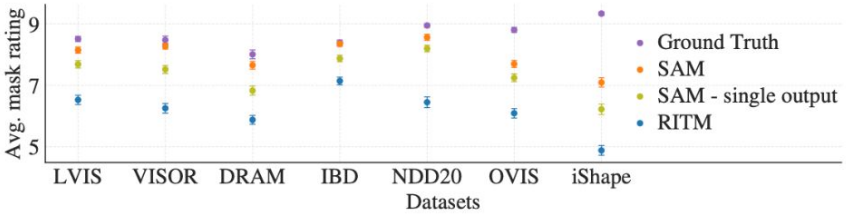
Zero-Shot Single Point Valid Mask Evaluation



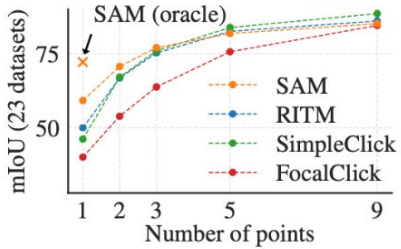
Zero-Shot Single Point Valid Mask Evaluation



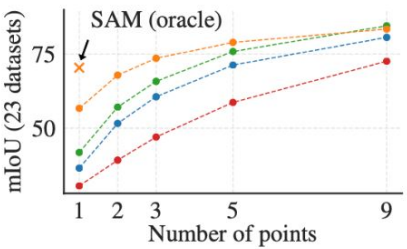
(a) SAM *vs.* RITM [92] on 23 datasets



(b) Mask quality ratings by human annotators



(c) Center points (default)

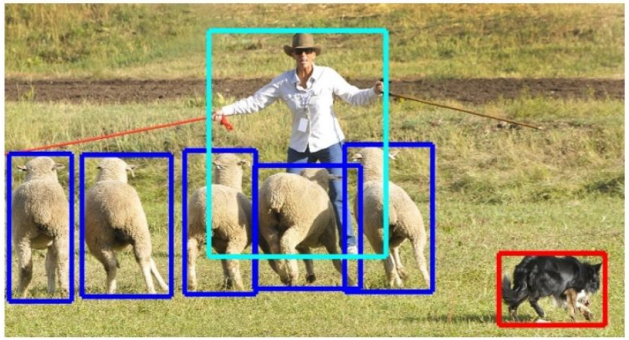


(d) Random points

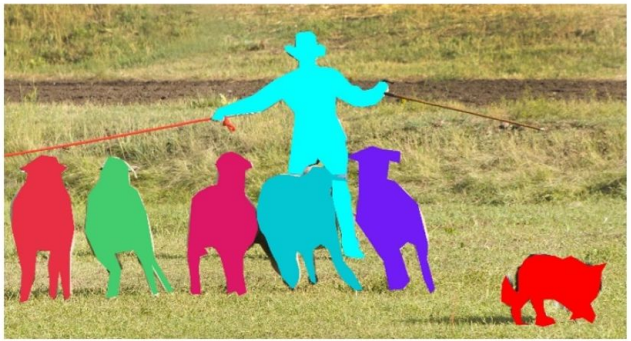
Zero-Shot Edge Detection

Generalization

Zero-Shot Object Proposals (Boxes)



Zero-Shot Instance Segmentation



Microsoft COCO: Common Objects in Context
https://arxiv.org/abs/1405.0312

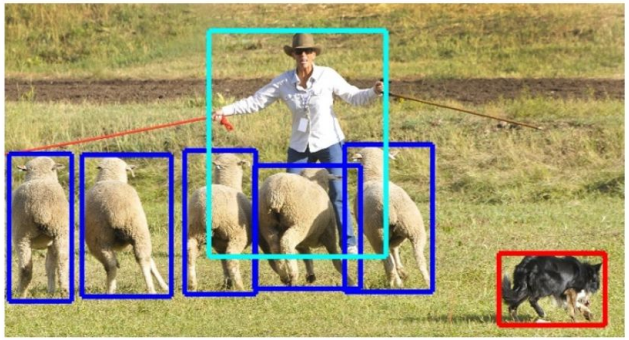
Generalization

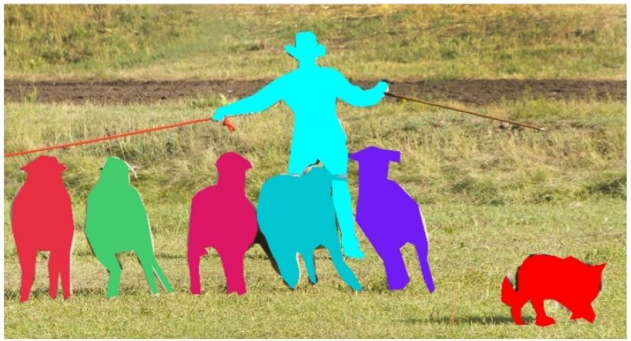
Zero-Shot Object Proposals (Boxes)

✅ 👍

Zero-Shot Instance Segmentation

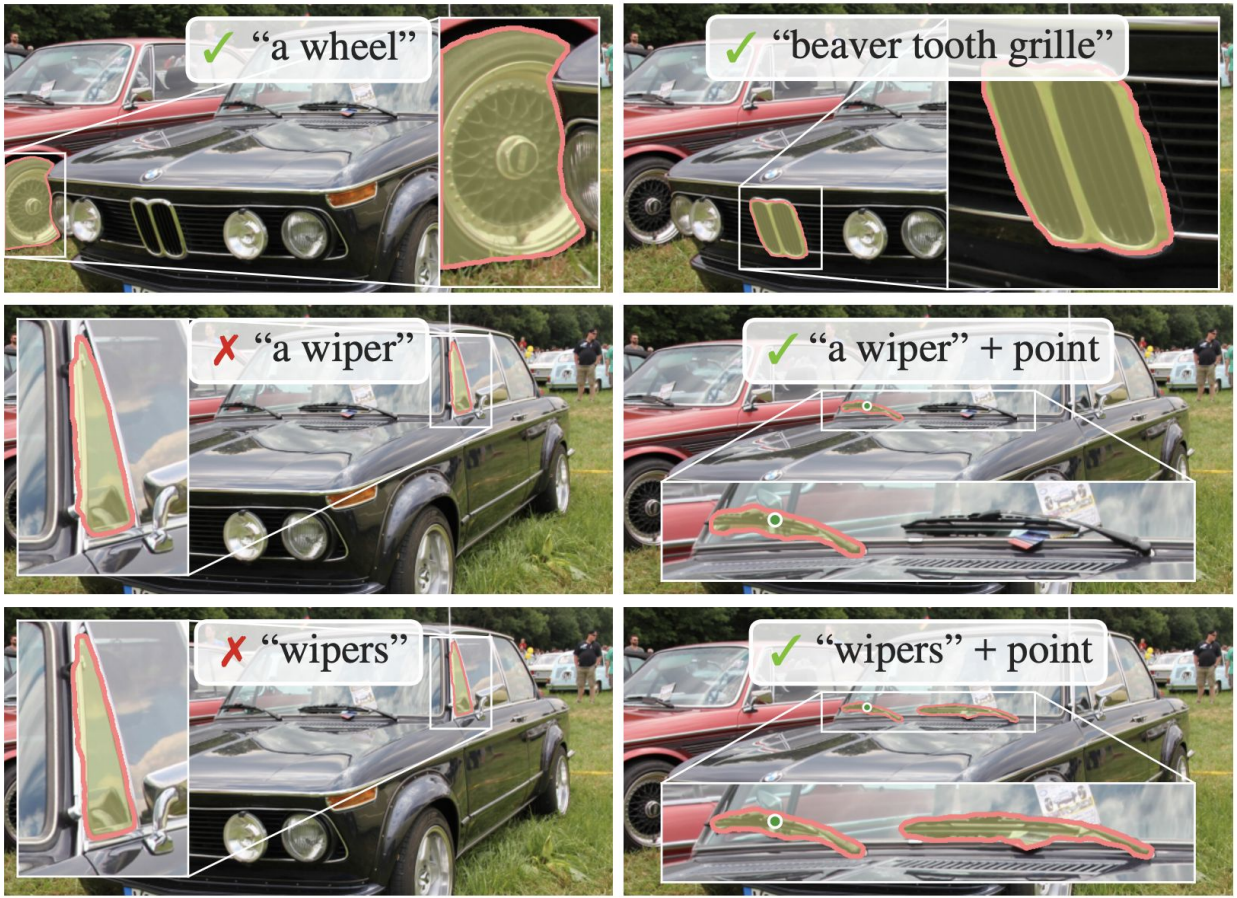
✅ 👍





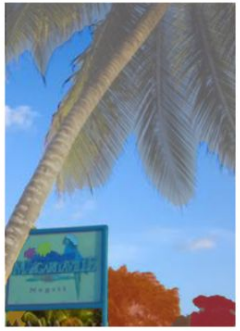
Microsoft COCO: Common Objects in Context
https://arxiv.org/abs/1405.0312

Zero-shot text-to-mask









THINH THANH



contact@mva.ai

Segment Anything and the Rise of Foundation Models

MEDICAL VOLUME ANNOTATOR

René Donner