Foundation Model for Image Segmentation

at Visgraf by Irving Badolato

About Me

Irving Badolato

Assistant Professor

CARTO – FEN – UERJ

- DSc. Computational Sciences (UERJ, in progress)
- MSc. Computer & Systems Engineering (UFRJ, 2014)
- Grad. in Electrical Engineering (UERJ, 2010)

- Researcher at LFSR – UERJ
- Developer at the E-FOTO Project
- Worked at FIOCRUZ, CPRM, PUC-RJ, CEFEN, FUNCATE and PTV Tecnologia

The Presentation Agenda

1. What is a Foundation Model?
2. Segment Anything Model
3. Diving into architecture
4. Model pre-training
5. What came next?

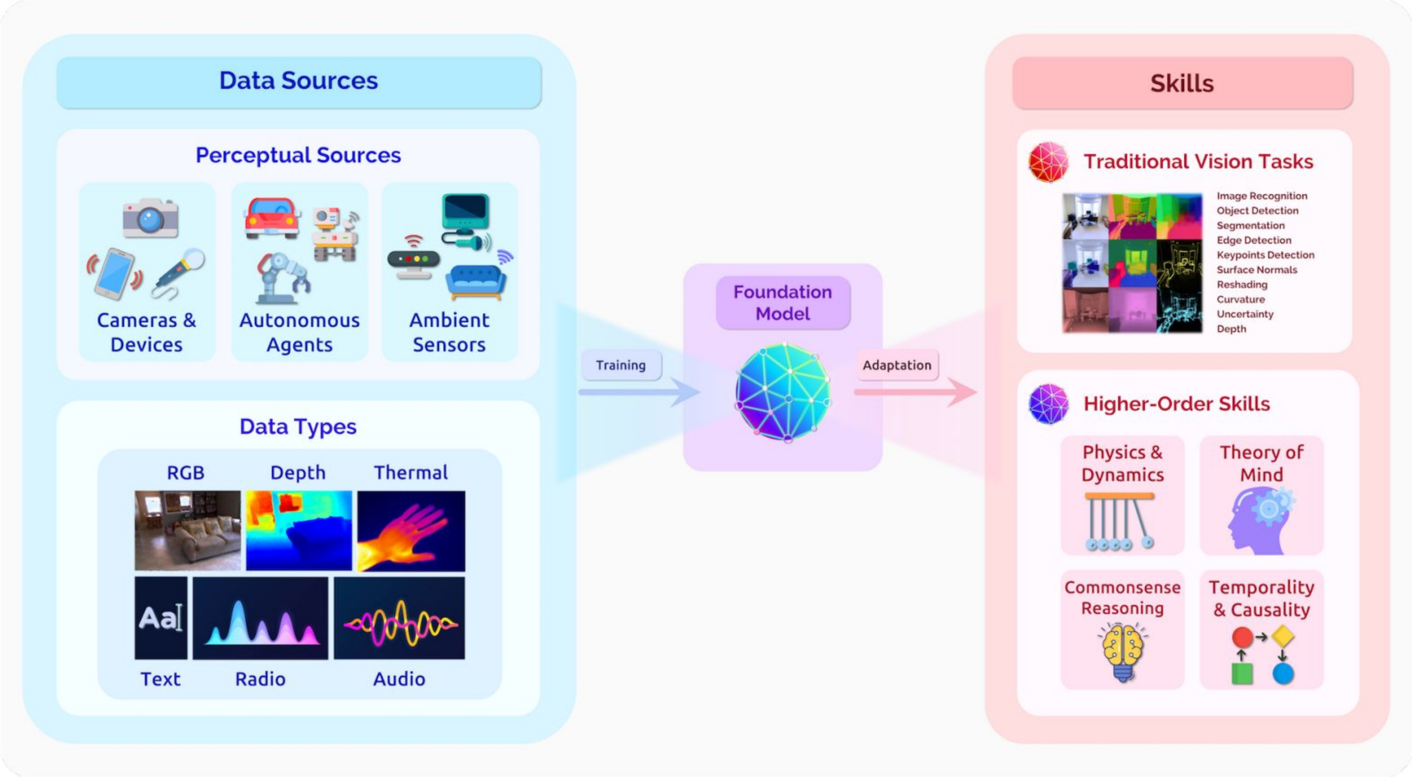
Center for Research on Foundation Models

Stanford University Human-Centered Artificial Intelligence

– *What is a Foundation Model?*

In recent years, a new successful paradigm for building AI systems has emerged: Train one model on a huge amount of data and adapt it to many applications. We call such a model a foundation model. (CRFM, 2021)

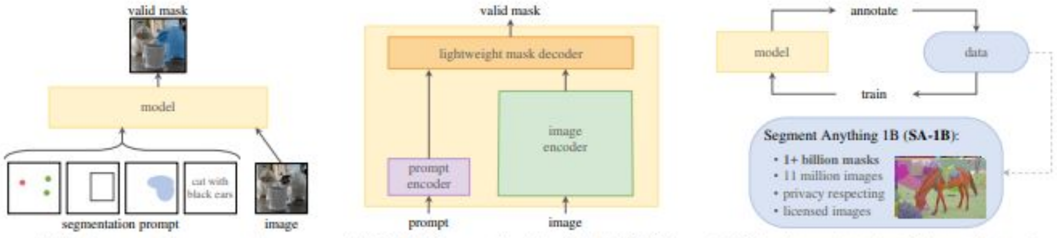
- E.g.: BERT, GPT, CLIP, DALL-E, Stable Diffusion, Copilot, HLS Geospatial FM
- Capabilities: Language, Vision, Interaction, Robotics, Search and reasoning
- Challenges: sudden faults, biases, lack of understanding and ethics of scale



Data from diverse sources and types is trained into visual knowledge and after adapted for a wide variety of tasks, like image segmentation. (CRFM, 2022)

Segment Anything

Alexander Kirillov[1,2,4] Eric Mintun[2] Nikhila Ravi[1,2] Hanzi Mao[2] Chloe Rolland[3] Laura Gustafson[3]
Tete Xiao[3] Spencer Whitehead Alexander C. Berg Wan-Yen Lo Piotr Dollár[4] Ross Girshick[4]
[1]project lead [2]joint first author [3]equal contribution [4]directional lead
Meta AI Research, FAIR



(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

Figure 1: We aim to build a foundation model for segmentation by introducing three interconnected components: a promptable segmentation *task*, a segmentation *model* (SAM) that powers data annotation and enables zero-shot transfer to a range of tasks via prompt engineering, and a *data* engine for collecting SA-1B, our dataset of over 1 billion masks.

Abstract

We introduce the Segment Anything (SA) project: a new task, model, and dataset for image segmentation. Using our efficient model in a data collection loop, we built the largest segmentation dataset to date (by far), with over 1 ***billion*** *masks on 11M licensed and privacy respecting images. The model is designed and trained to be promptable, so it can transfer zero-shot to new image distributions and tasks. We evaluate its capabilities on numerous tasks and find that its zero-shot performance is impressive – often competitive with or even superior to prior fully supervised results. We*

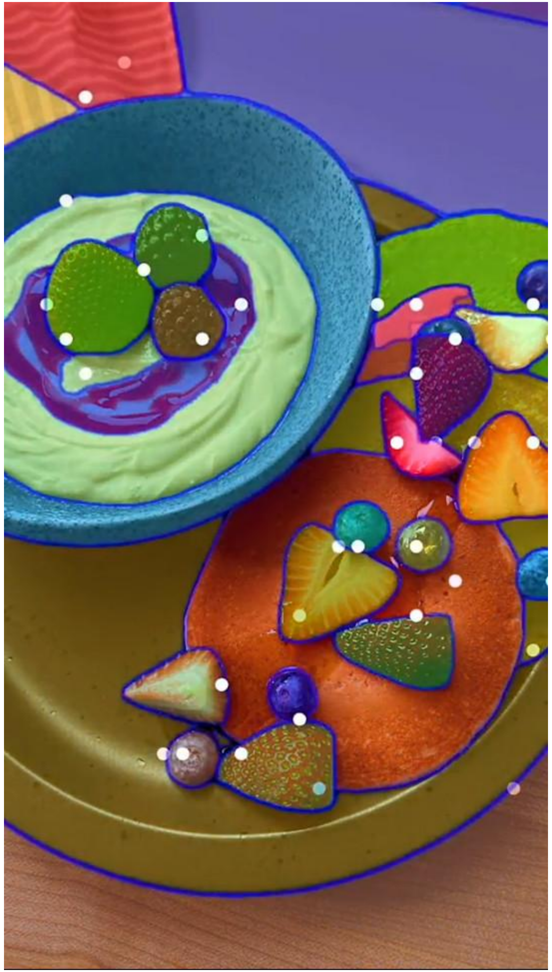
matching in some cases) fine-tuned models [10, 21]. Empirical trends show this behavior improving with model scale, dataset size, and total training compute [56, 10, 21, 51].

Foundation models have also been explored in computer vision, albeit to a lesser extent. Perhaps the most prominent illustration aligns paired text and images from the web. For example, CLIP [82] and ALIGN [55] use contrastive learning to train text and image encoders that align the two modalities. Once trained, engineered text prompts enable zero-shot generalization to novel visual concepts and data distributions. Such encoders also compose effectively with

SAM → A *Foundation model*

A step toward the first foundation model for image segmentation, capable of one-click segmentation of any object from photos or videos + zero-shot transfer to other segmentation tasks. (Meta AI, 2023)

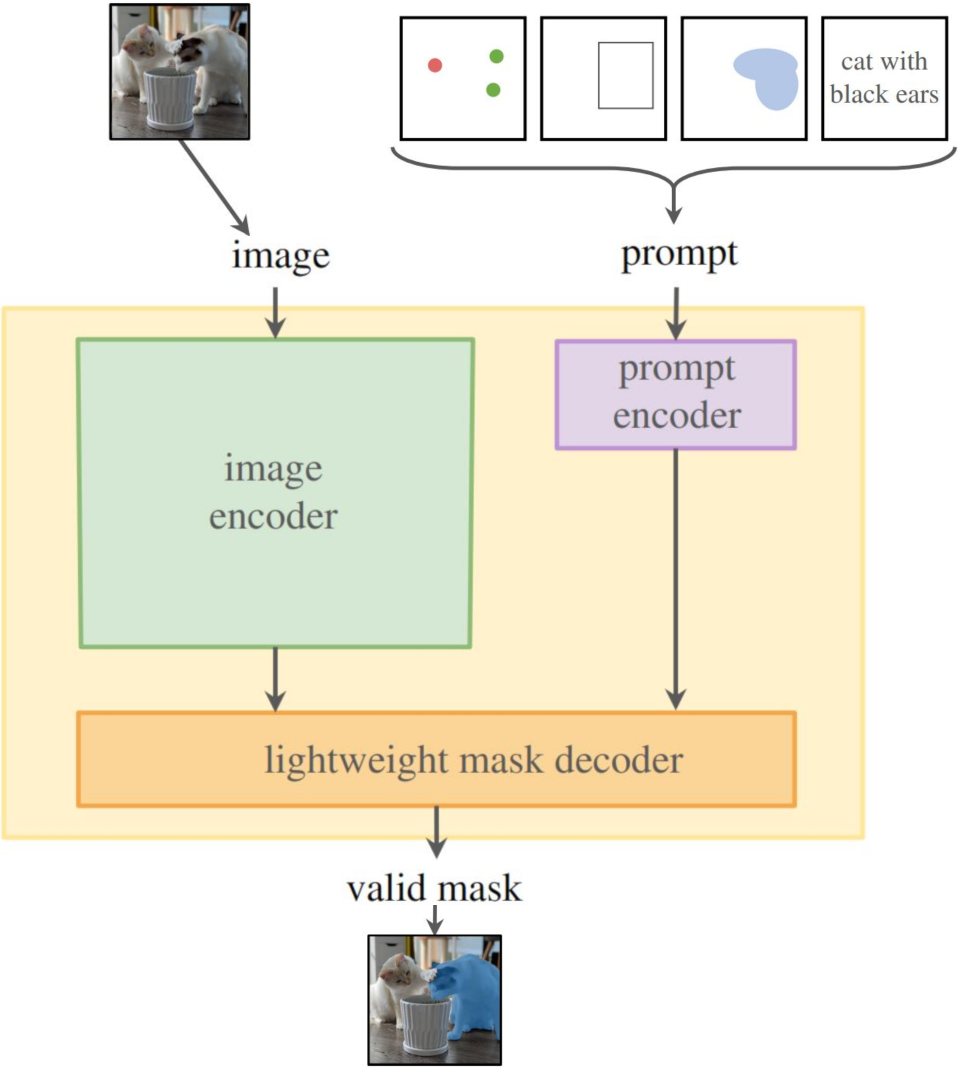
The SAM 1 billion mask (SA-1B) dataset is the largest labeled segmentation dataset to date. It is specifically designed for the development and evaluation of advanced segmentation models. (Buhl, 2023)



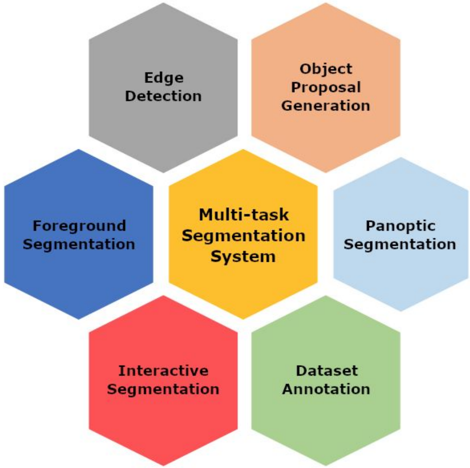
Conceptual design

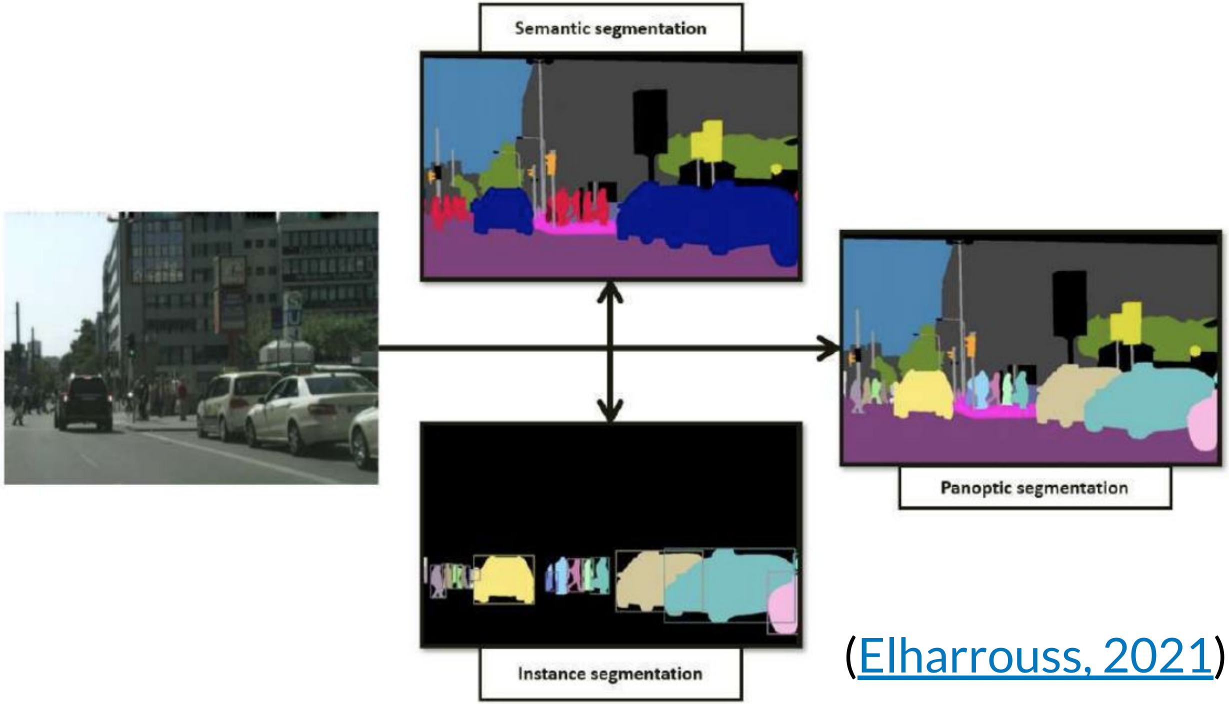
The goal impose main constraints:

1. Support flexible prompts;
2. Compute masks in real-time to allow interactive use;
3. Must be ambiguity-aware.



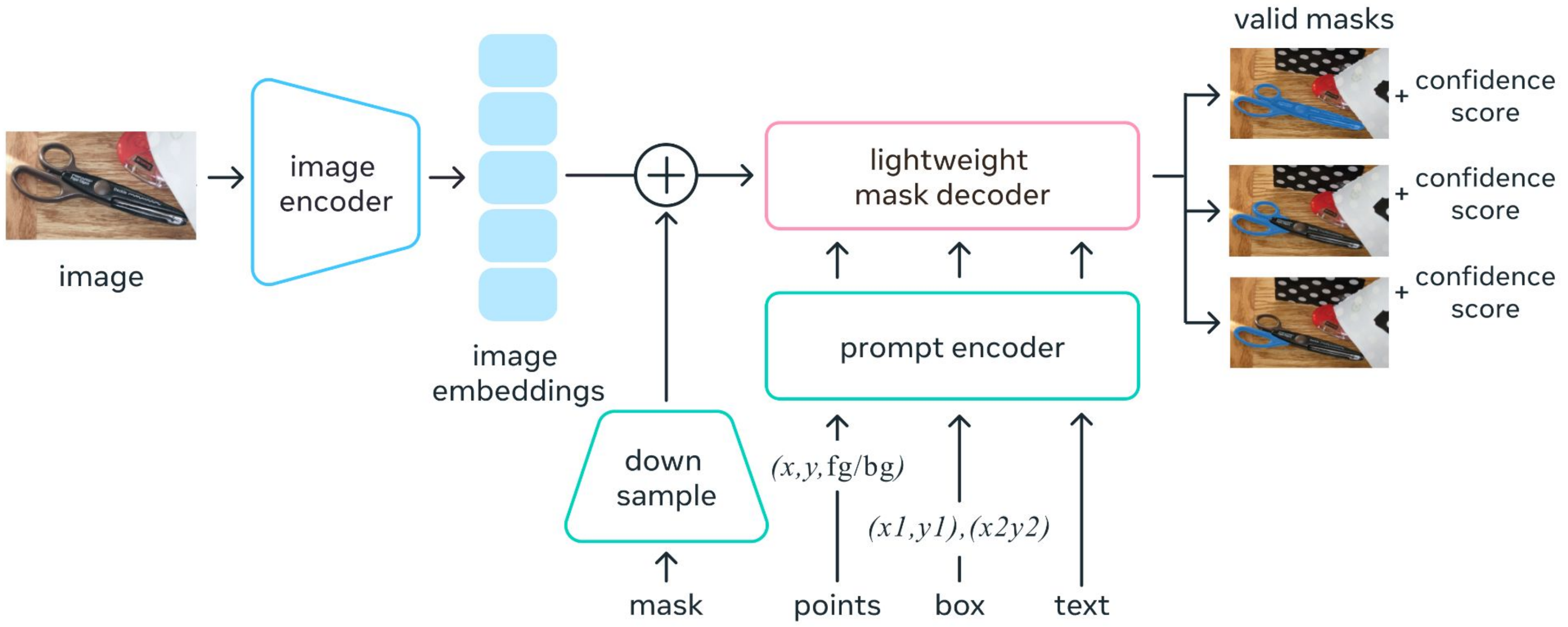
Related tasks



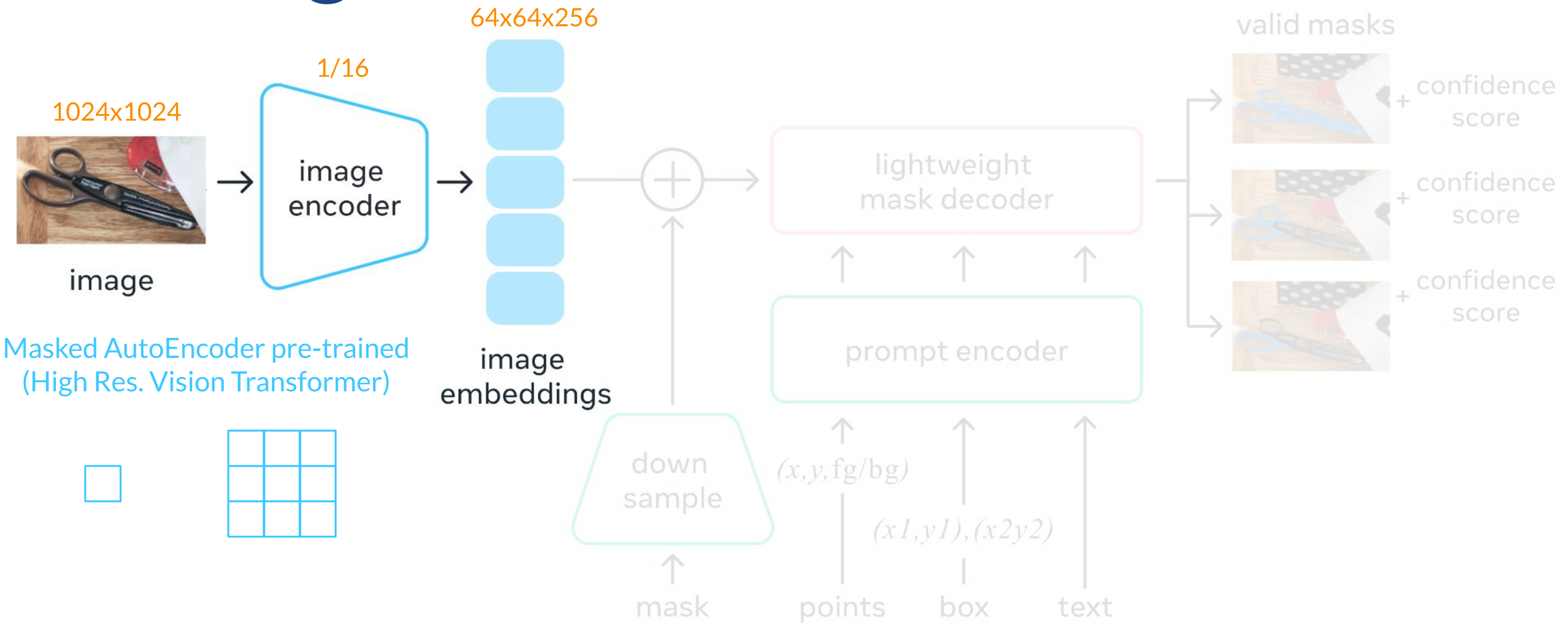


(Elharrouss, 2021)

Architecture overview

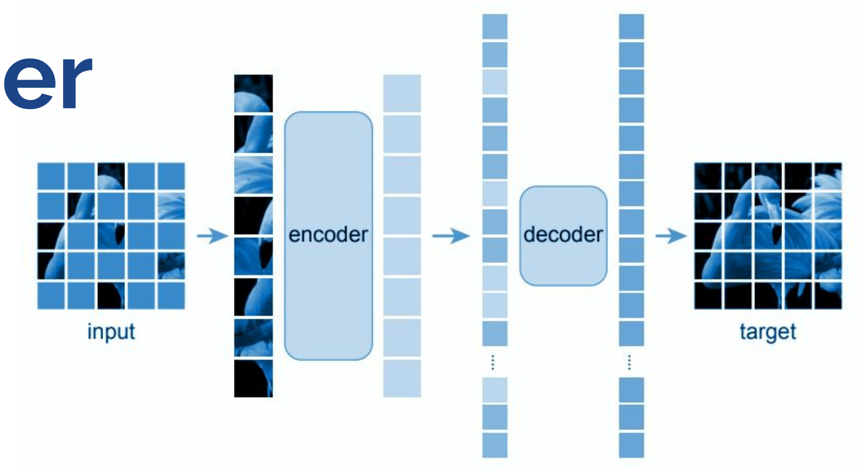


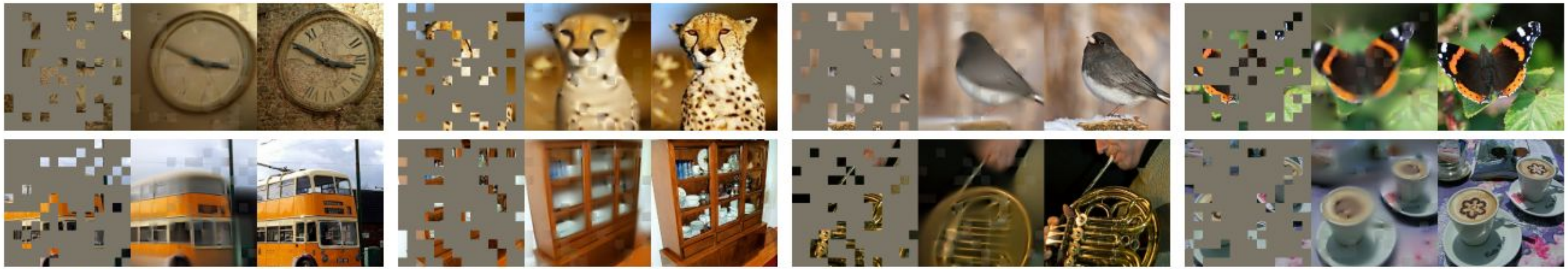
The image encoder



A Masked AutoEncoder

Visible patches are encoded, mask tokens are introduced after the encoder, and processed by a small decoder that reconstructs the original image in pixels. (He, 2022)

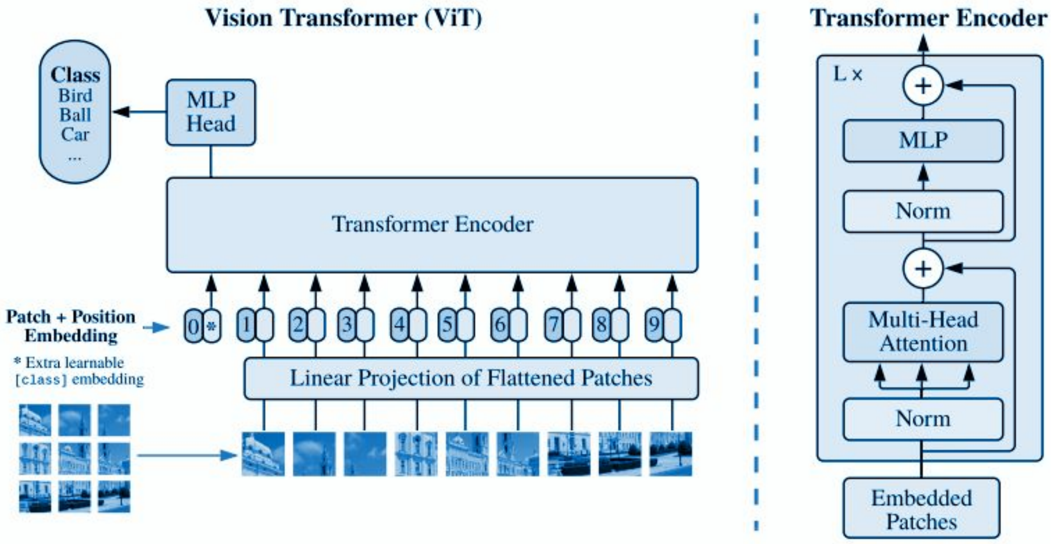




Like a Vision Transformer

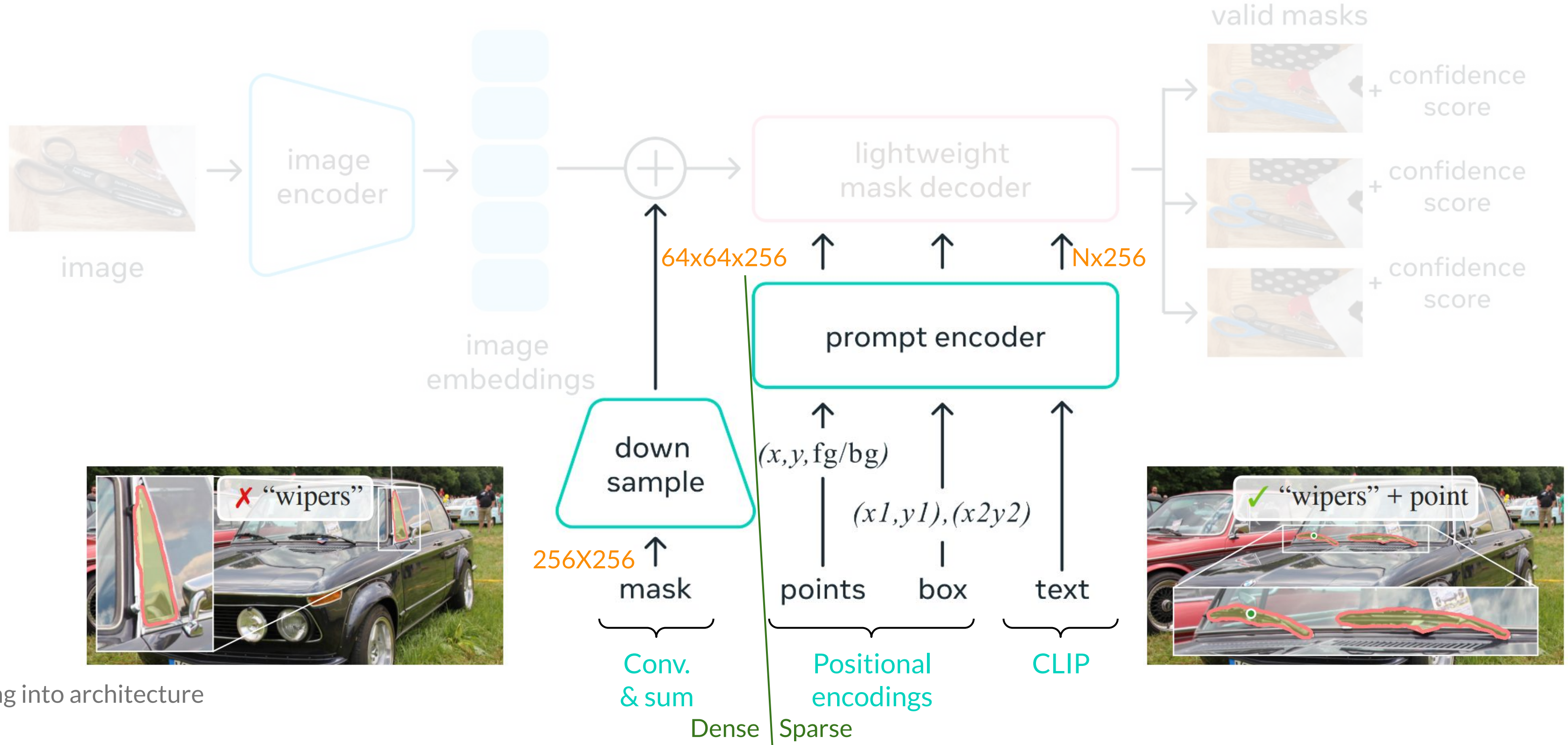
- ➔ *Split image into fixed-size patches;*
- ➔ *Embed each of them with positions;*
- ➔ *Feed the resulting sequence to a standard transformer encoder;*
- ➔ *Add an extra learnable token to the sequence to perform classification.*

([Dosovitskiy, 2020](#))

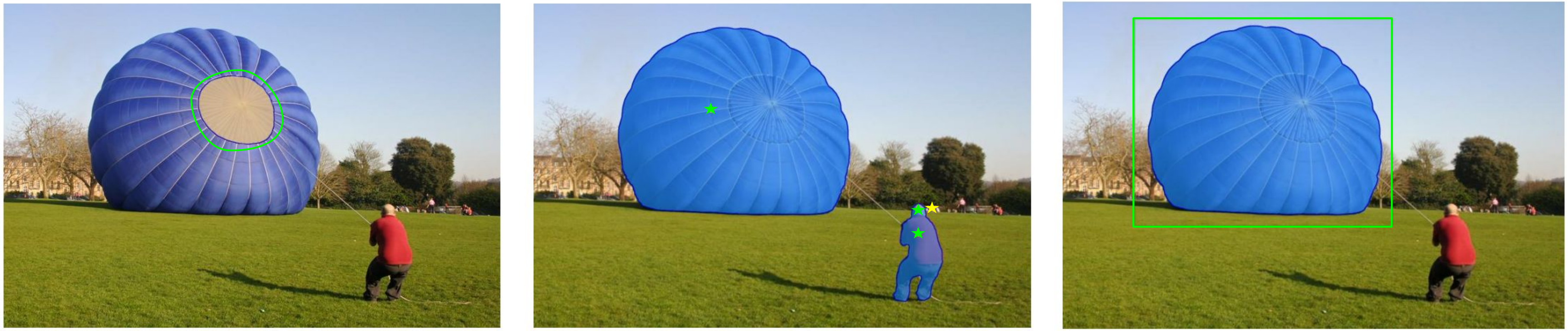


Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

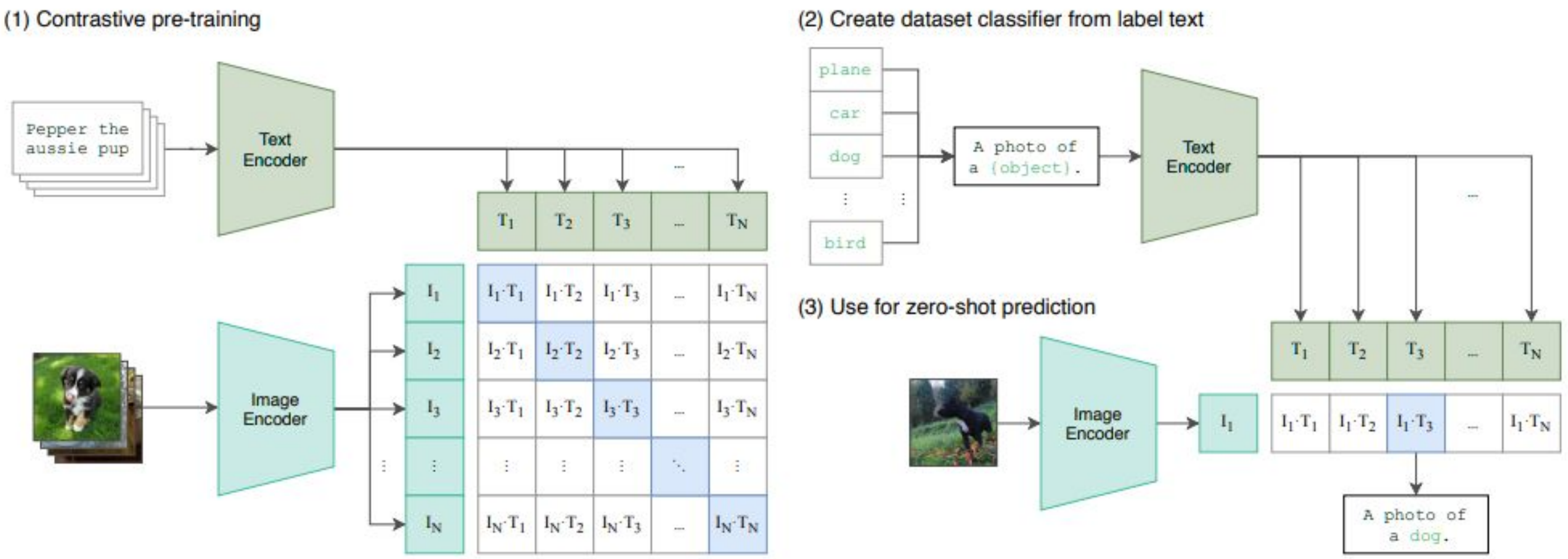
The prompt encoder



Positional encoder

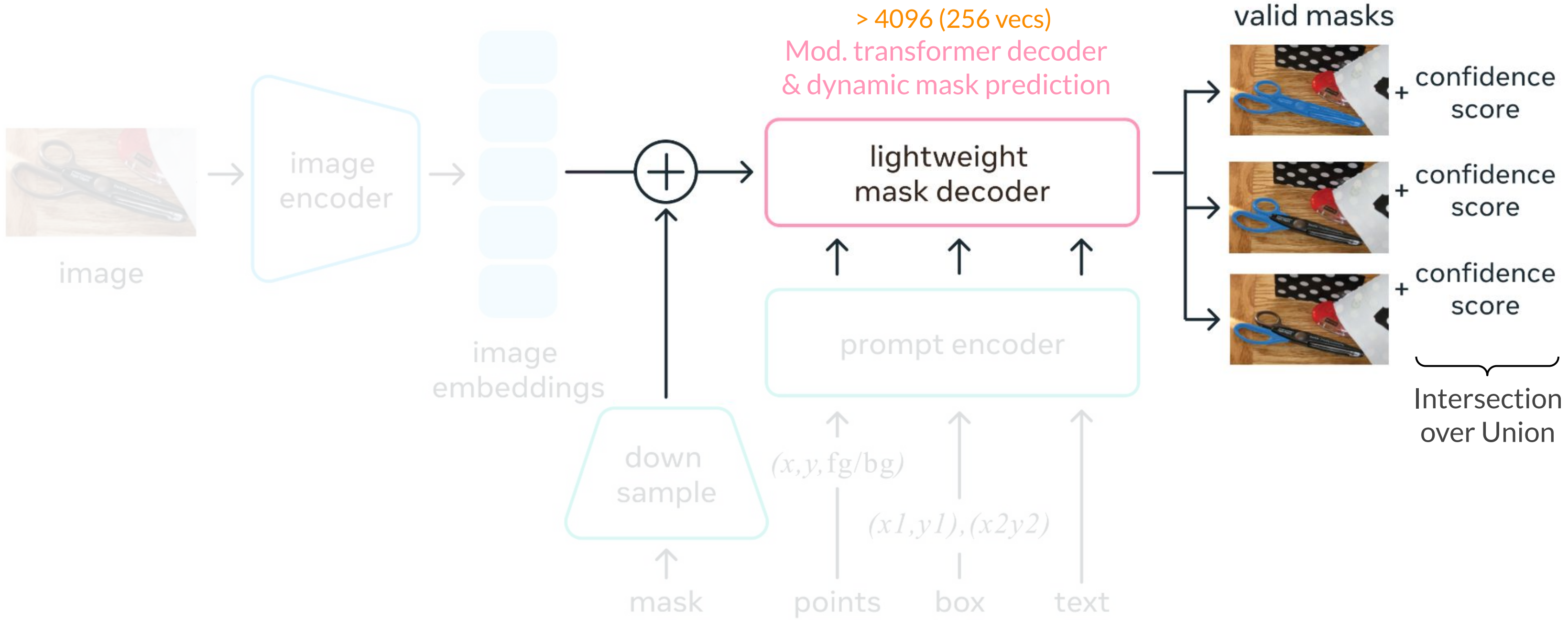


CLIP



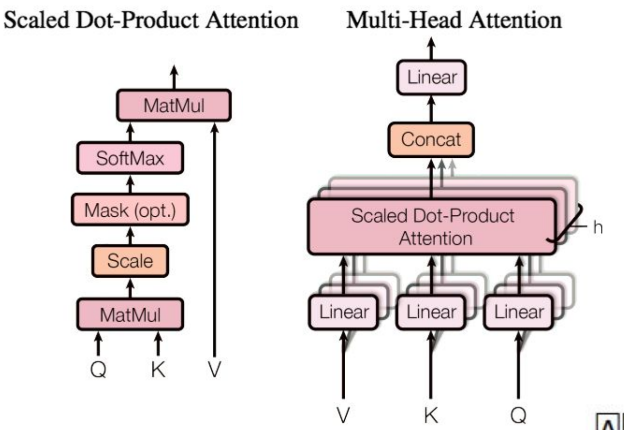
([Radford, 2021](#))

The lightweight mask decoder

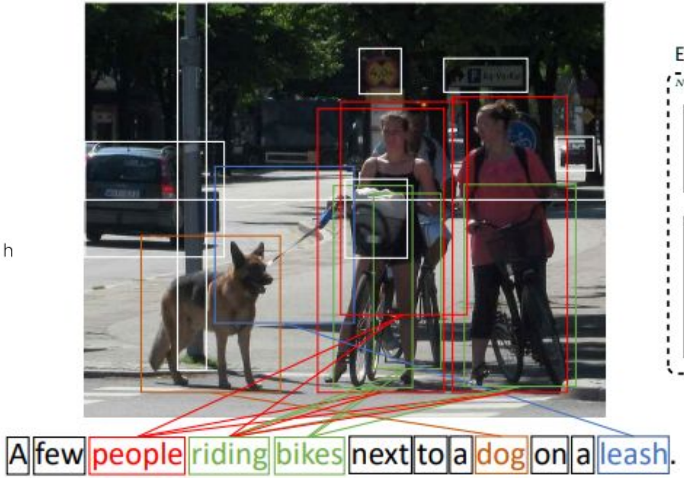


Attention is all you need!

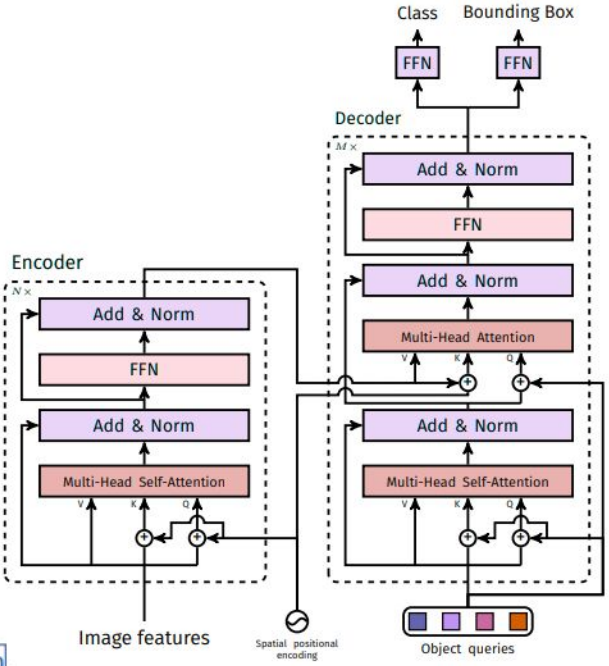
$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



(Vaswani, 2017)

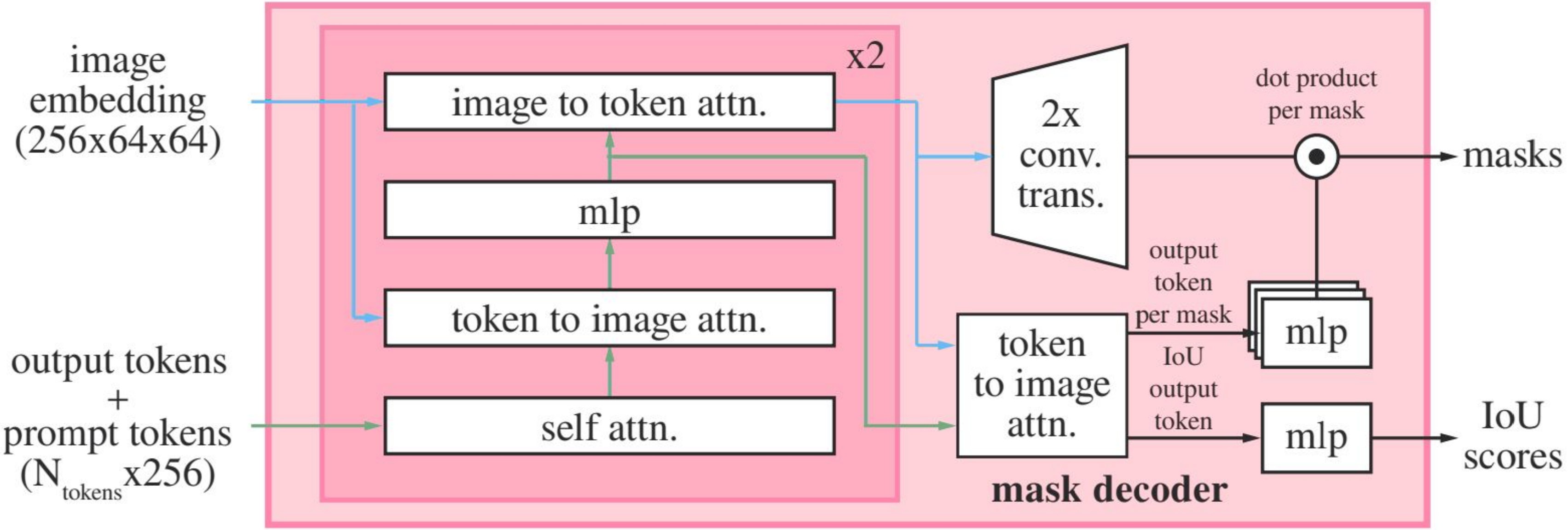


(Lee, 2018)



(Carion-Massa, 2020)

Details of this decoder

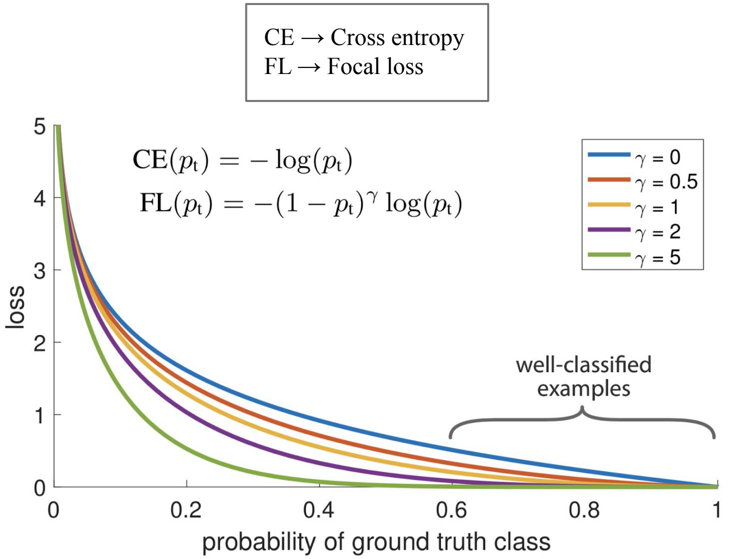


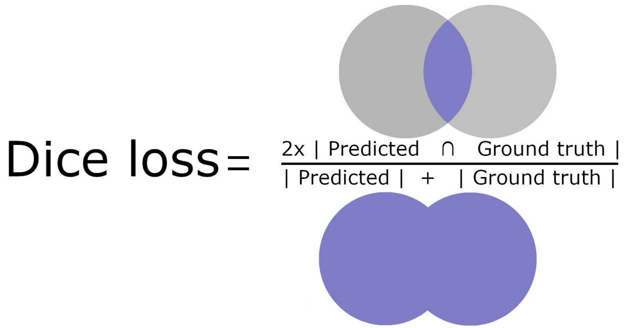
The pre-training algorithm

- *This simulates a sequence of prompts (e.g., points, boxes, masks) for each training sample and compare model's mask predictions against the ground truth.*
- *This is modified from interactive segmentation, the goal is to always predict a valid mask for any prompt, even when the prompt is ambiguous.*
- *Due to ambiguity, during training, only the minimal losses in the masks are back-propagated. To classify the masks, the model predicts a confidence score (i.e., estimated IoU) for each mask.*

Confidence scores

Use a linear combination of losses



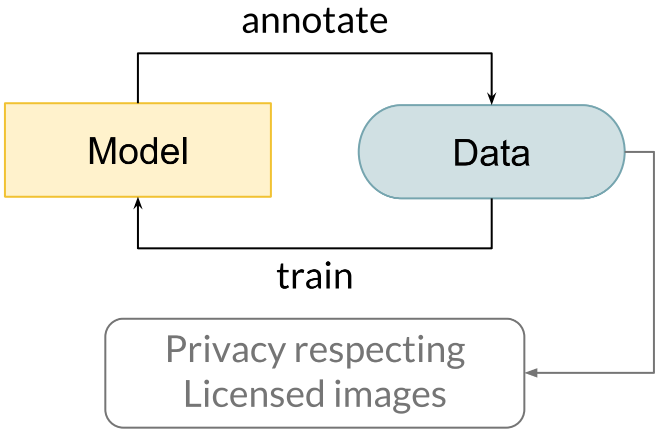


20:1

How to go beyond existing datasets?

we co-develop our model and dataset annotation in a loop with three stages:

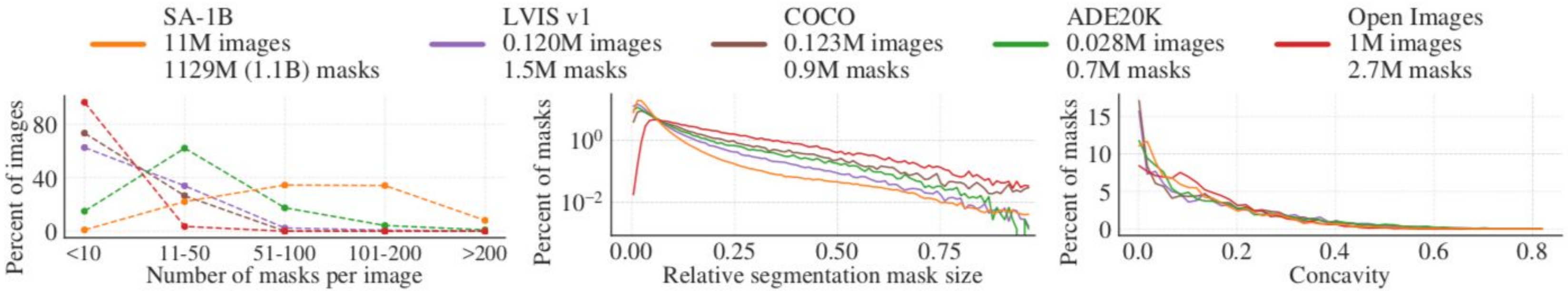
- ➔ **Assisted-manual**
  - ◆ **4.3M masks, 120k images**
- ➔ **Semi-automatic**
  - ◆ **10.2M masks, 180k images**
- ➔ **Fully automatic**
  - ◆ **1B masks, 11M images**

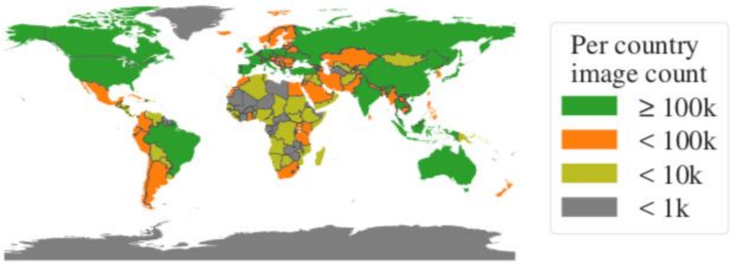


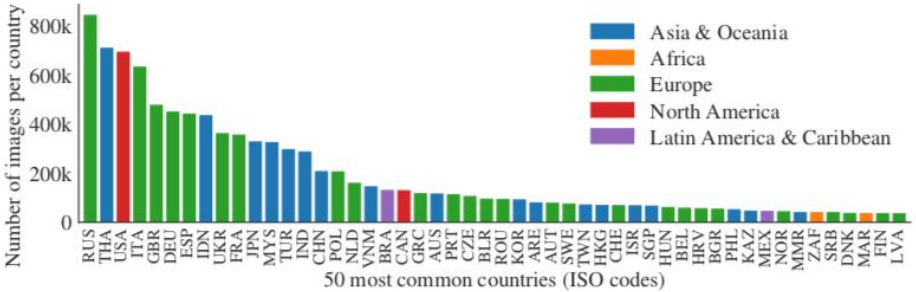
Computational effort

- *SAM is initialized with pre-trained ViT-H (both, ViT-L and ViT-B, can be used too)*
- *The training required approx. 100K iterations using the AdamW optimizer, a linear learning rate warm-up, and a step-wise learning rate decay schedule*
- *Batch size is 256 images, distributed across 256 GPUs, limited to 64 masks/GPU*
- *Points are sampled uniformly from the ground truth mask. Boxes are taken as the ground truth mask's bounding box, with random noise added in each coordinate*
- *After making a prediction from this first prompt, subsequent points are selected uniformly from the error region between the previous mask prediction*
- *Text-to-mask using CLIP, data augmentation and batch size of 128 images*
- *SAM was trained on 256 A100 for 68 hours (energy cons. is appr. 450 MW)*

Resulting dataset metrics



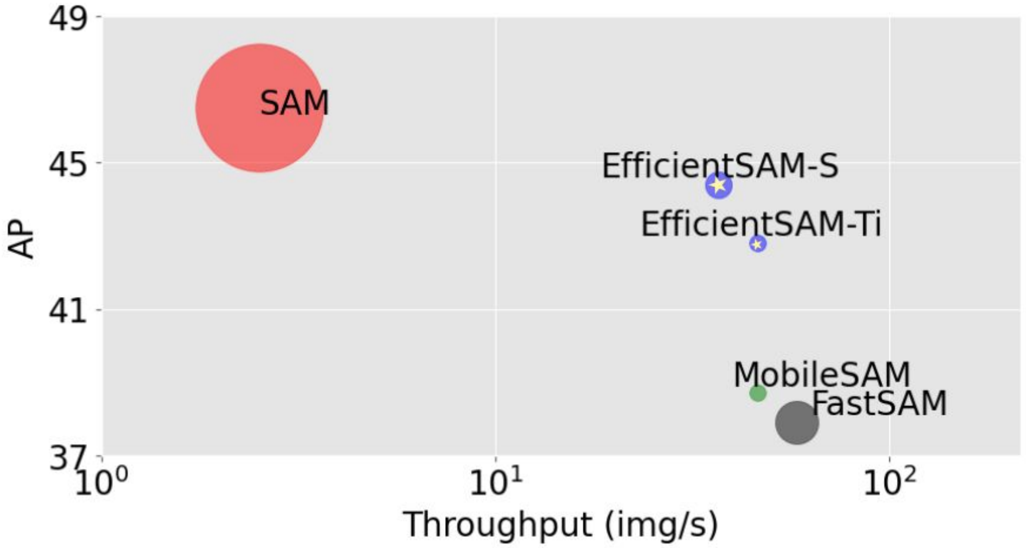




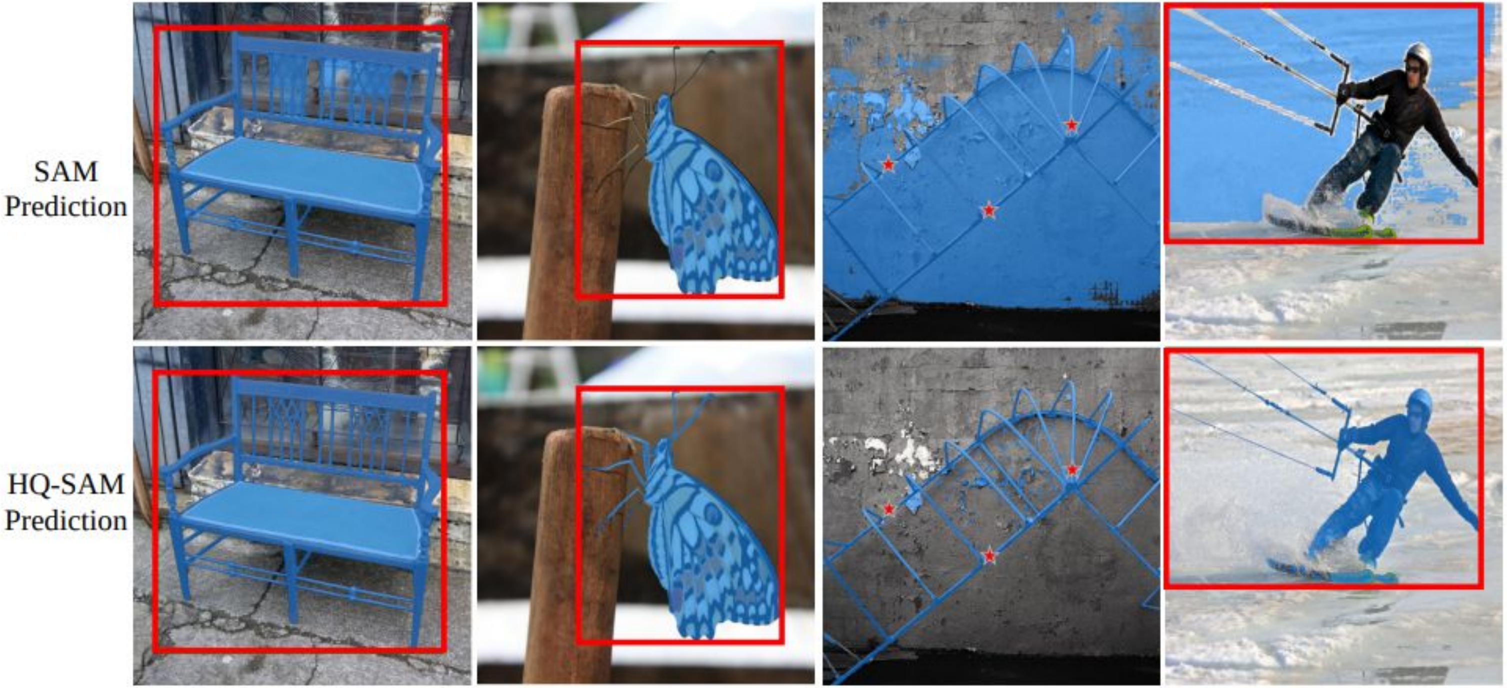
What came next?

- *Speedup (FastSAM, MobileSAM, EfficientSAM, EdgeSAM)*
- *High quality masks (HQ-SAM, Stable-SAM)*
- *Tracking (TAM, SAM-Track, SAM-PT, HQTrack, FAn, DEVA)*
- *Annotation (Region captioning, Grounded SAM)*
- *Geospatial (SAM-DA, Geo SAM, samgeo, SAMRS, SAM-CD)*
- *SAM 3D (RGB-D, Volumetric medical images, LiDAR to object selection)*

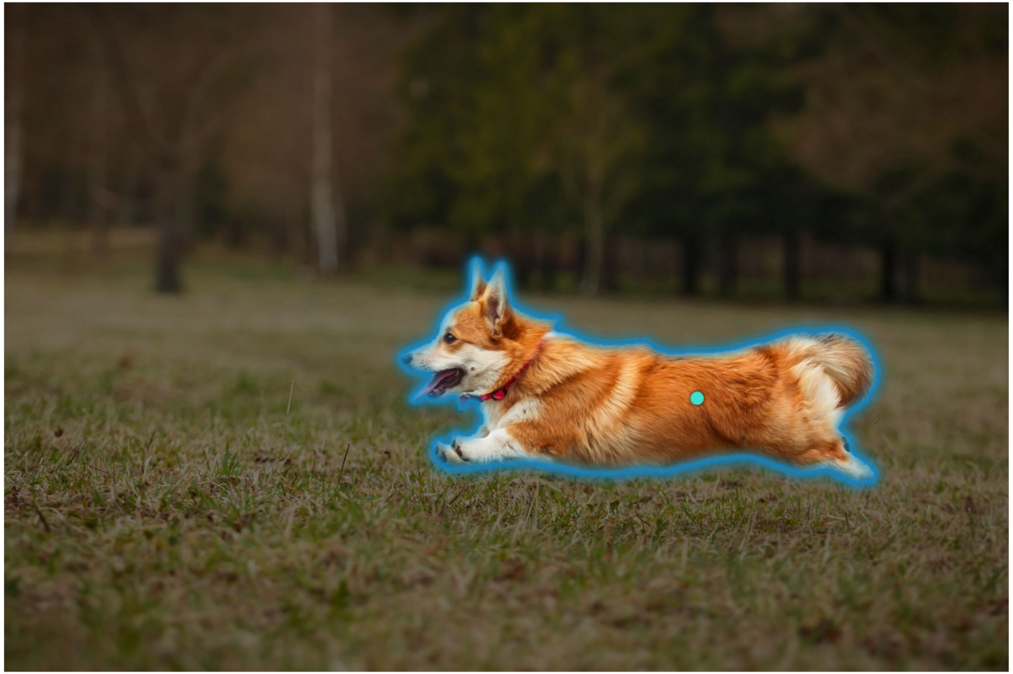
Speedup


49
45
41
37
AP
SAM
EfficientSAM-S
EfficientSAM-Ti
MobileSAM
FastSAM
10^0
10^1
10^2
Throughput (img/s)

High quality



Go to 3D (Anything)



Considerations about work

- *We note that a foundation model for image segmentation is an inherently limited scope, since it represents an important, yet fractional, subset of computer vision;*
- *A central objective is to simplify the interface for composition with other components, enabling new applications;*
- *The model's performance will be good in general, but less than models specializing in their own domains.*

The community itself responded very well.

This work received almost 1.5k citations in ten months.


Thanks for your attention!