Segment Anything

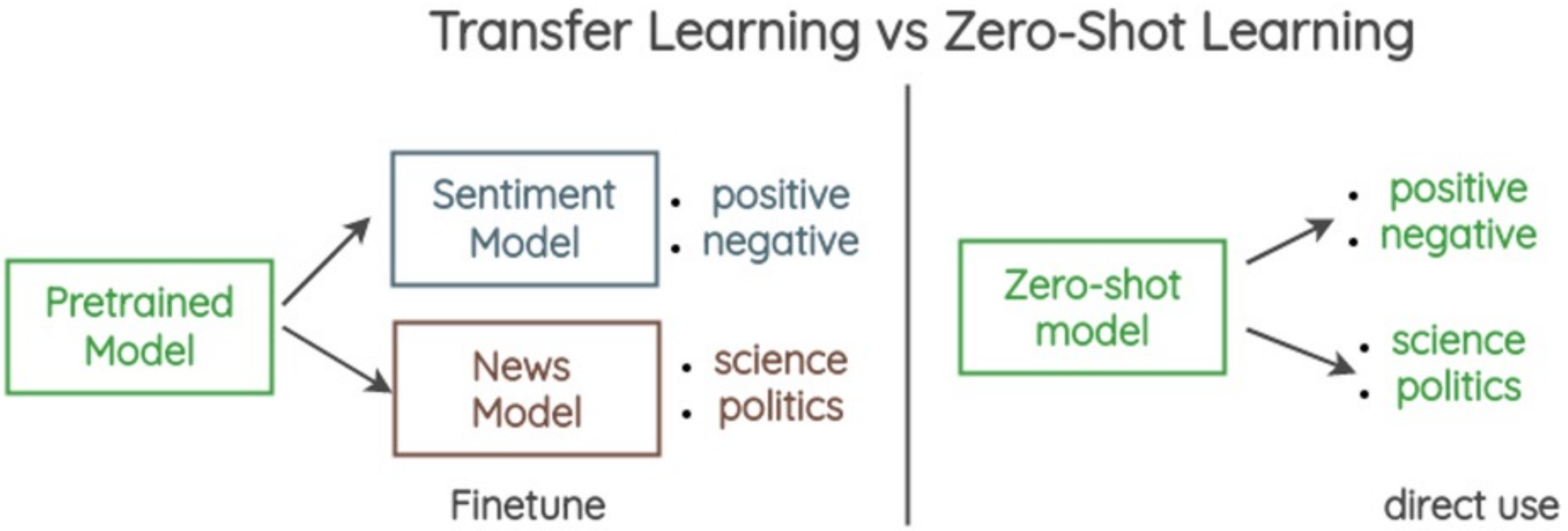
Anton Gorokhov

Table of contents

1. Intro
2. Training & Data
3. Model architecture
4. Experiments
5. Results

1. Intro – NLP inspiration

The goal is to build *a foundation model for image segmentation*



1. Intro

The model is designed and trained to be promptable, so it can transfer zero-shot to new image distributions and tasks

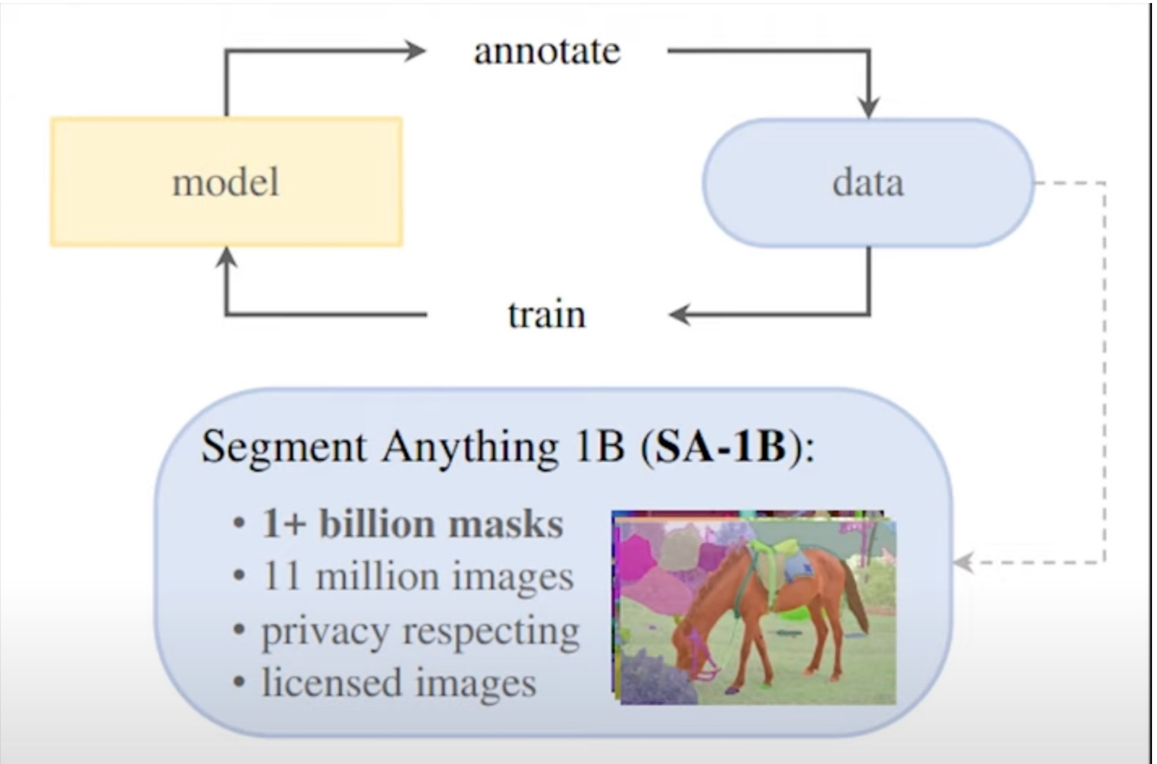
Promt can be **anything** – point, area, bounding box or text, …

2. Data collection problem

Not much high-quality segmentation data in web

2. Data collection solution

build a "data engine"
***i.e.*, we co-develop our model with**
model-in-the-loop dataset annotation



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

2. Data engine 3 stages

(1) assisted-manual

Model helps the assesor with data annotation

https://habr.com/ru/companies/sberdevices/articles/739352/

Speed: 14 seconds per image

collected 4.3M masks from 120k images

(2) semi-automatic

Subset of objects + possible location

SAM

Segmentation masks with high diversity

Speed: 35 seconds per image

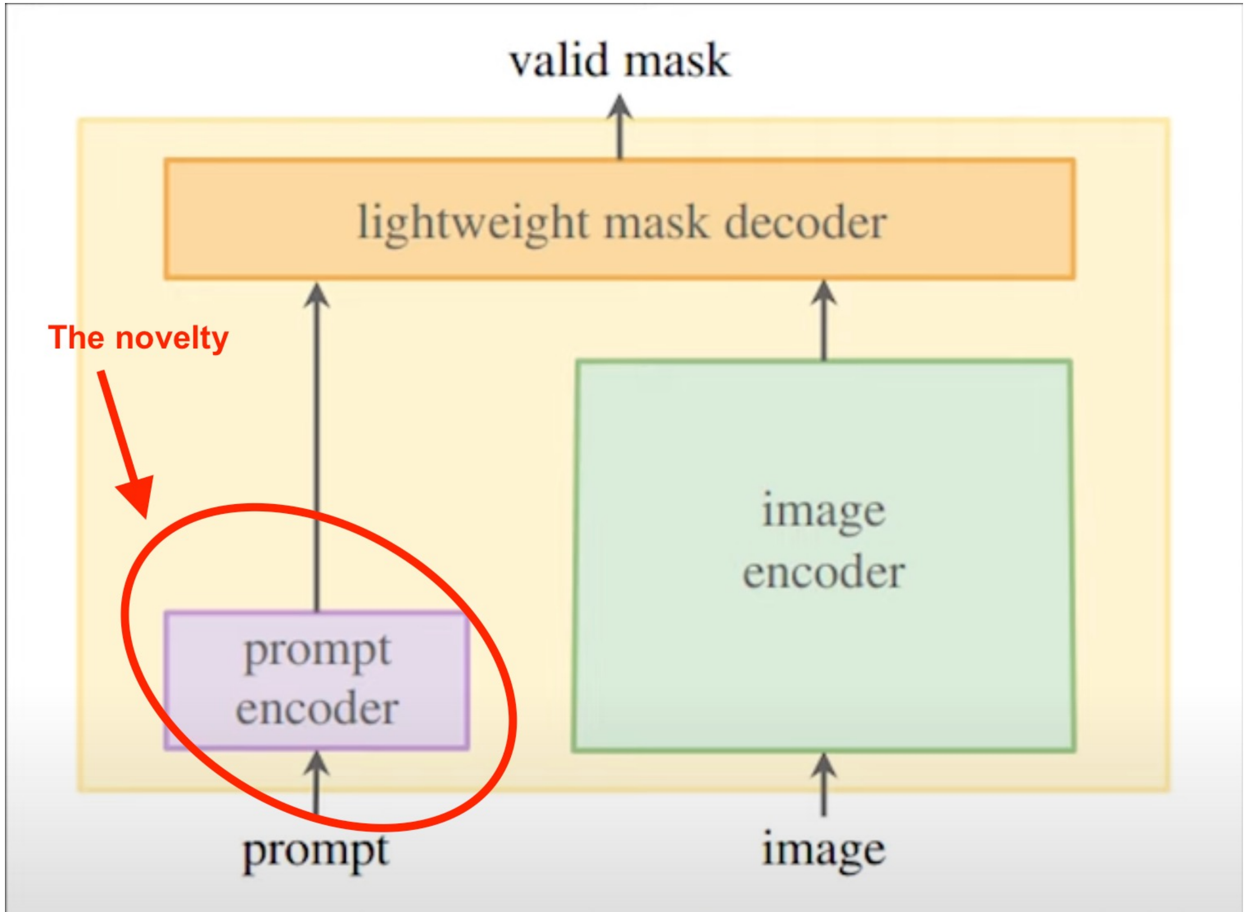
collected an additional 5.9M masks in 180k images

(3) fully automatic

Foreground points

SAM

3. Model



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

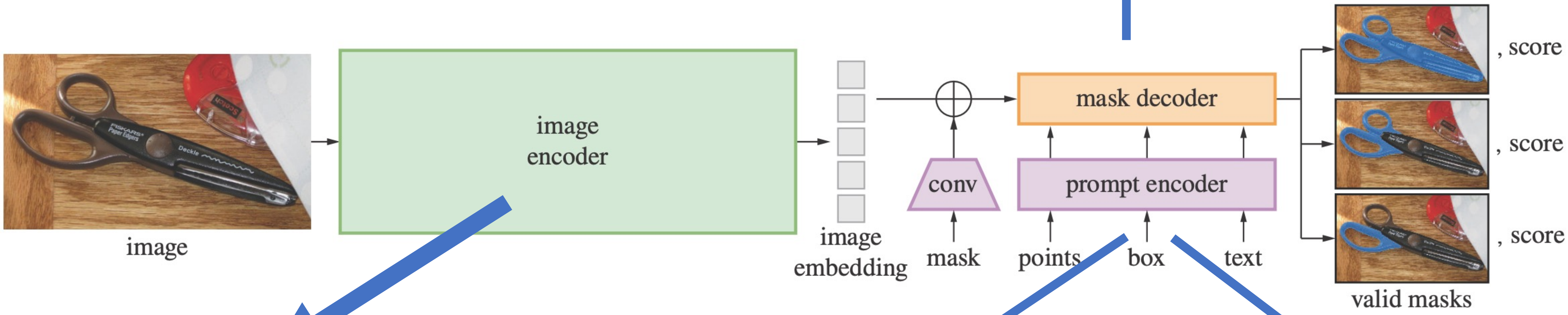
SAM uses **focal loss (1) and dice loss (2)** for training the model.

The **focal loss** is simply a variation of the cross-entropy loss function

On the other hand, the **dice loss** aims to increase the overlap (i.e., the intersection over the union area, to be more precise) between the predicted and ground truth mask.

3. Model structure

modification of a Transformer decoder block (prompt self-attention and cross-attention in two directions)



Authors use a **pre-trained ViT** minimally adapted to process high resolution inputs.

sparse (points, boxes, text)

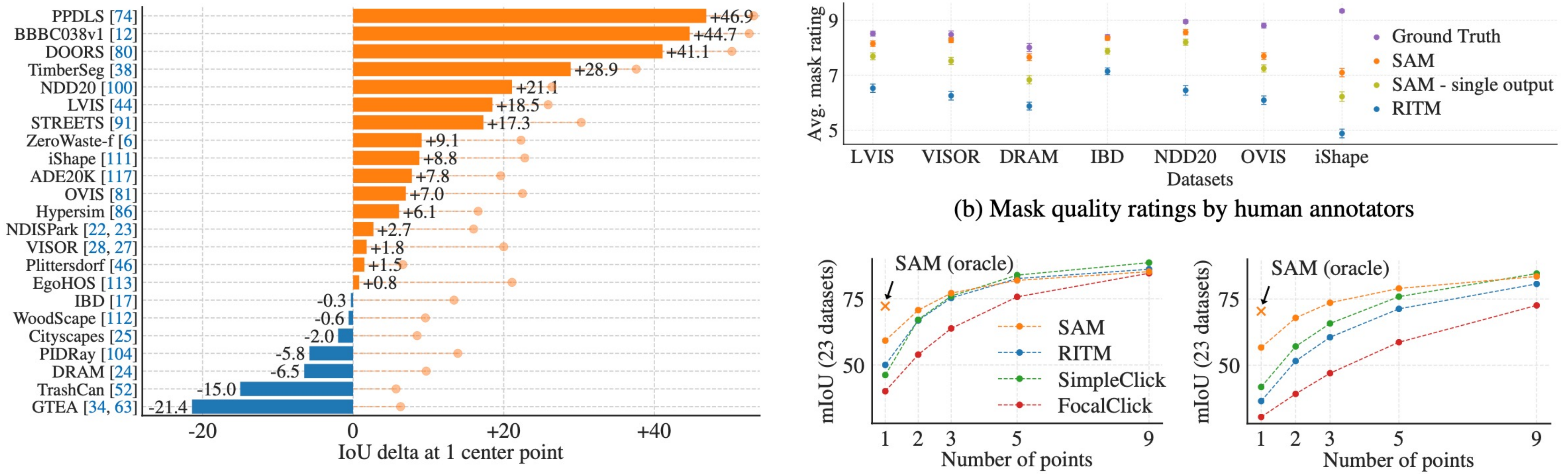
positional encodings + learned embeddings for each prompt

CLIP

dense (masks)

convolutions + image embeddings

4. Experiments - 1 point promt



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

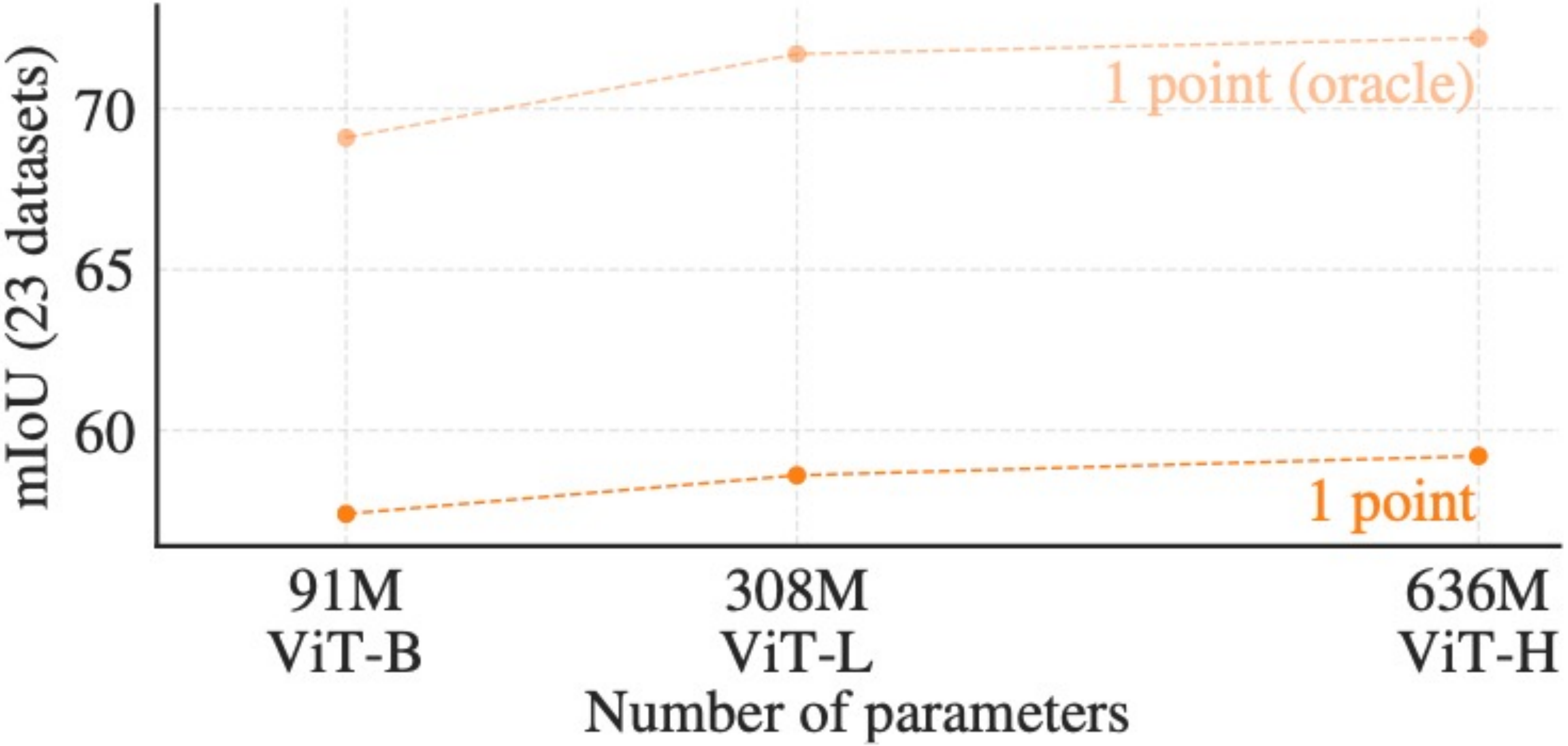
(b) Mask quality ratings by human annotators

(c) Center points (default)

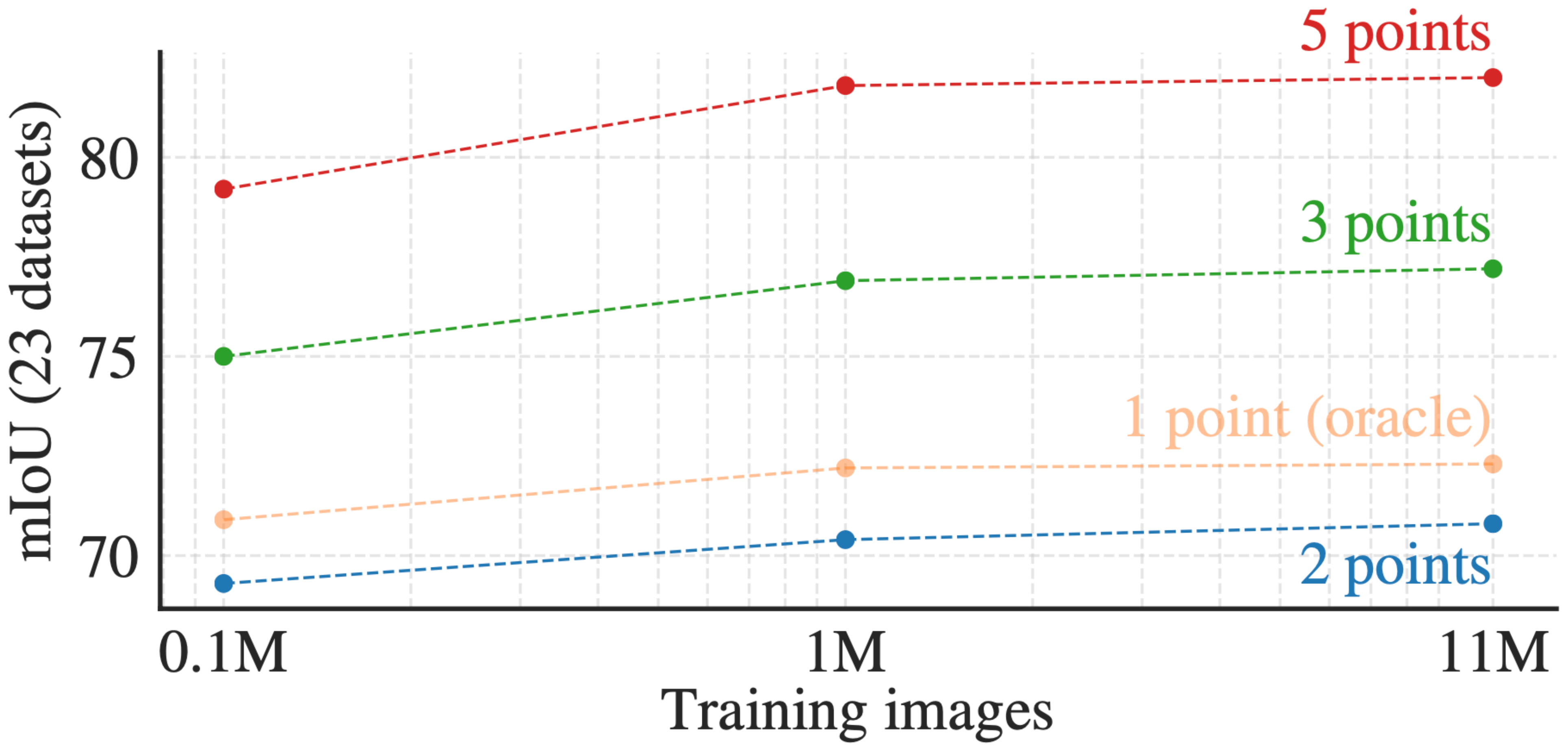
(d) Random points

Figure 9: Point to mask evaluation on 23 datasets. (a) Mean IoU of SAM and the strongest single point segmenter, RITM [92]. Due to ambiguity, a single mask may not match ground truth; circles show "oracle" results of the most relevant of SAM's 3 predictions. (b) Per-dataset comparison of mask quality ratings by annotators from 1 (worst) to 10 (best). All methods use the ground truth mask center as the prompt. (c, d) mIoU with varying number of points. SAM significantly outperforms prior interactive segmenters with 1 point and is on par with more points. Low absolute mIoU at 1 point is the result of ambiguity.

4. Experiments – VIT size



4. Experiments – SAM trains fast



5. Dataset

SA-1B Has much more mask diversity per image with more precision than other segmentation datasets

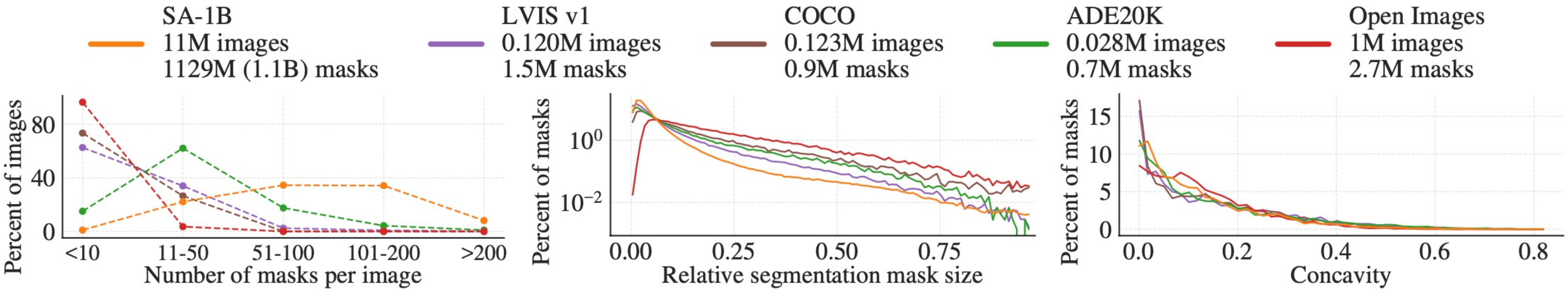


Figure 6: Dataset mask properties. The legend references the number of images and masks in each dataset. Note, that SA-1B has 11× more images and 400× more masks than the largest existing segmentation dataset Open Images [60].

5. Edge detection

5. Geogpaphical diversity of SA-1B

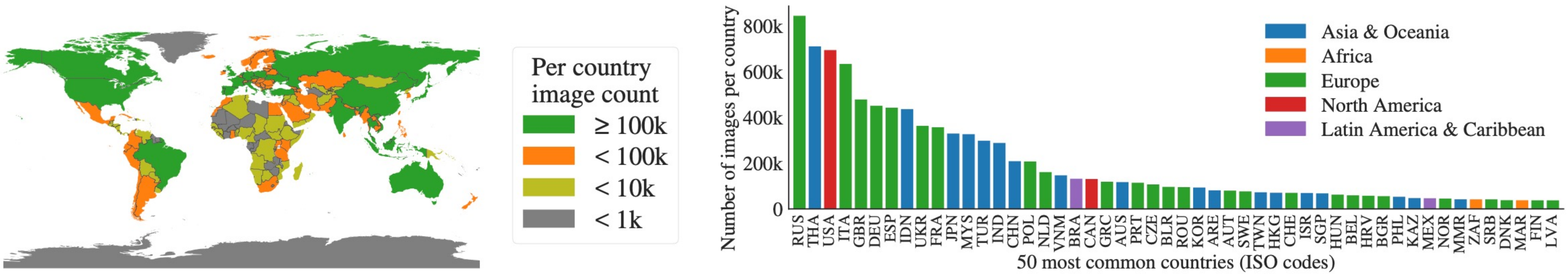


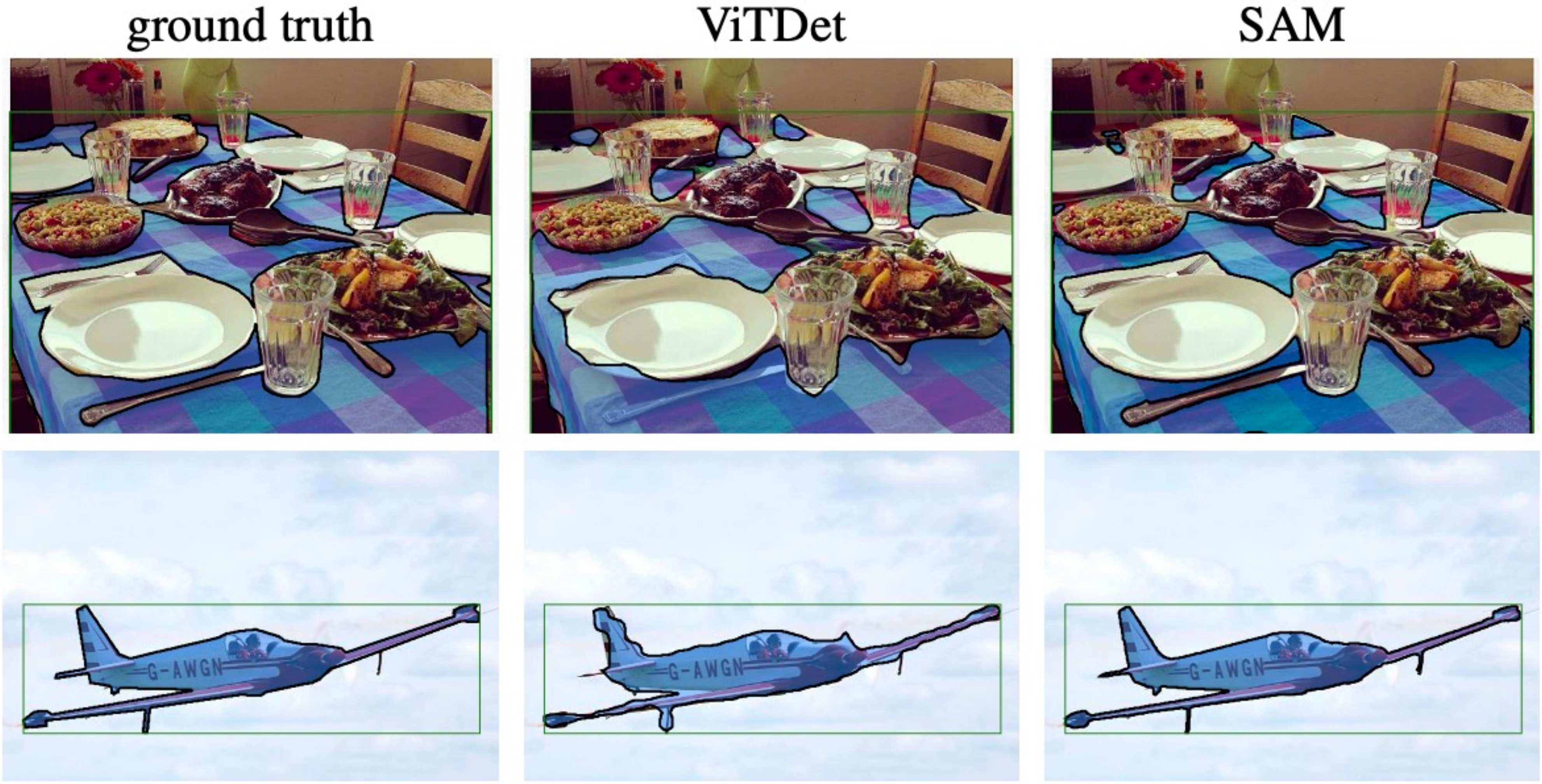
Figure 7: Estimated geographic distribution of SA-1B images. Most of the world's countries have more than 1000 images in SA-1B, and the three countries with the most images are from different parts of the world.

5. People detection

	mIoU at	
	1 point	3 points
perceived gender presentation		
feminine	54.4 ±1.7	90.4 ±0.6
masculine	55.7 ±1.7	90.1 ±0.6
perceived age group		
older	62.9 ±6.7	92.6 ±1.3
middle	54.5 ±1.3	90.2 ±0.5
young	54.2 ±2.2	91.2 ±0.7

	mIoU at	
	1 point	3 points
perceived skin tone		
1	52.9 ±2.2	91.0 ±0.9
2	51.5 ±1.4	91.1 ±0.5
3	52.2 ±1.9	91.4 ±0.7
4	51.5 ±2.7	91.7 ±1.0
5	52.4 ±4.2	92.5 ±1.4
6	56.7 ±6.3	91.2 ±2.4

5. SAM vs. ViTDet



5. SAM vs. ViTDet