SEGMENTATION Is All You Need

1. SegGPT: Segmenting Everything In Context
2. Segment Anything

Presenter: Jiwon Jeong (Data Science 21)

jjwon4086@korea.ac.kr

Agenda

- SegGPT: Segmenting Everything In Context
  - Generalist Model
  - Paper Link
- Segment Anything
  - Meta AI Research, FAIR
  - Foundation Model
  - Paper Link

SegGPT

SegGPT: Segmenting Everything In Context

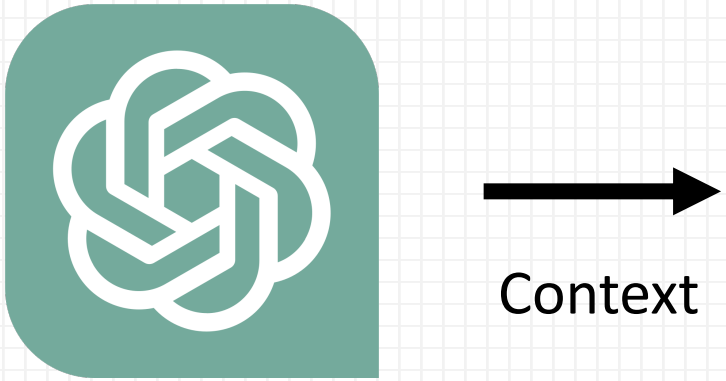
Xinlong Wang[1*] Xiaosong Zhang[1*] Yue Cao[1*] Wen Wang[2] Chunhua Shen[2] Tiejun Huang[1,3]

[1] Beijing Academy of Artificial Intelligence [2] Zhejiang University [3] Peking University

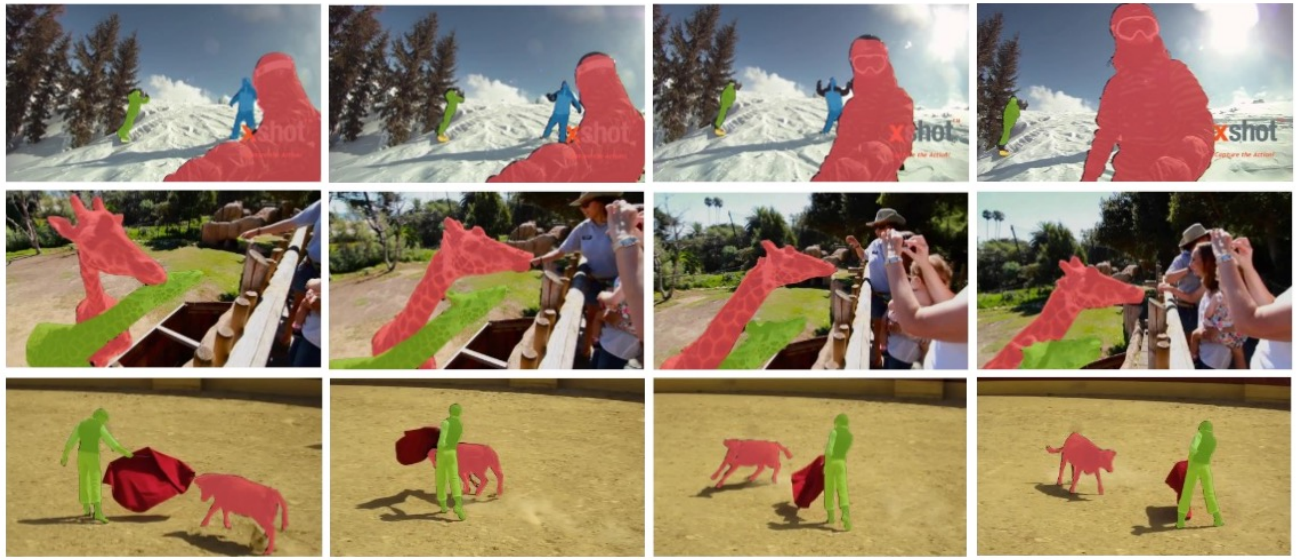
Code & Demo: https://github.com/baaivision/Painter

GPT???

- Seg**GPT**: Segmenting Everything in **Context**



Context



GPT???

- Seg**GPT**: Segmenting Everything in **Context**

GPT???

- **Seg**ment everything with a **G**eneralist **P**ain**T**er
  -> **SegGPT**
  - Not Generative Pre-trained Transformer
- DRAGON…?

representations of text and KG. Here we propose **DRAGON** (**D**eep Bidi**r**ectional **L**a**ng**uage-Kn**o**wledge Graph Pretrai**n**ing), a self-supervised method to pretrain

- Then, What is Generalist Painter?

Generalist Painter

- SegGPT is a modification of Generalist Painter

Images Speak in Images: A Generalist Painter for In-Context Visual Learning

Xinlong Wang[1*] Wen Wang[2*] Yue Cao[1*] Chunhua Shen[2] Tiejun Huang[1,3]

[1] Beijing Academy of Artificial Intelligence [2] Zhejiang University [3] Peking University

https://github.com/baaivision/Painter

Introduction

- Previous specialist segmentation models are limited to specific tasks, classes, granularities, data types, etc.

- Aim to train single model that is capable of solving diverse and unlimited segmentation tasks
  -> Generalist model !

Introduction

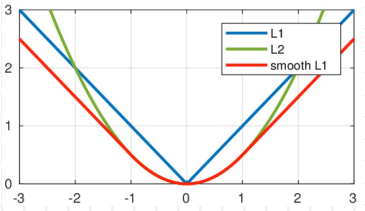
- Main Challenge
  - Incorporate very different data types in training
    - part, semantic, instance, panoptic, person, medical image, aerial image, etc.
  - Design a generalizable training scheme
    - differs from conventional multi-task learning
    - flexible on task definition and capable of handling out-of-domain tasks
- Propose SegGPT,
  a generalist model for segmenting everything in context

Introduction

- SegGPT
  - Generalist Painter + “Random Coloring Scheme”
    -> more flexible and generalizable
  - Same architecture with generalist painter
    - Vanilla ViT and a simple smooth-l_1 loss

Faster R-CNN: smooth-l_1 loss

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases}$$



Introduction

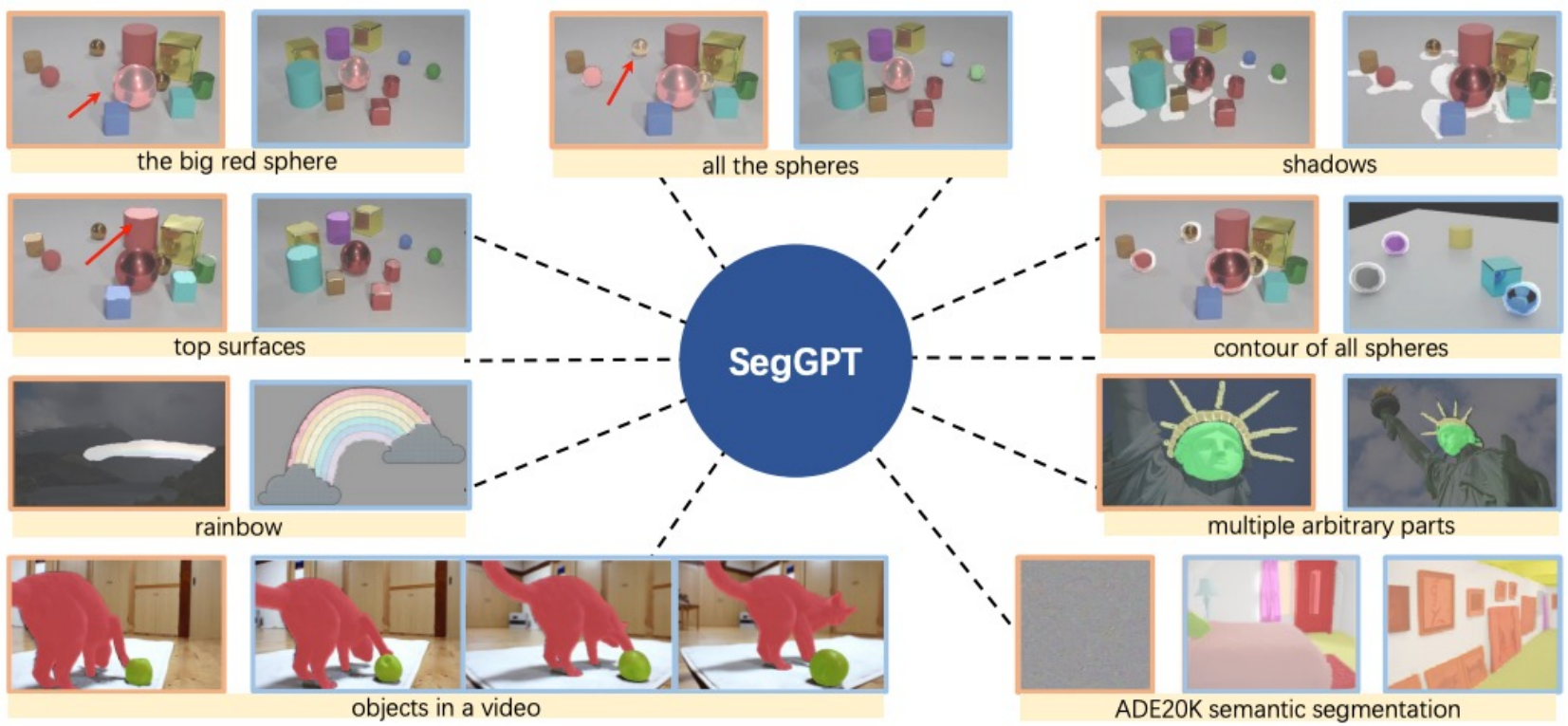
- Main Contributions
  - For the first time, we demonstrate a single generalist model capable of performing a diverse set of segmentation tasks automatically
  - We evaluate the pre-trained SegGPT on a broad range of tasks directly without fine-tuning
  - Our results show strong capabilities in segmenting in-domain and out-domain targets

Introduction

- However…
  - No new state-of-the-art results or outperform existing specialist methods
  - Believe that this may not be the responsibility of a general-purpose model

Introduction

- SegGPT is capable of segmenting everything in context with only **one single** model



Related Work

- Vision Generalist
  - DETR, Pix2Seq, Unified-IO, OFA, UViM, …
  - They actually accomplish each task through some form of hard indicators, such as a special token
    -> difficult to generalize to new tasks
  - This work uses an in-context framework that maintains flexibility on task definition

Related Work

Visual Prompting via Image Inpainting (arxiv)

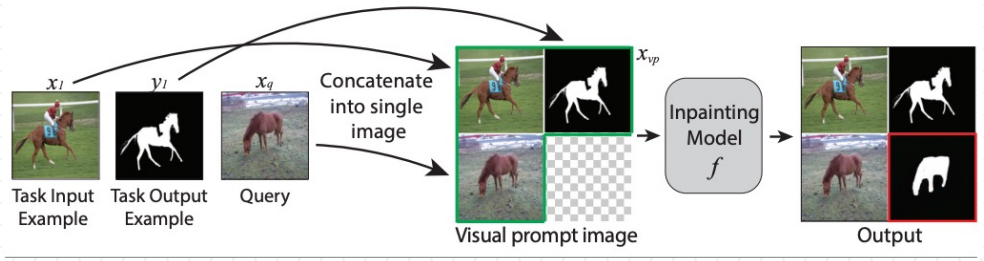
- In-context Visual Learning
  - GPT
    - text completion problems given prompts and examples

```
Je suis désolé                  I'm sorry
J'adore la glace
```

will prompt the model [5] to perform the task of French-to-English translation, returning:

```
I love ice cream
```

  - In computer vision,
    - Inpainting with discrete tokens on figures and infographics from vision articles (like MAE)



Related Work

- Generalist Painter

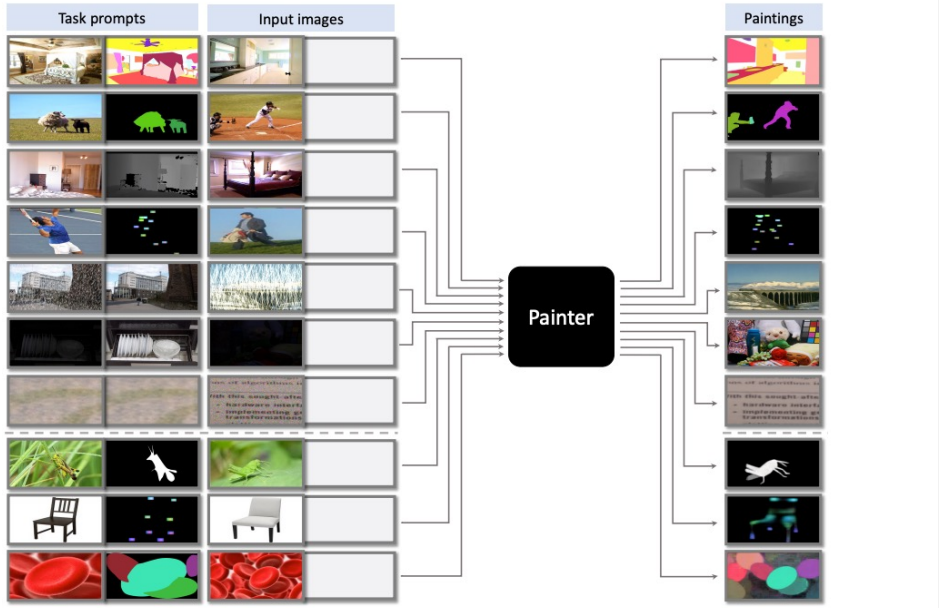


Figure 1. An illustration of the in-context inference of Painter. Painter is a generalist vision model, which can automatically perform vision tasks according to the input task prompts without the task specific heads. Painter can not only perform in-domain tasks with highly competitive performance, such as semantic segmentation (Row 1), instance segmentation (Row 2), depth estimation (Row 3), keypoint detection (Row 4), denoising (Row 5), deraining (Row 6), and image enhancement (Row7), but also be able to rapidly adapt to various out-of-domain vision tasks using simple prompts, such as open-category object segmentation, keypoint detection, and instance segmentation (Row 8-10).

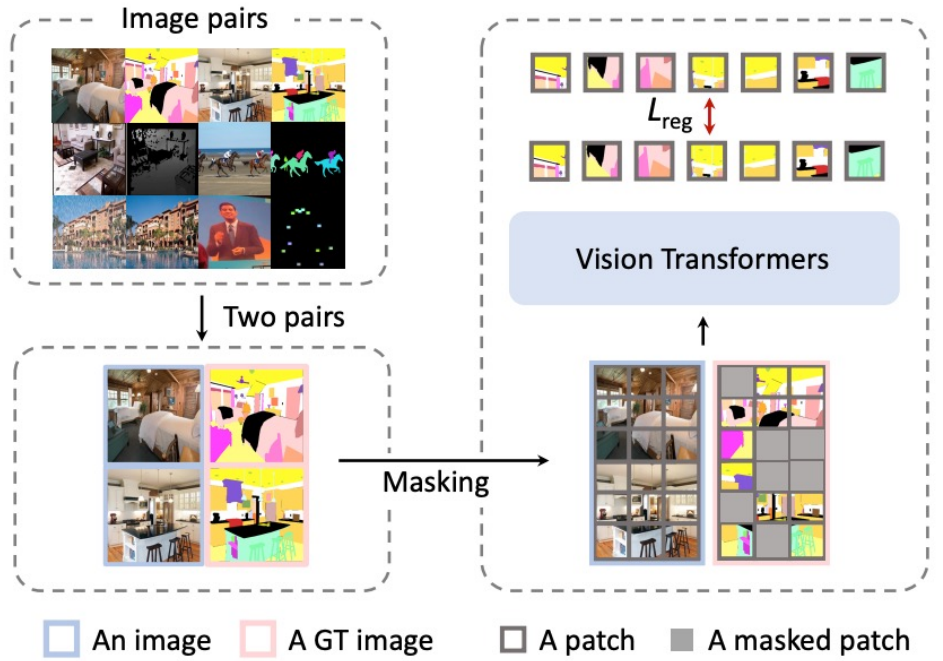
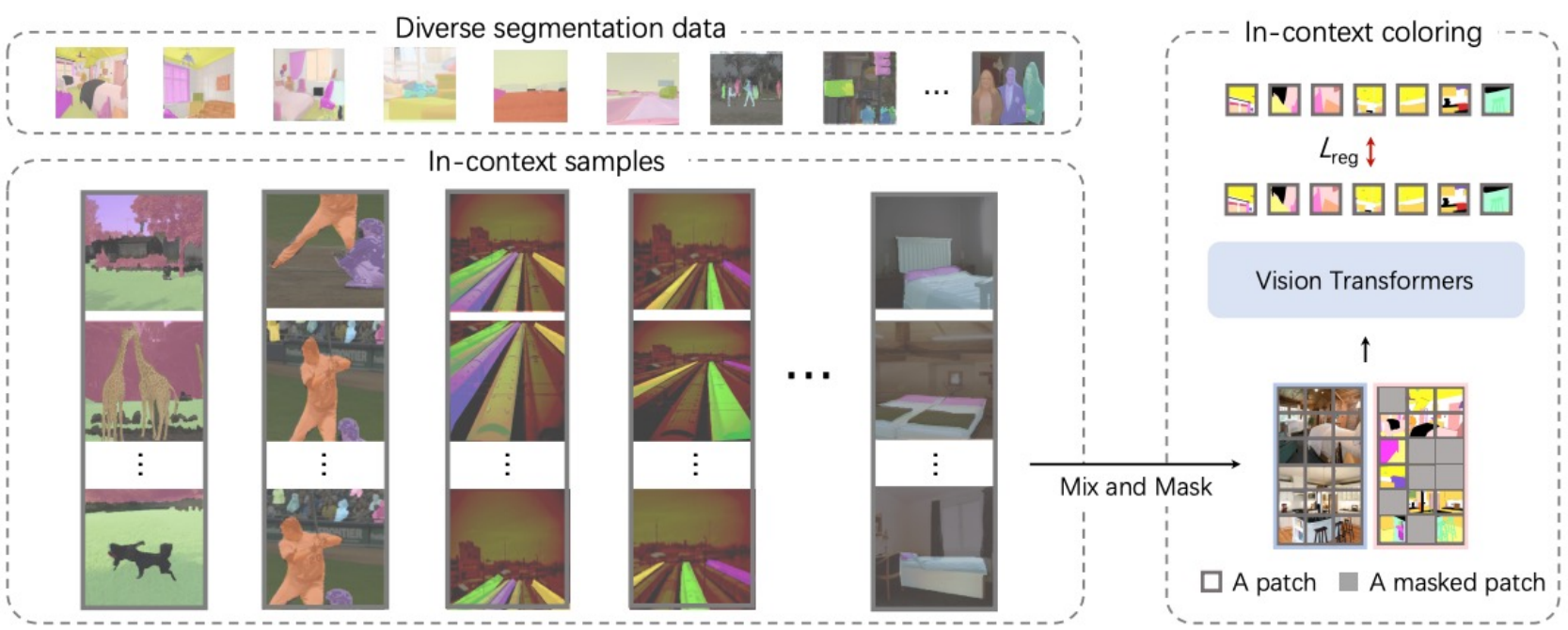


Figure 2. The training pipeline of the masked image modeling (MIM) framework.

Approach

- Incorporate diverse segmentation data and transform them into the same format of images
- Adopt a general Painter framework with in-context coloring as the training objective

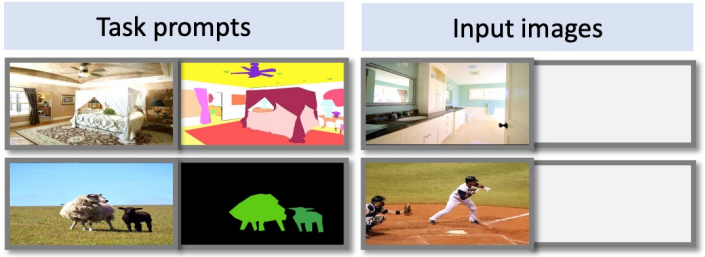


Approach

- Training Framework
  - Redefines the output space of vision tasks as “images”
    - NLP: sequences of discrete language tokens (same as input)
    - DETR, Pix2Seq (OD): discrete space (points of bounding box)
  - Unifies different tasks into the same image inpainting problem
    - randomly mask the task output images and reconstruct the missing pixels
  - No modifications to the architecture and loss function
    - Vanilla ViT and a simple smooth-l_1 loss (same as Painter)

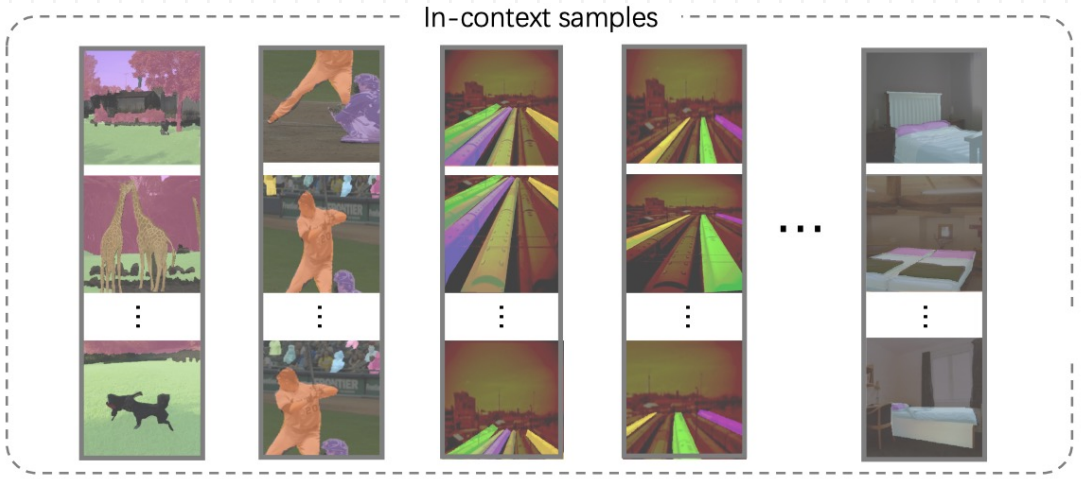
Random Coloring Scheme

- Limitations of Painter
  - Color space for each task is pre-defined
  - model rely on the color itself to determine the task rather than using the relationships between segments
- Random Coloring Scheme
  - In-context coloring
  - Mix-context training method



Random Coloring Scheme

- In-context Coloring
  1. Randomly sampling another image that shares a similar context
  2. Randomly sample a set of colors from the target image and map each color to a random one
  3. Two pairs image -> in-context pair

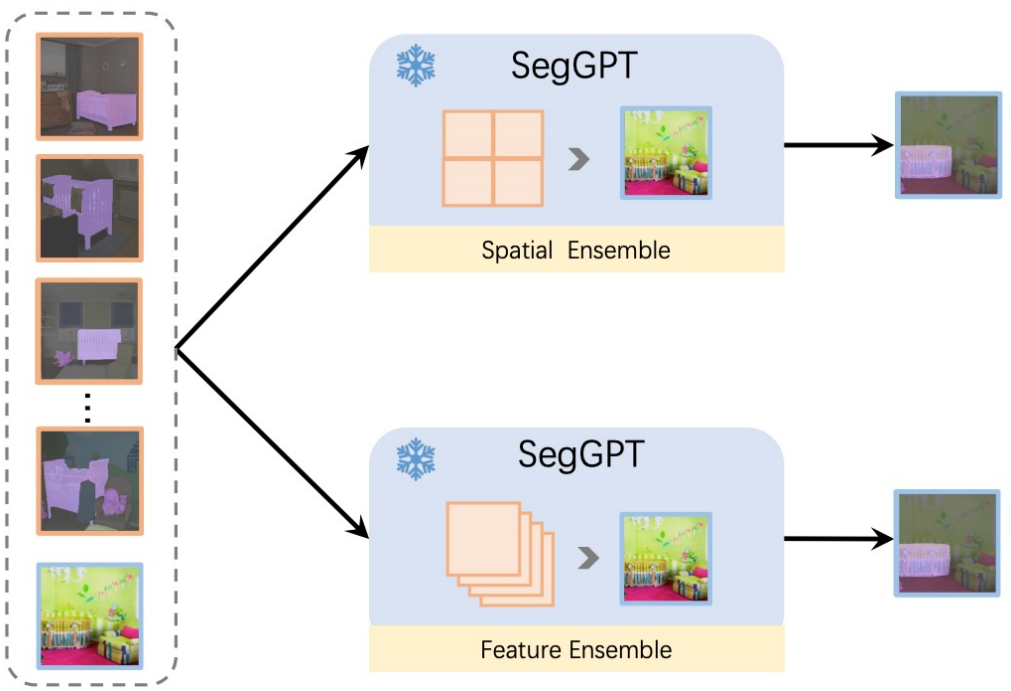


Random Coloring Scheme

- Mix-context training method
  1. Stitching together multiple images with the same color mapping
  2. Randomly cropped and resized to form a mixed-context training sample

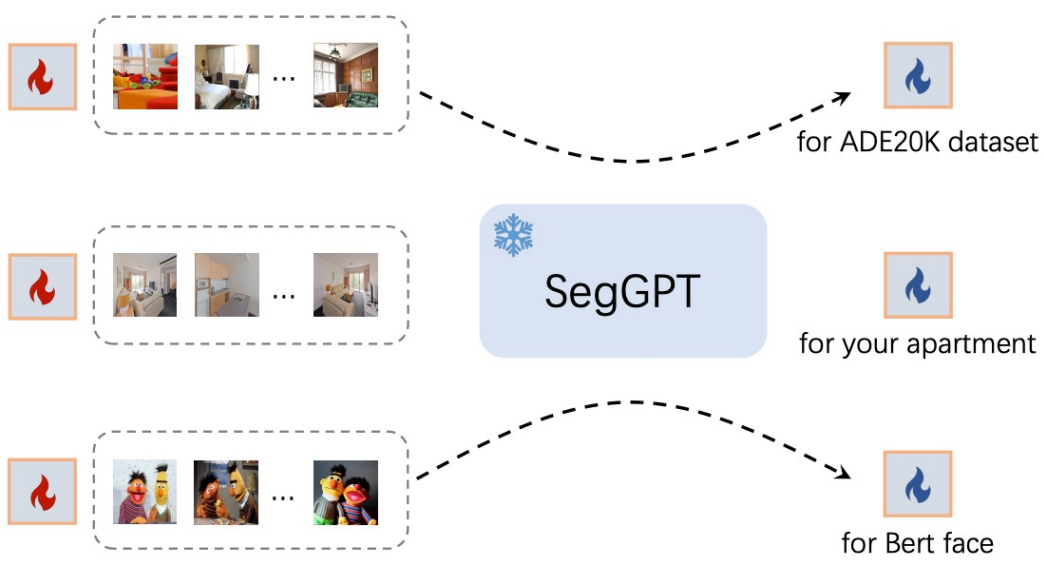
Context Ensemble

- To serve a more accurate and concrete context, multiple examples can be used
  - Spatial Ensemble
    - concatenated in $n \times n$ grid
    - subsampled to the same size as single
  - Feature Ensemble
    - combined in the batch dimension
    - averages features of the query image after each attention layer



In-Context Tuning

- Painter: Redefine the per pixel ground-truth for each task as a 3-channel tensor, similar to the RGB space
- During training, freeze the whole model and train learnable image tensor

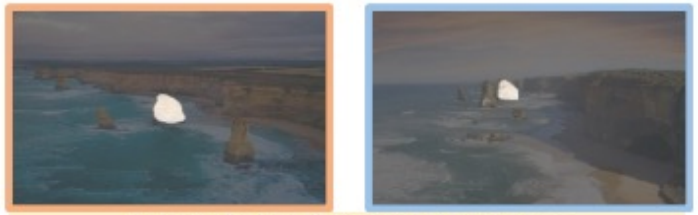


More Visualizations

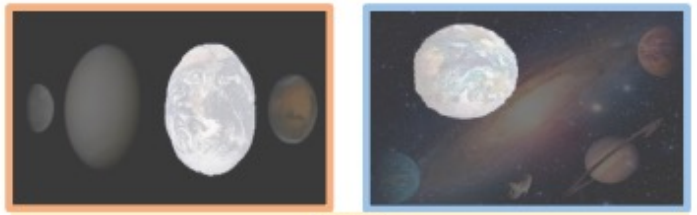
cubes

yellow cubes

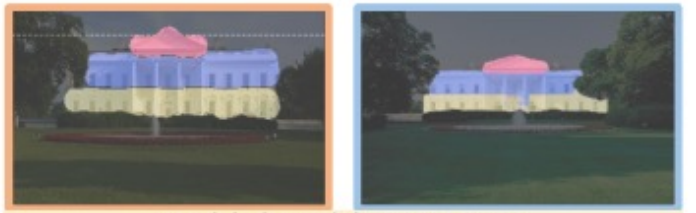
Ernie



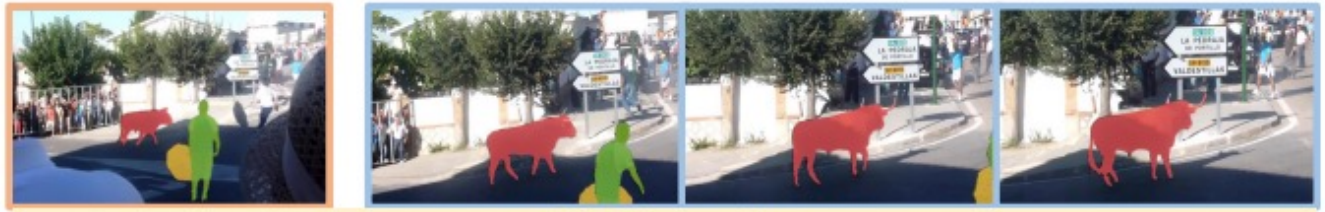
one of the Twelve Apostles



Earth



multiple arbitrary parts



objects in a video

COCO instance segmentation

Experiment

- Training Data
  - ADE20K, COCO
  - PASCAL VOC
  - Cityscapes (street view)
  - LIP (person)
  - PACO
  - CHASE_DB1, DRIVE, HRF, STARE (retinal vessel)
  - iSAID, loveDA (aerial images)

Experiment

- Qualitative Results on YouTube-VOS

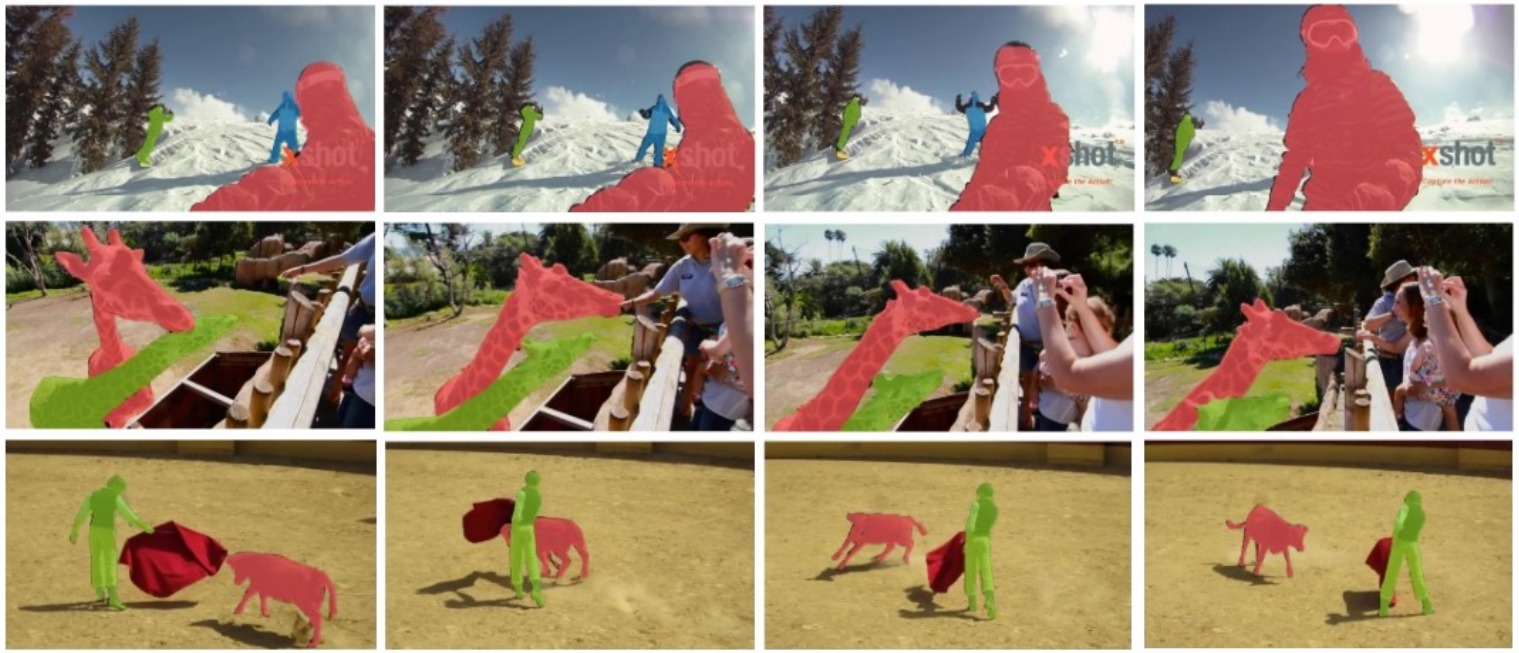


Figure 6: Qualitative results of video object segmentation on YouTube-VOS 2018.

Experiment

- Quantitative Results

method	venue	COCO-20^i one-shot	COCO-20^i few-shot	PASCAL-5^i one-shot	PASCAL-5^i few-shot
specialist model					
HSNet [35]	ICCV'21	41.2	49.5	66.2	70.4
HSNet*		41.7	50.7	68.7	73.8
VAT [19]	ECCV'22	41.3	47.9	67.9	72.0
VAT*		42.9	49.4	72.4	76.3
FPTrans [53]	NeurIPS'22	47.0	58.9	68.8	78.0
FPTrans*		56.5	65.5	77.7	83.2
generalist model					
Painter	CVPR'23	32.8	32.6	64.5	64.6
SegGPT	this work	56.1	67.9	83.2	89.8

method	venue	mIoU one-shot	mIoU few-shot
trained on FSS-1000			
DAN [43]	ECCV'20	85.2	88.1
HSNet [35]	ICCV'21	86.5	88.5
SSP [15]	ECCV'22	87.3	88.6
VAT [19]	ECCV'22	90.3	90.8
DACM [50]	ECCV'22	90.8	91.7
not trained on FSS-1000			
Painter	CVPR'23	61.7	62.3
SegGPT	this work	85.6	89.3

Experiment

- Quantitative Results on “unseen” video data

method	venue	YouTube-VOS 2018 [52]					DAVIS 2017 [37]			MOSE [12]		
		G	J_s	F_s	J_u	F_u	$J\&F$	J	F	$J\&F$	J	F
with video data												
AGAME [21]	CVPR’19	66.0	66.9	-	61.2	-	70.0	67.2	72.7	-	-	-
AGSS [29]	ICCV’19	71.3	71.3	65.5	75.2	73.1	67.4	64.9	69.9	-	-	-
STM [36]	ICCV’19	79.4	79.7	84.2	72.8	80.9	81.8	79.2	84.3	-	-	-
AFB-URR [27]	NeurIPS’20	79.6	78.8	83.1	74.1	82.6	74.6	73.0	76.1	-	-	-
RDE [25]	CVPR’22	83.3	81.9	86.3	78.0	86.9	86.1	82.1	90.0	48.8	44.6	52.9
SWEM [31]	CVPR’22	82.8	82.4	86.9	77.1	85.0	84.3	81.2	87.4	50.9	46.8	54.9
XMem [9]	ECCV’22	86.1	85.1	89.8	80.3	89.2	87.7	84.0	91.4	57.6	53.3	62.0
without video data												
Painter	CVPR’23	24.1	27.6	35.8	14.3	18.7	34.6	28.5	40.8	14.5	10.4	18.5
SegGPT	this work	74.7	75.1	80.2	67.4	75.9	75.6	72.5	78.6	45.1	42.2	48.0

Ablation Study

- (a) ensemble strategy
  - spatial and feature ensemble approach performs well
- (b) the number of frames (multiple examples)
  - optimal when using 8 frames

examples	ensemble	DAVIS 2017			FSS-1000	
		$J\&F$	J	F	mIoU	FB-IoU
1	-	70.0	66.4	73.7	85.5	90.8
4	Spatial	61.9	58.0	65.8	89.3	93.5
4	Feature	74.7	71.6	77.7	87.8	92.4
8	Feature	75.6	72.5	78.6	89.8	93.8

(a)

frames	DAVIS 2017				
	1	4	8	12	16
$J\&F$	70.0	74.7	75.6	74.8	74.6
J	66.4	71.6	72.5	71.6	71.4
F	73.7	77.7	78.6	77.9	77.8

(b)

In-Context Tuning

- Enables to customize a unique application with a set of data samples
- Tune a prompt for a specific dataset, scene, or even a person

method	venue	mIoU
specialist model		
FCN [32]	CVPR'15	29.4
RefineNet [28]	CVPR'17	40.7
DPT [39]	ICCV'21	49.2
Mask2Former [8]	CVPR'22	57.7
generalist model		
Painter	CVPR'23	49.9
SegGPT	this work	39.6

Table 5: Results on ADE20K semantic segmentation.

method	venue	PQ
specialist model		
PanopticFPN [23]	CVPR'19	40.3
SOLOv2 [47]	NeurIPS'20	42.1
Mask2Former [8]	CVPR'22	57.8
UViM [24]	NeurIPS'22	45.8
generalist model		
Painter	CVPR'23	43.4
SegGPT	this work	34.4

Table 6: Results on COCO panoptic segmentation.

- Random color scheme makes it more challenging for the model to use color as a simple indicator of in-domain tasks

Discussion and Conclusion

- Present a generalist segmentation model
- Strong capabilities in handling both in-domain and out-of-domain
- Random coloring regime for better generalization makes the training task more difficult, resulting in inferior performance in in-domain tasks
- Potential to serve as a powerful tool for enabling more diverse applications
- Future work: Large model / get more data using self-supervised?
- Remain optimistic that the best GPT-3 moment in the vision field is yet to come

Questions?

- A lot of details in previous paper

Images Speak in Images: A Generalist Painter for In-Context Visual Learning

Xinlong Wang[1*] Wen Wang[2*] Yue Cao[1*] Chunhua Shen[2] Tiejun Huang[1,3]

[1] Beijing Academy of Artificial Intelligence [2] Zhejiang University [3] Peking University

https://github.com/baaivision/Painter

Segment Anything

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

30 pages…

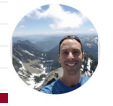
Segment Anything

Alexander Kirillov

FOLLOW

Research Scientist, Facebook AI Research (FAIR)
Verified email at fb.com - Homepage
computer vision machine learning deep learning

TITLE	CITED BY	YEAR
End-to-end object detection with transformers N Carion, F Massa, G Synnaeve, N Usunier, A Kirillov, S Zagoruyko European conference on computer vision, 213-229	6887	2020
Detectron2 Y Wu, A Kirillov, F Massa, WY Lo, R Girshick	2073 *	2019
Panoptic segmentation A Kirillov, K He, R Girshick, C Rother, P Dollár Proceedings of the IEEE/CVF conference on computer vision and pattern …	1150	2019
Panoptic Feature Pyramid Networks A Kirillov, R Girshick, K He, P Dollár Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern …	990	2019



Ross Girshick

FOLLOW

Research Scientist, Facebook AI Research (FAIR)
Verified email at eecs.berkeley.edu - Homepage
computer vision machine learning

TITLE	CITED BY	YEAR
Faster R-CNN: Towards real-time object detection with region proposal networks S Ren, K He, R Girshick, J Sun Advances in neural information processing systems, 91-99	62258	2015
Microsoft coco: Common objects in context TY Lin, M Maire, S Belongie, J Hays, P Perona, D Ramanan, P Dollár, … Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland …	37632	2014
You only look once: Unified, real-time object detection J Redmon, S Divvala, R Girshick, A Farhadi Proceedings of the IEEE conference on computer vision and pattern …	37623	2016
Rich feature hierarchies for accurate object detection and semantic segmentation R Girshick, J Donahue, T Darrell, J Malik Proceedings of the IEEE conference on computer vision and pattern …	34332	2014
Fast R-CNN R Girshick Proceedings of the IEEE International Conference on Computer Vision, 1440-1448	28720	2015
Mask R-CNN K He, G Gkioxari, P Dollár, R Girshick arXiv preprint arXiv:1703.06870	28355	2017
Focal loss for dense object detection TY Lin, P Goyal, R Girshick, K He, P Dollár Proceedings of the IEEE international conference on computer vision, 2980-2988	22820	2017
Feature pyramid networks for object detection TY Lin, P Dollár, R Girshick, K He, B Hariharan, S Belongie Proceedings of the IEEE conference on computer vision and pattern …	20402	2017

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

Piotr Dollár

팔로우

FAIR
fb.com의 이메일 확인됨 - 홈페이지
computer vision machine learning

제목	인용	연도
Microsoft coco: Common objects in context TY Lin, M Maire, S Belongie, J Hays, P Perona, D Ramanan, P Dollár, … Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland …	37632	2014
Mask r-cnn K He, G Gkioxari, P Dollár, R Girshick Computer Vision (ICCV), 2017 IEEE International Conference on, 2980-2988	28191	2017
Focal loss for dense object detection TY Lin, P Goyal, R Girshick, K He, P Dollár Proceedings of the IEEE international conference on computer vision, 2980-2988	22680	2017
Feature pyramid networks for object detection TY Lin, P Dollár, R Girshick, K He, B Hariharan, S Belongie Proceedings of the IEEE conference on computer vision and pattern …	20251	2017
Aggregated residual transformations for deep neural networks S Xie, R Girshick, P Dollár, Z Tu, K He Proceedings of the IEEE conference on computer vision and pattern …	10136	2017

30 pages…

Introduction

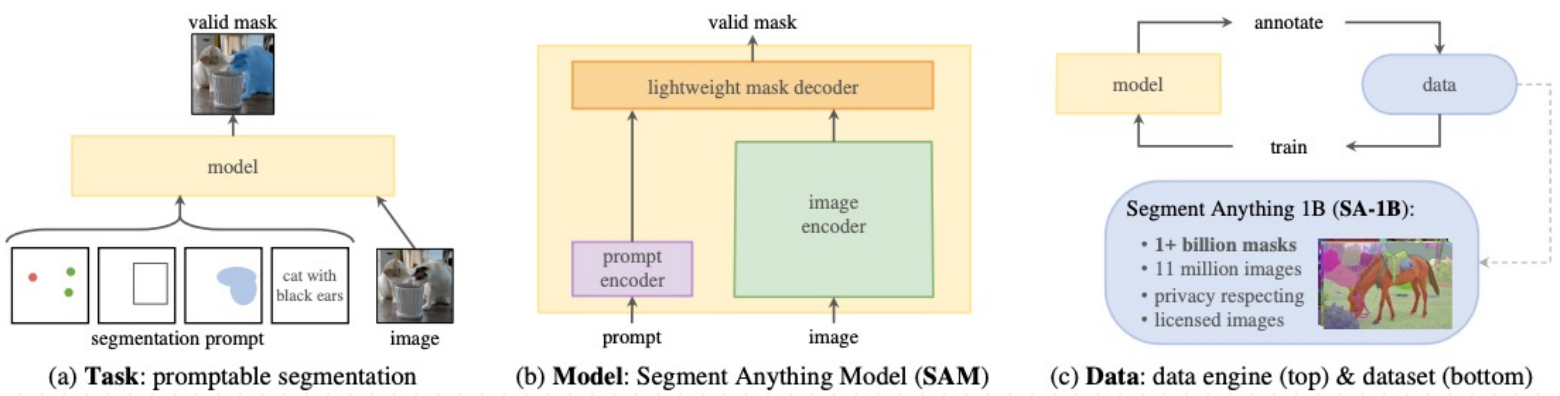
- Large Language Models (GPT-3)
  - has strong zero-shot and few-shot generalization capability
  - Then, How to build a foundation model for image segmentation?
- Three components
  - Task: What task will enable zero-shot generalization?
  - Model: What is the corresponding model architecture?
  - Data: What data can power this task and model?

Introduction

- Task
  - Promptable Segmentation
- Model
  - Image encoder + Prompt encoder + Mask decoder
- Data
  - Data engine: co-develop our model with model-in-the-loop dataset annotation
  - SA-1B (1B+ masks from 11M images)

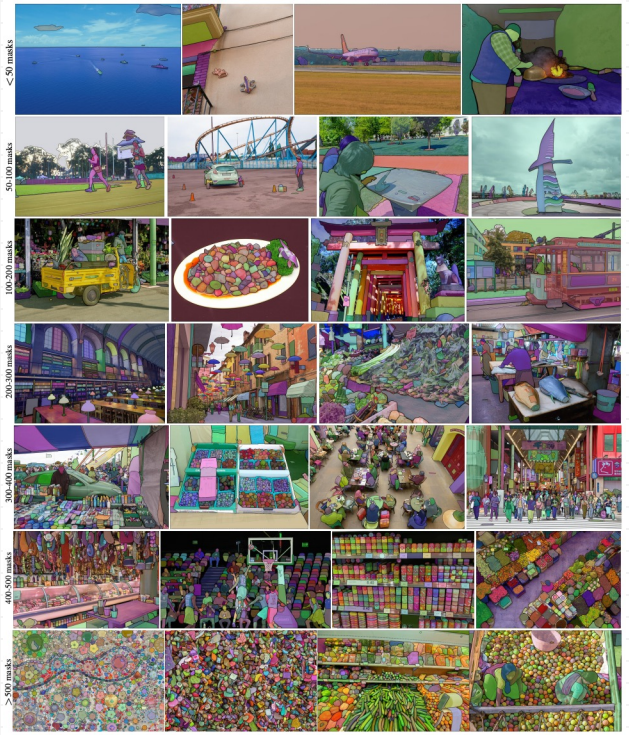
Introduction

Three Components



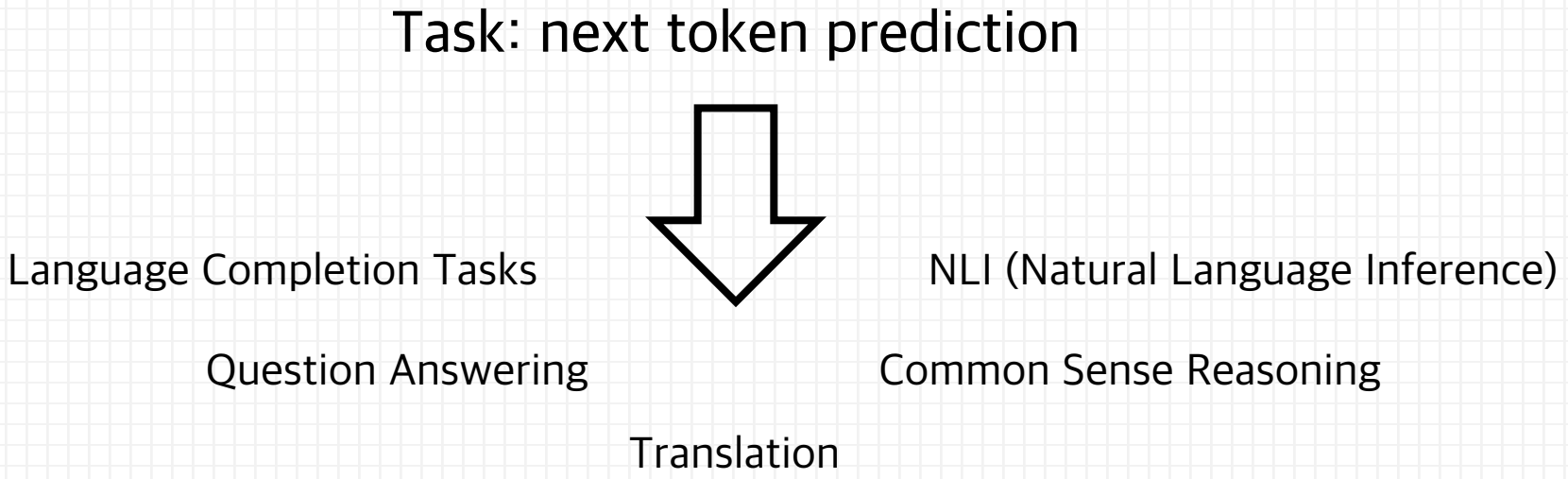
(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

SA-1B



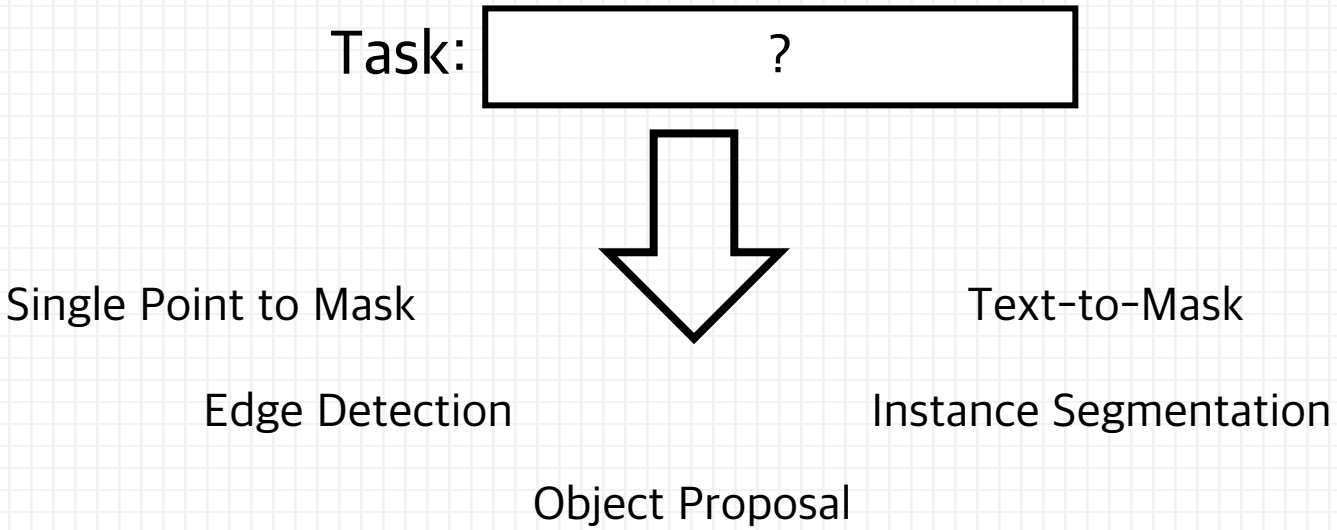
Task

- GPT-3 trained to The next token prediction
  -> solve diverse downstream tasks via prompt engineering



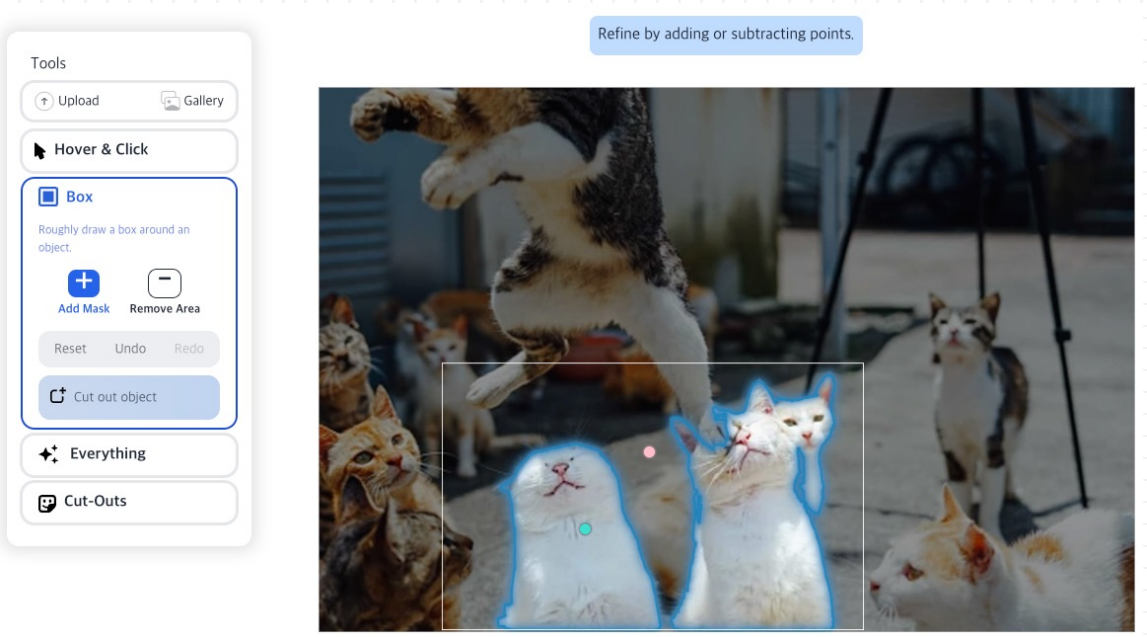
Task

- What task is will enable zero-shot segmentation?
  - like GPT-3



Task

- Promptable Segmentation
  - return a valid segmentation mask given any prompt
    - prompt: points, boxes, masks, texts

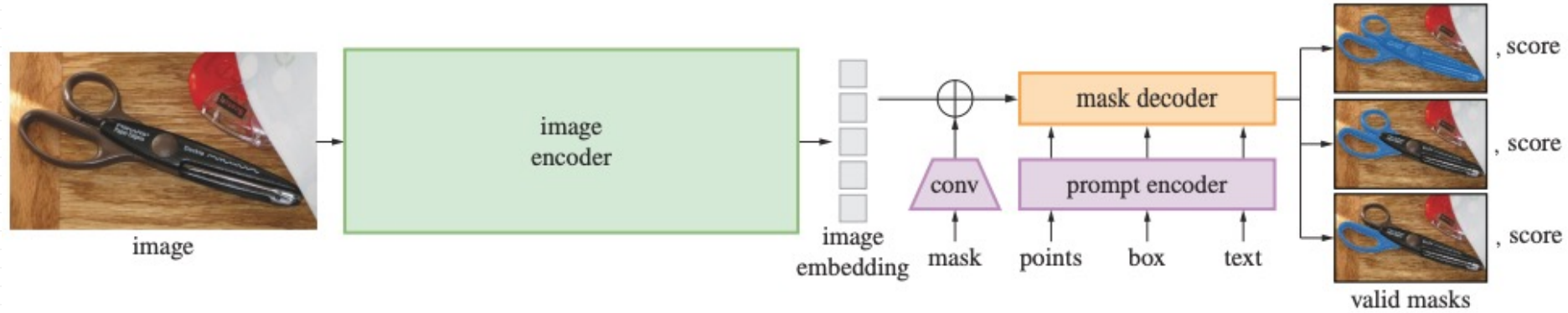


Task

- Promptable Segmentation
  - leads to a natural pre-training algorithm
    and a general method for zero-shot transfer to downstream tasks
- Pre-training Algorithm
  - simulates a sequence of prompts
  - compares the model's mask predictions against the ground truth
  - + predicts a valid mask for any prompt even ambiguous

Model

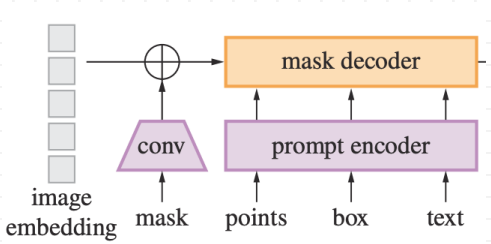
- Segment Anything Model (SAM)



  - heavyweight image encoder outputs an image embedding
  - prompt encoder embeds prompts
  - lightweight mask decoder predicts segmentation masks

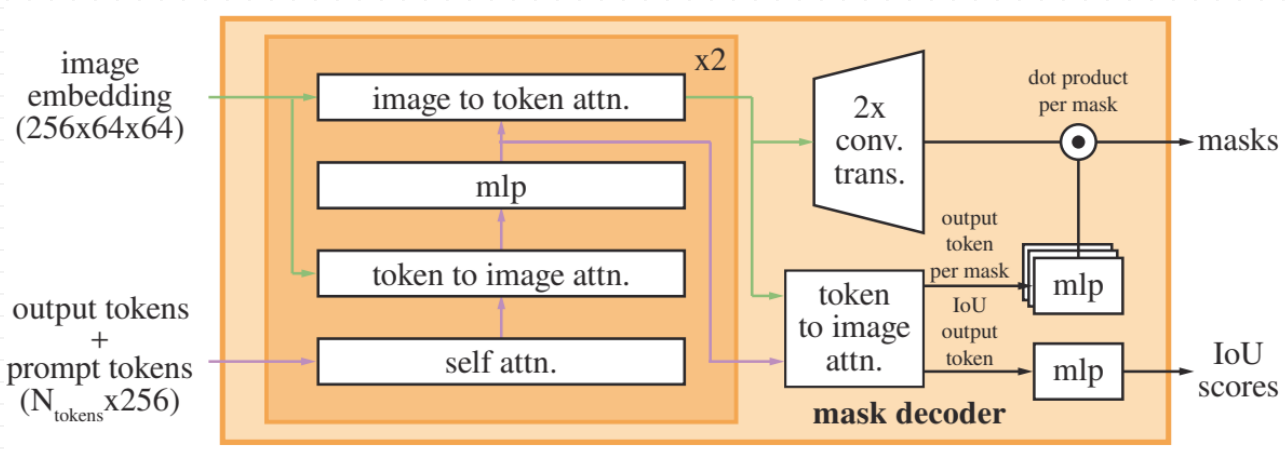
Model

- Image encoder
  - MAE pre-trained Vision Transformer (ViT)
- Prompt encoder
  - sparse (points, boxes, text)
    - represent by positional encodings summed with learned embeddings
    - free-from text with an off-the-shelf text encoder from CLIP
  - dense (masks)
    - embedded using convolutions and summed element-wise with the image embedding



Model

- Mask decoder
  - Modification of a Transformer decoder block
    - prompt self-attention
    - cross-attention in two direction (prompt-to-image embedding and vice-versa)



Model

- Resolving ambiguity
  - Average multiple valid masks
  - 3 masks output
    - whole, part, and subpart
    - backprop only the minimum loss over masks
    - To rank masks, predicts a confidence score (IoU)

Model

- Efficiency
  - Prompt encoder and mask decoder run in a web browser on CPU, in ~50ms
- Losses and training
  - focal loss $\mathrm{FL}(p_t) = -(1-p_t)^{\gamma}\log(p_t)$
  - dice loss $1 - \frac{2 * |A \cap B|}{|A| + |B|} = 1 - \frac{2 * \mathrm{TP}}{(\mathrm{TP}+\mathrm{FP}) + (\mathrm{TP}+\mathrm{FN})}$
  - interactive segmentation setup
    - by randomly sampling prompts in 11 rounds per mask

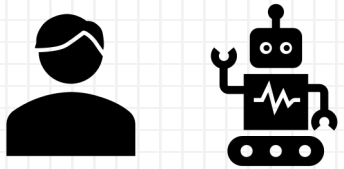
Model

- Interactive segmentation setup
  - 11 total iterations:
  1. one sampled initial input prompt
  2. 8 iteratively sampled points
  3. two iterations where no new external information is supplied

Data Engine

- Three stages

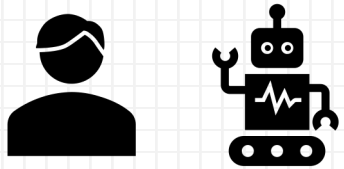
1. Assisted-manual



Main Assist

4.3M masks from 120k images
20 to 44 masks / image

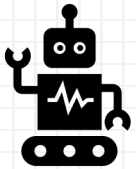
2. Semi-automatic



Assist Main

5.9M masks from 180k images
44 to 72 masks / image

3. Fully-automatic

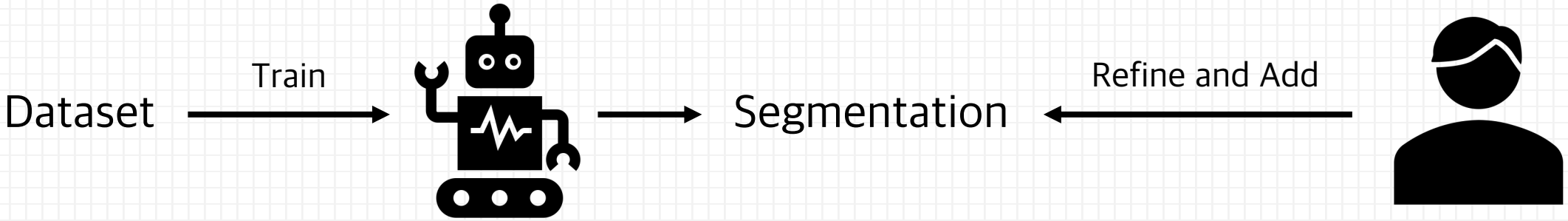


Main

1.1B masks from 11M images

Data Engine

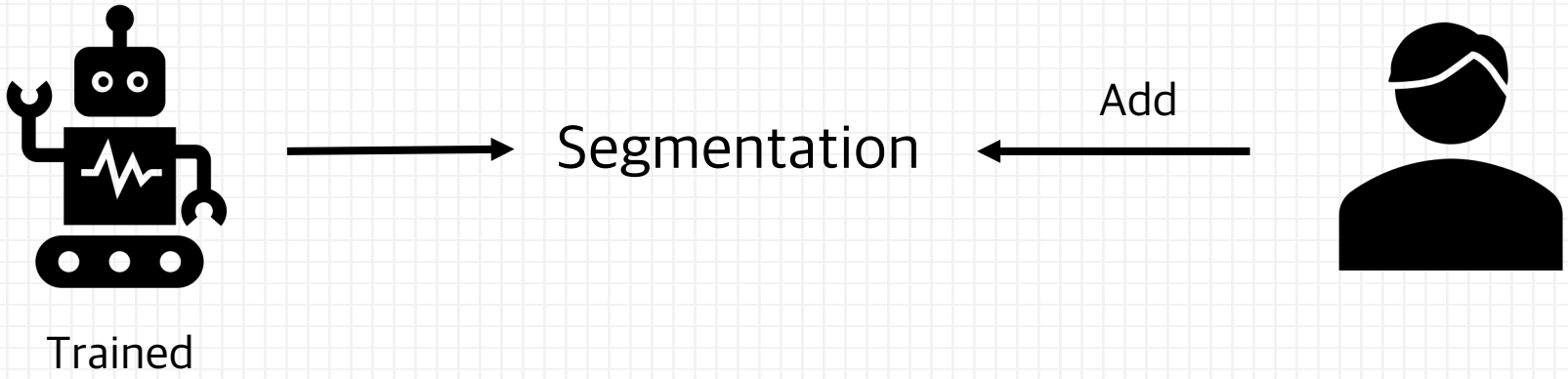
- Assisted-manual
  - in browser



4.3M masks from 120k images

Data Engine

- Semi-automatic
  - in browser



5.9M masks from 180k images

Data Engine

- Fully-automatic

1.1B masks from 11M images

Dataset

- Images
  - 11M high resolution (3300x4950 on average)
  - Releasing downsampled images (1500x~)
- Masks
  - 1.1B masks, 99.1% of which were generated fully automatically
- Mask quality
  - 94% of sampled pairs have greater than 90% IoU

Dataset

- Mask properties
  - Spatial distributions on object centers

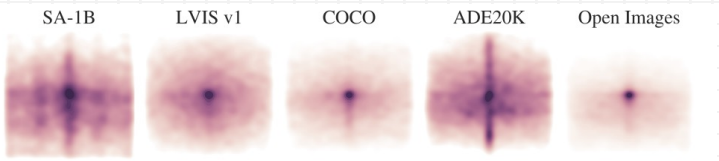


Figure 5: Image-size normalized mask center distributions.

  - Comparisons

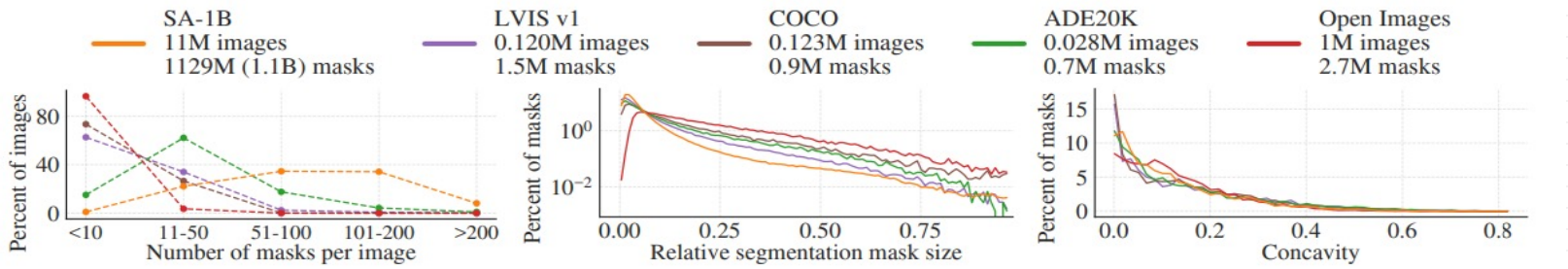


Figure 6: Dataset mask properties. The legend references the number of images and masks in each dataset. Note, that SA-1B has 11× more images and 400× more masks than the largest existing segmentation dataset Open Images [60].

RAI Analysis

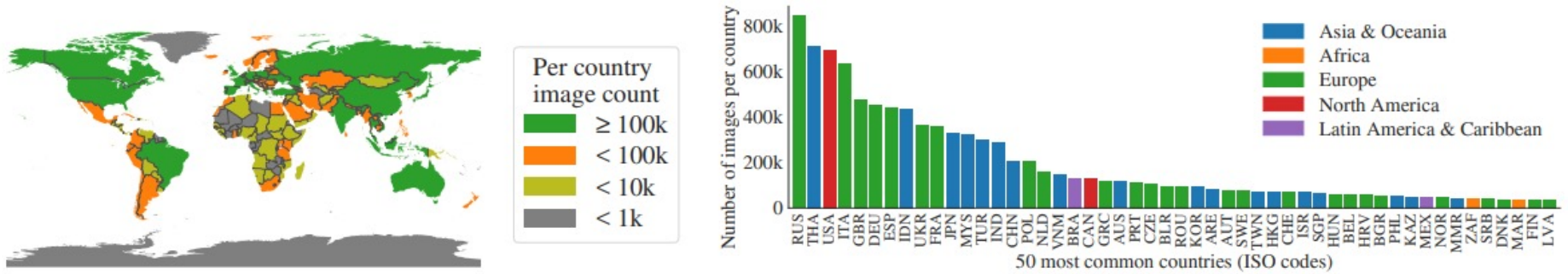


Figure 7: Estimated geographic distribution of SA-1B images. Most of the world's countries have more than 1000 images in SA-1B, and the three countries with the most images are from different parts of the world.

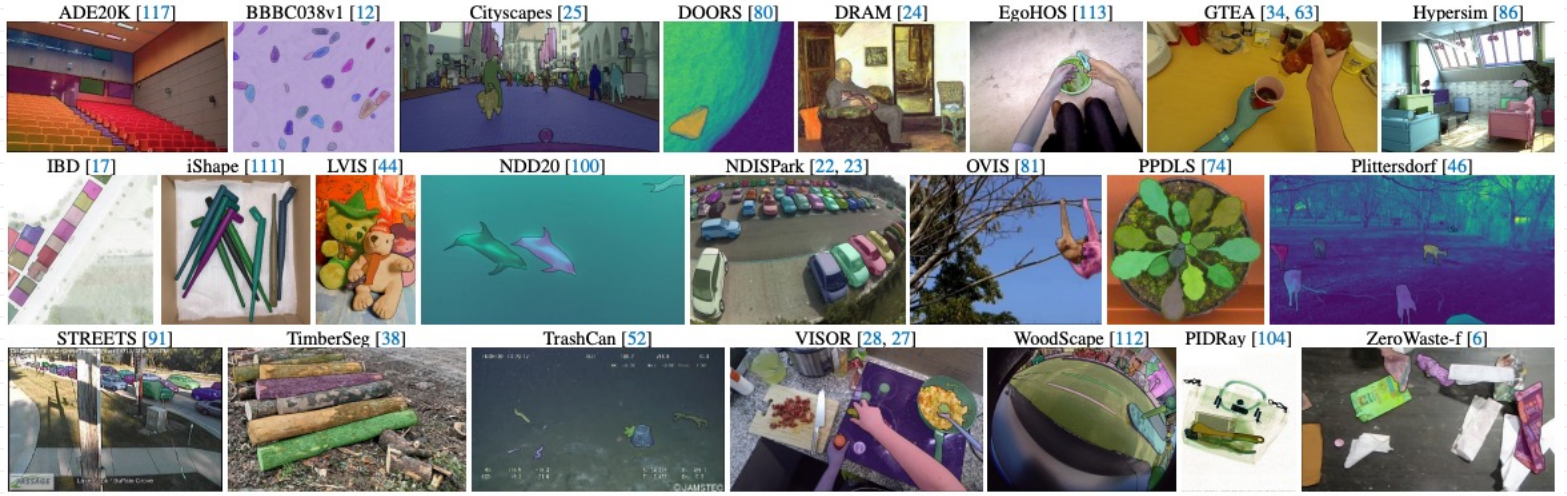
	# countries	SA-1B #imgs	SA-1B #masks	% images SA-1B	% images COCO	% images O.I.
Africa	54	300k	28M	2.8%	3.0%	1.7%
Asia & Oceania	70	3.9M	423M	36.2%	11.4%	14.3%
Europe	47	5.4M	540M	49.8%	34.2%	36.2%
Latin America & Carib.	42	380k	36M	3.5%	3.1%	5.0%
North America	4	830k	80M	7.7%	48.3%	42.8%
high income countries	81	5.8M	598M	54.0%	89.1%	87.5%
middle income countries	108	4.9M	499M	45.0%	10.5%	12.0%
low income countries	28	100k	9.4M	0.9%	0.4%	0.5%

	mIoU at 1 point	mIoU at 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	mIoU at 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

Zero-shot Transfer Experiments

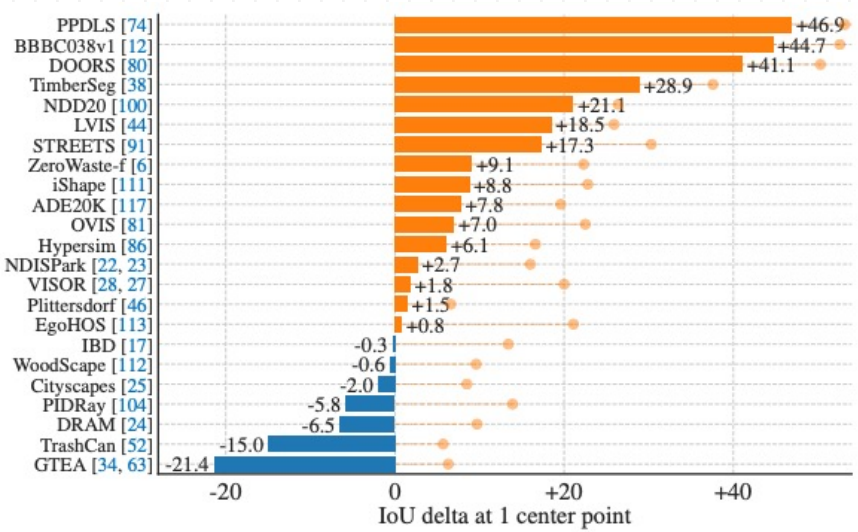
- Testing the core goal of promptable segmentations
  - Five tests
  - 23 diverse datasets



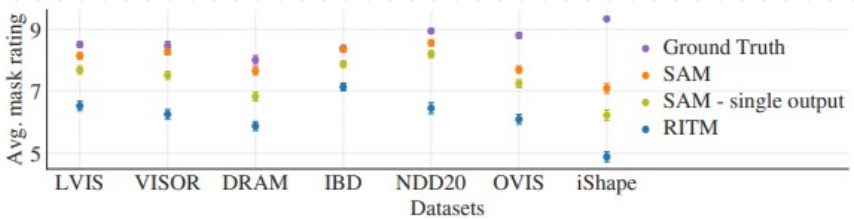
Zero-shot Transfer Experiments

- Zero-shot Single Point Valid Mask Evaluation

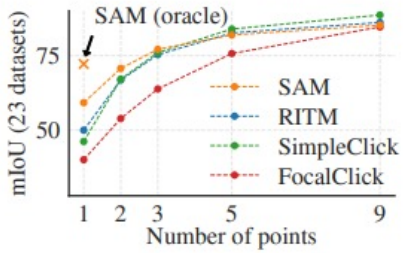
Task



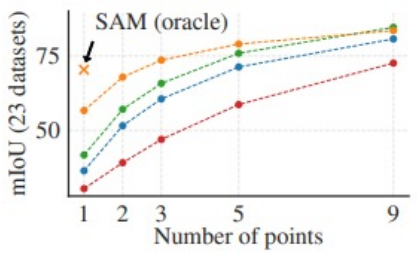
(a) SAM *vs.* RITM [92] on 23 datasets



(b) Mask quality ratings by human annotators



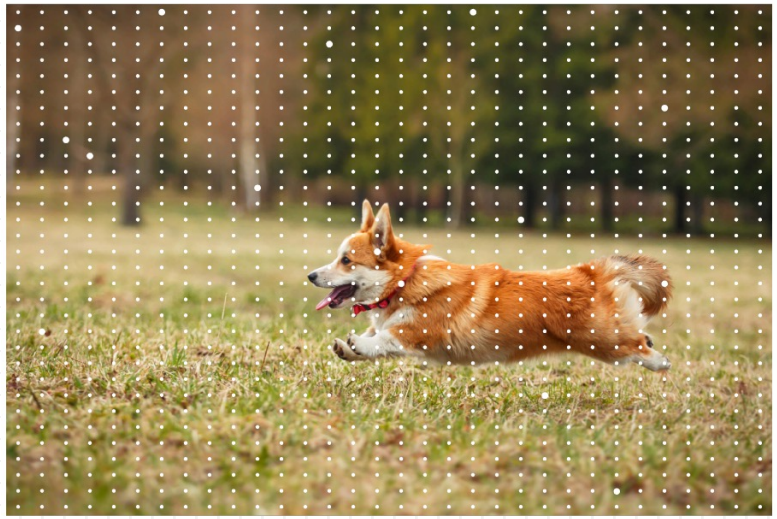
(c) Center points (default)



(d) Random points

Zero-shot Transfer Experiments

- Zero-shot Edge Detection
  1. Prompt SAM 16x16 regular grid of foreground points
  2. 3 masks per point -> 16x16x3=768 predicted masks
  3. Redundant masks are removed by NMS
  4. Sobel filtering



Zero-shot Transfer Experiments

- Zero-shot Edge Detection

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

Table 3: Zero-shot transfer to edge detection on BSDS500.

Zero-shot Transfer Experiments

- Zero-shot Object Proposals
  - Modify the number of prompts -> 1000+ masks
  - Based on confidence score and stability, test top 1000 masks

	mask AR@1000						
method	all	small	med.	large	freq.	com.	rare
ViTDet-H [62]	63.0	51.7	80.8	87.0	63.1	63.3	58.3
zero-shot transfer methods:							
SAM – single out.	54.9	42.8	76.7	74.4	54.7	59.8	62.0
SAM	59.3	45.5	81.6	86.9	59.1	63.9	65.8

Table 4: Object proposal generation on LVIS v1. SAM is applied zero-shot, *i.e.* it was not trained for object proposal generation nor did it access LVIS images or annotations.

Zero-shot Transfer Experiments

- Zero-shot Instance Segmentation
  - Run object detector (ViTDet) -> prompt SAM with its output box

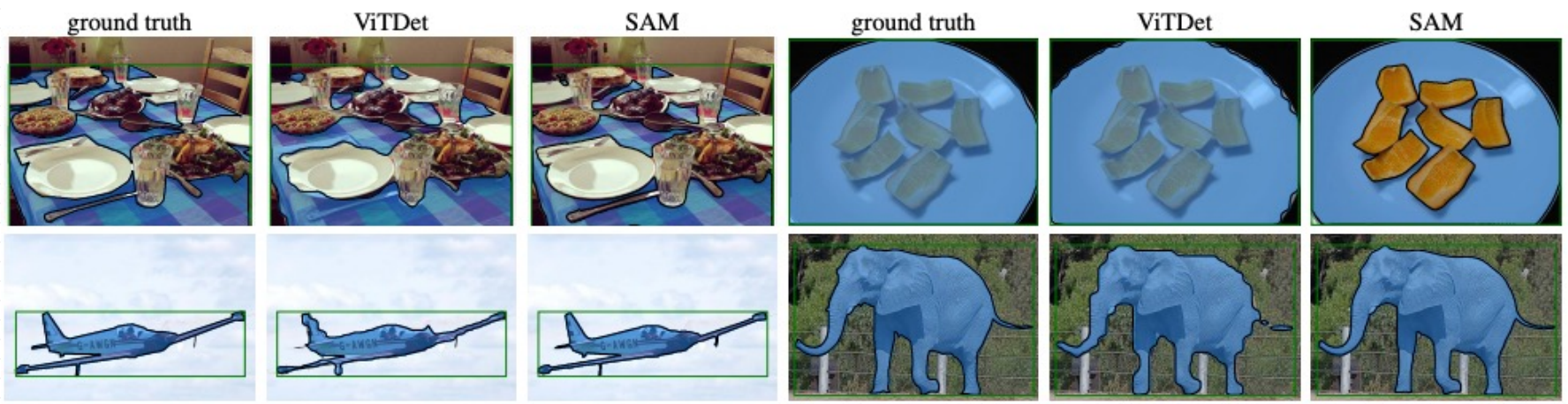
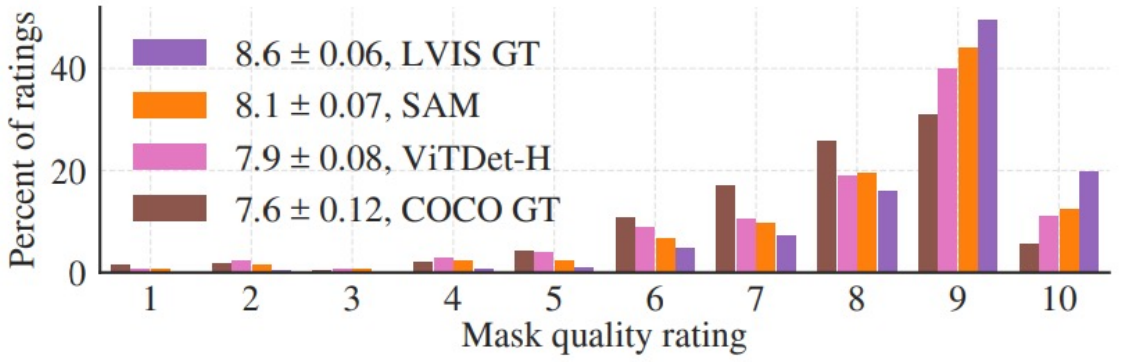


Figure 16: Zero-shot instance segmentation on LVIS v1. SAM produces higher quality masks than ViTDet. As a zero-shot model, SAM does not have the opportunity to learn specific training data biases; see top-right as an example where SAM makes a modal prediction, whereas the ground truth in LVIS is amodal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

Zero-shot Transfer Experiments

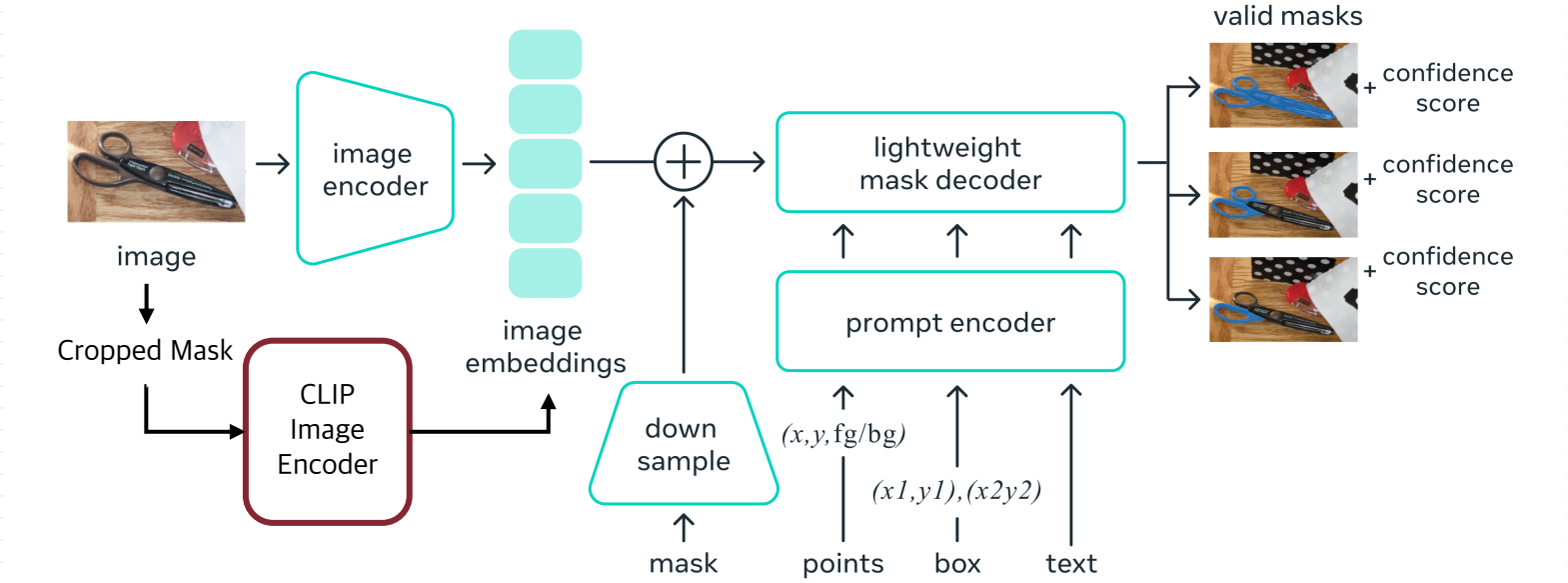
- Zero-shot Instance Segmentation

	COCO [66]				LVIS v1 [44]			
method	AP	AP^S	AP^M	AP^L	AP	AP^S	AP^M	AP^L
ViTDet-H [62]	51.0	32.0	54.3	68.9	46.6	35.0	58.0	66.3
zero-shot transfer methods (segmentation module only):								
SAM	46.5	30.8	51.0	61.7	44.7	32.5	57.6	65.5



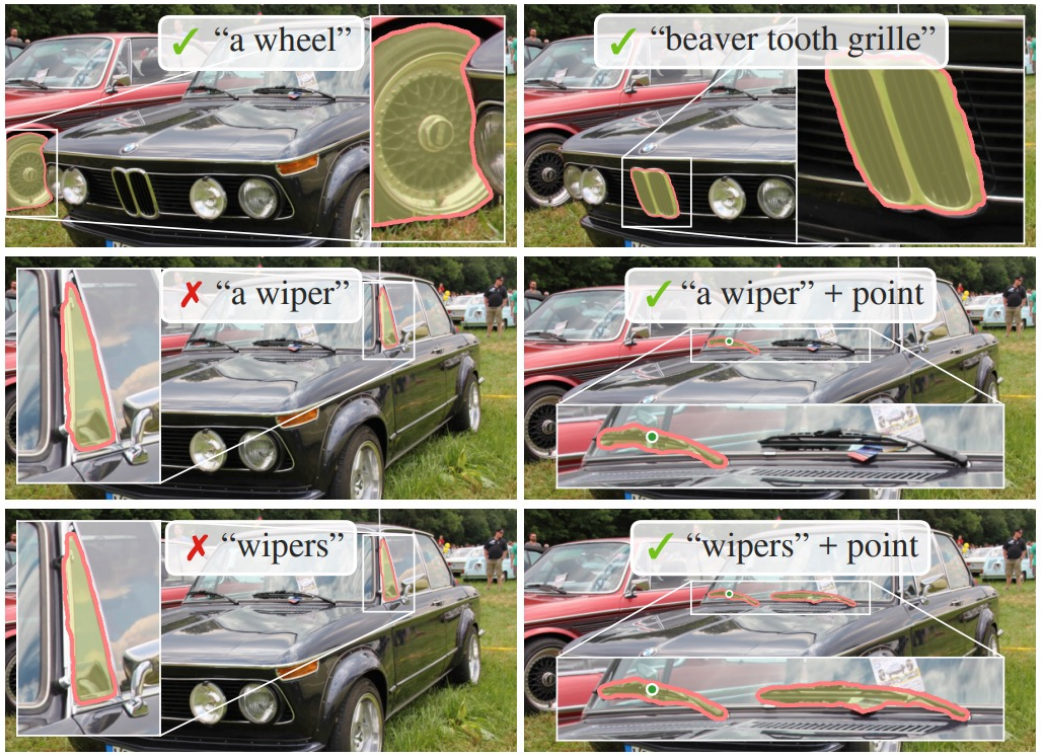
Zero-shot Transfer Experiments

- Zero-shot Text-to-Mask
  - SAM's training procedure is modified to make it text-aware



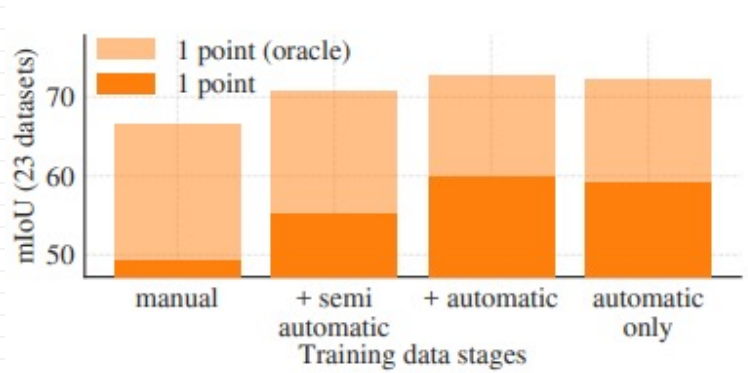
Zero-shot Transfer Experiments

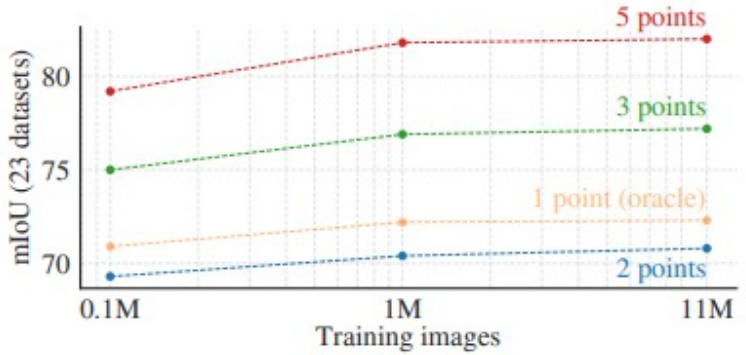
- Zero-shot Text-to-Mask

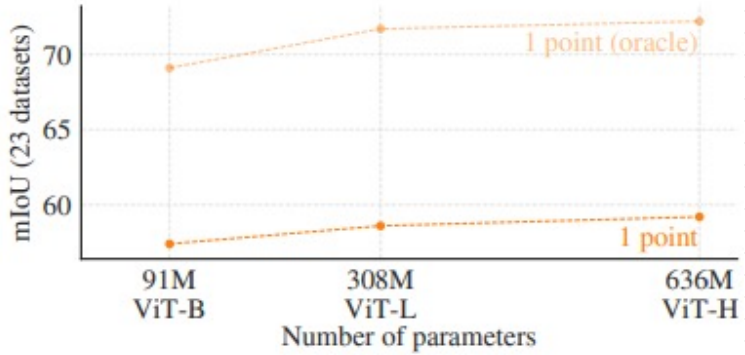


Ablations

- Training data stages
  - SAM has systematic biases between top-ranked masks vs. GT
  - Oracle: best mask with respect to the ground truth







Discussion

- Foundation Models
  - pre-trained models -> foundation models
    - “trained on broad data at scale and are adaptable to wide range of downstream tasks”
  - initialized with a self-supervised technique (MAE)
  - large-scale supervised learning -> effective
- Compositionality
  - SAM predicts a valid mask for a wide range of segmentation prompts
  - generalization capability -> without need for additional training

Discussion

- Limitations
  - Not perfect (due to focusing on generality and breadth of use)
  - Not real-time when using a heavy image encoder
    - Unclear to design simple prompts that implement semantic and panoptic segmentation
  - Domain-specific tools expect to outperform SAM
- Conclusion
  - New task (promptable segmentation), model (SAM), and dataset (SA-1B)
  - Our huge dataset and model will help pave the path ahead

Thank you for listening !

- More details in Appendix

Table of contents:

- Questions?