Transformers, Vision Transformers and SAMJ

Carlos Javier García López de Haro (IP)
Caterina Fuster Barceló (UC3M)
Daniel Sage (EPFL)

INSTITUT PASTEUR
uc3m Universidad Carlos III de Madrid
EPFL

Content

- Transformers and Vision Transformers
- Segment Anything Model (SAM) and SAM-like models
- SAMJ
- Hands on activities

Transformers and Vision Transformers

The Transformer

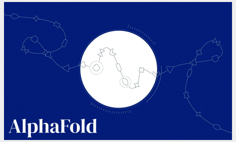
Introduced in 2017 by Vaswani et al, from Google

New architecture “just” for language translation

Currently is the cornerstone of the Artificial Intelligence revolution🤖

ChatGPT

Music generation



Protein folding

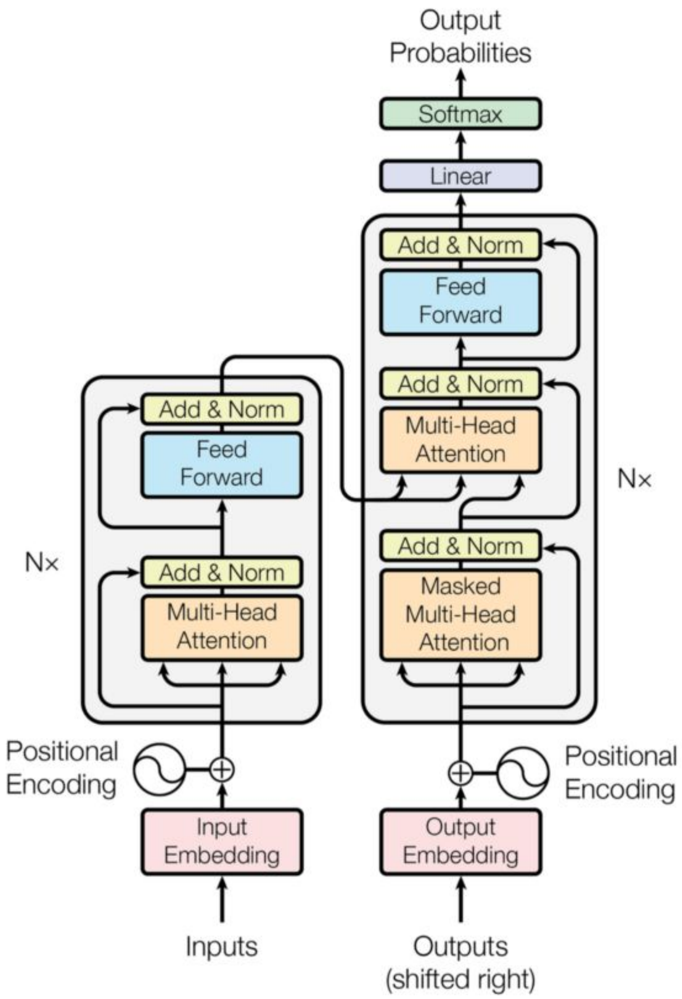


Figure 1: The Transformer - model architecture.

Attention is all you need

3 key contributions

Sef-attention

Multi-head attention

Positional encoding

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Attention is all you need

Tokenization → letters to numbers

My big black dog is called Harry.

↓

My big black dog is called Harry.

4 chars ~ 1 token

Attention is all you need

Embedding

tokens (numbers) to tensors

My big black dog is called Harry.

My big black dog is called Harry.

(12228 x 8) tensor

Tries to represent tokens as “ideas”

Attention is all you need

Embeddings locate similar ideas together

My big black dog is called Harry.

Harry Kane

Prince Harry

Harry Potter

Attention is all you need

Attention blocks

change the "meaning" of words given the context

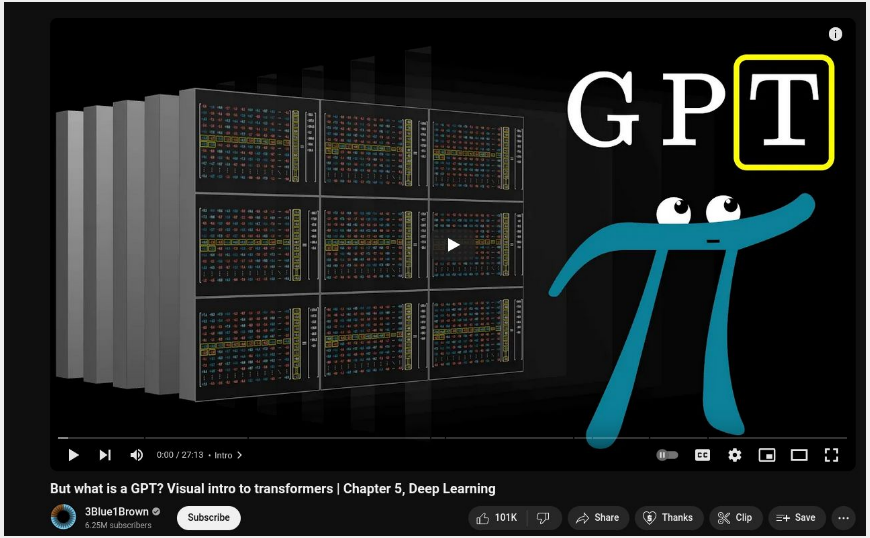
self-attention + multihead attention

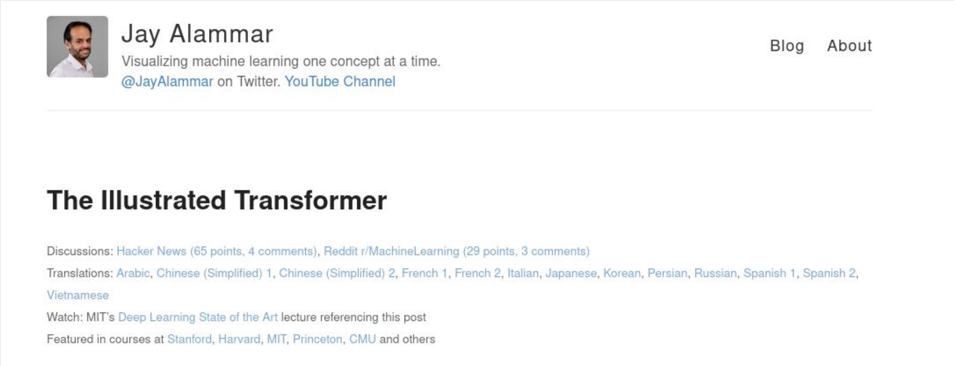
My big black dog is called Harry.

My big black dog is called Harry.

Harry
(after the last attention block)

Attention is all you need

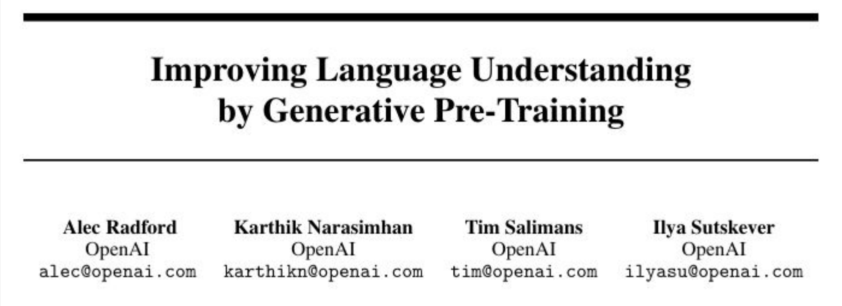




[3blue1brown videos on Transformers]

[The Illustrated Transformer]

Generative Pre-trained Transformer (GPT)


Improving Language Understanding by Generative Pre-Training

Alec Radford
OpenAI
alec@openai.com

Karthik Narasimhan
OpenAI
karthikn@openai.com

Tim Salimans
OpenAI
tim@openai.com

Ilya Sutskever
OpenAI
ilyasu@openai.com

Decoder-only

Self-supervised → No need for annotated data(!!)

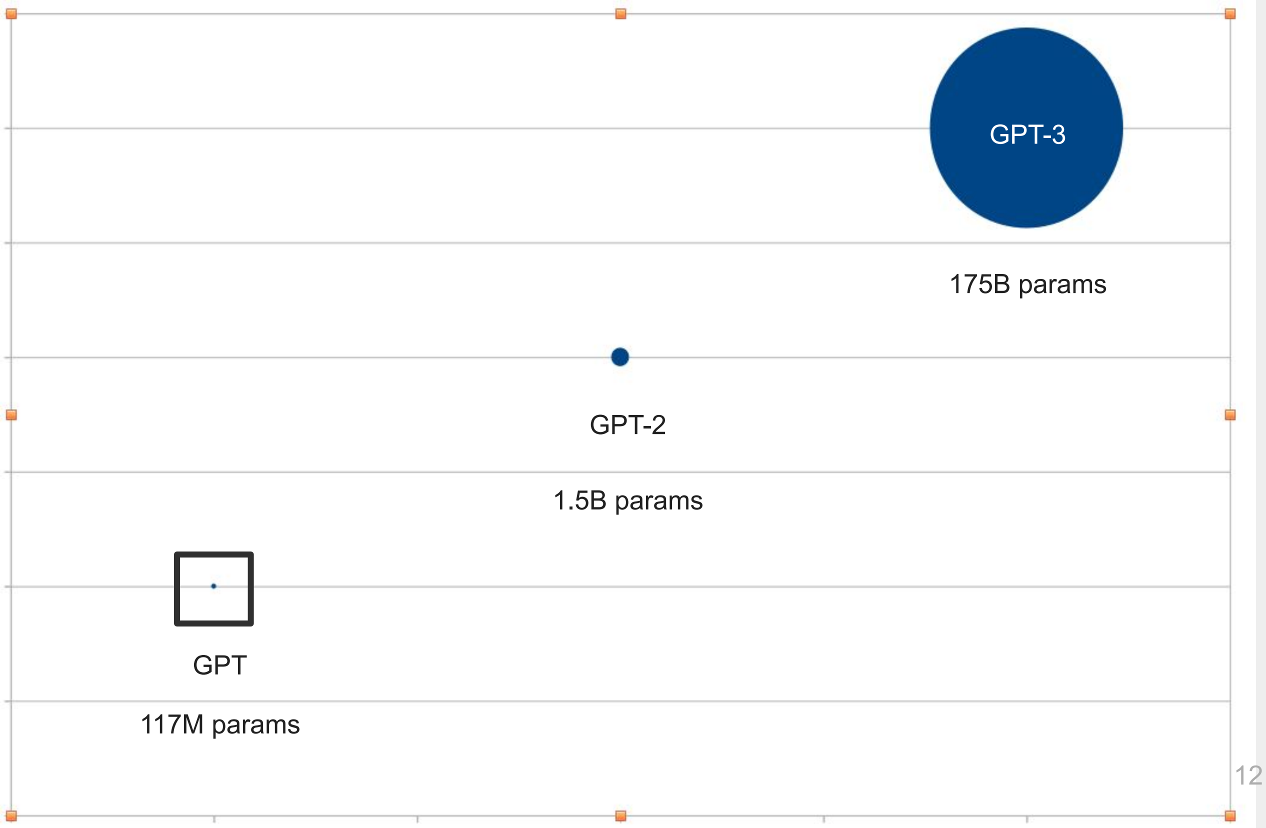
Emerging capabilities

Trained for next token prediction → Works for translation, question answering… (!!!!)

Generative Pre-trained Transformer

Scaling the model works

Both in number of **params** and **training data**



Generative Pre-trained Transformer

Scaling the model works

Prompt : *Python code to find the smallest factor of a number*

GPT-1:

the lack of a body in the room before me. after several moments of silence, he spoke again. " you are my daughter. the two of us are one. and in time you will

GPT-2

let p = &[5 - 3] => (1, 2, 3) The example above uses the "P" module to print the smallest factor of the number:

GPT-3

import math def lfact (n): factors = [1] factors . append (n) while n % factor <> 0: factor = factor * 2 - 1 while int (factor) > n % factor : factors . append (factor) return factors

What the code does:

Imports math module for math functions. [..]

Vision Transformer (ViT)

AN IMAGE IS WORTH 16X16 WORDS:
TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy[*,†], Lucas Beyer[*], Alexander Kolesnikov[*], Dirk Weissenborn[*], Xiaohua Zhai[*], Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby[*,†]

[*]equal technical contribution, [†]equal advising

Google Research, Brain Team

{adosovitskiy, neilhoulsby}@google.com

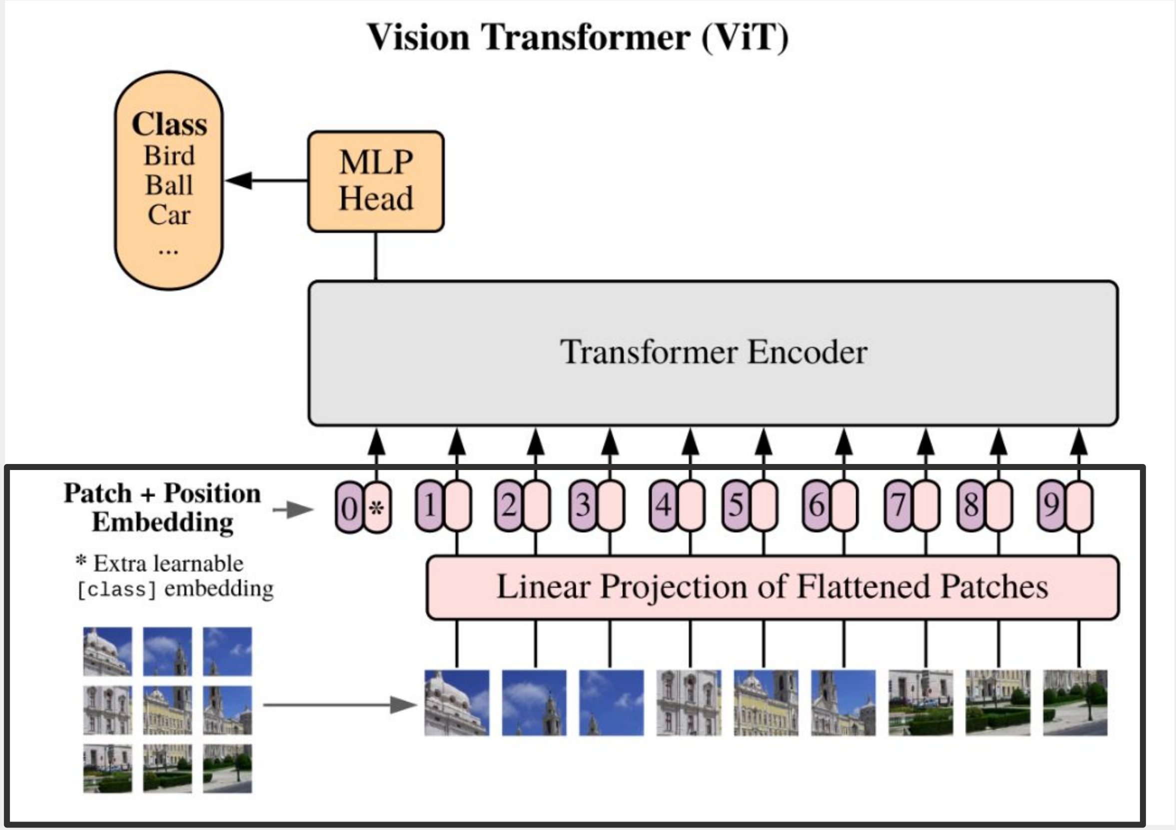
Using transformers for vision

Required **huge amounts of data and params** to outperform CNNs

Vision Transformer (ViT)

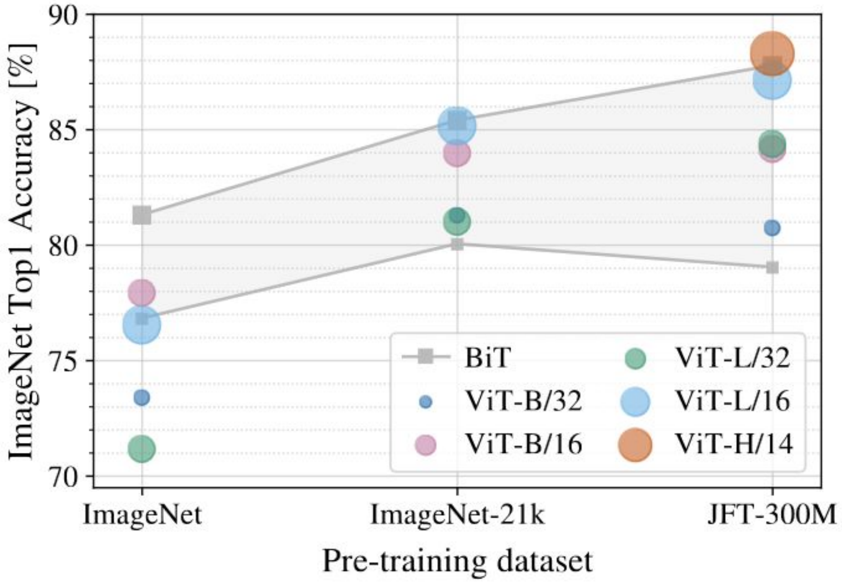
Divide the image into patches

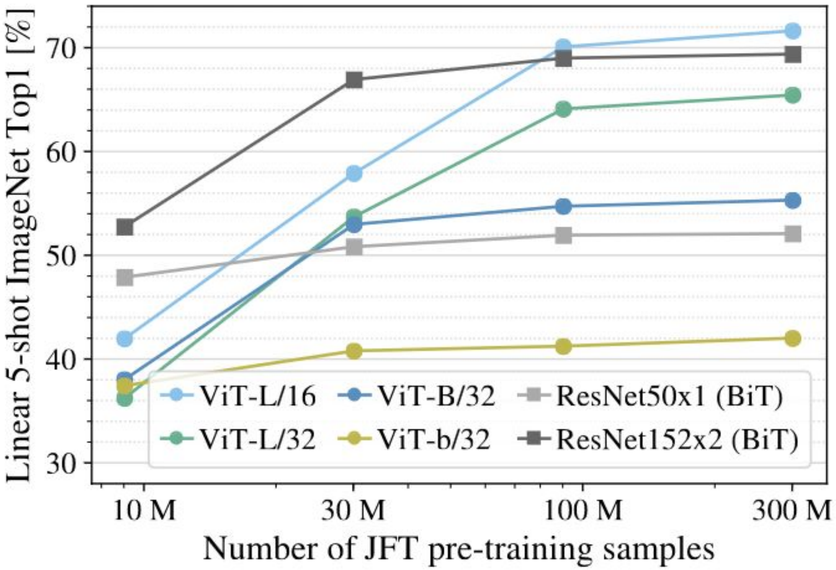
Find relations between patches



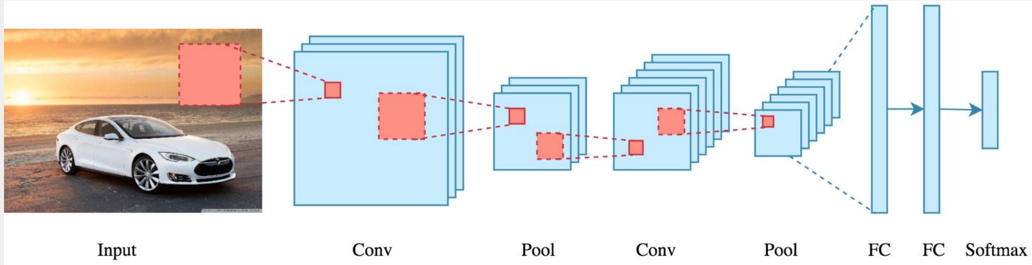
Transformers vs CNNs

BiT -Big Transfer (CNN)
ImageNet - 1.2m images
ImageNet21k - 14m images
ViT-H > ViT-L > ViT-B





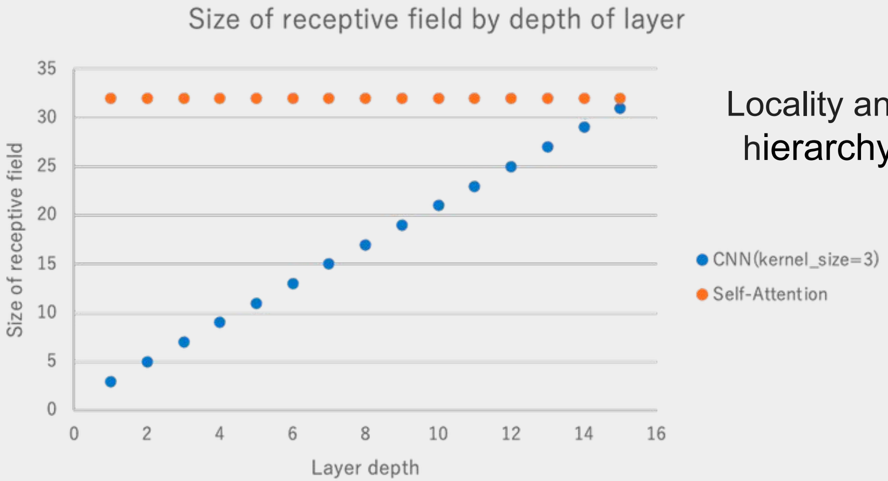
Transformers vs CNNs



CNNs enforce inductive biases → Useful assumptions for image data ← ViTs have to learn them

CNNs enforce:

- Locality
- Translational equivariance
- Hierarchy



Locality and hierarchy

Transformers in Vision - Useful resources

Overview of ViTs with one of the authors

ViT explanation with code

ViTs for small datasets

ViTs for small datasets (the whole channel is quite good)

Foundational models for Vision: SAM and Dino

Extra:
ConvNexts

Transformers are data hungry

Transformers need a LOT of data

The bigger they are and the more data they see the more they learn about it

Transformers **can learn relationships between anything**

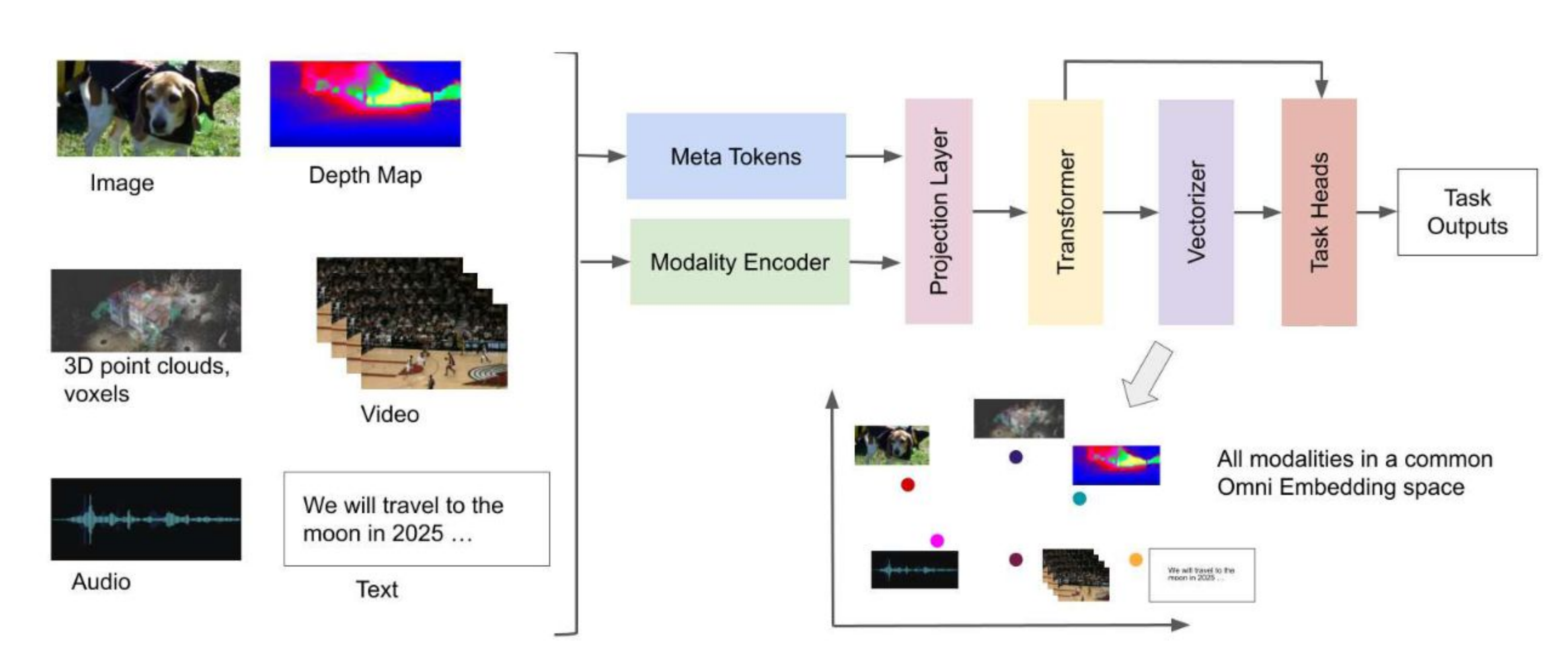
aminoacids -> protein folding

text

images

audio

Everything to everything models -Multimodal transformers

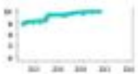
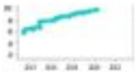


https://paperswithcode.com/sota

State of the art in Computer Vision

Image classification

Natural images

Trend	Dataset	Best Model	
	ImageNet	OmniVec(ViT)	Multi-modal transformer
	CIFAR-10	ViT-H/14	ViT
	CIFAR-100	EffNet-L2 (SAM)	CNN

https://paperswithcode.com/sota

State of the art in Computer Vision

Natural images

Semantic segmentation

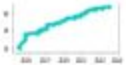
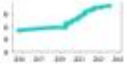
Trend	Dataset	Best Model	
	ADE20K	ONE-PEACE	Multi-modal transformer
	NYU Depth v2	GeminiFusion (Finetune-Swin-Large)	ViT
	Cityscapes test	VLTSeg	Vision-Language (multimodal) transformer

https://paperswithcode.com/sota

State of the art in Computer Vision

Object detection

Natural images

Trend	Dataset	Best Model	
	COCO test-dev	Co-DETR	ViT
	COCO minival	Co-DETR	ViT
	COCO-O	EVA	ViT
	PASCAL VOC 2007	Cascade Eff-B7 NAS-FPN (Copy Paste pre-training, single-scale)	CNN

https://paperswithcode.com/sota

State of the art in Computer Vision

Medical Image segmentation

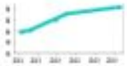
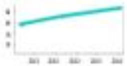
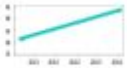
Colonoscopy images

Trend	Dataset	Best Model	
	Kvasir-SEG	DUCK-Net	CNN
	CVC-ClinicDB	DUCK-Net	CNN
	CVC-ColonDB	DUCK-Net	CNN
	ETIS-LARIBPOLYPDB	DUCK-Net	CNN

State of the art in Computer Vision

https://paperswithcode.com/sota

Medical Image segmentation

CT scans		Synapse multi-organ CT	Swin UNETR	CNN
MRI cardiac images		Automatic Cardiac Diagnosis Challenge (ACDC)	FCT	CNN
Tissue images		MoNuSeg	Hi-gMISnet	CNN
Nuclei images		2018 Data Science Bowl	DuAT	ViT
Gland segmentation in Colon Histology images		GlaS	Hi-gMISnet	CNN

State of the art in Computer Vision

Tasks with **millions of images available** are dominated by **transformers**

Specific tasks with **more difficult data acquisition** are still dominated by **CNNs**

Transformers in Microscopy - Cell segmentation

Transformers still **underperform** CNN methods for cell segmentation

Cellpose (CNN) method is still the king

Cellpose with transformer backbone underperforms CNN backbone

ViT

Analysis | Published: 26 March 2024

The multimodality cell segmentation challenge: toward universal solutions

Jun Ma, Ronald Xie, Shamini Ayyadhury, Cheng Ge, Anubha Gupta, Ritu Gupta, Song Gu, Yao Zhang, Gihun Lee, Joonkee Kim, Wei Lou, Haofeng Li, Eric Upschulte, Timo Dickscheid, José Guilherme de Almeida, Yixin Wang, Lin Han, Xin Yang, Marco Labagnara, Vojislav Gligorovski, Maxime Scheder, Sahand Jamal Rahi, Carly Kempster, Alice Pollitt, … Bo Wang + Show authors

Nature Methods (2024) | Cite this article

13k Accesses | 65 Altmetric | Metrics

Debunked by

Transformers do not outperform Cellpose

Carsen Stringer[†], **Marius Pachitariu**[†]

HHMI Janelia Research Campus, Ashburn, VA, USA

[†] correspondence to (stringerc, pachitarium) @ janelia.hhmi.org

CNN

Transformers in Microscopy - Cell segmentation

Article | Published: 14 December 2020

Cellpose: a generalist algorithm for cellular segmentation

Carsen Stringer, Tim Wang, Michalis Michaelos & Marius Pachitariu

Nature Methods 18, 100–106 (2021) | Cite this article

Transformers do not outperform Cellpose

Carsen Stringer[†], Marius Pachitariu[†]
HHMI Janelia Research Campus, Ashburn, VA, USA
[†] correspondence to (stringerc, pachitarium) @ janelia.hhmi.org

Cellpose authors claim that **ViTs success may not translate to biological images**

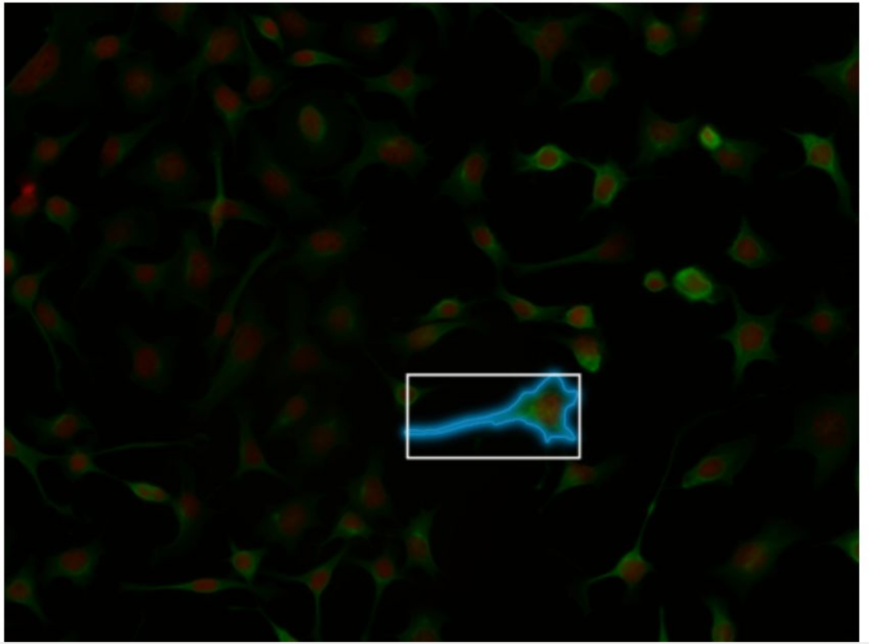
It may be impossible to collect millions of diverse biological images for training

Transformers in Microscopy - Cell segmentation

SAM (Segment Anything model) **performs well** on cell data

Training data of natural images, cell images were a small percentage

There might be hope for ViTs in cell images



https://segment-anything.com/demo#

The story of Uncle SAM

Rule-based

Model-based

Machine Learning

Deep Learning

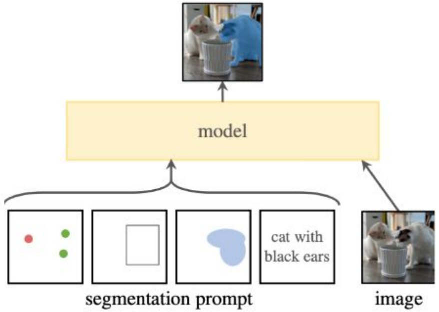
Trained Models

Foun-dation Models

Segment-Anything Model (SAM)

Foundation model from MetaAI

Transform: encoding / decoding



PROMPT

The ChatGPT of the Computer Vision

BIG DATA

Model SA-1B

- Natural photographies
- Huge model (~1GB)
- 11M diverse, high-res. images
- 1.1B segmentation masks
- Open, privacy

Alexander Kirillov et al. IEEE/CVF, 2023, 2700 citations

Rule-based

Model-based

Machine Learning

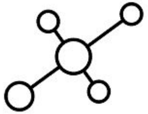
Deep Learning

Trained Models

Foun-dation Models

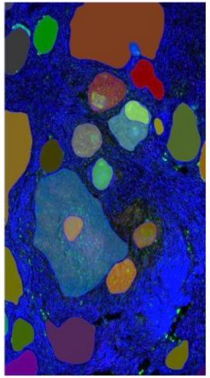
SAM for Science?

Web interface
Python package
QuPath
Napari
Fiji

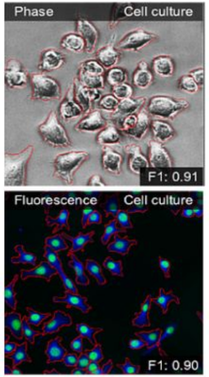


Variants of SAM Models

- MicroSAM
- MedSAM
- CellSAM
- EfficientSAM
- MobileSAM
- ...

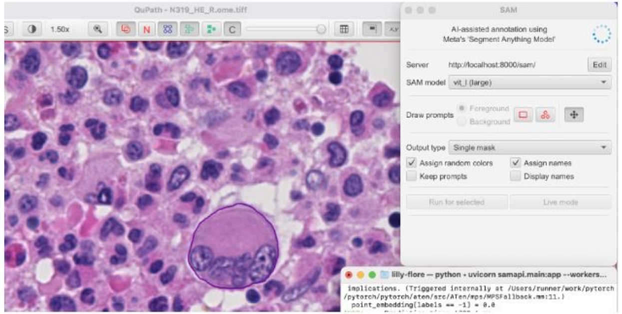


MicroSAM
C. Pape



CellSAM

ARCGIS



Acceleration of annotations

Megakaryocotes on human biopsis

SAM Large model

SAM extension of WuPath

SAM on server

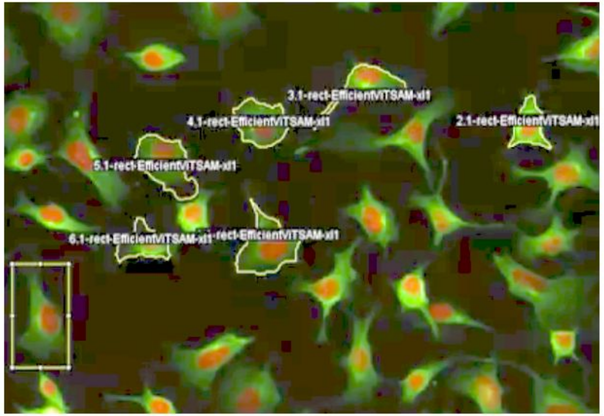
R. Sarkis, CHUV, L-F. Celma, EPF

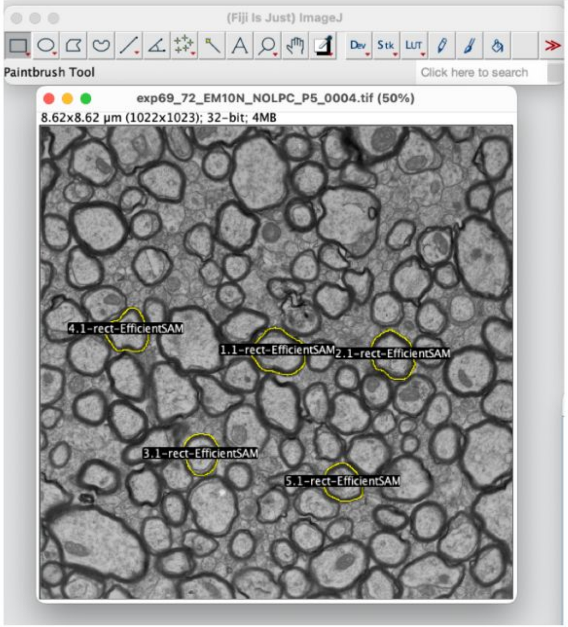
April 2024

SAMJ Annotation with SAM on FIJI (CPU)

SAMJ

- FIJI Plugin and ICY plugin
- Model Efficient SAM (run on CPU)
- Automatic installation of the Python environment
- Smart strategy for tiling





SAMJ Team: Carlos, Caterina, Arrate, Vladimir Ulman, Adrian Ines, Jonathan Heras, Curtis Rueden, Jean-Christophe, Daniel

Segment Anything Model (SAM) and SAM-like models

Segment Anything Model

https://segment-anything.com

by Meta AI

Segment Anything Model

- Promptable Segmentation (bounding box and points)
- Real-time interaction (~50 miliseconds)
- 1 Billion masks, 11 Million images
- Diverse and high-resolution images
- Manual to automatic annotation process
- Vision Transformer-based Architecture (ViT)
- Real-time web browser interaction
- Zero-Shot Capabilities
- Real-world scenarios
- Ethical and fairness focus

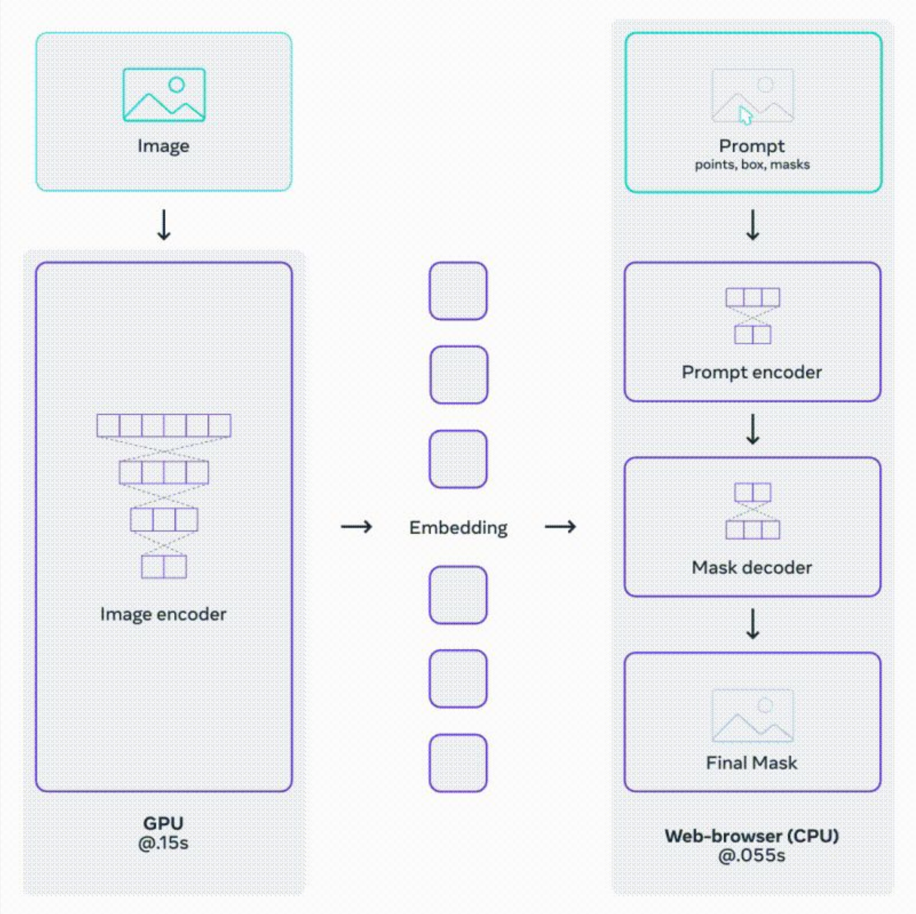
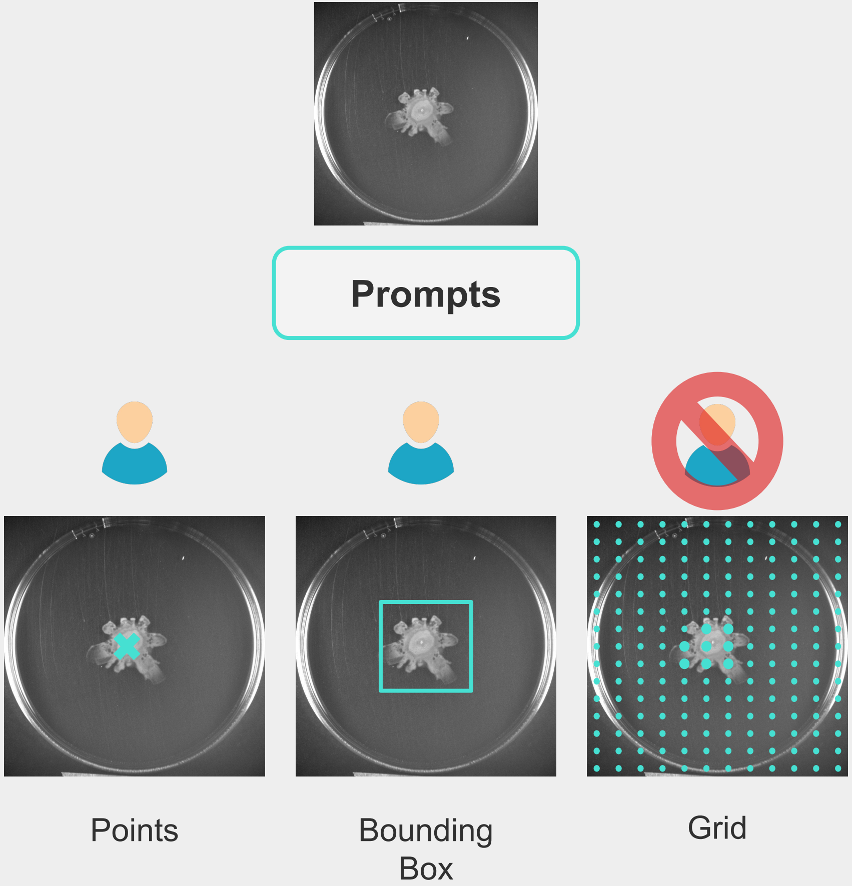
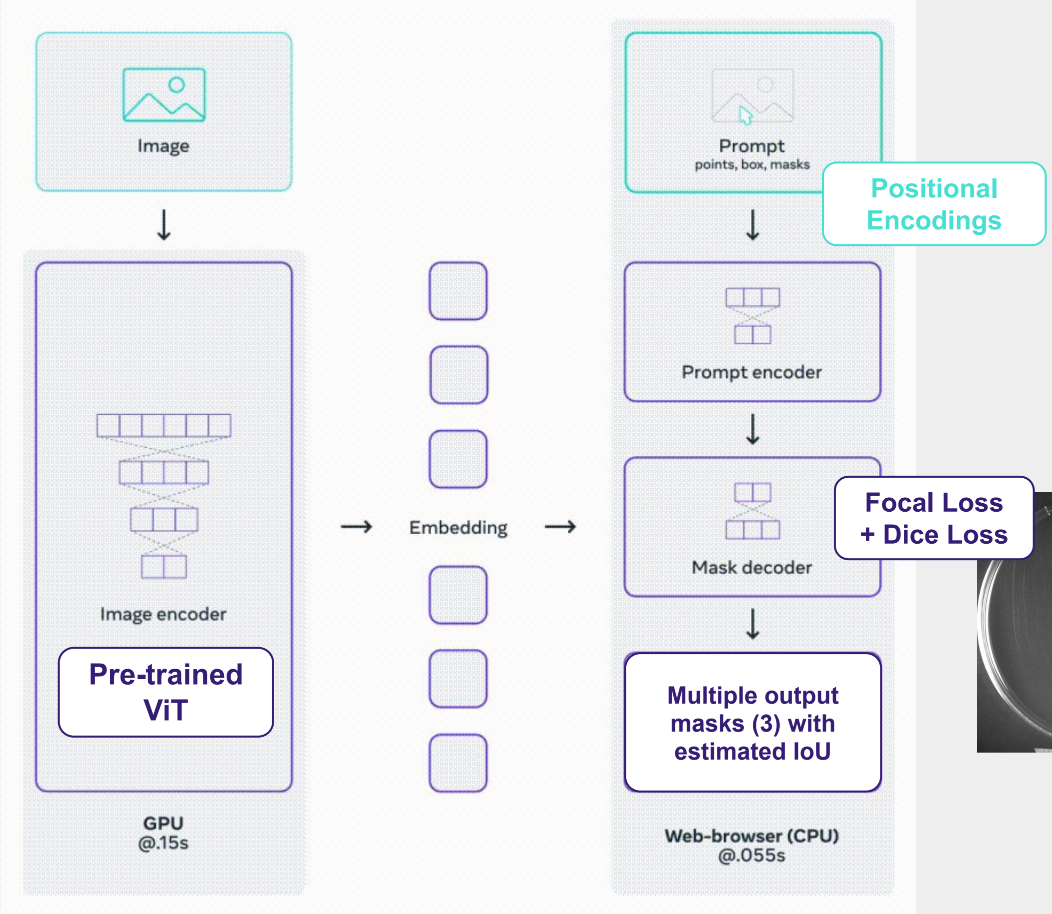
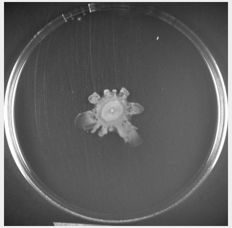


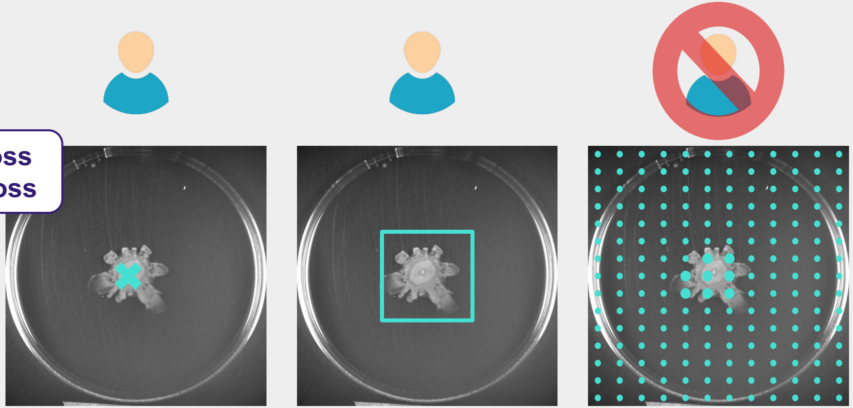
Image
Prompt
points, box, masks
Image encoder
Embedding
Prompt encoder
Mask decoder
Final Mask
GPU
@.15s
Web-browser (CPU)
@.055s


Prompts
Points
Bounding Box
Grid





Prompts



Points

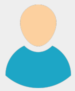
Bounding Box

Grid

Segment Anything Data Engine

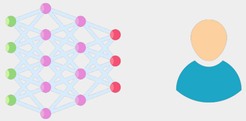
1

Model-assisted **manual annotation** stage



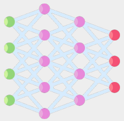
2

Semi-automatic stage with a mix of **automatically predicted masks** and **model-assisted annotation**

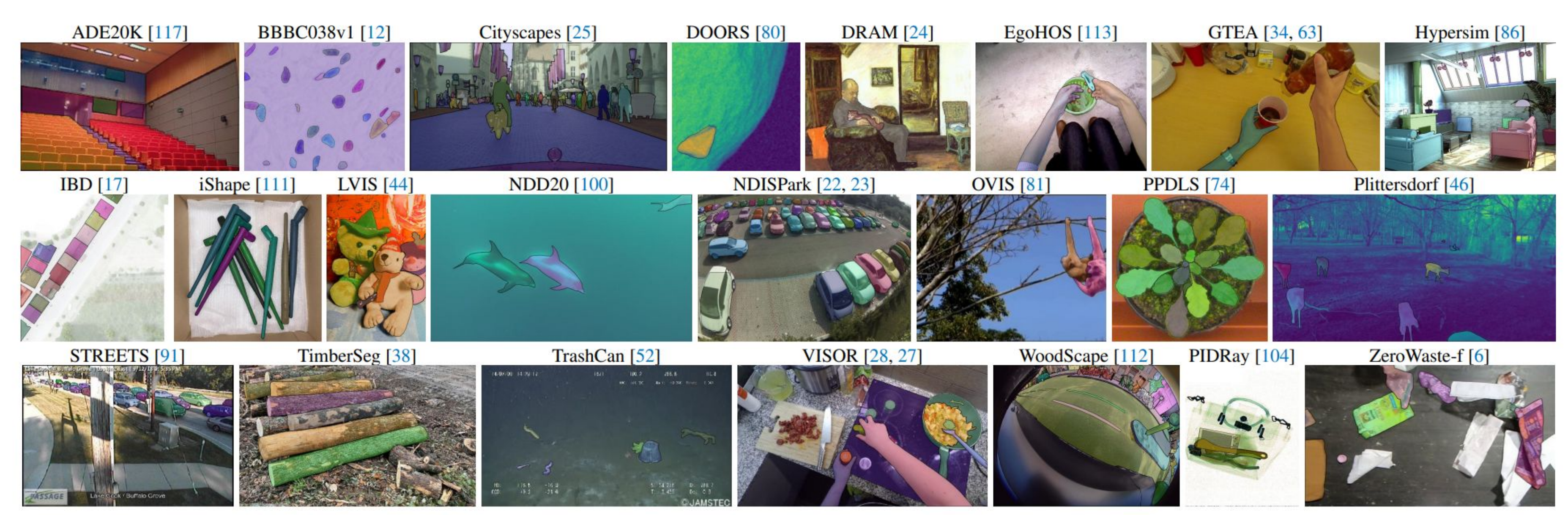


3

Fully automatic stage, model generates masks **without annotator** input

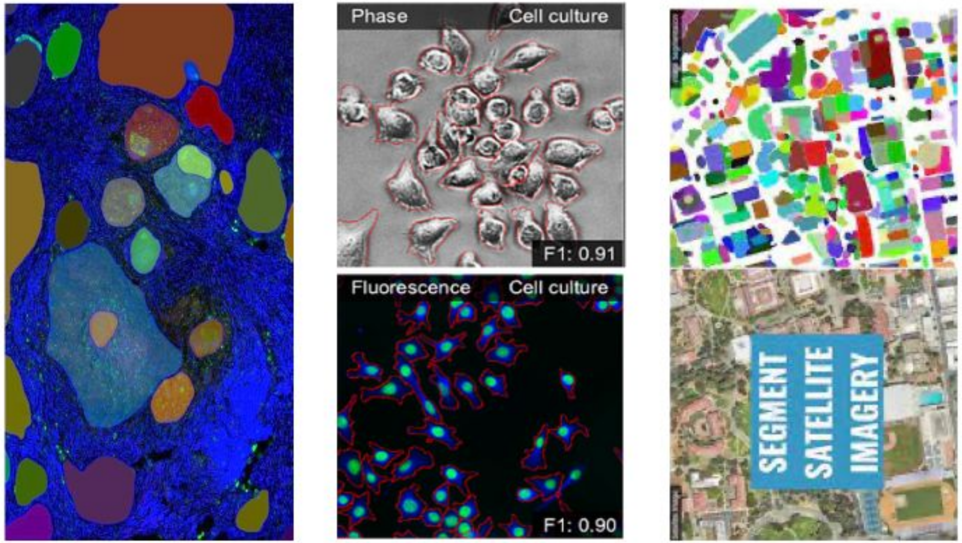


SAM’s Zero-Shot **transfer capabilities on image types**



SAM for Science

- MicroSAM
- CellSAM
- MedSAM
- …



MicroSAM
C. Pape

CellSAM

ARCGIS

EfficientSAM

https://yformer.github.io/efficient-sam/

by Y. Xiong et al.

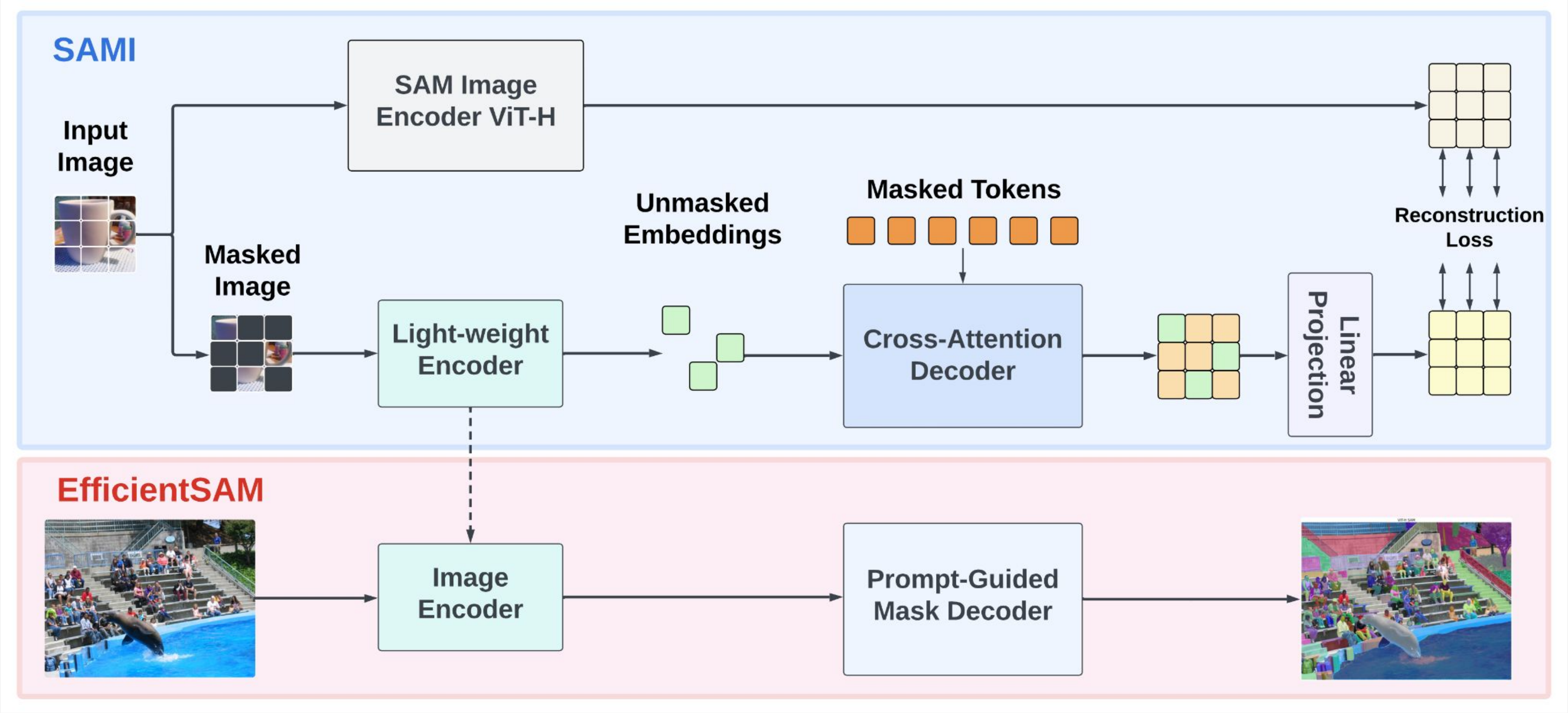
EfficientSAM

Develop SAMI, a **masked image pretrained framework** to reconstruct features from SAM ViT-H image encoder

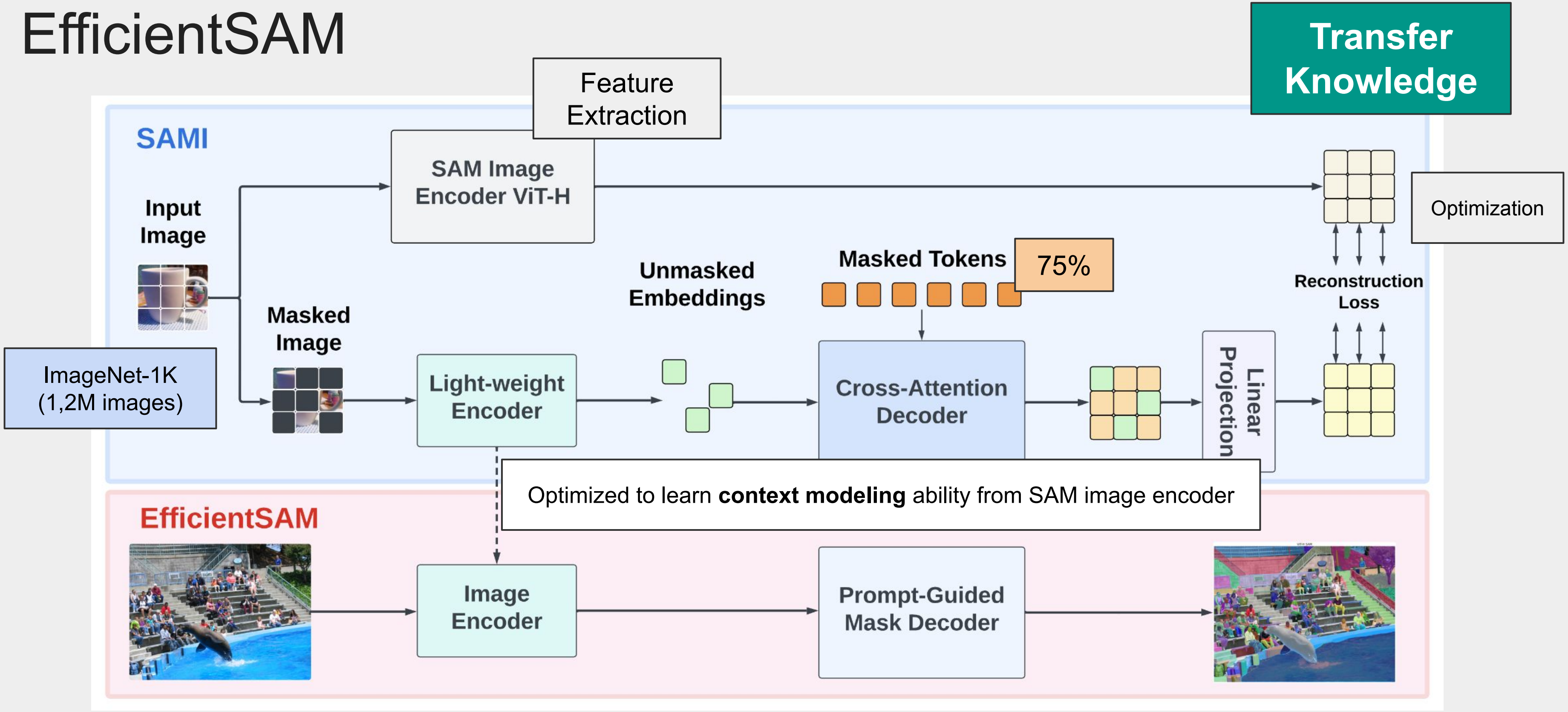
SAMI-pretrained backbone generalize to **many tasks** including classification

Deliver EfficientSAM, a **light-weight SAM** model for practical deployment

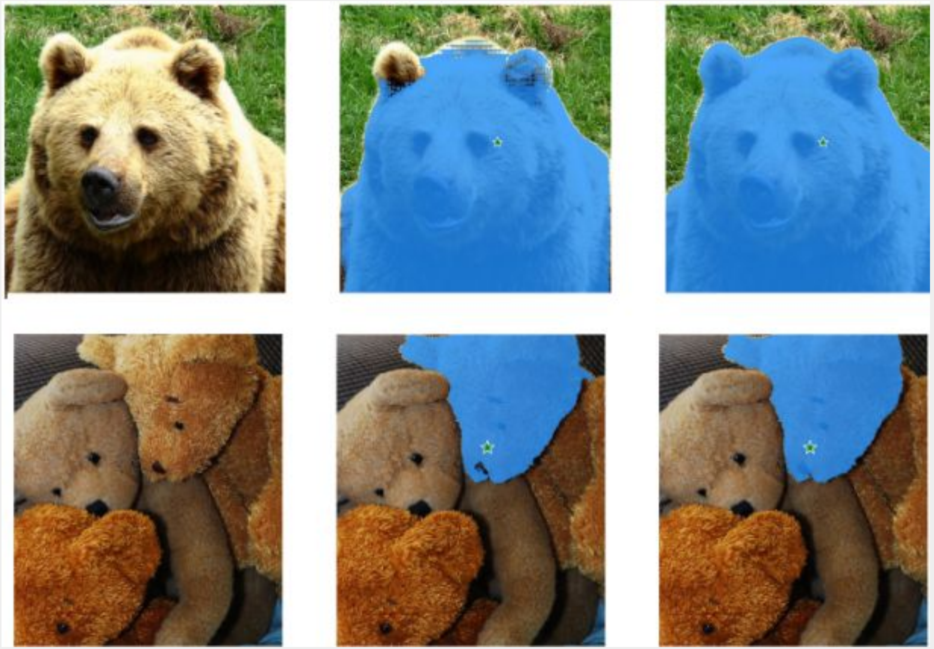
EfficientSAM



EfficientSAM

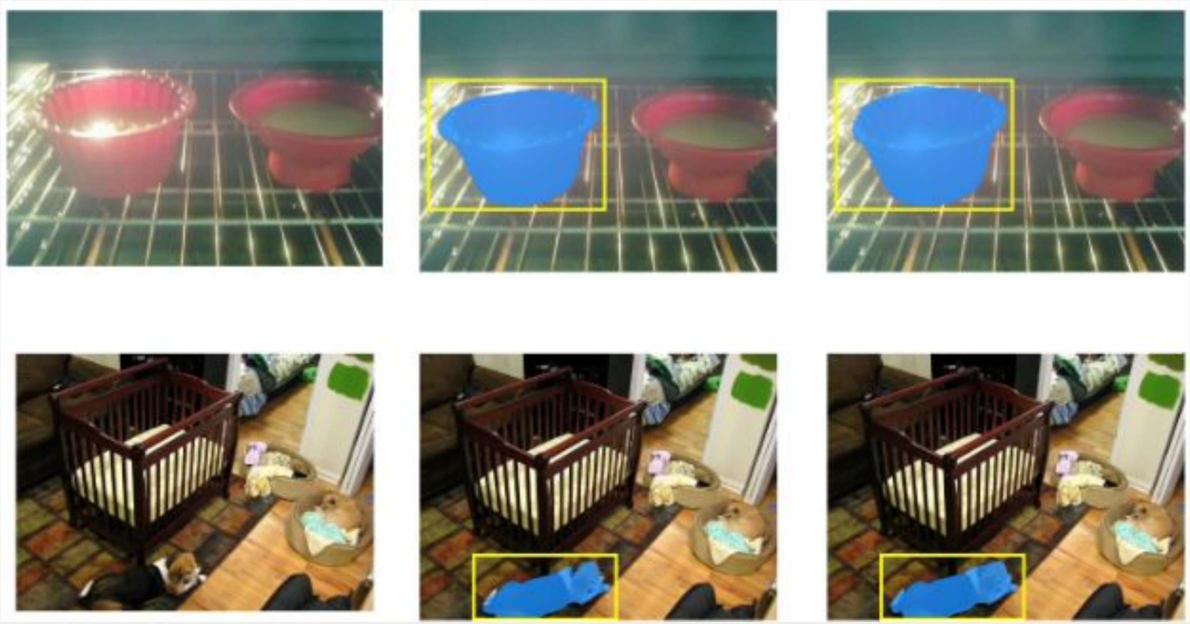


EfficientSAM: points



Input (left), SAM (middle), EfficientSAM (right)

EfficientSAM: ROIs



Input (left), SAM (middle), EfficientSAM (right)

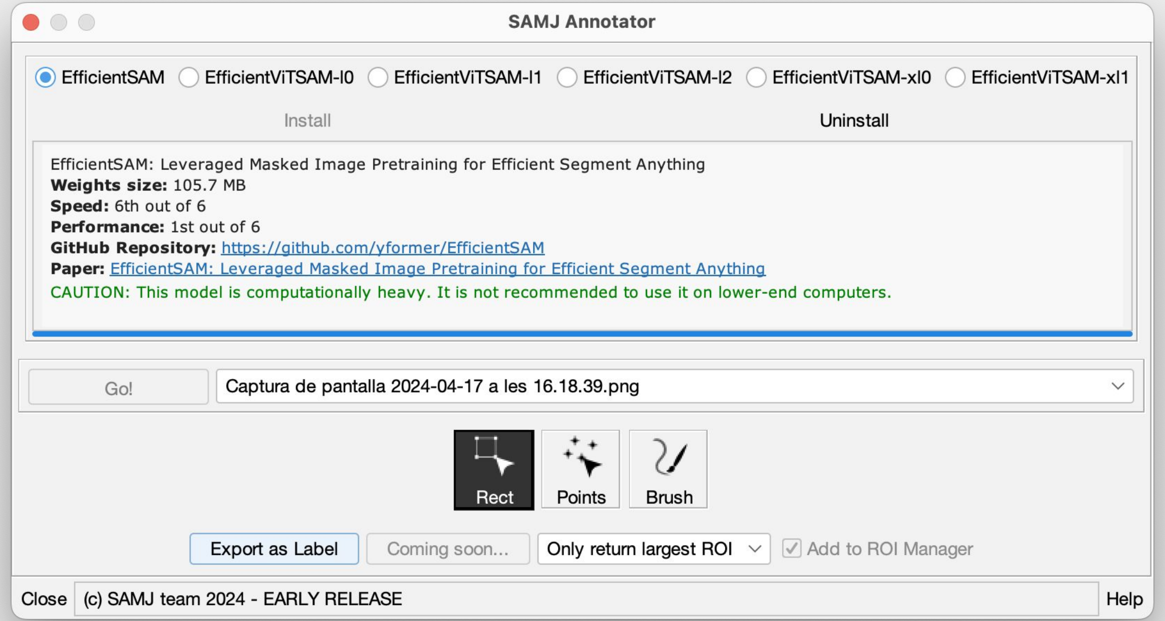
EfficientSAM: Everything

Input (left), SAM (middle), EfficientSAM (right)

SAMJ

https://github.com/segment-anything-models-java/SAMJ-IJ

(Fiji Is Just) ImageJ
Dev
Stk
LUT
(Fiji Is Just) ImageJ 2.14.0/1.54f; Java 1.8.0_322 [64-bit];
Click here to search



SAMJ Annotator
EfficientSAM
EfficientViTSAM-l0
EfficientViTSAM-l1
EfficientViTSAM-l2
EfficientViTSAM-xl0
EfficientViTSAM-xl1
Install
Uninstall
EfficientSAM: Leveraged Masked Image Pretraining for Efficient Segment Anything
Weights size: 105.7 MB
Speed: 6th out of 6
Performance: 1st out of 6
GitHub Repository: https://github.com/yformer/EfficientSAM
Paper: EfficientSAM: Leveraged Masked Image Pretraining for Efficient Segment Anything
CAUTION: This model is computationally heavy. It is not recommended to use it on lower-end computers.
Go!
Captura de pantalla 2024-04-17 a les 16.18.39.png
Rect
Points
Brush
Export as Label
Coming soon...
Only return largest ROI
Add to ROI Manager
Close
(c) SAMJ team 2024 - EARLY RELEASE
Help

SAMJ Functionalities

SAM-like models implementation

Different models based on SAM (e.g. EfficientSAM) available for your annotations

Semi-automatic annotation

Annotation of objects through **different prompts** (bounding box, points, etc) in seconds

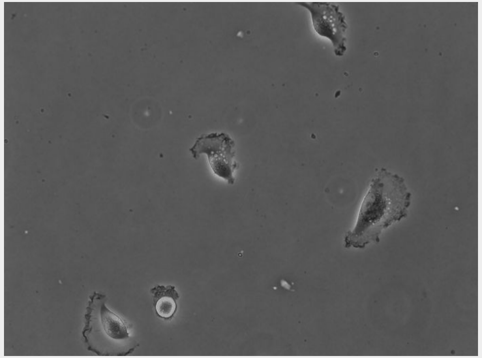
Adaptable to different use cases

Capable of performing annotations over **different images**, cell types, morphologies, etc

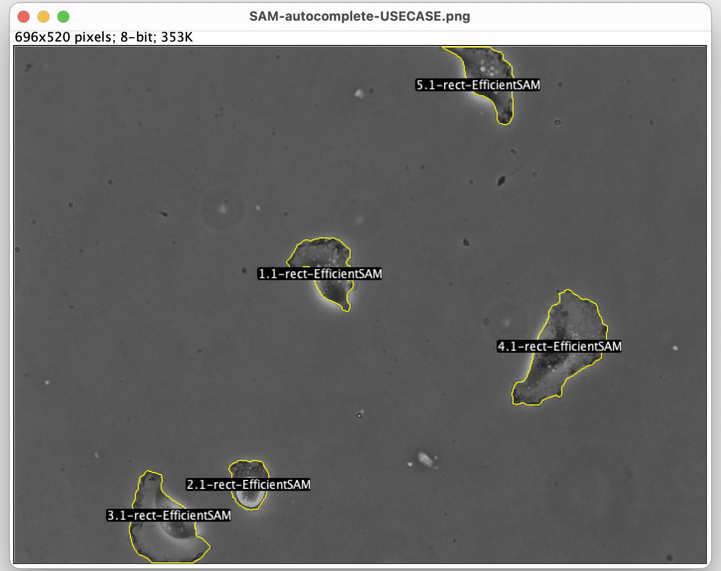
No need of GPU

Using different **SAM-like models with a CPU** as using lighter versions

SAMJ usage example



Original Image*

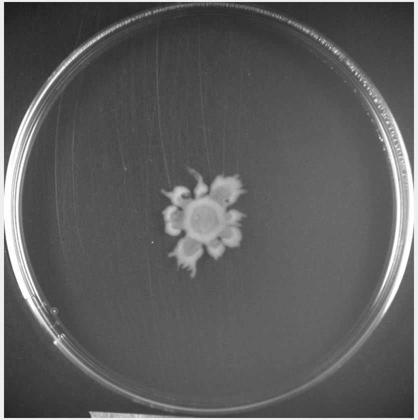


SAMJ Annotations

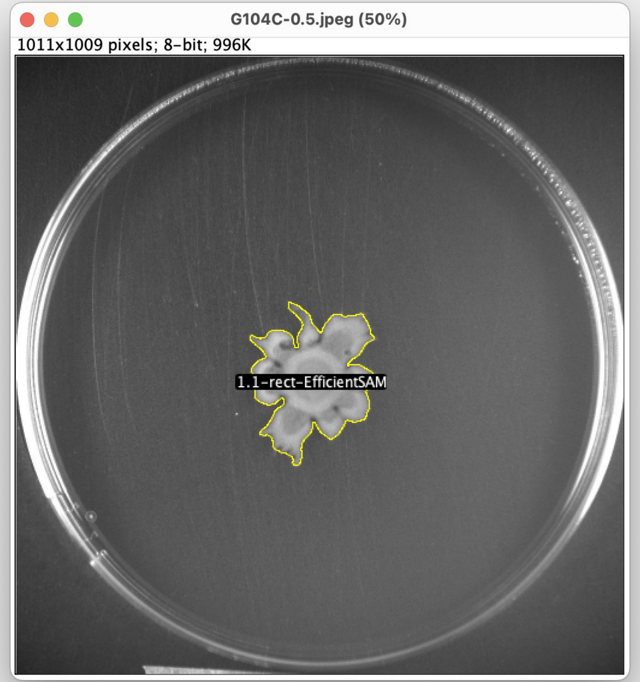
Generated Mask

*Original Image obtained from the Cell Tracking Challenge

SAMJ usage example



Original Image*



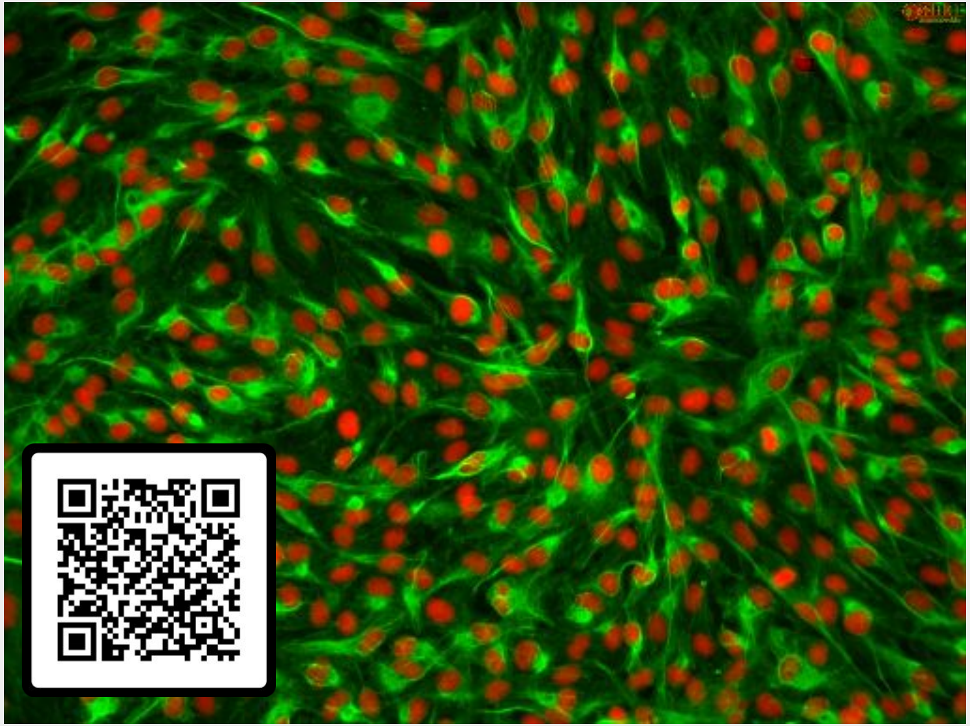
SAMJ Annotations

Generated Mask

*Original Image obtained from doi.org/10.1016/j.compbiomed.2021.104673

Hands on activities

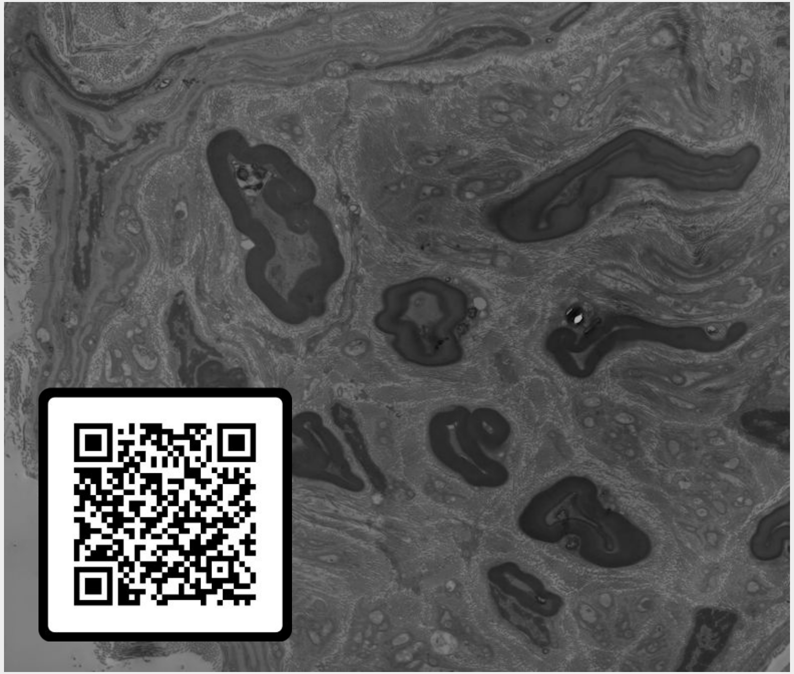
Counting nuclei!



1. Download the image from the GitHub repository.
2. Open Fiji and SAMJ plugin.
3. Open the image and encode it with the SAMJ plugin.
4. Start annotating nuclei for 20 seconds.

Who can annotate more?

Annotation of myelin sheaths on Fiji!



1. Segment by using few preprocessing set and then threshold
2. Annotate by hand (mouse).
3. Annotate using the magic wand of Fiji (select the tolerance) and then interpolate the selection (menu Edit>Selection>Interpolate)
4. Annotate using SAMJ
5. Comment these 4 methods in term of speed of annotation, accuracy of segmentation, and required resources.

Data: Marta Di Fabrizio, Dubochet Imaging Center EPFL and Daniel Sage, Center for Imaging, EPFL

Bibliography

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4015-4026).

Xiong, Y., Varadarajan, B., Wu, L., Xiang, X., Xiao, F., Zhu, C., ... & Chandra, V. (2023). Efficientsam: Leveraged masked image pretraining for efficient segment anything. arXiv preprint arXiv:2312.00863.