DLinCV Lecture 2.2

Image Segmentation: Auto-ML and Foundation models

08.06.2023

Aasa Feragen
afhar@dtu.dk

This lecture's learning goals

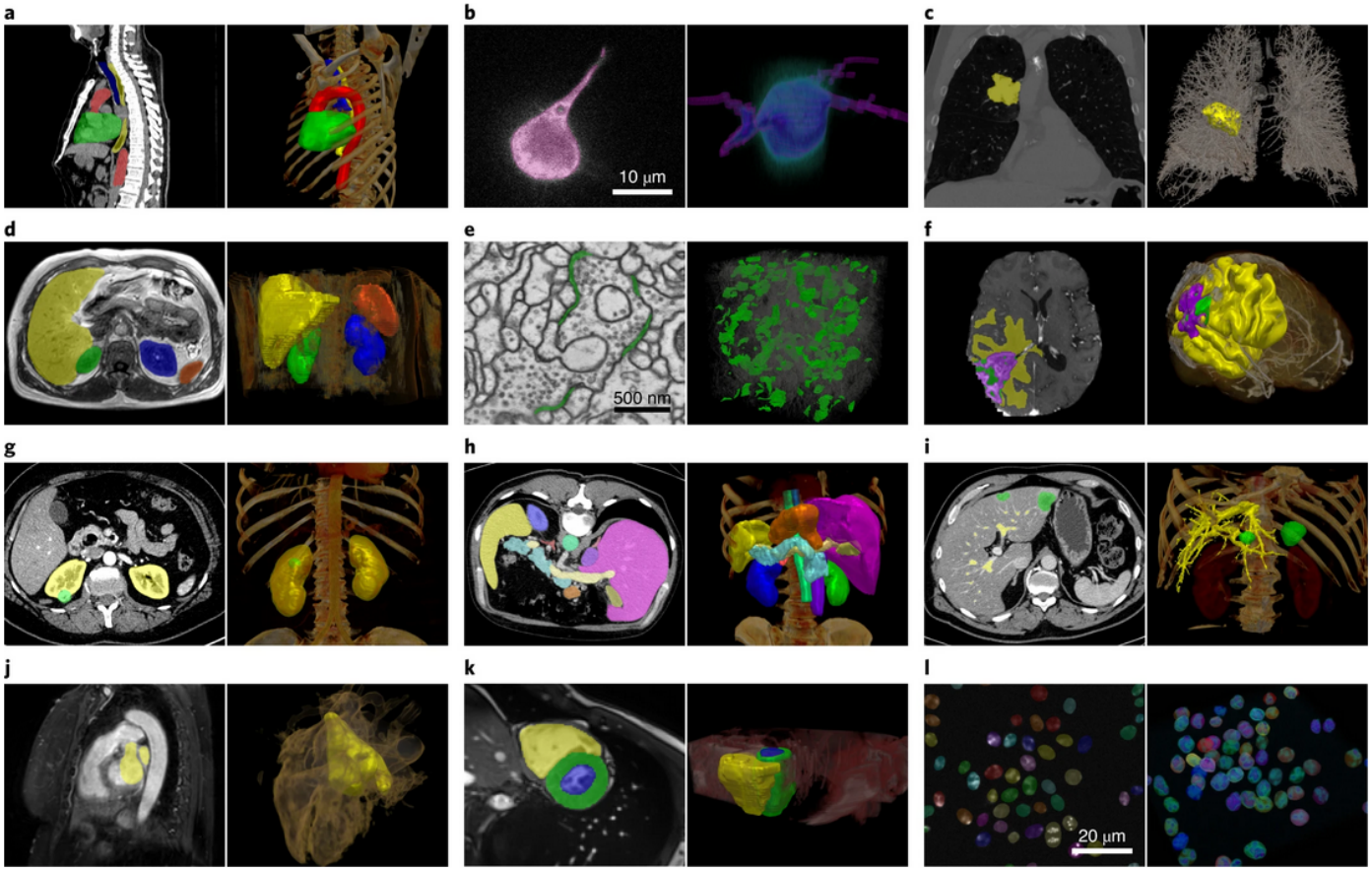
After this lecture you should

- Be familiar with the overall modelling goal and design of the nnU-net: An automatic model configuration
- Be familiar with the overall goal and design of the Segment Anything Model (SAM)

Auto-ML for segmentation: The nnU-net [1]

Fig. 1: nnU-Net handles a broad variety of datasets and target image properties.

From: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation



[1] Isensee, Fabian, et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation." Nature methods 18.2 (2021): 203-211.

Auto-ML for segmentation: The nnU-net

Based on a range of different biomedical image segmentation datasets, define a recipe for model configuration:

1. Collect design decisions that do not require adaptation between datasets and identify a robust common configuration ('fixed parameters').

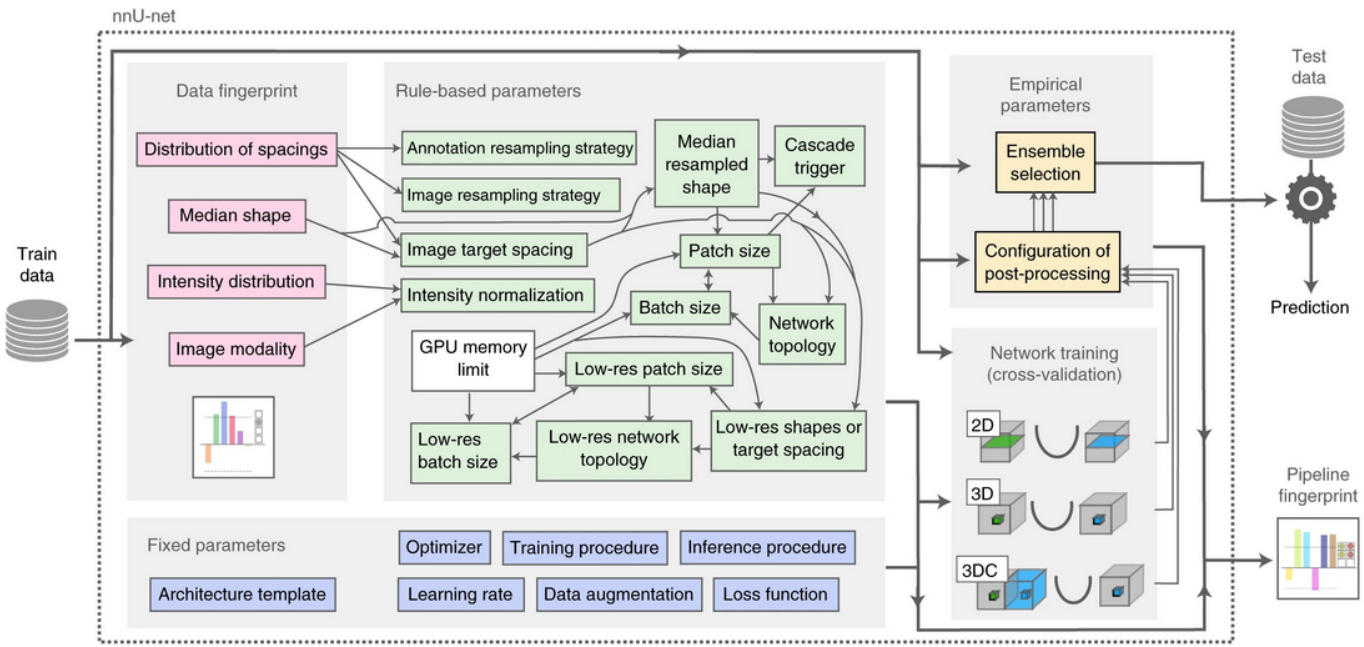
2. For as many of the remaining decisions as possible, formulate explicit dependencies between specific dataset properties ('dataset fingerprint') and design choices ('pipeline fingerprint') in the form of heuristic rules to allow for almost-instant adaptation on application ('rule-based parameters').

3. Learn only the remaining decisions empirically from the data ('empirical parameters').

Auto-ML for segmentation: The nnU-net

Fig. 2: Proposed automated method configuration for deep learning-based biomedical image segmentation.

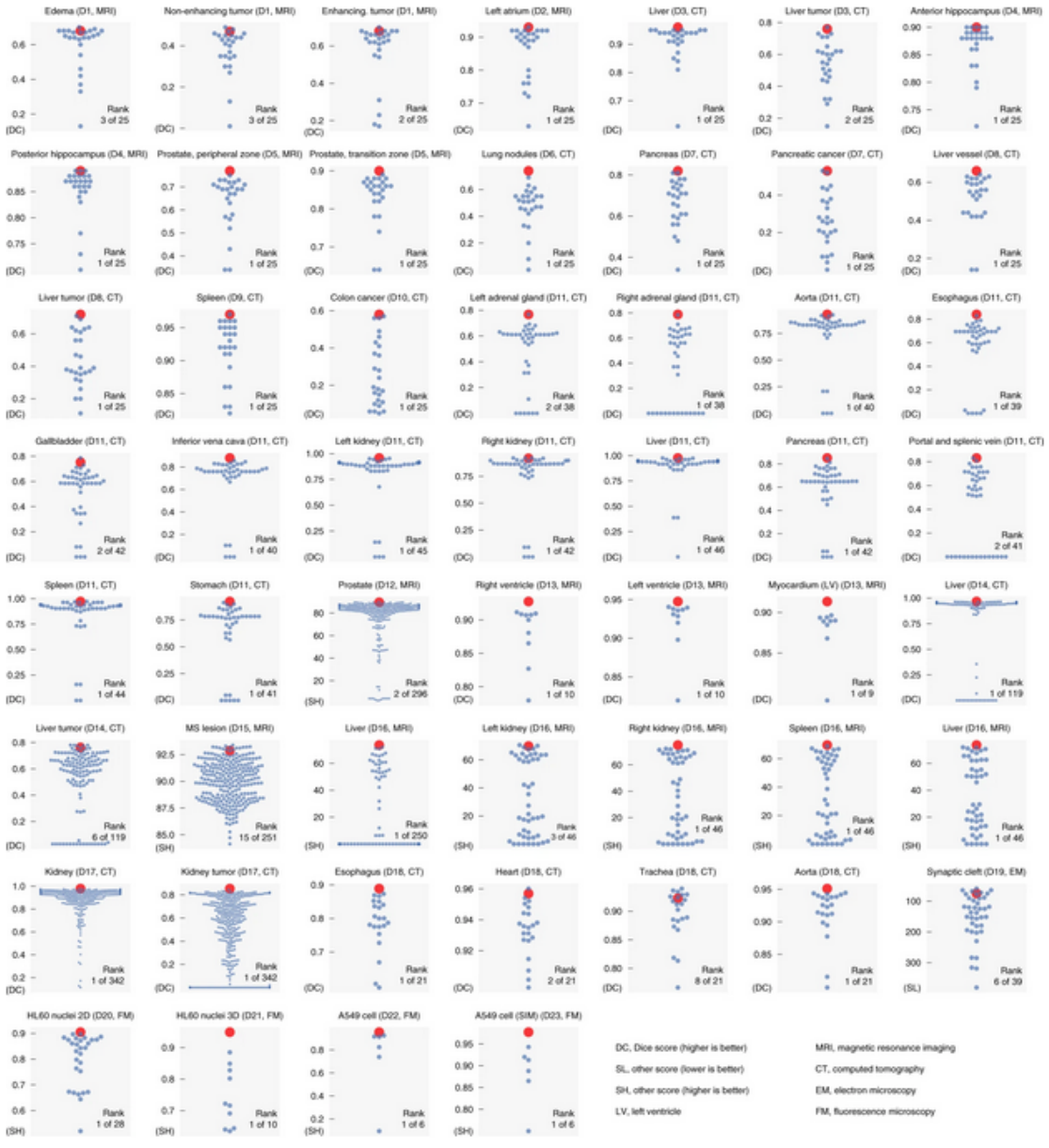
From: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation



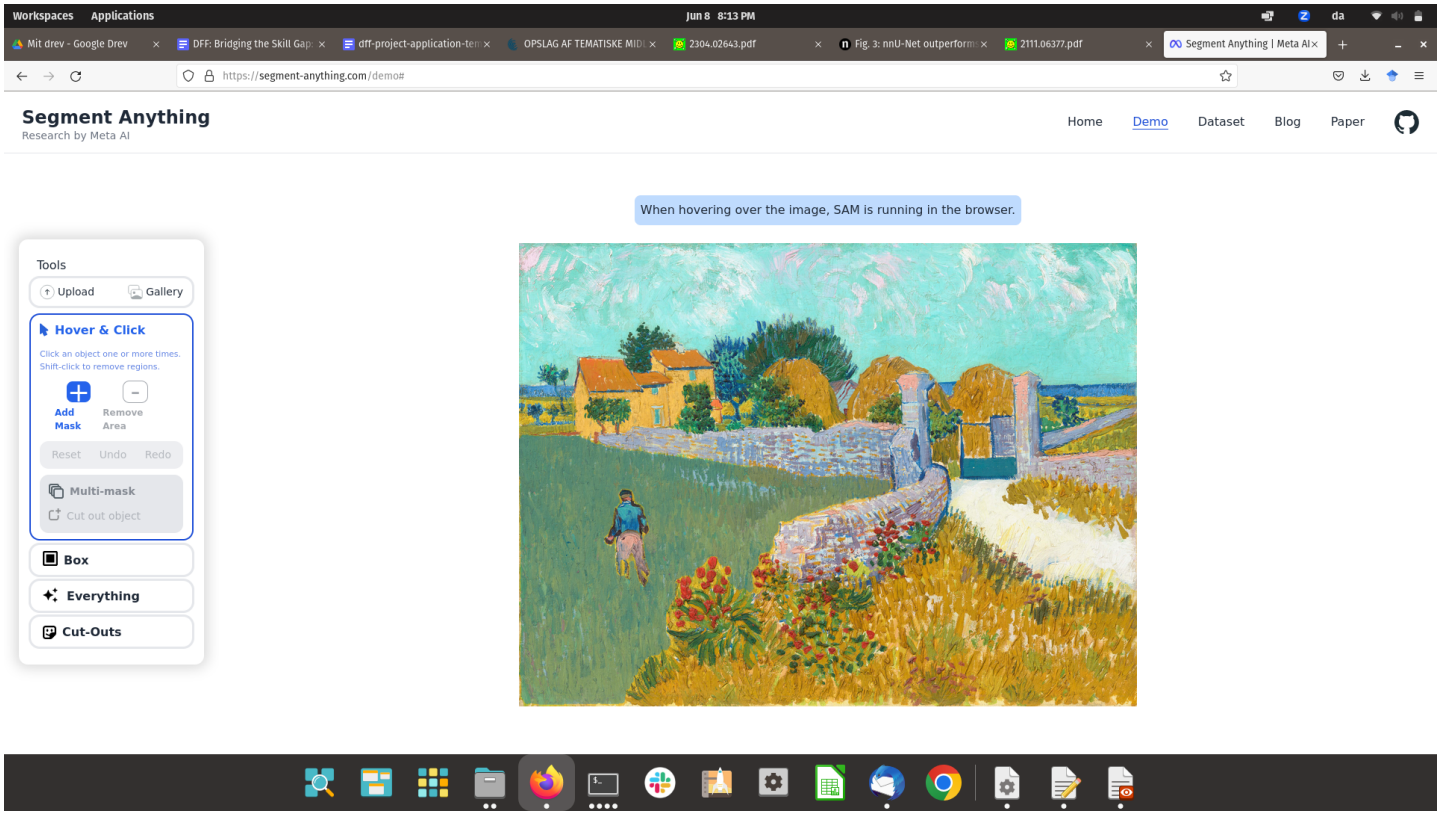
Auto-ML for segmentation: The nnU-net

Fig. 3: nnU-Net outperforms most specialized deep learning pipelines.

From: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation



Foundation models: The Segment Anything Model[2]

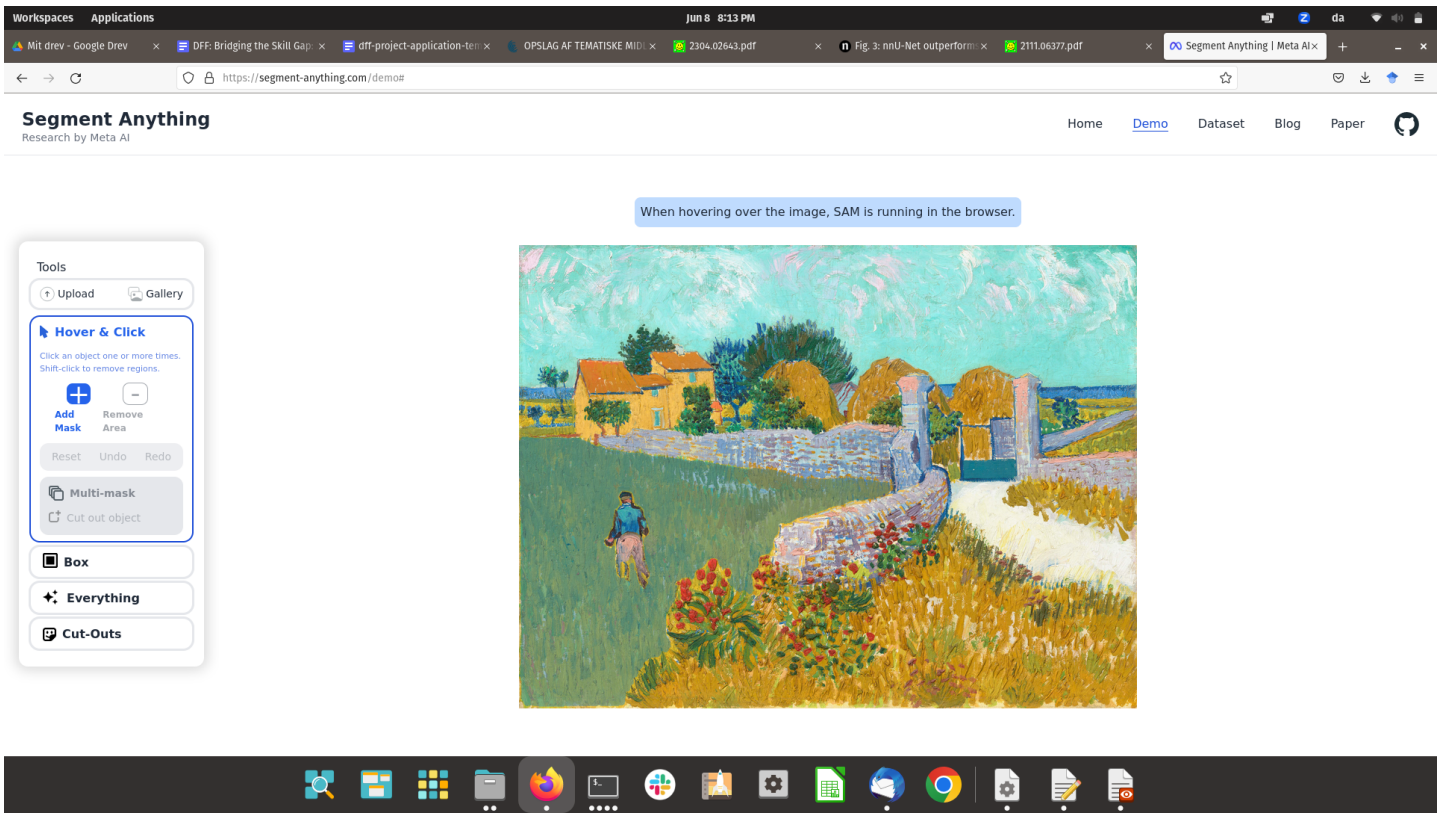


[2]Kirillov, Alexander, et al. "Segment anything." arXiv preprint arXiv:2304.02643 (2023).

Foundation models: The Segment Anything Model

SAM builds on principles known from large generative language models:

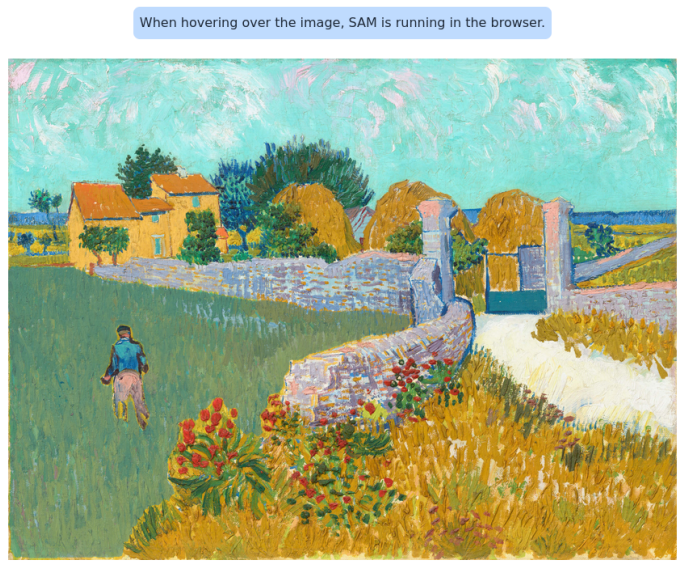
- Self-supervised pretraining for an object completion task
- Prompt-based inference
- Realtime interaction with a user, and using human interaction to refine models



Foundation models: The Segment Anything Model

SAM builds on principles known from large generative language models:

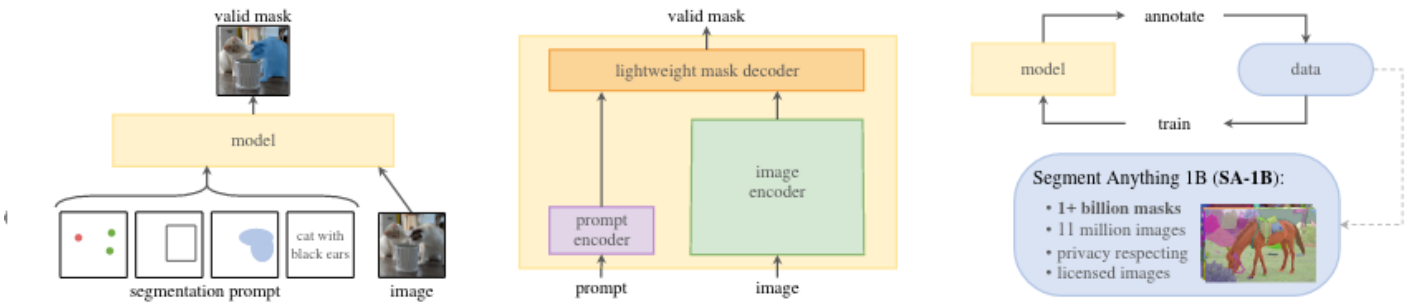
- Self-supervised pretraining for an object completion task
- Prompt-based inference
- Realtime interaction with a user, and using human interaction to refine models



What constitutes a segmentation foundation model?

SAM modelling choices:

- Redefining prompting for segmentation
- Model design for amortized cost for realtime human interaction
- The dataset



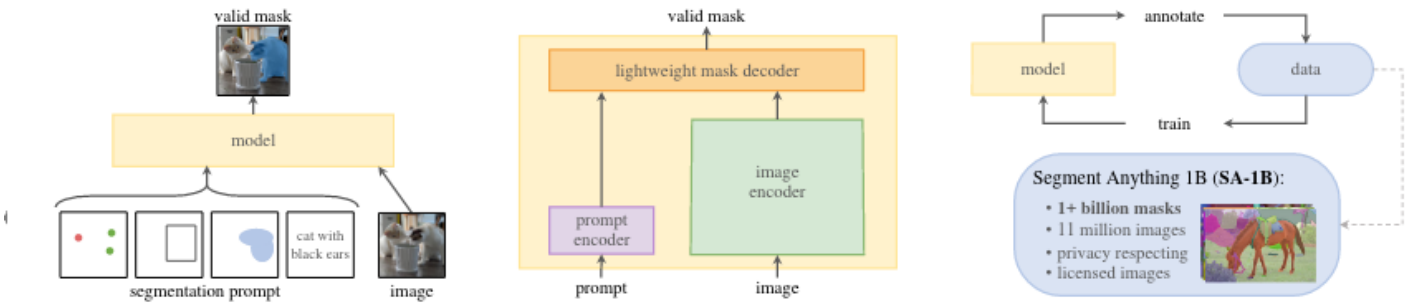
(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

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.

What constitutes a segmentation foundation model?

SAM modelling choices:

- Redefining prompting for segmentation
- Model design for amortized cost for realtime human interaction
- The dataset



(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.

The Segment Anything task

- Promptable segmentation task
- Handle ambiguous prompts – return multiple masks
- Pretrain with simulated prompts
- Enables creative prompt engineering –
  e.g. segmentation from bounding box detection

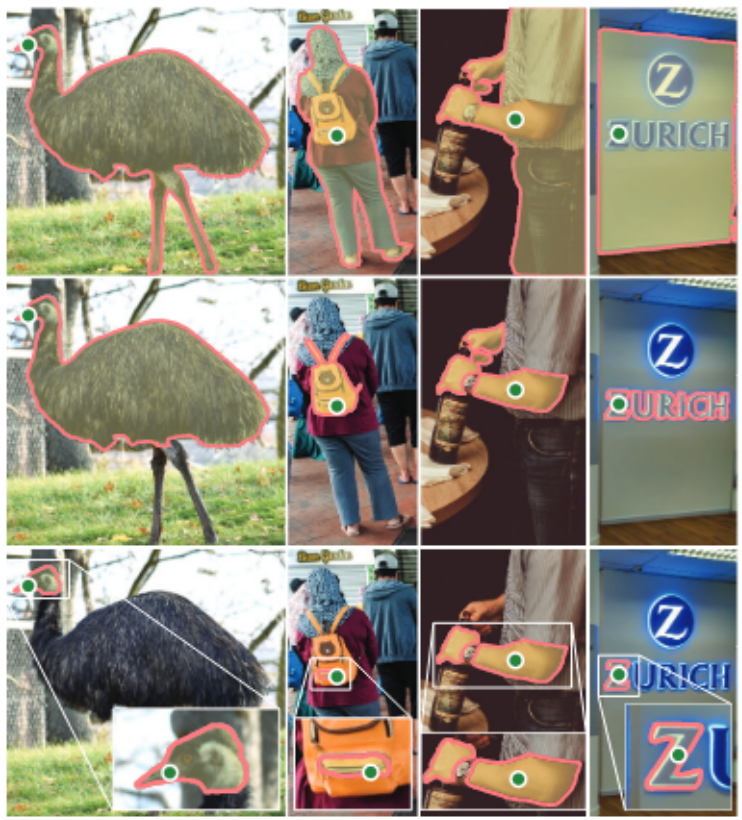


Figure 3: Each column shows 3 valid masks generated by SAM from a single ambiguous point prompt (green circle).

The Segment Anything model

- Model designed optimizing amortized cost – initial image loading and processing is heavy, the rest is lightweight

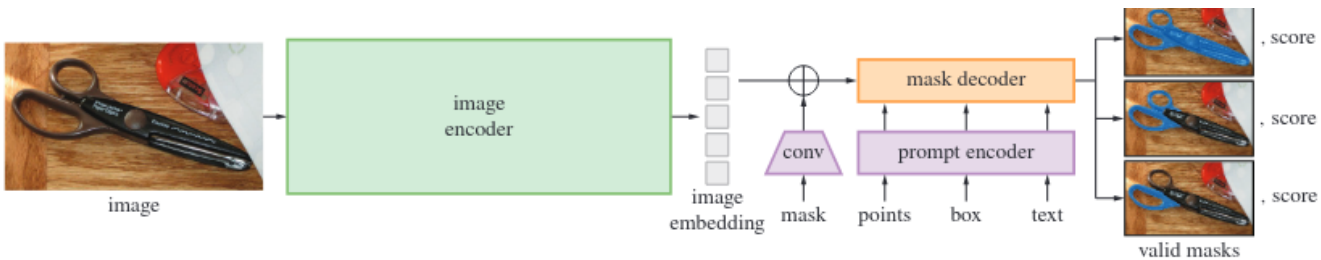
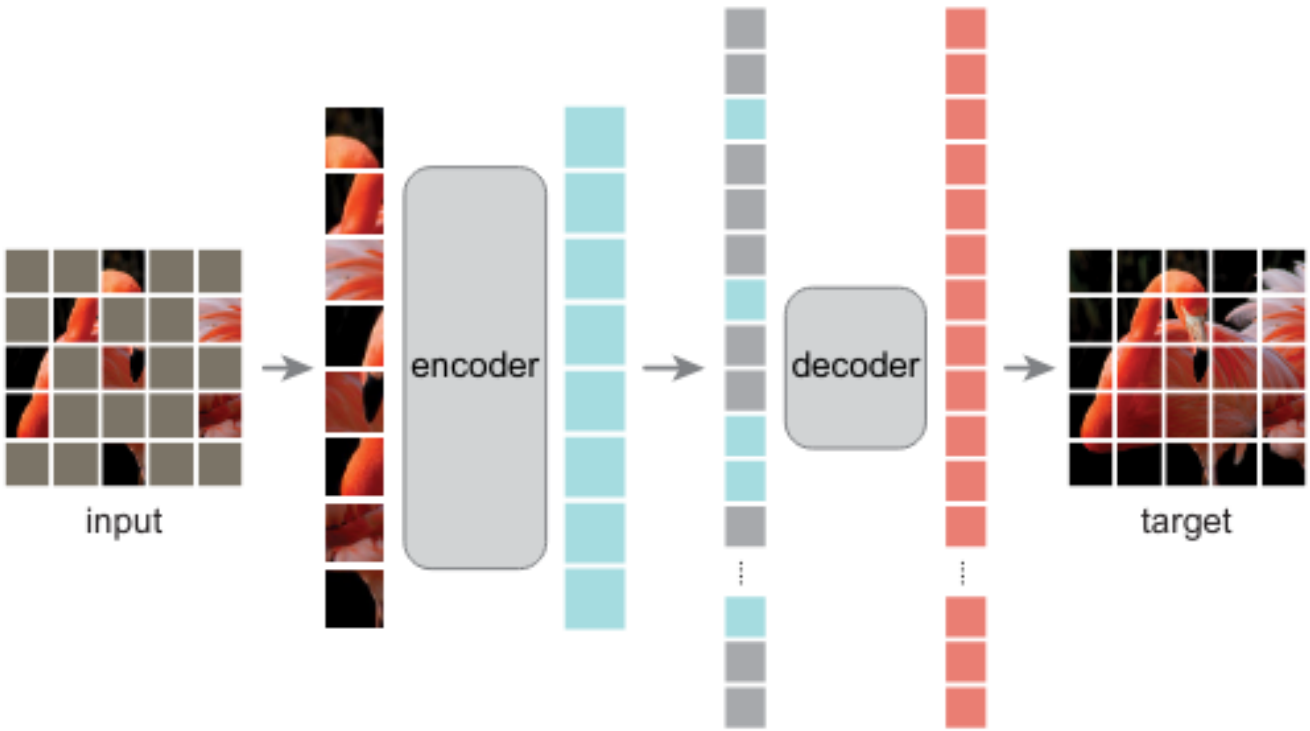


Figure 4: Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks at amortized real-time speed. For ambiguous prompts corresponding to more than one object, SAM can output multiple valid masks and associated confidence scores.

The Segment Anything model

- Pretrained for image completion (a masked autoencoder[3])

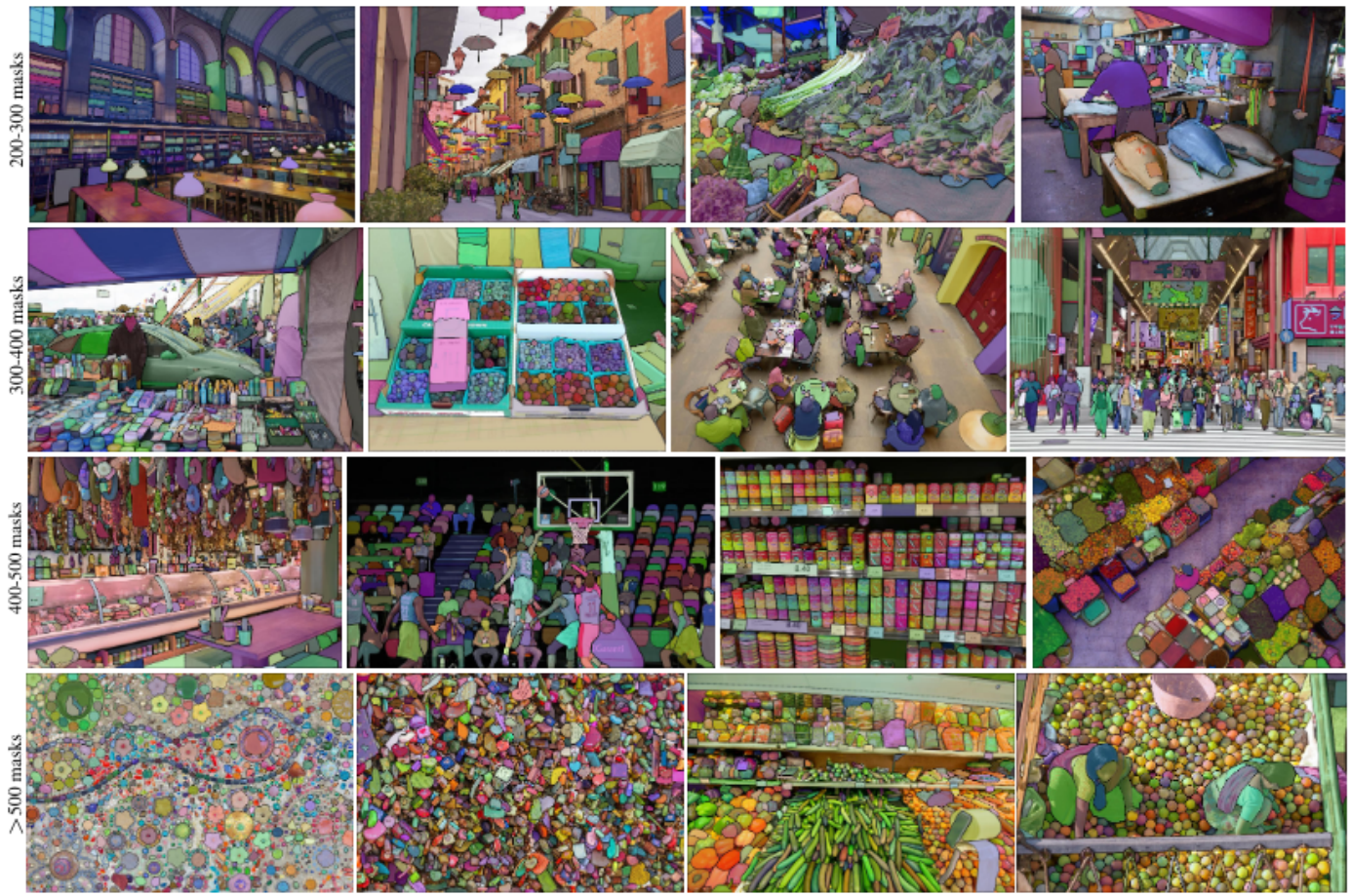


[3] He, Kaiming, et al. "Masked autoencoders are scalable vision learners." CVPR 2022.

The Segment Anything data engine

- Stage 1: Assisted manual stage – annotators refining model generated segmentations. First trained with public data – model improved over time using only newly collected data. Annotation time and quality both improved with the model.
- Stage 2: Semi-automatic stage – making more diverse masks by annotators refining confident masks
- Stage 3: Fully automatic stage – automatic refinement of existing masks to improve mask quality

The Segment Anything dataset (SA-1B)



The Segment Anything dataset (SA-1B)

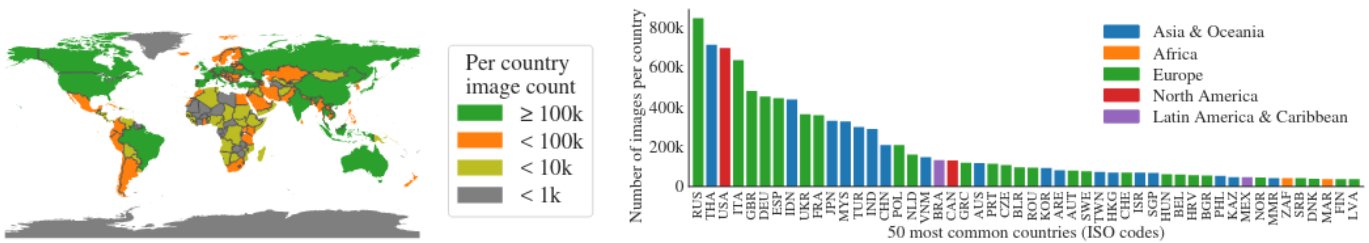


Figure 7: Estimated geographic distribution of SA-1B images. Most of the world's countries have more than 1000 images in SA-1B, and the three countries with the most images are from different parts of the world.

SAM performance

Zero-shot transfer

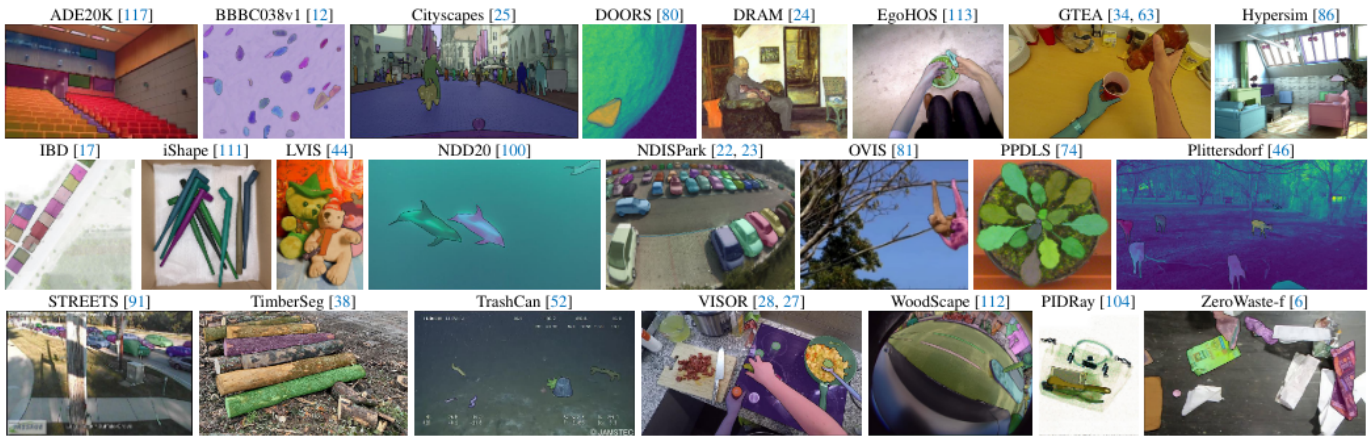


Figure 8: Samples from the 23 diverse segmentation datasets used to evaluate SAM's zero-shot transfer capabilities.

What's next? Text prompts?

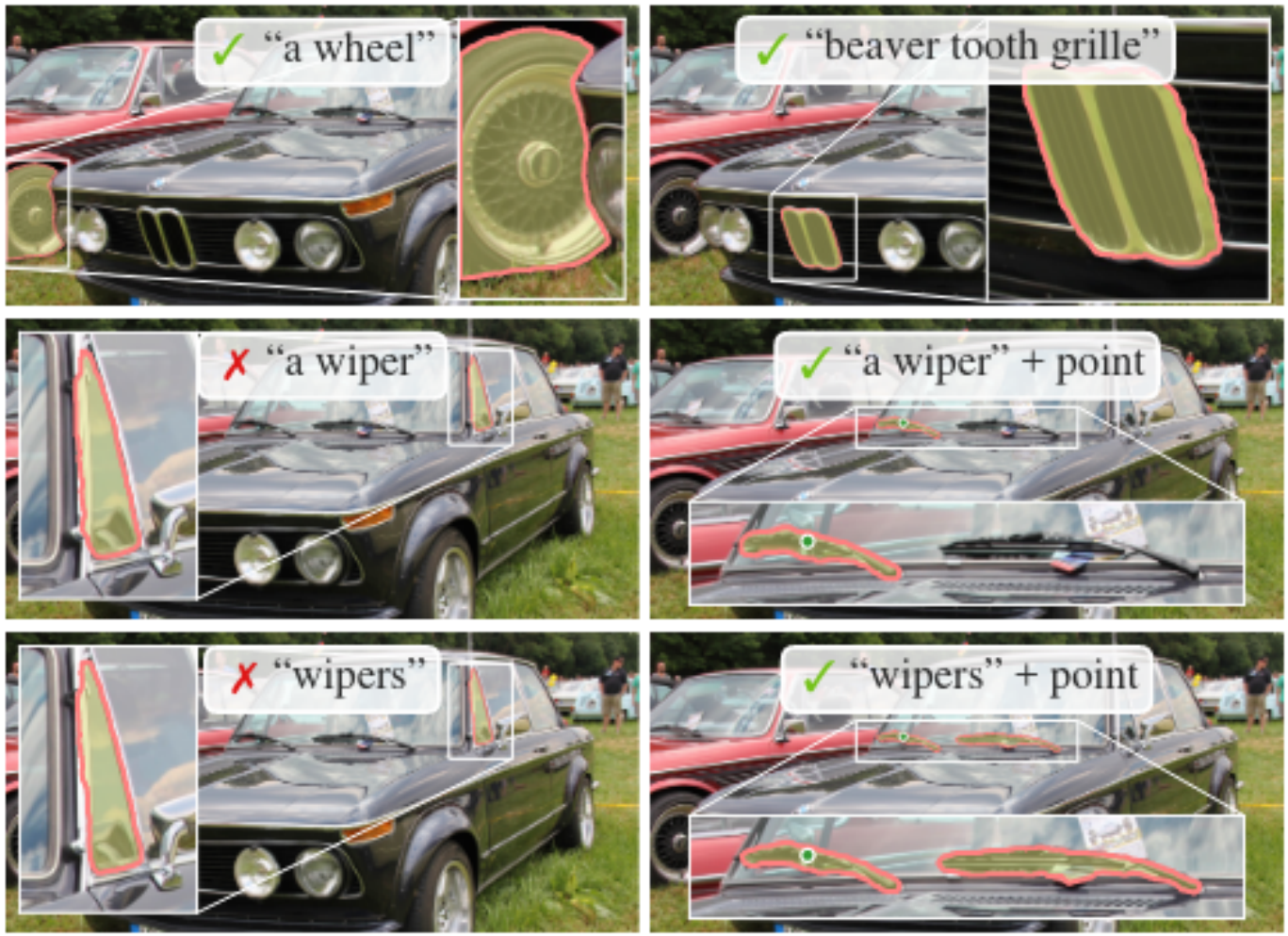


Figure 12: Zero-shot text-to-mask. SAM can work with simple and nuanced text prompts. When SAM fails to make a correct prediction, an additional point prompt can help.

Summary

By now you should

- Be familiar with the overall modelling goal and design of the nnU-net: An automatic model configuration
- Be familiar with the overall goal and design of the Segment Anything Model (SAM)

Some of the concepts that we touched on here, will come back next week:

- The fully automatic stage of the data engine used concepts from object detection
- You will visit joint text+image modelling in the GAN project
- ...and now, you will be playing around with the Segment Anything model in Project 2.