Segmentation

Philipp Krähenbühl, UT Austin

Task: Image segmentation

- Group pixels
  - Same object: Instance segmentation
  - Same “stuff”: Semantic segmentation
  - Same part: Part segmentation

Semantic segmentation

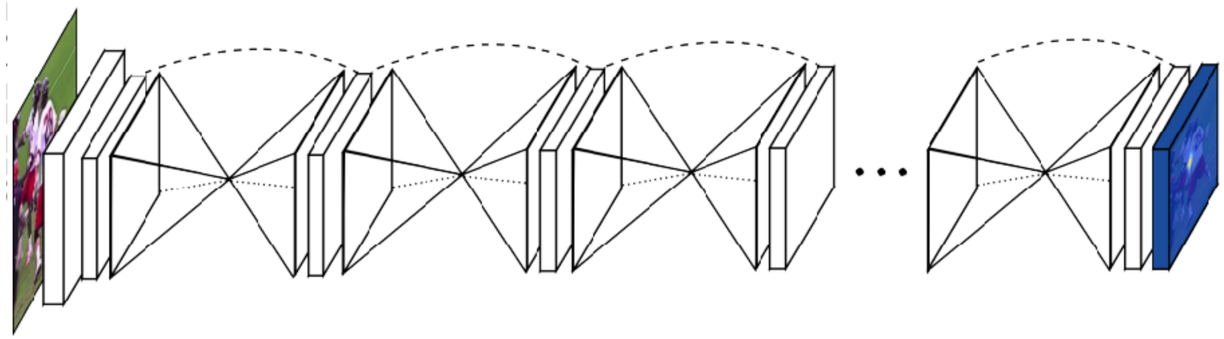
- Group pixels according to their semantic class
  - Does not distinguish identities
  - Segments objects and stuff
  - Easiest to train and evaluate

Fully convolutional networks

- Semantic segmentation by classifying each pixel
  - Fully convolutionally
  - Cross entropy loss at every pixel

Fully Convolutional Networks for Semantic Segmentation, Shelhamer et al., 2015

UNet, Hourglass



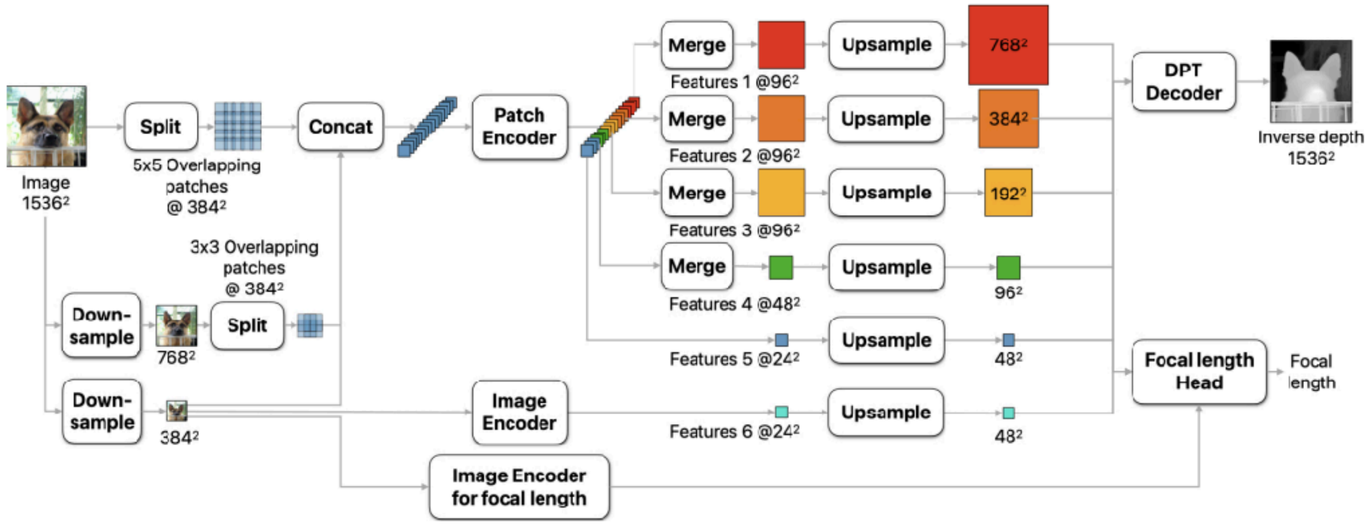
- Semantic segmentation by classifying each pixel
  - Down and up-sampling architectures
- More accurate
  - Higher output resolution
  - Still efficient

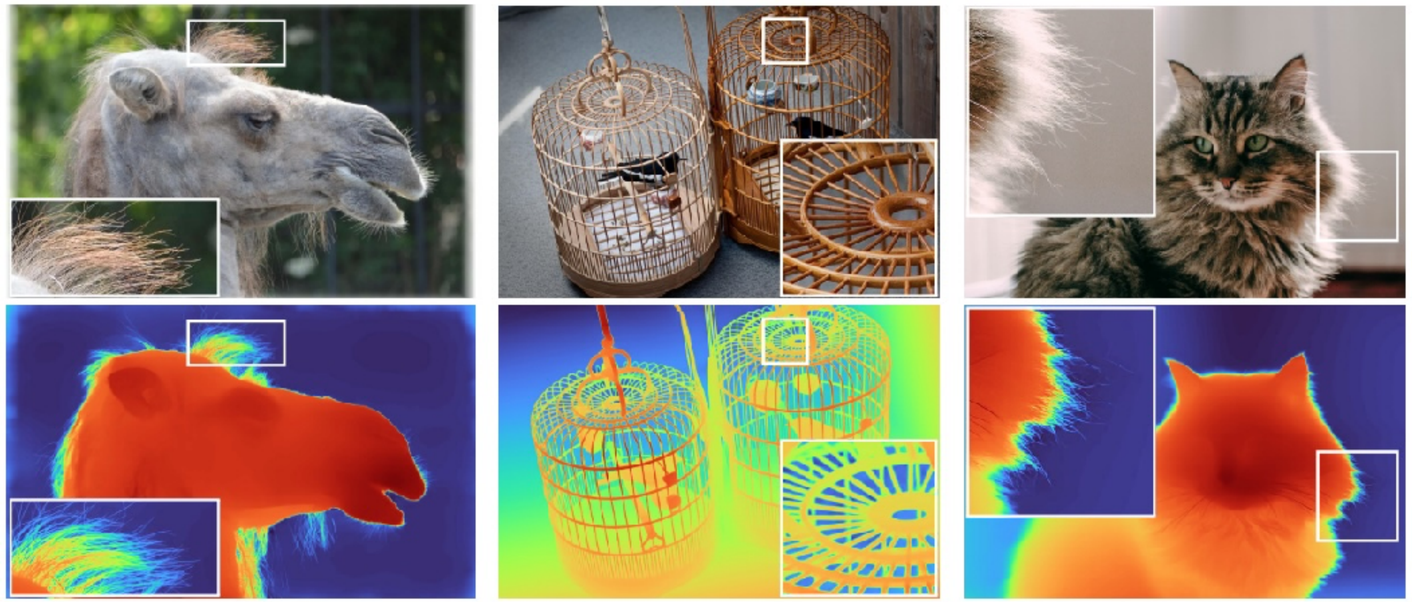
U-Net: Convolutional Networks for Biomedical Image Segmentation, Ronneberger etal 2016
Stacked Hourglass Networks for Human Pose Estimation, Newell etal 2016

Depth Estimation

DepthPro

- Transformer-based architecture
  - Fixed resolution, aspect ratio
- Trained on
  - Real-world depth data (Stage 1)
  - Synthetic data (Stage 1 + 2)
    - Supervise depth + gradients





Depth Pro: Sharp Monocular Metric Depth in Less Than a Second, Bochkovskii etal 2024

Task: Instance segmentation

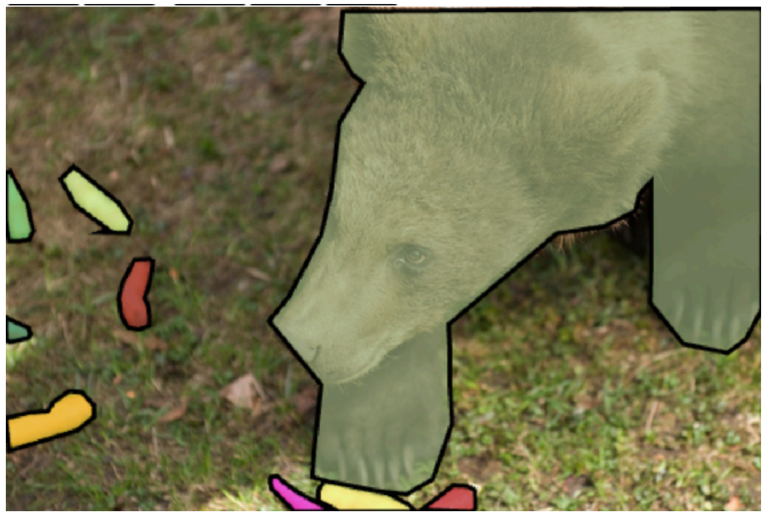
- Segment only objects
  - Segment and label each instance
  - Extension of object detection

Task: Panoptic segmentation

- Segment instances and stuff
  - Combines semantic and instance segmentation in a single task

Datasets: MS COCO

- 80 classes
- 200k images
  - 1.5M objects
  - Both instance and stuff labels



Microsoft coco: Common objects in context, Lin et al., 2014

Datasets: Driving

- Semantic segmention
  - Cityscapes
  - Mapillary

Image source : Cityscapes dataset

The Cityscapes Dataset for Semantic Urban Scene Understanding, Cordts et al., 2016
The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes, Neuhold et al., 2017

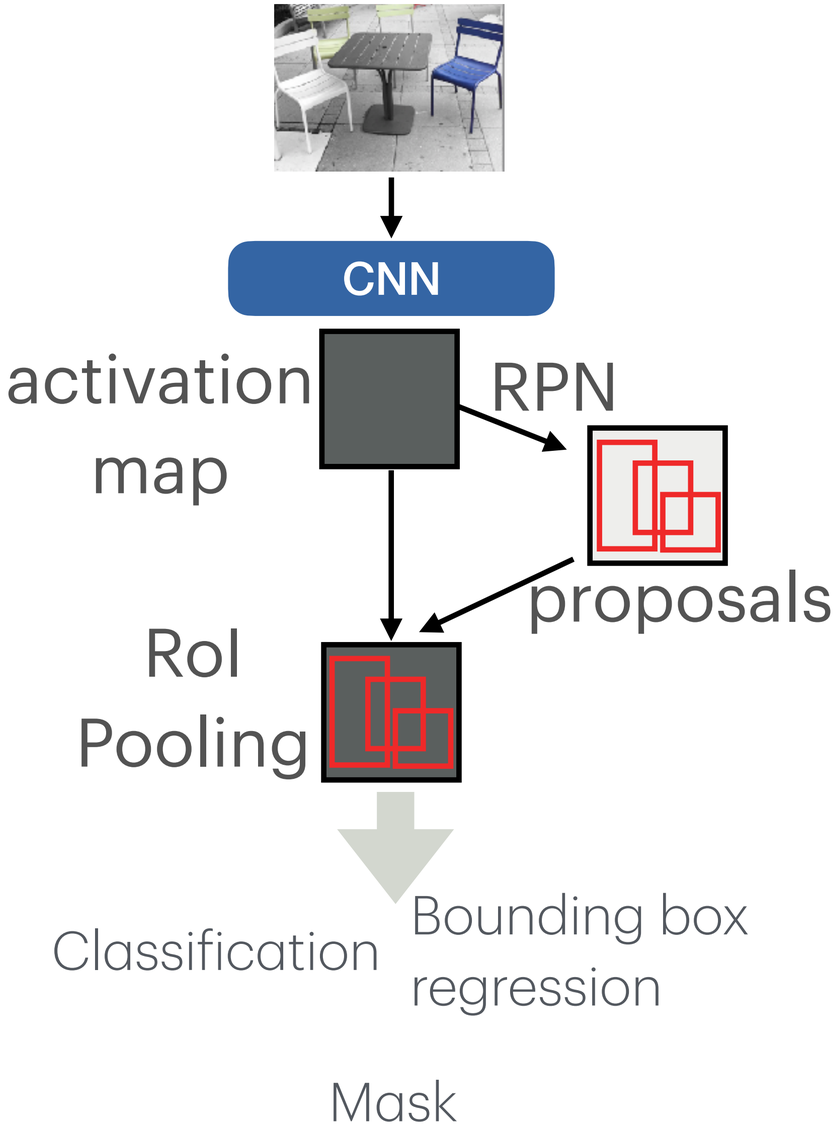
Datasets: Simulators

- Segmentation at no additional cost from rendering engine
- Examples
  - GTA V
  - Carla
  - Habitat

Playing for data: Ground truth from computer games, Richter et al., 2016
Free supervision from video games, Krähenbühl, 2018

MaskRCNN

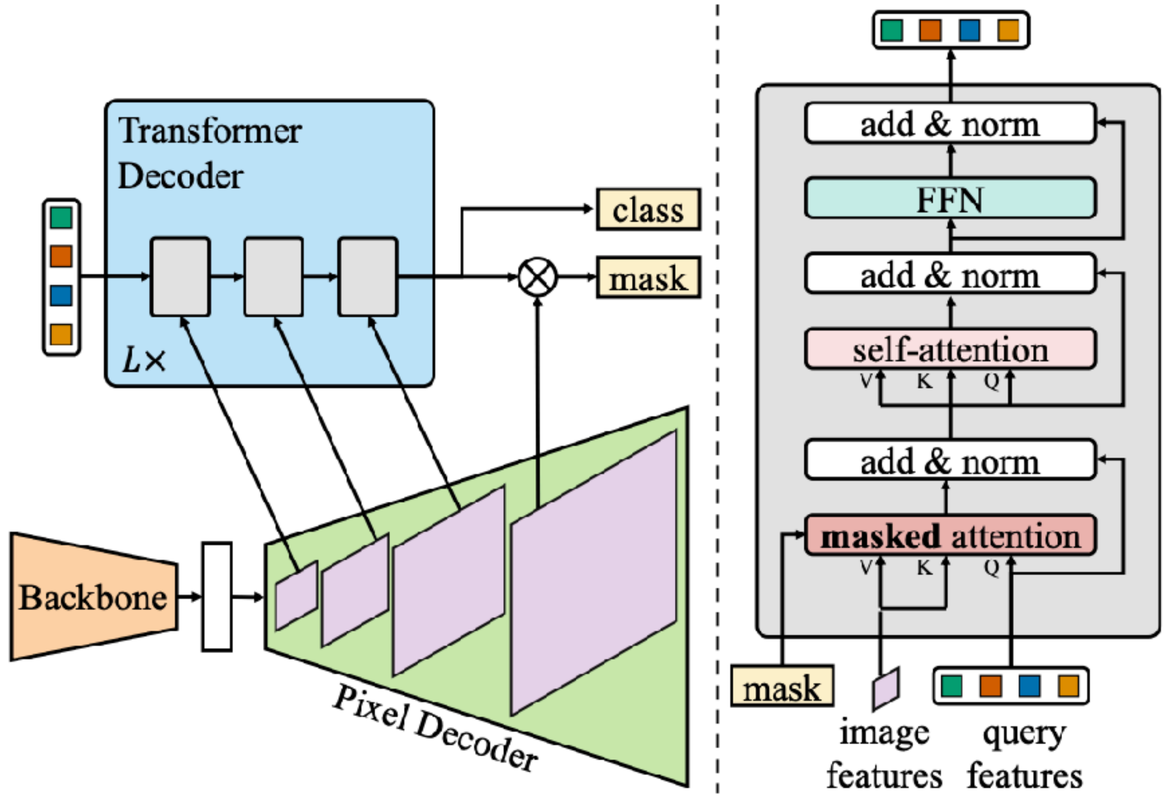
- Instance segmentation
- Segmentation by detection

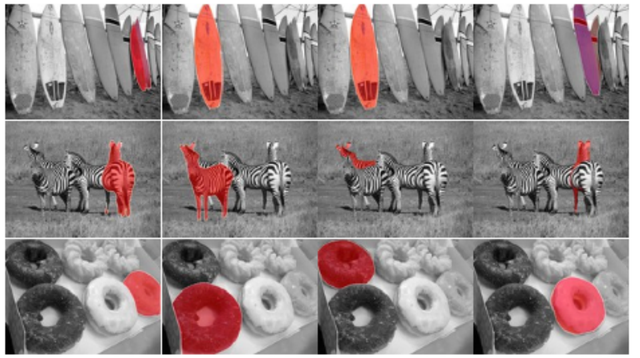


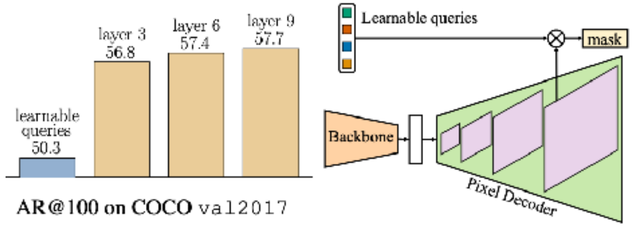
Mask R-CNN, He et al., ICCV 2017

Mask2Former

- DETR-style detector
- Predict an object mask at every layer
  - Use mask to shape attention
- Can do semantic, instance, panoptic segmentation with one architecture



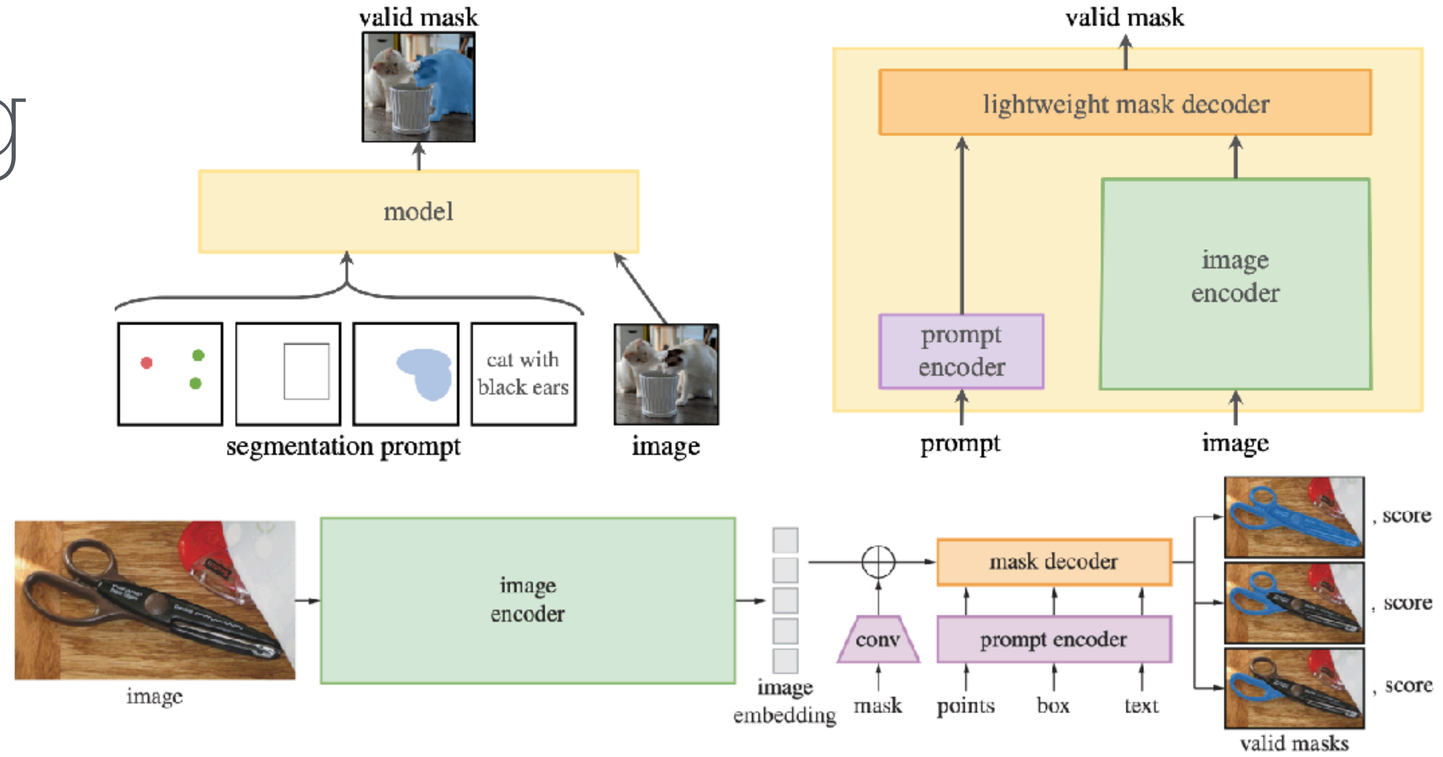




Masked-attention mask transformer for universal image segmentation, Cheng et al., CVPR 2022

Segment Anything

- MaskFormer-style segmentation network with queries from
  - Points
  - Boxes
  - Regions
  - Text descriptions



Segment Anything, Kirillov et al., ICCV 2023

Training

- Generate “synthetic” data
  - Use ground truth mask
  - Sample point
- Inference: Run SAM on dense grid and deduplicate
- Issue: Data

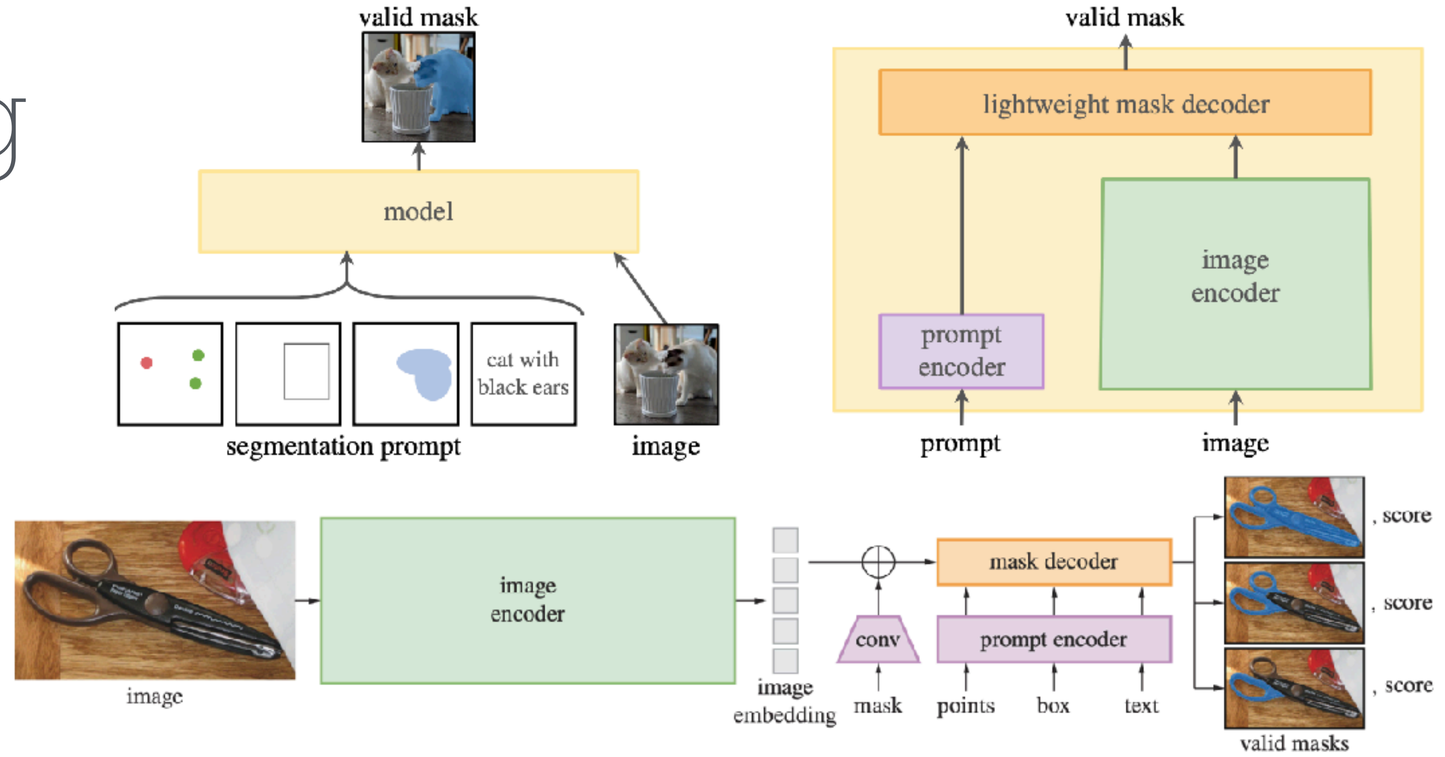
Data pipeline

11M high-res images

- Assisted-manual stage
  - Labelers select foreground / background
  - SAM produces masks
- Semi-automatic stage
  - Run SAM densely, annotate missing objects
- Fully automatic stage

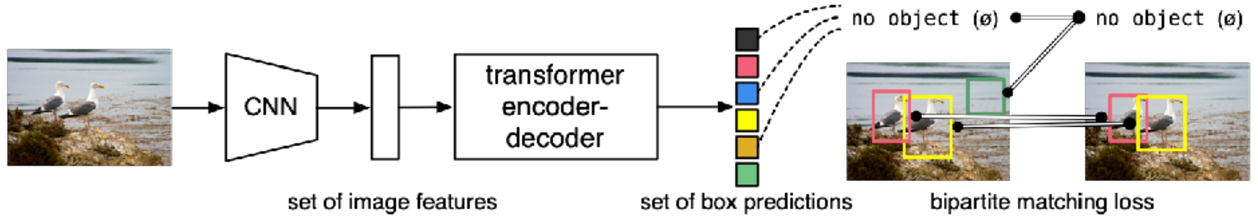
Segment Anything

- MaskFormer-style segmentation network with queries from
  - Points (from segmentation)
  - Boxes (from segmentation)
  - Regions (from segmentation)
  - Text descriptions (from CLIP model)



Segmentation

Trend 1



- **Semantic segmentation is done**
  - Task lost its popularity
  - Early successful applications:
    - Kinect, Lane boundary detection in AVs
  - Not clear if it’s worth the annotation cost
  - Dense prediction still very active (i.e. for monocular depth prediction)
  - Instance segmentation still very active (SAM)

Segmentation

Trend 2

- **Architectures become more general**
  - Instance -> Panoptic -> Anything
  - Segment Anything (and followup work) greatly simplify segmentation
    - No more class-specific dataset required
  - Off-the-shelf models are useful

Segmentation

Trend 3

- **Segmentation less tied to labels**
  - Instance segmentation seems to be label independent
  - Data can be masks only (at scale)

References

- [1] Fully Convolutional Networks for Semantic Segmentation, Shelhamer et al., 2015
- [2] U-Net: Convolutional Networks for Biomedical Image Segmentation, Ronneberger etal 2016
- [3] Stacked Hourglass Networks for Human Pose Estimation, Newell etal 2016
- [4] Depth Pro: Sharp Monocular Metric Depth in Less Than a Second, Bochkovskii etal 2024
- [5] The Cityscapes Dataset for Semantic Urban Scene Understanding, Cordts et al., 2016
- [6] The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes, Neuhold et al., 2017
- [7] Playing for data: Ground truth from computer games, Richter et al., 2016
- [8] Free supervision from video games, Krähenbühl, 2018
- [9] Mask R-CNN, He et al., ICCV 2017
- [10] Masked-attention mask transformer for universal image segmentation, Cheng et al., CVPR 2022
- [11] Segment Anything, Kirillov et al., ICCV 2023