Large Vision Models - SAM

Rowel Atienza, PhD
University of the Philippines
github.com/roatienza
2023

Why Foundation Model for Vision

- Broad benefits of LLM in NLP – Why not apply the same concept on vision
- A Large Vision Model can solve multiple vision tasks using a single model

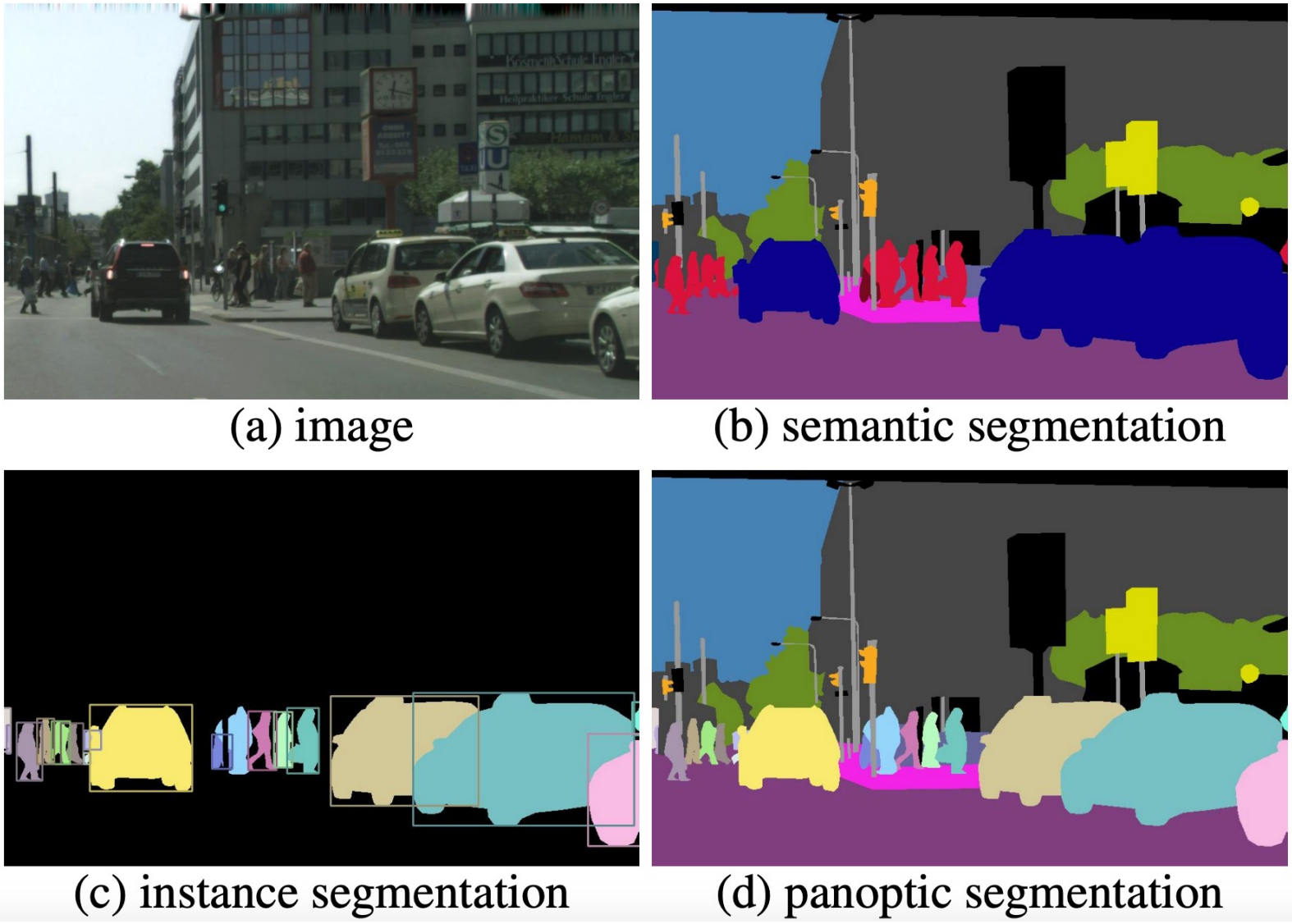
Why Segment Anything Model (SAM)

- SAM has demonstrated that it can segment almost anything

Segmentation as a Computer Vision Problem

Different Segmentation Tasks

Kirillov, Alexander, et al. "Panoptic segmentation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.



Different Segmentation Tasks

- Semantic - assign a class label to each pixel
- Instance - detect and segment each object instance
- Panoptic – per pixel class + instance label
  - Stuff – roads, sky, bodies of water, etc
  - Things – people, cars, dogs, cats, etc

SAM's Idea

- Use prompts to elicit the desired outcomes from a vision model

SAM's Task

- Given a prompt - INPUT
  - Point
  - Bounding box
  - Mask
  - Free-form text
- OUTPUT
  - Instance segmentation masks + confidence scores

SAM – the Model Architecture

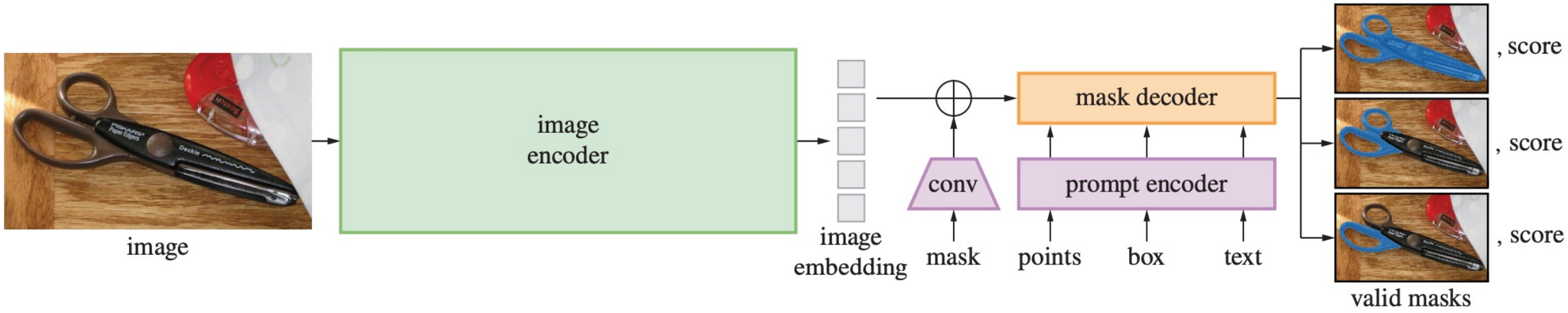


Figure 4: Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks at amortized real-time speed. For ambiguous prompts corresponding to more than one object, SAM can output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

Image Encoder is a Masked AutoEncoder (MAE)

He, Kaiming, et al. "Masked autoencoders are scalable vision learners." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

- ViT backbone

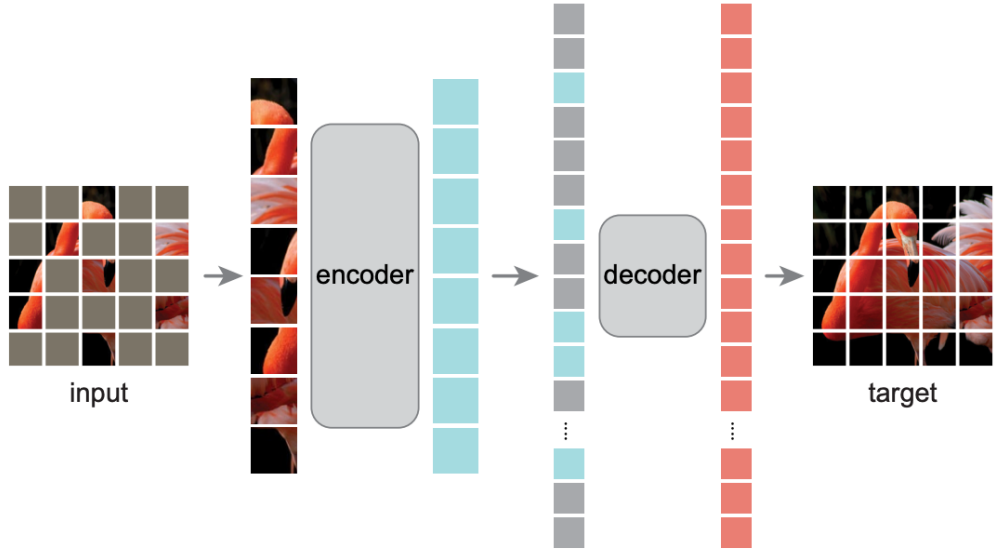


Figure 1. **Our MAE architecture**. During pre-training, a large random subset of image patches (*e.g.*, 75%) is masked out. The encoder is applied to the small subset of *visible patches*. Mask tokens are introduced *after* the encoder, and the full set of encoded patches and mask tokens is processed by a small decoder that reconstructs the original image in pixels. After pre-training, the decoder is discarded and the encoder is applied to uncorrupted images (full sets of patches) for recognition tasks.

Prompt Encoder

- Points and Bounding Boxes – Positional Encodings
- Masks - CNN
- Text - CLIP

Points and Boxes Embedding

Tancik, Matthew, et al. "Fourier features let networks learn high frequency functions in low dimensional domains." *Advances in Neural Information Processing Systems* 33 (2020): 7537-7547.

(a) Coordinate-based MLP | (b) Image regression $(x,y) \rightarrow$ RGB | (c) 3D shape regression $(x,y,z) \rightarrow$ occupancy | (d) MRI reconstruction $(x,y,z) \rightarrow$ density | (e) Inverse rendering $(x,y,z) \rightarrow$ RGB, density

Figure 1: Fourier features improve the results of coordinate-based MLPs for a variety of high frequency low-dimensional regression tasks, both with direct (b, c) and indirect (d, e) supervision We visualize an example MLP (a) for an image regression task (b), where the input to the network is a pixel coordinate and the output is that pixel's color. Passing coordinates directly into the network (top) produces blurry images, whereas preprocessing the input with a Fourier feature mapping (bottom) enables the MLP to represent higher frequency details.

Text Embedding using CLIP

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

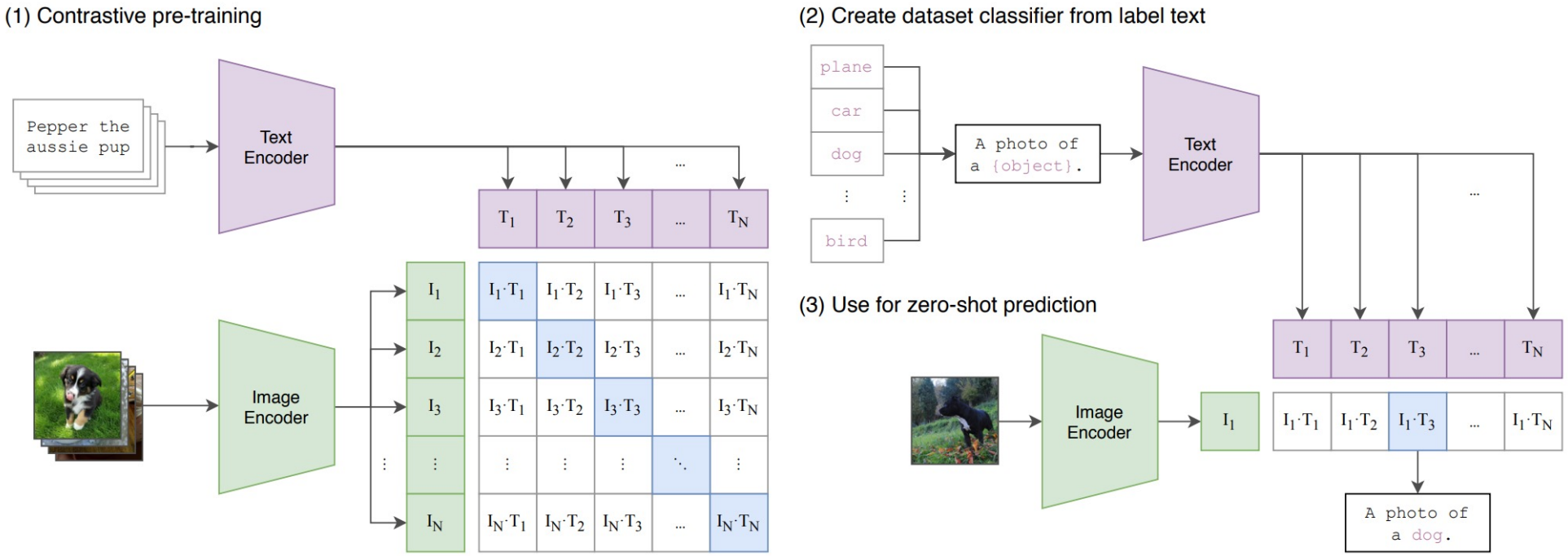
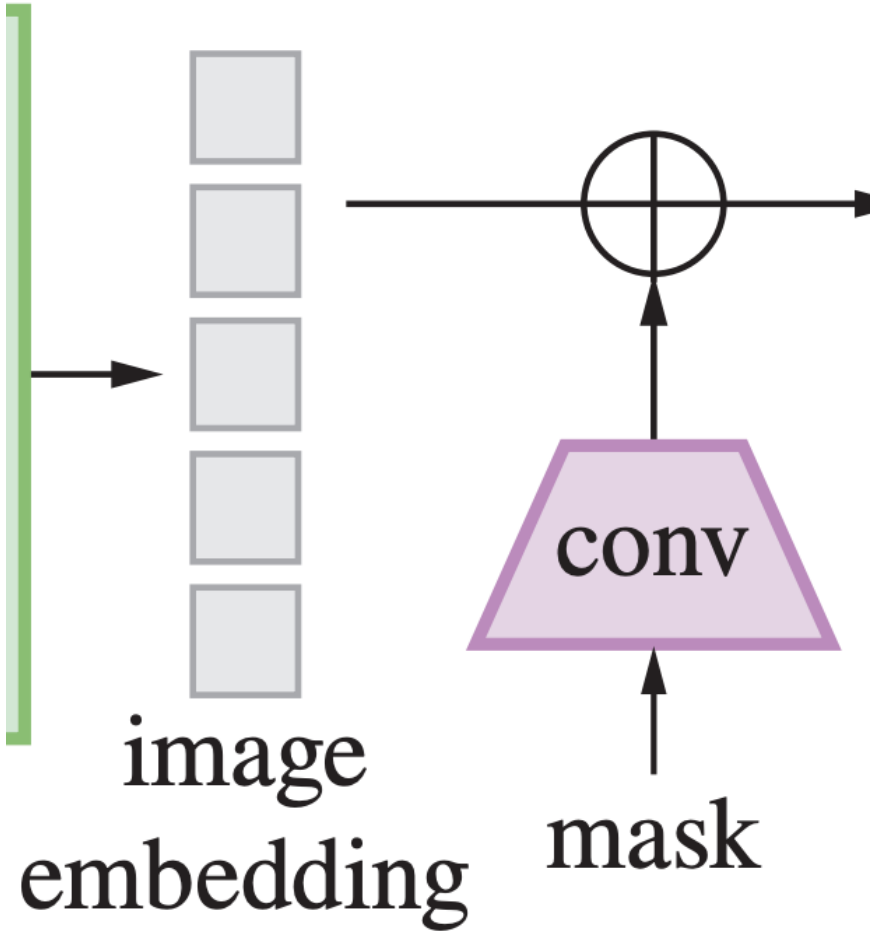


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

Mask Embedding



MaskFormer as Mask Decoder

Cheng, Bowen, Alex Schwing, and Alexander Kirillov. "Per-pixel classification is not all you need for semantic segmentation." *Advances in Neural Information Processing Systems* 34 (2021): 17864-17875.

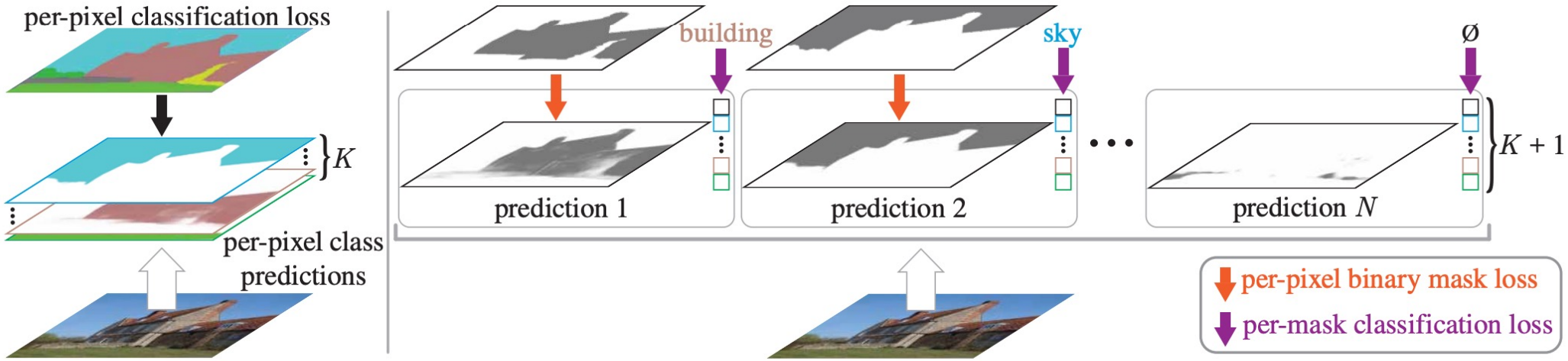


Figure 1: **Per-pixel classification *vs*. mask classification. (left)** Semantic segmentation with per-pixel classification applies the same classification loss to each location. **(right)** Mask classification predicts a set of binary masks and assigns a single class to each mask. Each prediction is supervised with a per-pixel binary mask loss and a classification loss. Matching between the set of predictions and ground truth segments can be done either via *bipartite matching* similarly to DETR [3] or by *fixed matching* via direct indexing if the number of predictions and classes match, *i.e.*, if $N = K$.

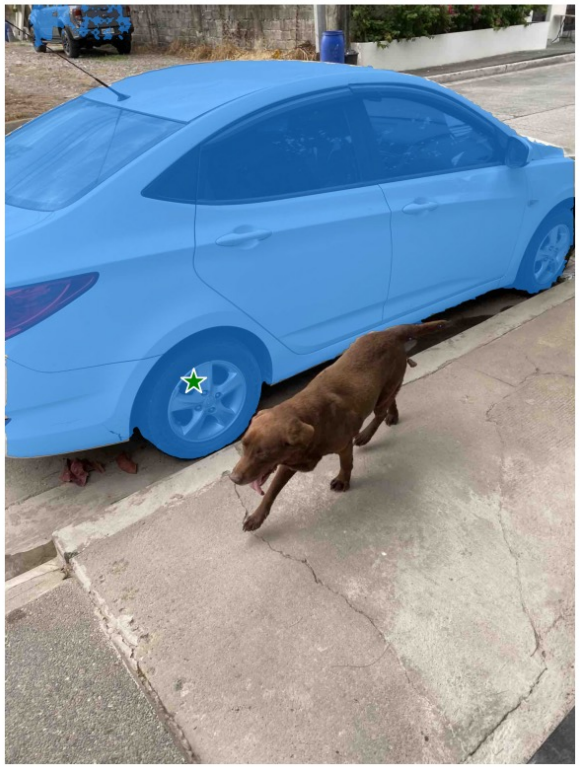
Ambiguous Masks

- When the prompt is ambiguous (eg point on the rim may refer to the rim, tire, wheel or car)
- 3 candidate masks are generated

Mask 1, Score: 1.015

Mask 2, Score: 1.011

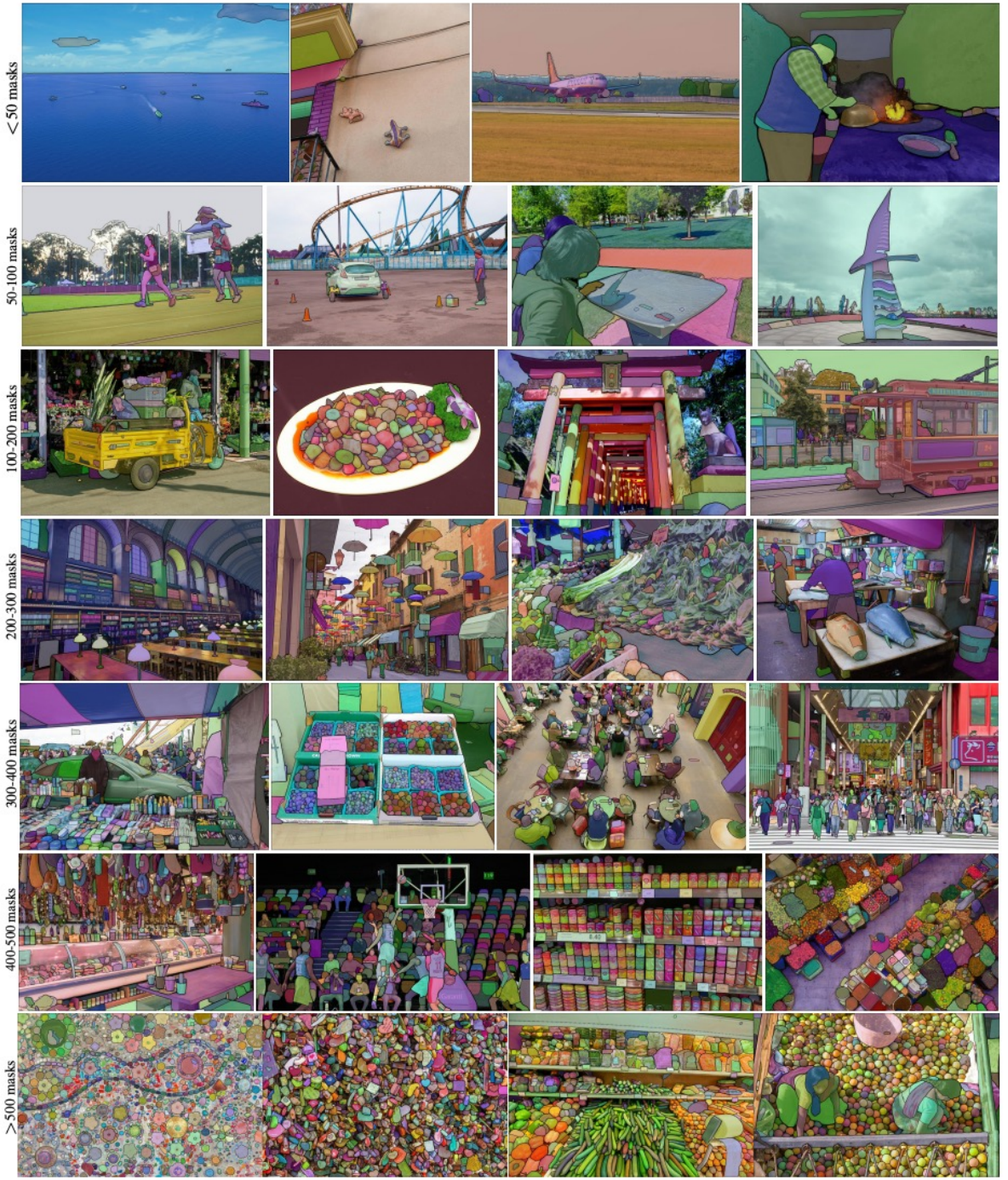
Mask 3, Score: 0.970



Training Details

Data

- 11M Images
- 1.1B Masks
- Use of data engine
  - Assisted-manual
  - Semi-automatic
  - Fully-automatic



Loss Function

- Dice Loss – measure of dissimilarity between the predicted and ground truth masks
  - $\mathcal{L}_{dice} = 1.0 - \frac{2\cap}{\cup+\cap}$
- Focal Loss – cross-entropy with focus on the hard examples
  - $\mathcal{L}_{focal} = -(1 - p_t)^{\gamma} \log p_t$

Training

- ViT-H encoder pre-trained using MAE
- Dataset is the from the fully-automatic stage (SA-1B)

Zero-Shot Experiments

Experiments

1. Zero-Shot Single Point Valid Mask Evaluation
2. Zero-Shot Edge Detection
3. Zero-Shot Object Proposals
4. Zero-Shot Instance Segmentation
5. Zero-Shot Text-to-Mask

Zero-Shot Single Point Valid Mask Evaluation

Sofiiuk, Konstantin, Ilya A. Petrov, and Anton Konushin. "Reviving iterative training with mask guidance for interactive segmentation." *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022.

- Choose a point (eg image center)
- Evaluate mask prediction using interactive segmentation

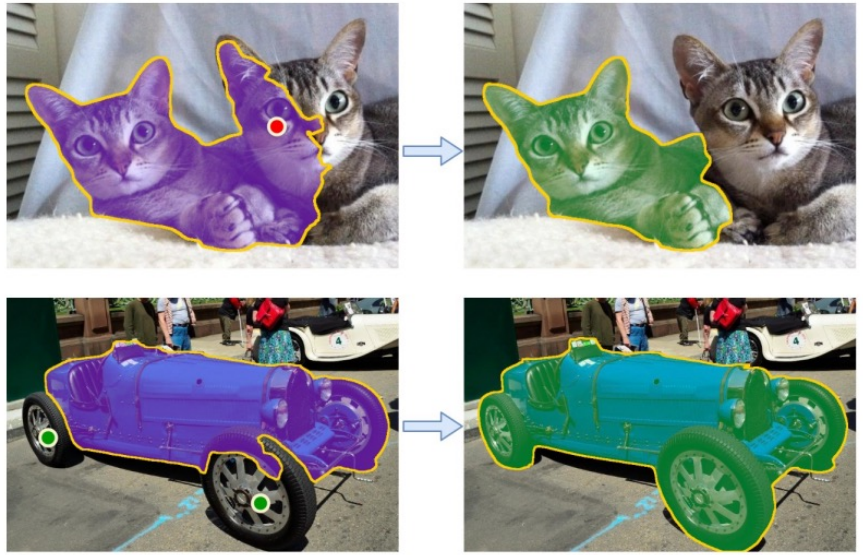
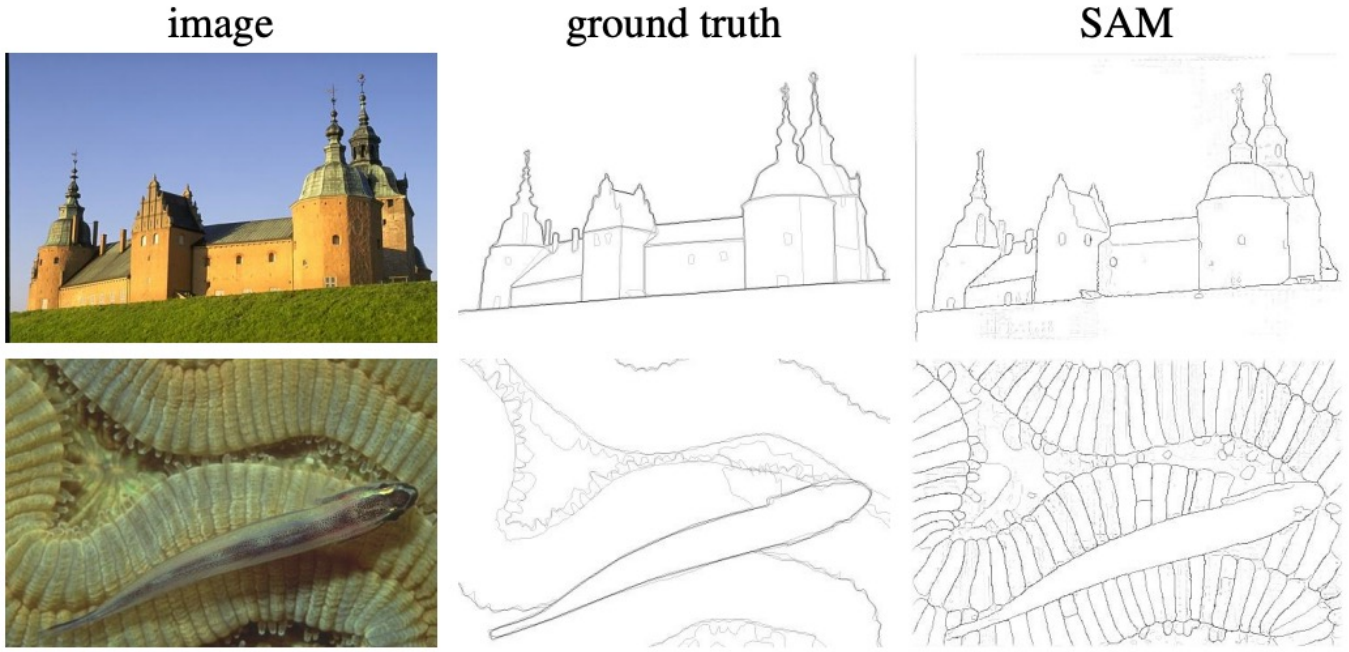


Fig. 1. Besides segmenting new objects, the proposed method allows to correct masks *e.g.* produced by other instance or semantic segmentation models. A user can fix erroneous regions with positive (green) and negative (red) clicks.

Zero-Shot Edge Detection

- Apply Sobel filter on masks' unthresholded probability maps
- Set values to zero if they do not intersect with the outer boundary pixels of a mask.
- Pixel-wise max over all the predictions
- Normalize the result to [0,1]
- Apply edge NMS to thin the edges



Non-Maximum Suppression (NMS)

Bodla, Navaneeth, et al. "Soft-NMS--improving object detection with one line of code." *Proceedings of the IEEE international conference on computer vision*. 2017.

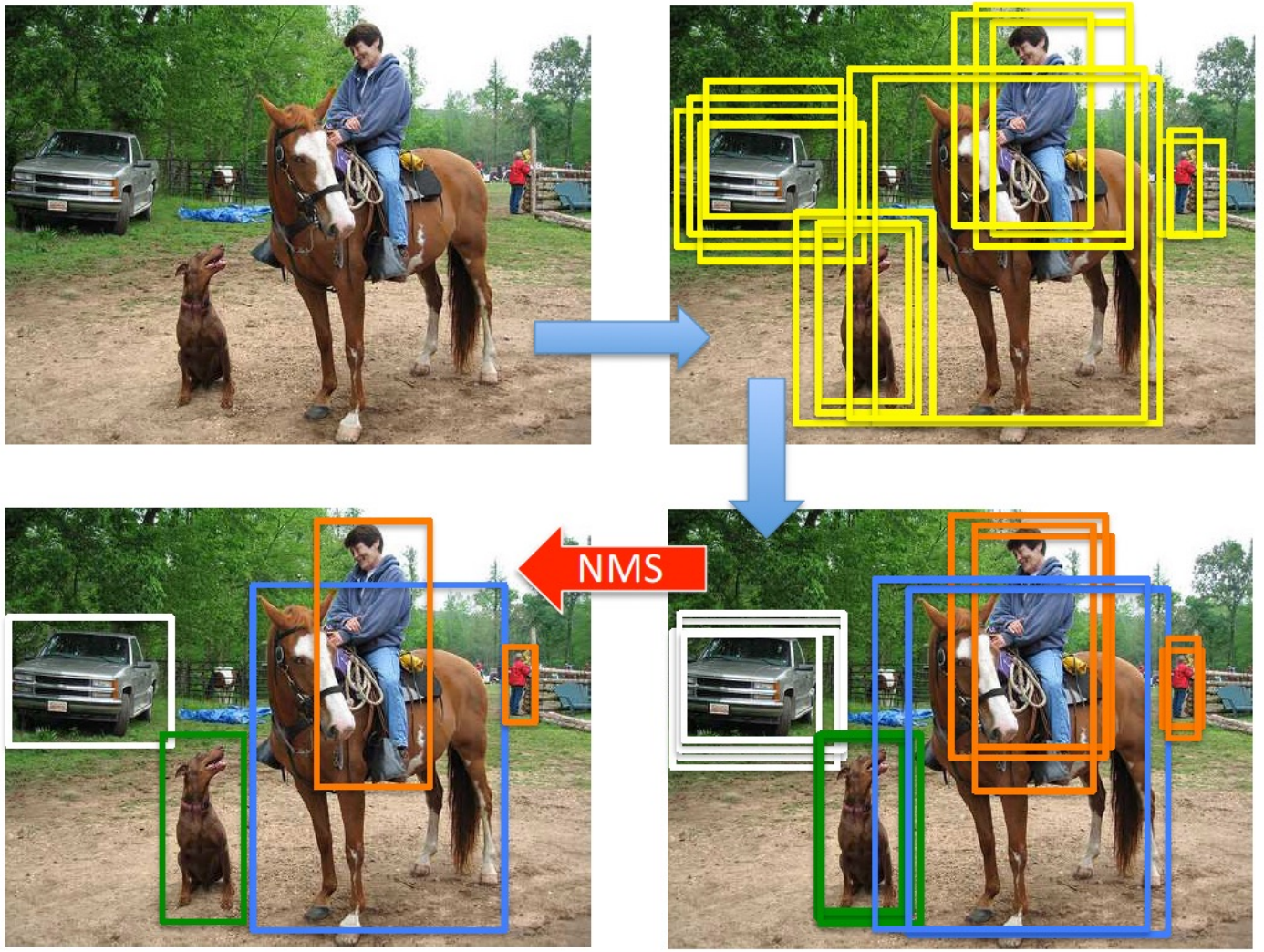


Figure 3. In object detection, first category independent region proposals are generated. These region proposals are then assigned a score for each class label using a classification network and their positions are updated slightly using a regression network. Finally, non-maximum-suppression is applied to obtain detections.

Zero-Shot Object Proposals

Van de Sande, Koen EA, et al. "Segmentation as selective search for object recognition." *2011 international conference on computer vision*. IEEE, 2011.

- Modify the masks as object proposals

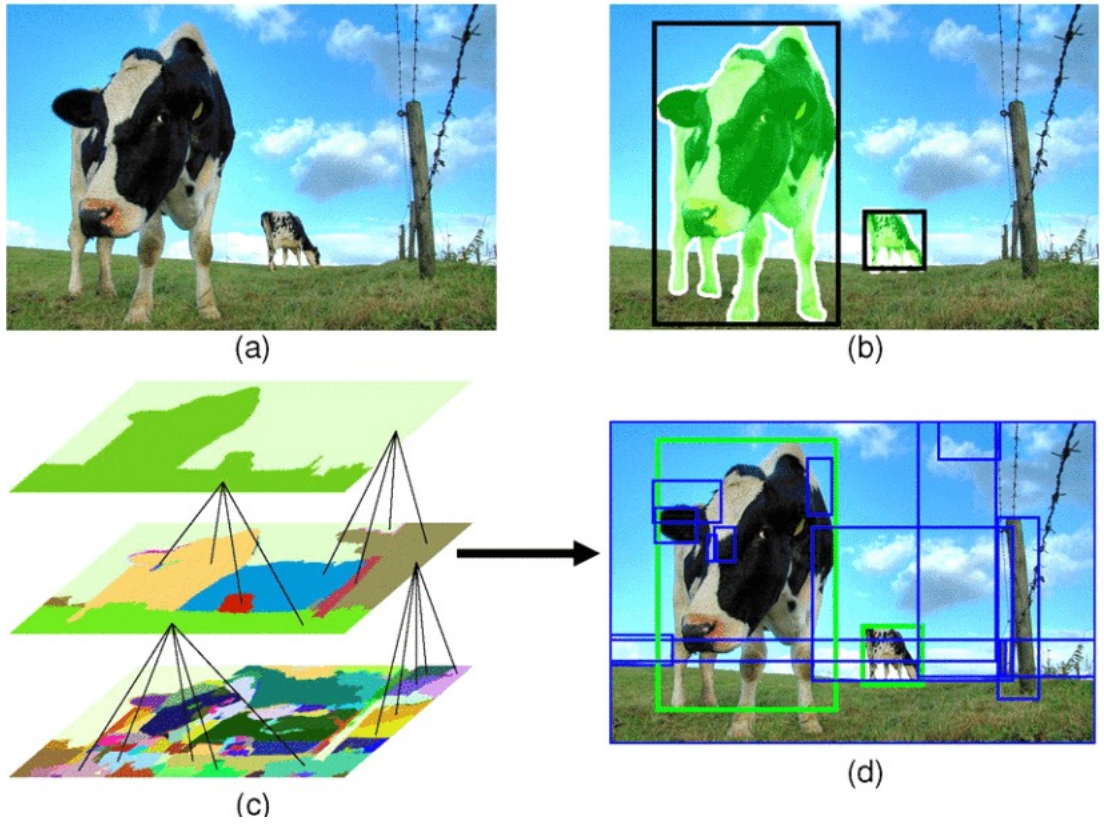


Figure 1. Given an image (a) our aim is to find its objects for which the ground truth is shown in (b). To achieve this, we adapt segmentation as a selective search strategy: We aim for high recall by generating locations at all scales and account for many different scene conditions by employing multiple invariant colour spaces. Example object hypotheses are visualised in (d)

Zero-Shot Instance Segmentation

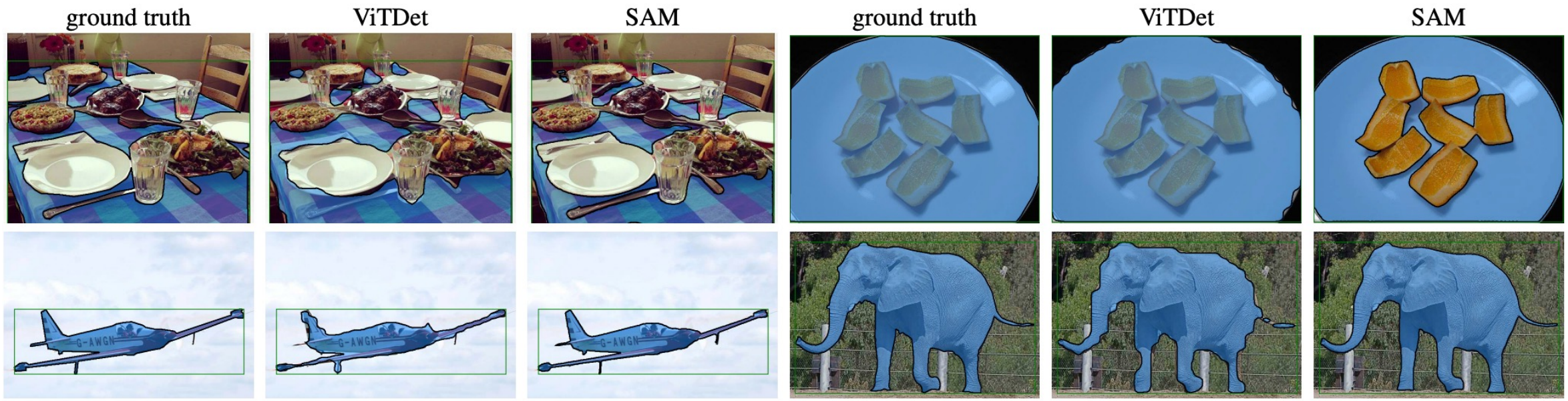


Figure 16: Zero-shot instance segmentation on LVIS v1. SAM produces higher quality masks than ViTDet. As a zero-shot model, SAM does not have the opportunity to learn specific training data biases; see top-right as an example where SAM makes a modal prediction, whereas the ground truth in LVIS is amodal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

Text-to-Mask

- Use CLIP image embedding as prompt during training
- Use CLIP text embedding as input during inference

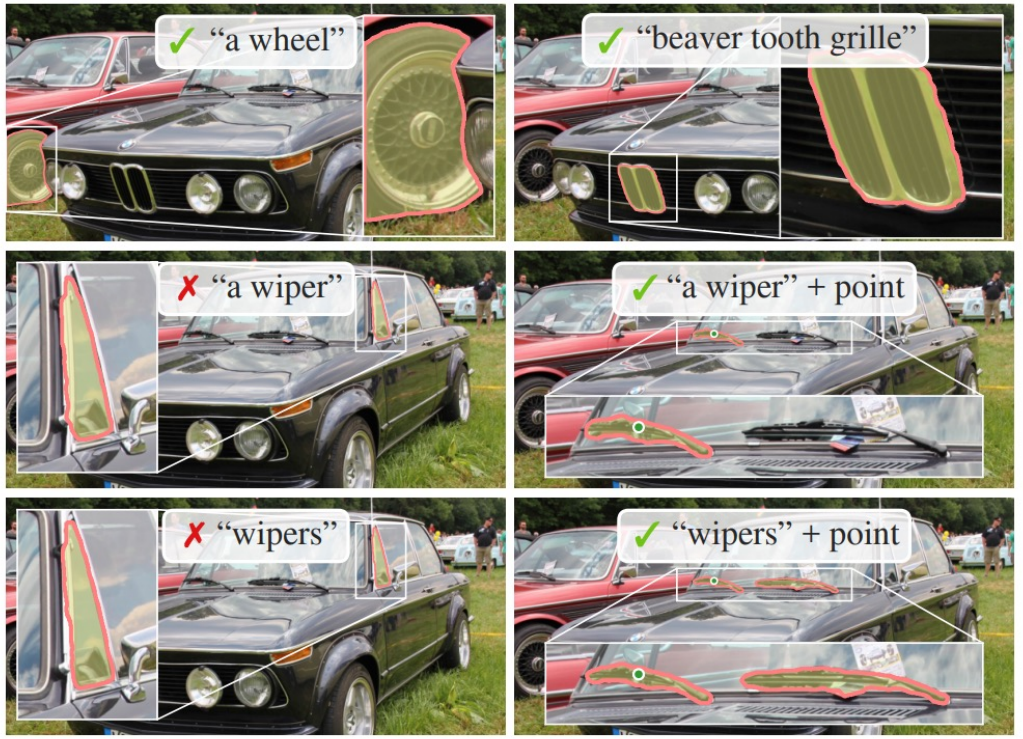


Figure 12: Zero-shot text-to-mask. SAM can work with simple and nuanced text prompts. When SAM fails to make a correct prediction, an additional point prompt can help.

Code demo