Friedrich-Alexander-Universität
Erlangen-Nürnberg

FAU

Orange Segmentation With Segment Anything

Evaluating zero-shot, fine-tuning, and comparing with UNet.

Ramkrishna Acharya[1]

[1]Friedrich-Alexander Universität Erlangen-Nürnberg, Department Data Science

September 12, 2024

Contents

Introduction

Segment Anything [Kir+23]

Research from Meta AI Research addresses the following questions about image segmentation:

- What **task** will enable zero-shot generalization?
- What is the corresponding **model** architecture?
- What **data** can power this task and model?

A. Kirillov et al. *Segment Anything*. 2023

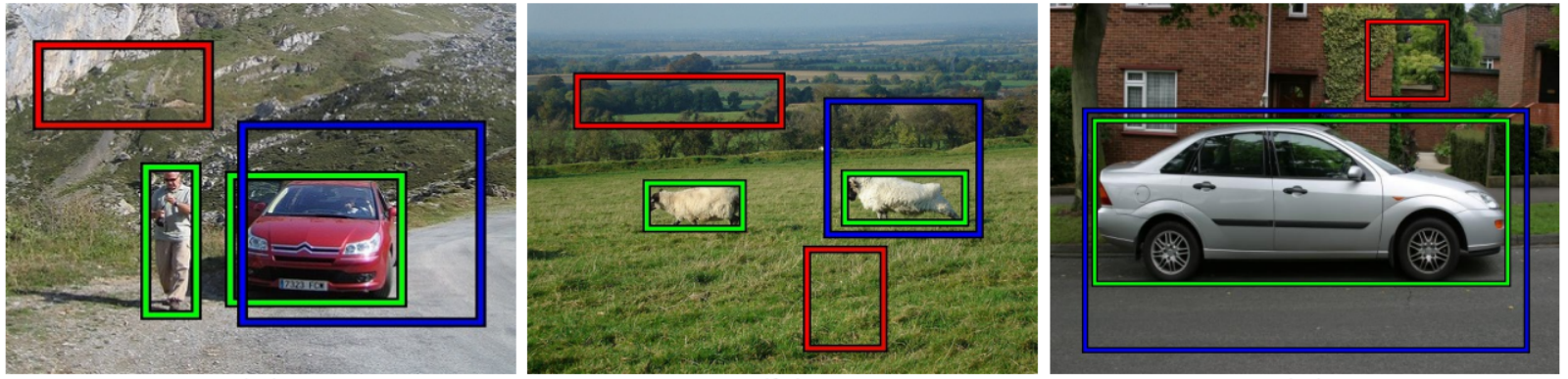
Introduction

Contributions From Segment Anything

- A **single model that can segment stuff** and things on various visual data.
- Pre-trained **model with Apache 2.0 license**.
- Large scale segmentation dataset i.e. SA-1B **dataset with a valid license**.
- **Inference code** with Apache 2.0 license.

Related Work

Object Proposal Generation (2010)



(a) (b) (c)

Fig. 1: **Desired behavior of an objectness measure.** *The desired objectness measure should score the blue windows, partially covering the objects, lower than the ground truth windows (green), and score even lower the red windows containing only stuff or small parts of objects.*

Figure: [ADF10]

How **likely** is the image to contain an object?

B. Alexe, T. Deselaers, and V. Ferrari. "What is an object?" In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2010, pp. 73–80.
DOI: 10.1109/CVPR.2010.5540226

Related Work

Semantic Segmentation (2015)

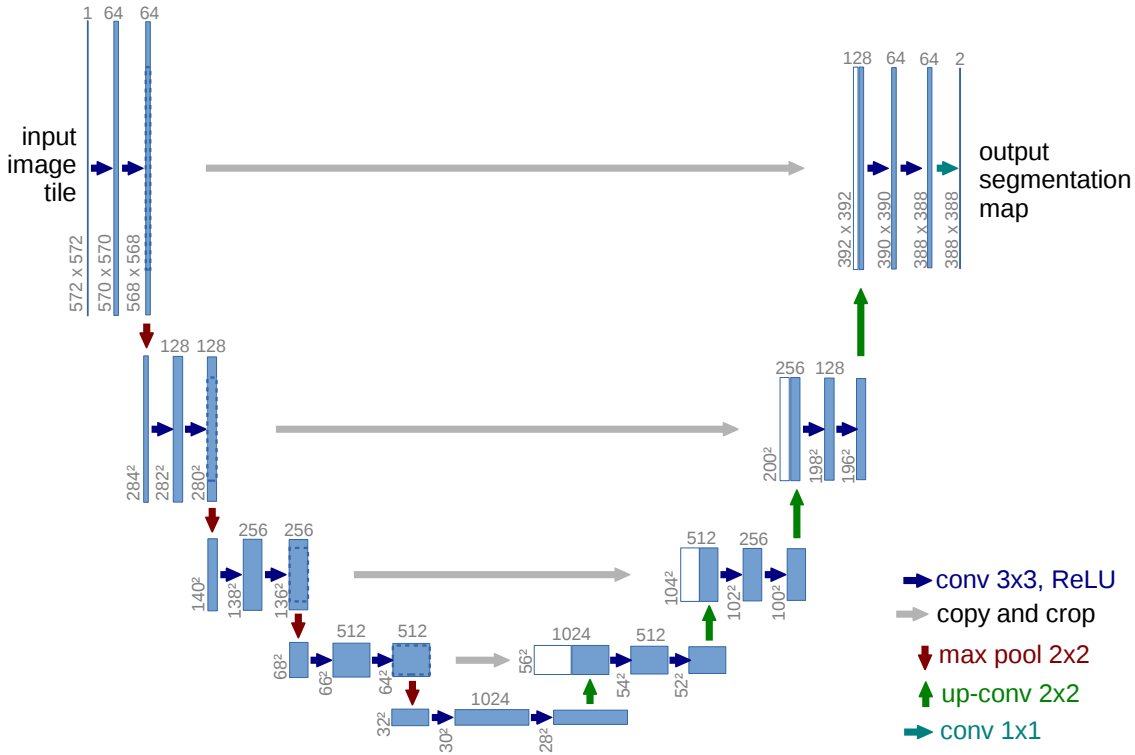


Figure: U-Net Architecture [RFB15]

Segments each pixel in one of the classes.

O. Ronneberger, P. Fischer, and T. Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015

Related Work

Instance Segmentation (2015)

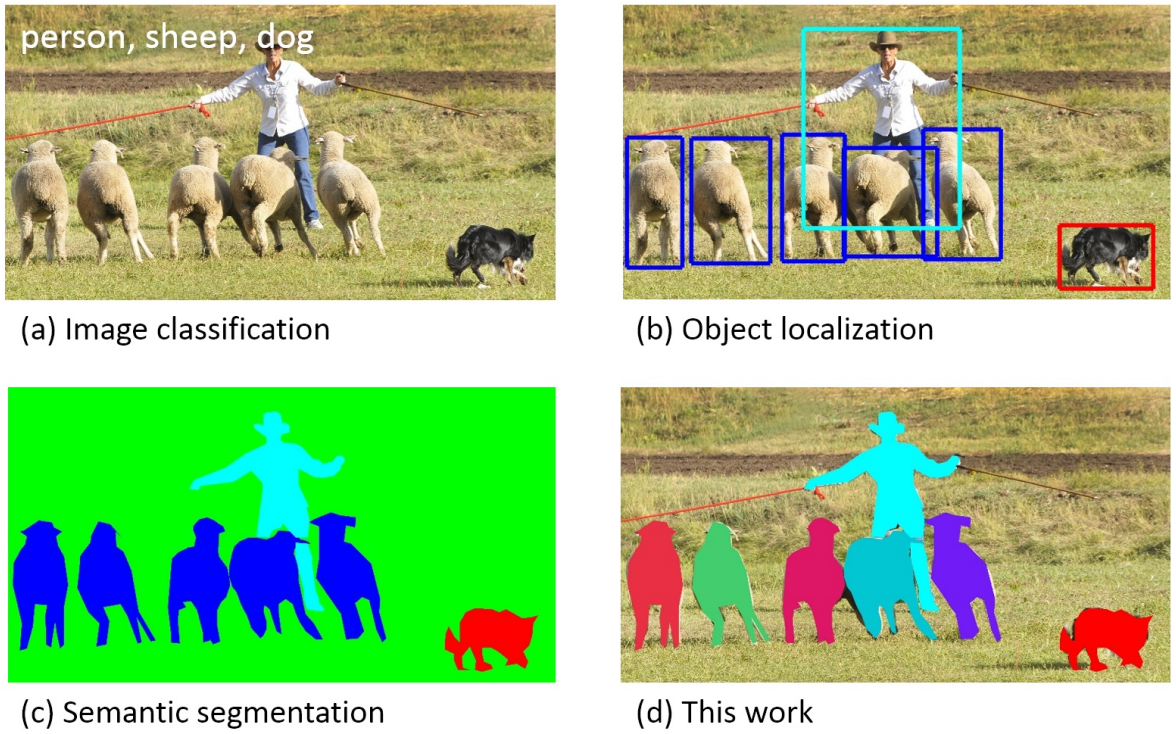


Figure: Source: [Lin+15]

Segmentation is **based on instances** but not based on classes.

T.-Y. Lin et al. *Microsoft COCO: Common Objects in Context*. 2015

Related Work

Interactive Segmentation (2016)

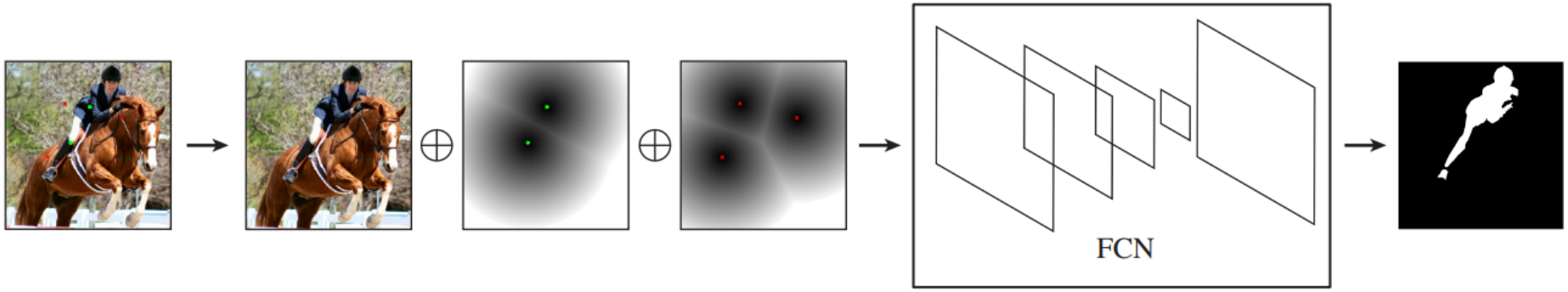


Figure 1: The framework of learning our FCN models. Given an input image and user interactions, our algorithm first transforms positive and negative clicks (denoted as green dots and red crosses respectively) into two separate channels, which are then concatenated (denoted as $\oplus$) with the image's RGB channels to compose an input pair to the FCN models. The corresponding output is the ground truth mask of the selected object.

Figure: Source: [Xu+16]

Segment based on interaction from the user i.e. some prompt.

N. Xu, B. Price, S. Cohen, J. Yang, and T. Huang. *Deep Interactive Object Selection*. 2016

Related Work

Panoptic Segmentation (2019)

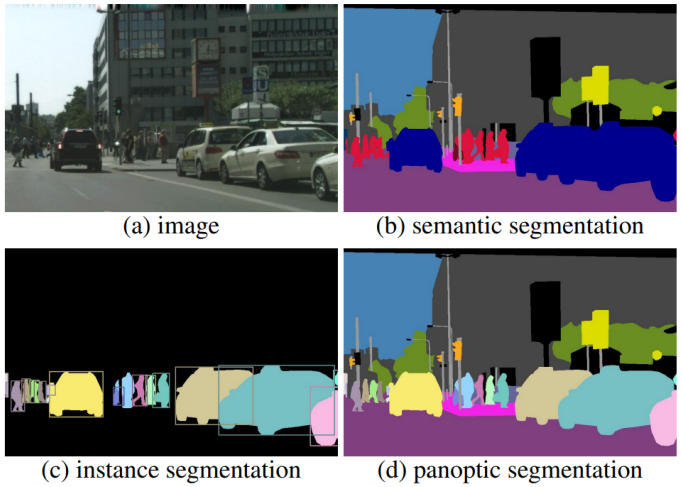


Figure 1: For a given (a) image, we show *ground truth* for: (b) semantic segmentation (per-pixel class labels), (c) instance segmentation (per-object mask and class label), and (d) the proposed *panoptic segmentation* task (per-pixel class+instance labels). The PS task: (1) encompasses both stuff and thing classes, (2) uses a simple but general format, and (3) introduces a uniform evaluation metric for all classes. Panoptic segmentation generalizes both semantic and instance segmentation and we expect the unified task will present novel challenges and enable innovative new methods.

Figure: Source: [Kir+19]

Panoptic segmentation unifies the typically distinct tasks of semantic segmentation (assign a class label to each pixel) and instance segmentation (detect and segment each object instance).

A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár. *Panoptic Segmentation*. 2019

Background

Transformer[Vas+23] (2017)

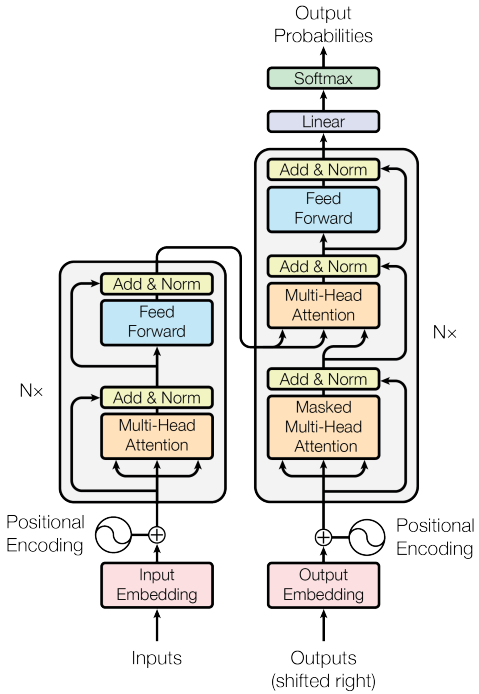


Figure: Transformer Architecture

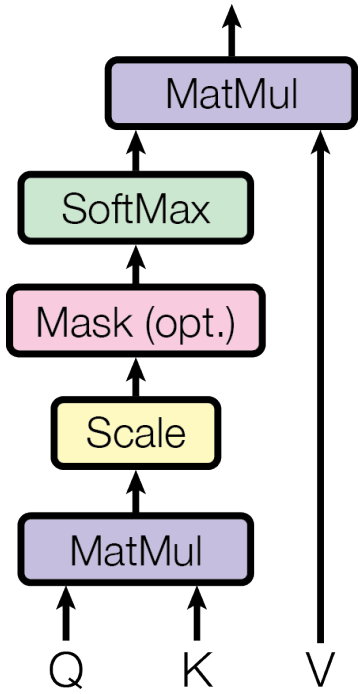


Figure: Scaled Dot-Product Attention = $\frac{softmax(Q.K^T)}{\sqrt{dim(K)}}V$

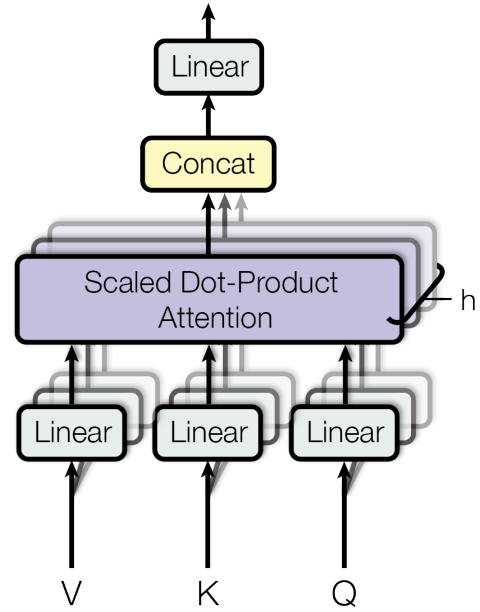


Figure: Multi-Headed Attention = $concat(Atn(QW_i^K, KW_i^K, VW_i^V))$.

Transformers are the foundational model for many state-of-the-art language models.

A. Vaswani et al. *Attention Is All You Need*. 2023

Background

Vision Transformer (2021)

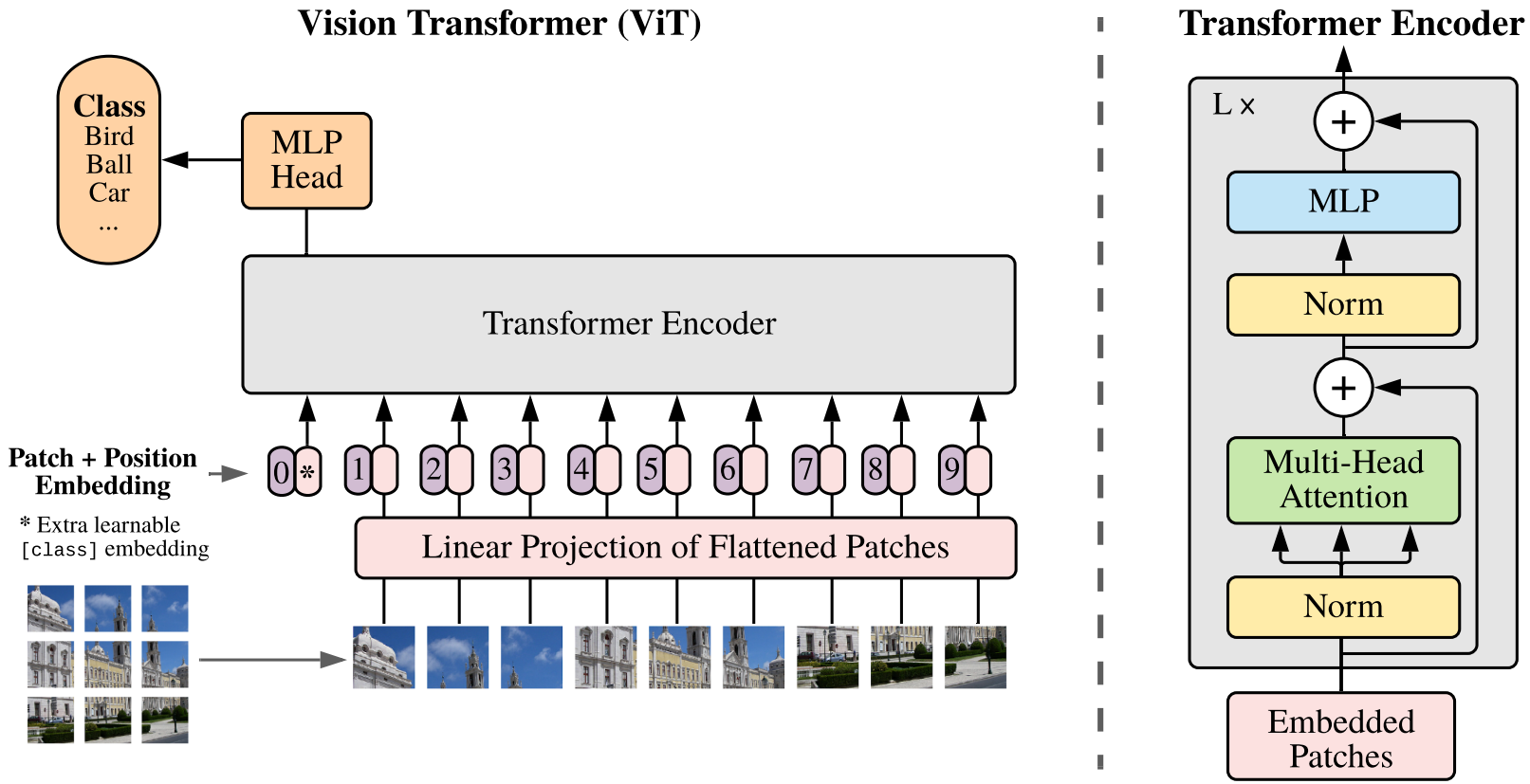


Figure: Vision Transformer [Dos+21]

ViT (Vision Transformer) is the foundational model for many state-of-the-art vision models.

A. Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021

Background

Masked Auto-Encoder (2021)

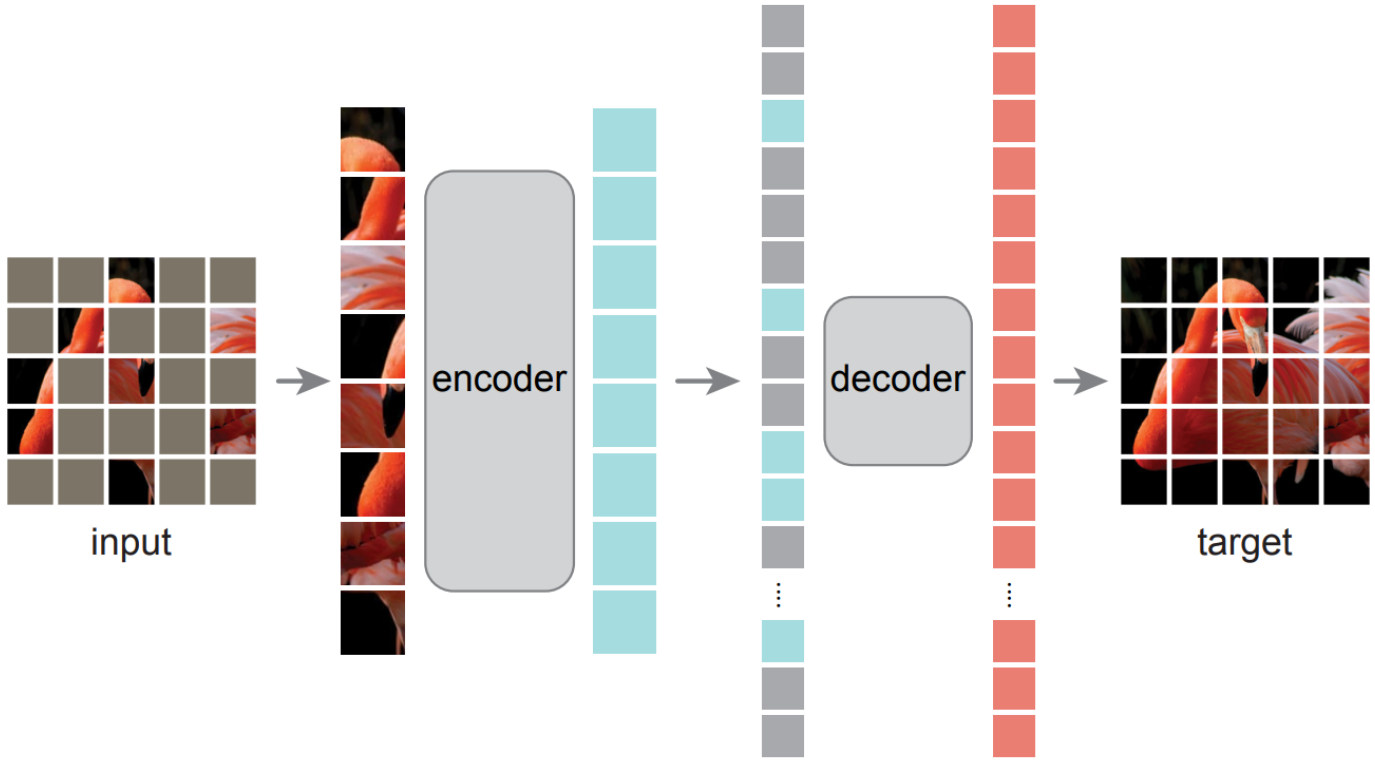


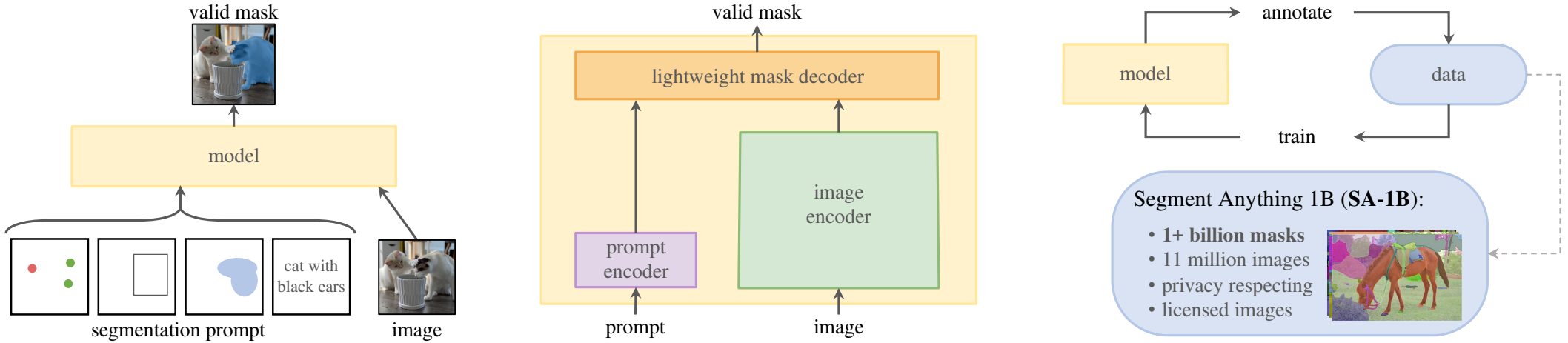
Figure: Masked Auto Encoder [He+21]

MAE (Masked Auto Encoder) uses ViT as the encoder and Segment Anything uses MAE as an image encoder.

K. He et al. *Masked Autoencoders Are Scalable Vision Learners*. 2021

Methodology

Task, Model and Data: Model



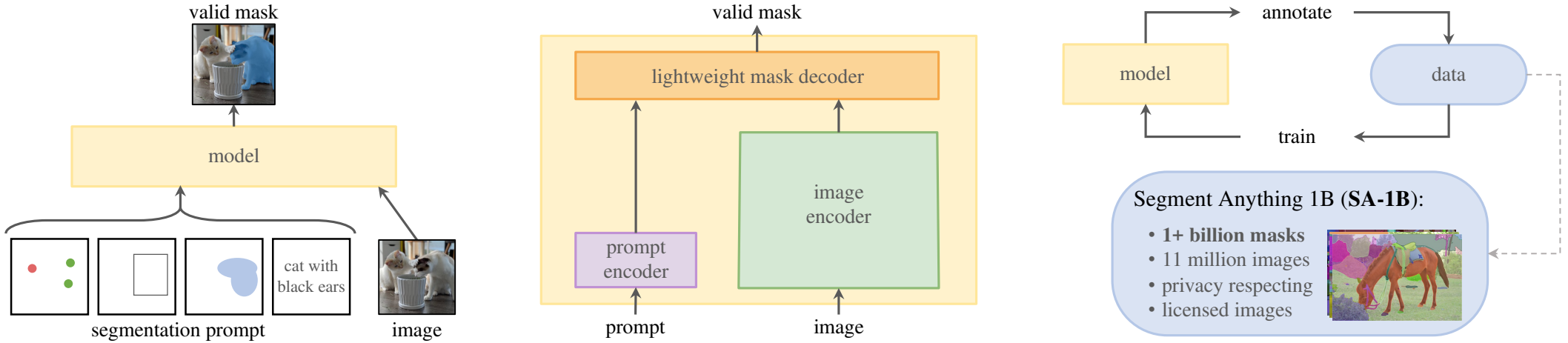
(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

Figure: Source: Original Authors

- **Image Encoder**: Pretrained MAE with ViT-Huge.
- **Prompt Encoder**: Points, boxes by positional encodings[Tan+20](passing input points through Fourier feature mapping for MLP) and text by encoder of CLIP[Rad+21] (jointly trains encoders in text to image task)
- **Mask decoder**: Modified transformer block with mask prediction head. It uses prompt self-attention and cross-attention to update all embeddings in prompt-to-image embedding and vice versa. After two blocks, unsampled image embedding is passed to MLP to get foreground probability at each image location.

Methodology

Task, Model and Data: Data



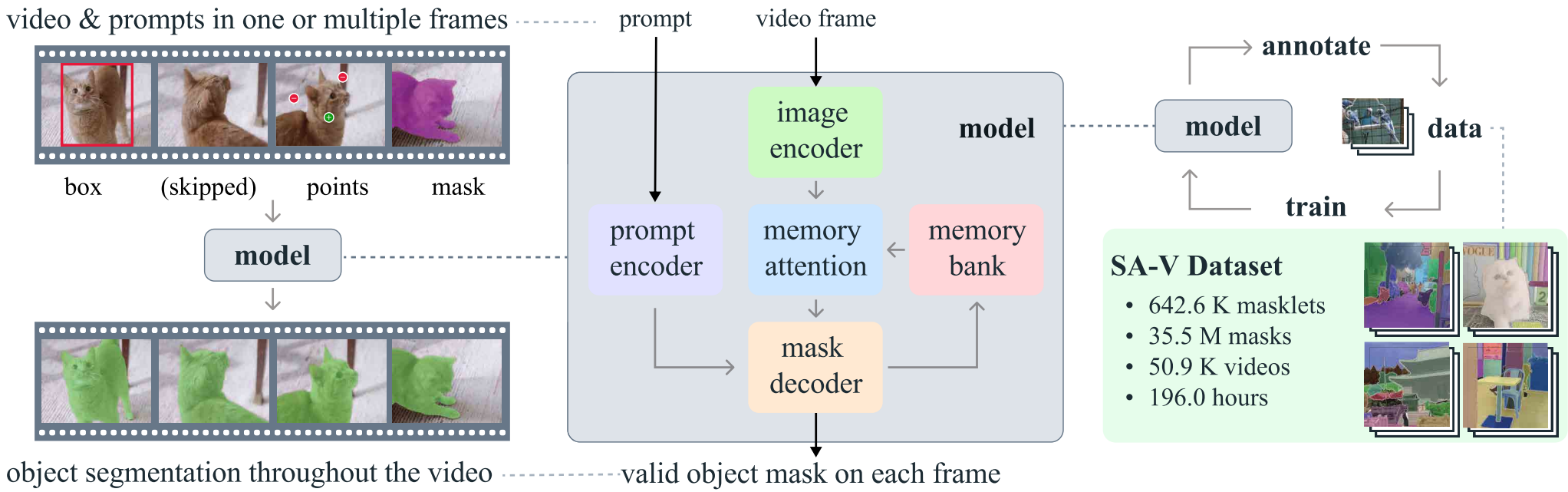
(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

Figure: Source: Original Authors

- **Data Engine**: Assisted manual annotation, semi-automatic and fully automatic stages.

Methodology

SAM2[Rav+24]



(a) Task: promptable visual segmentation **(b) Model:** Segment Anything Model 2 **(c) Data:** data engine and dataset

Figure: Source: Original Authors

- SAM2 can perform segmentation on video as well.

Methodology

SAM: Training Procedure

- **Ambiguity handling**: When one output, average the masks. Only backdrop minimum loss.
- **Loss Function**: 20 * Focal loss [Lin+18] + Dice loss [Sud+17] + MSE loss
- **Optimizer**: AdamW [LH19] and learning rate warm up for 250 iterations and step-wise learning rate decay.
- **Training iterations**: 90k iterations (2 SA-1B epochs)
- **Batch size**: 256
- **Input Image**: 1024 by 1024 and no augmentations applied.
- **Training Time**: 65 hours with 256 Nvidia A100 80 GB each.

$$\text{FL}(p_t) = -(1-p_t)^{\gamma}\log(p_t) \tag{1}$$

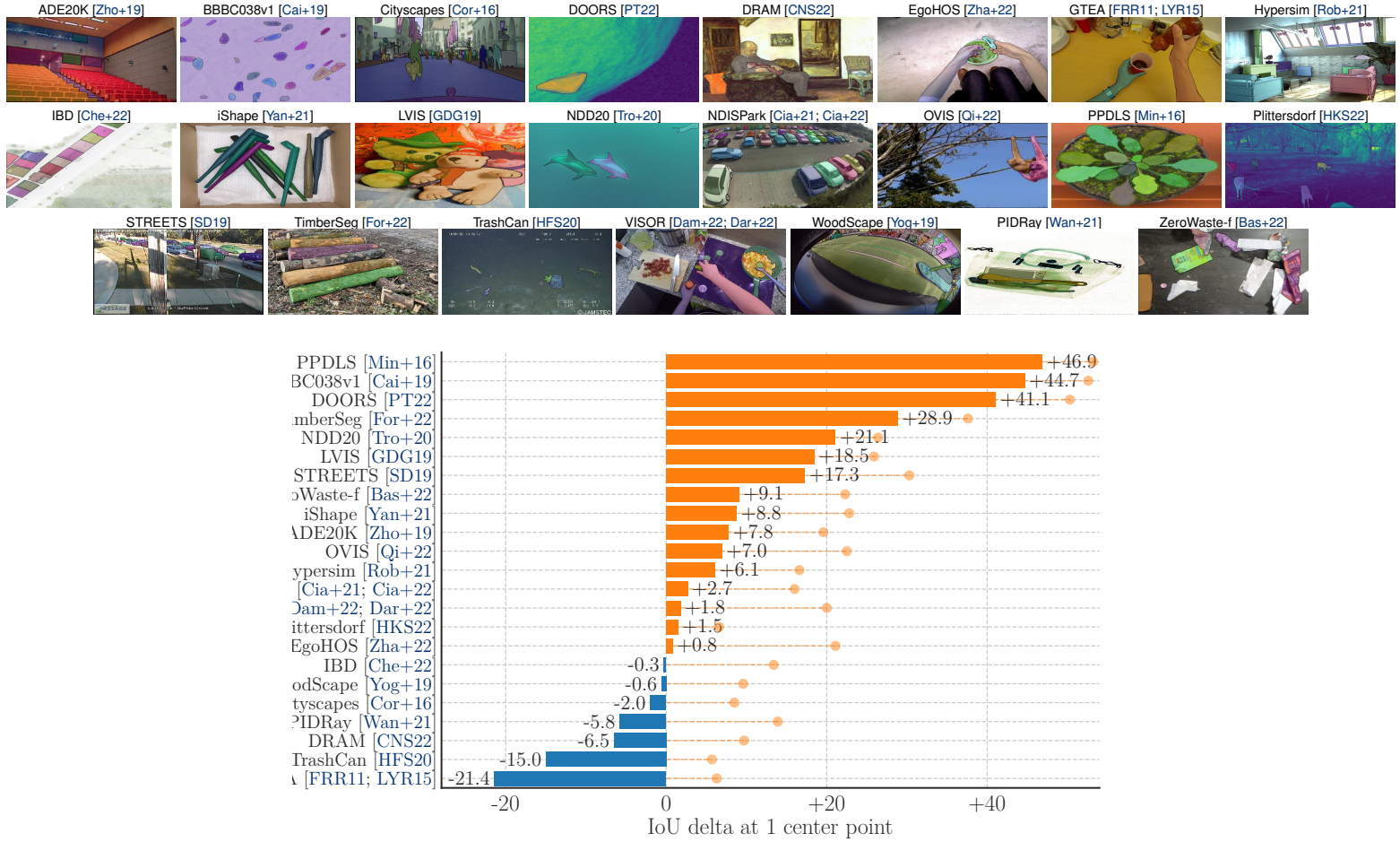
$$DiceLoss(y,\overline{p}) = 1 - \frac{(2y\overline{p}+1)}{(y+\overline{p}+1)} \tag{2}$$

AdamW

$$\theta_{t+1,i} = \theta_{t,i} - \eta\left(\frac{1}{\sqrt{\hat{v}_t+\epsilon}}\cdot\hat{m}_t + w_{t,i}\theta_{t,i}\right), \forall t \tag{3}$$

Results and Analysis

SAM vs RITM on Zero-Shot Single Point Mask Detection



(a) SAM vs RITM [SPK21] on 23 datasets

Figure: Mean IoU (58.1) of SAM and the strongest single point segmenter, RITM [SPK21]. Source: SAM Authors

Results and Analysis

SAM Zero-Shot Edge Detection

Figure: Zero-shot edge prediction on BSDS500. SAM was not trained to predict edge maps nor had access to BSDS images or annotations during training.
Source: SAM Authors

method	year	ODS	OIS	AP	R50
HED [XT15]	2015	.788	.808	.840	.923
EDETR [Pu+22]	2022	.840	.858	.896	.930
zero-shot transfer methods:					
Sobel filter	1968	.539	-	-	-
Canny [Can86]	1986	.600	.640	.580	-
Felz-Hutt [FH04]	2004	.610	.640	.560	-
SAM	2023	.768	.786	.794	.928

Table: Zero-shot transfer to edge detection on BSDS500.
Source: SAM Authors

Results and Analysis

SAM Zero-Shot Object Proposal

Method	mask AR (average recall) @1000						
	all	small	med.	large	freq.	com.	rare
ViTDet-H [Li+22]	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: 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. Source: SAM Authors

SAM outperforms ViTDet-H on medium and large objects, as well as rare and common objects.

Results and Analysis

SAM Zero-Shot Instance Segmentation

Method	COCO [Lin+14]				LVIS v1 [GDG19]			
	AP	AP^S	AP^M	AP^L	AP	AP^S	AP^M	AP^L
ViTDet-H [Li+22]	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

Table: SAM is prompted with ViTDet boxes to do zero-shot segmentation. The fully-supervised ViTDet outperforms SAM, but the gap shrinks on the higher-quality LVIS masks. Interestingly, SAM outperforms ViTDet according to human ratings.

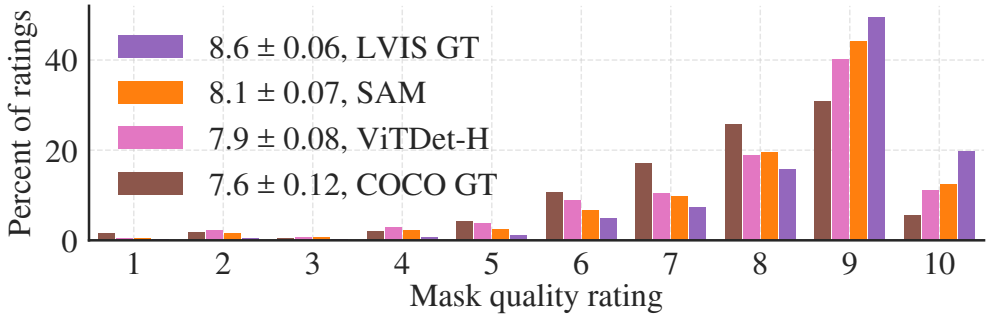


Figure: Mask quality rating distribution from human study for ViTDet and SAM, both applied to LVIS ground truth boxes. The legend shows rating means and 95% confidence intervals. Source: SAM Authors

Results and Analysis

SAM Zero-Shot Text to Mask

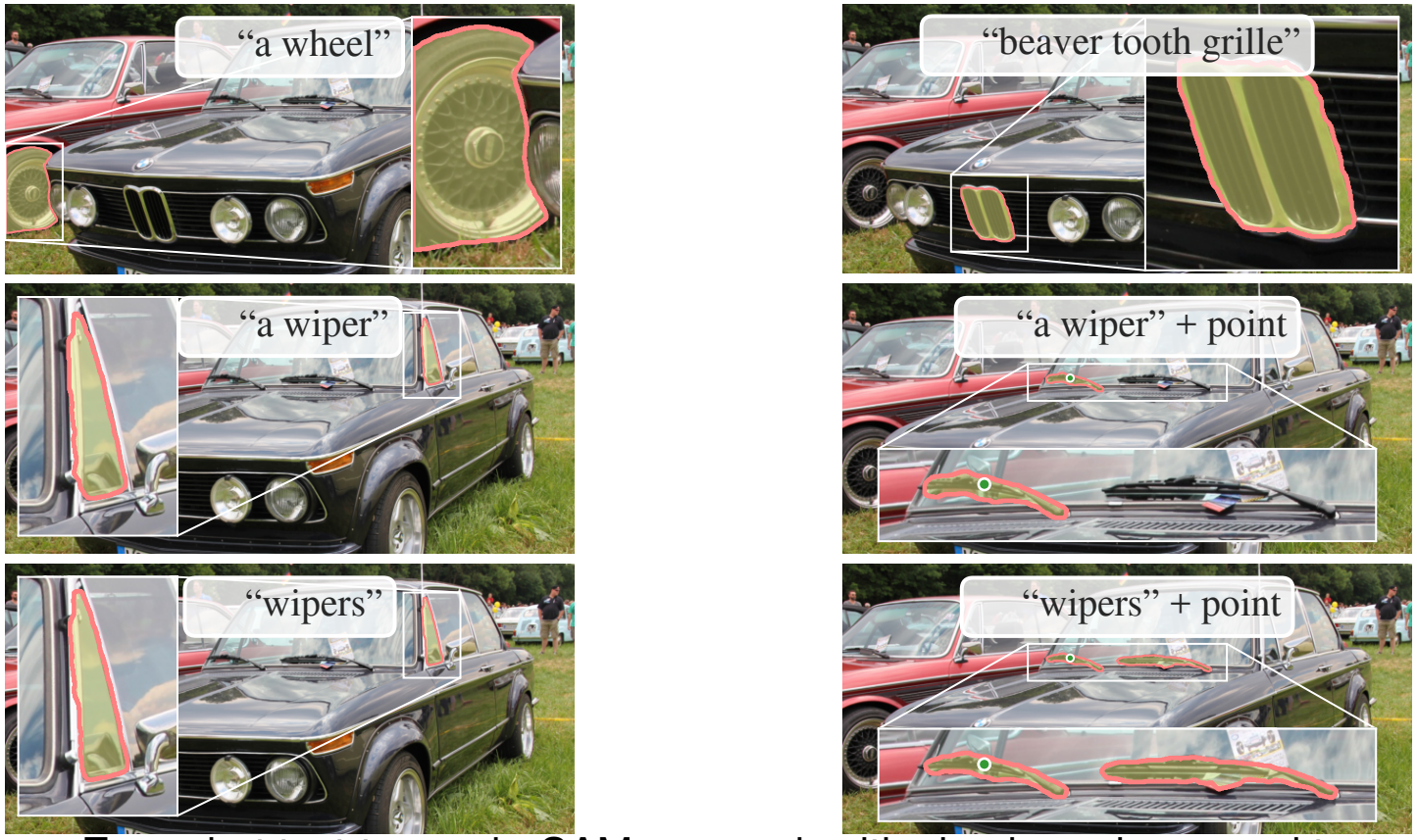


Figure: Zero-shot text-to-mask. SAM can work with simple and nuanced text prompts. When SAM fails to make a correct prediction, an additional point prompt can help. Source: SAM Authors

Results and Analysis

SAM Zero-Shot on Orange Dataset [PA24]

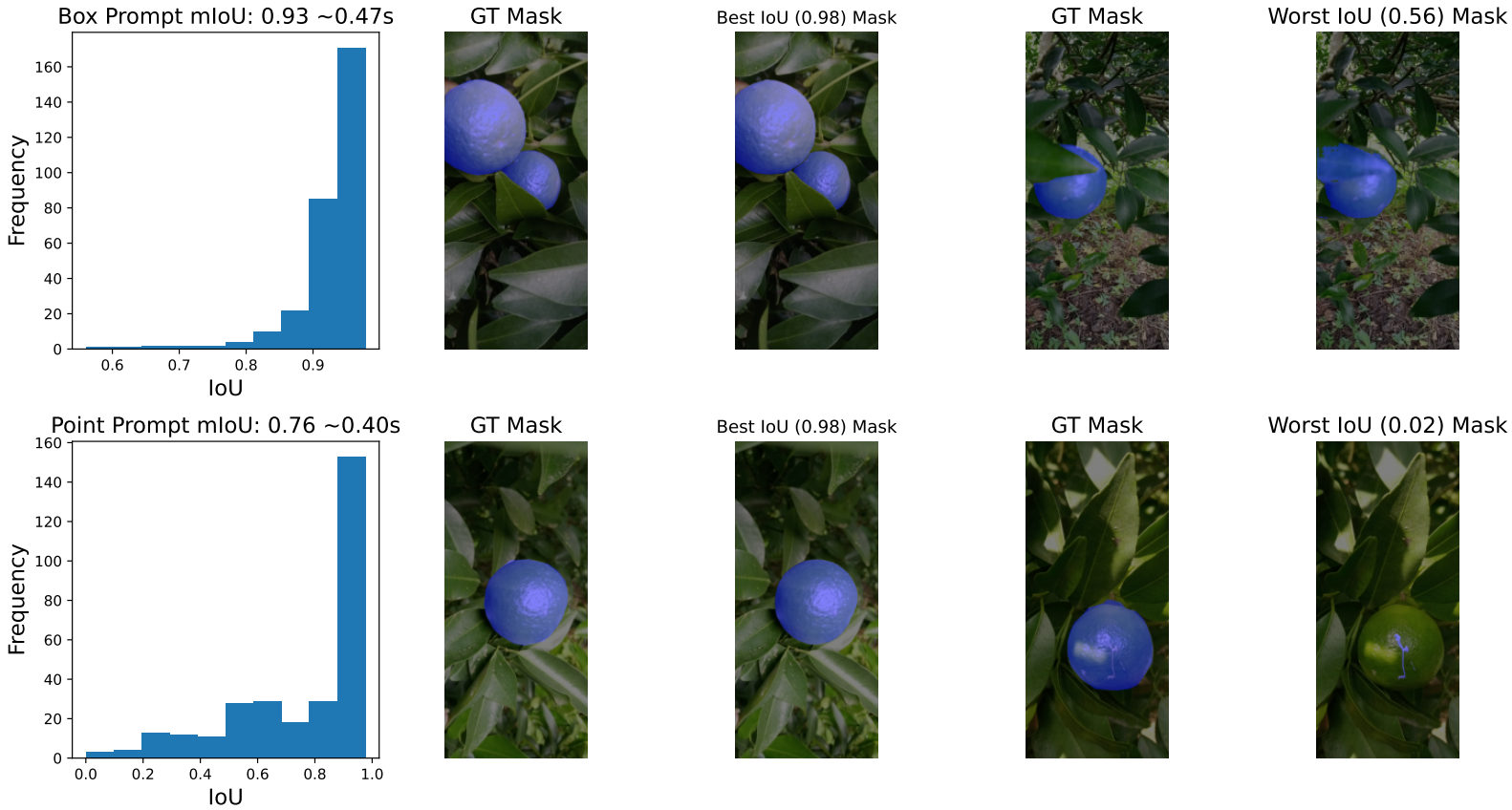


Figure: SAM's Zero-Shot Ability on Test Orange Dataset

The image's original size was Avg. (4160, 1800).

D. Pokharel and R. Acharya. *Orange Infection Mask Dataset*. 2024. DOI: 10.34740/KAGGLE/DSV/7742783

Results and Analysis

SAM2 [Rav+24] Zero-Shot on Orange Dataset

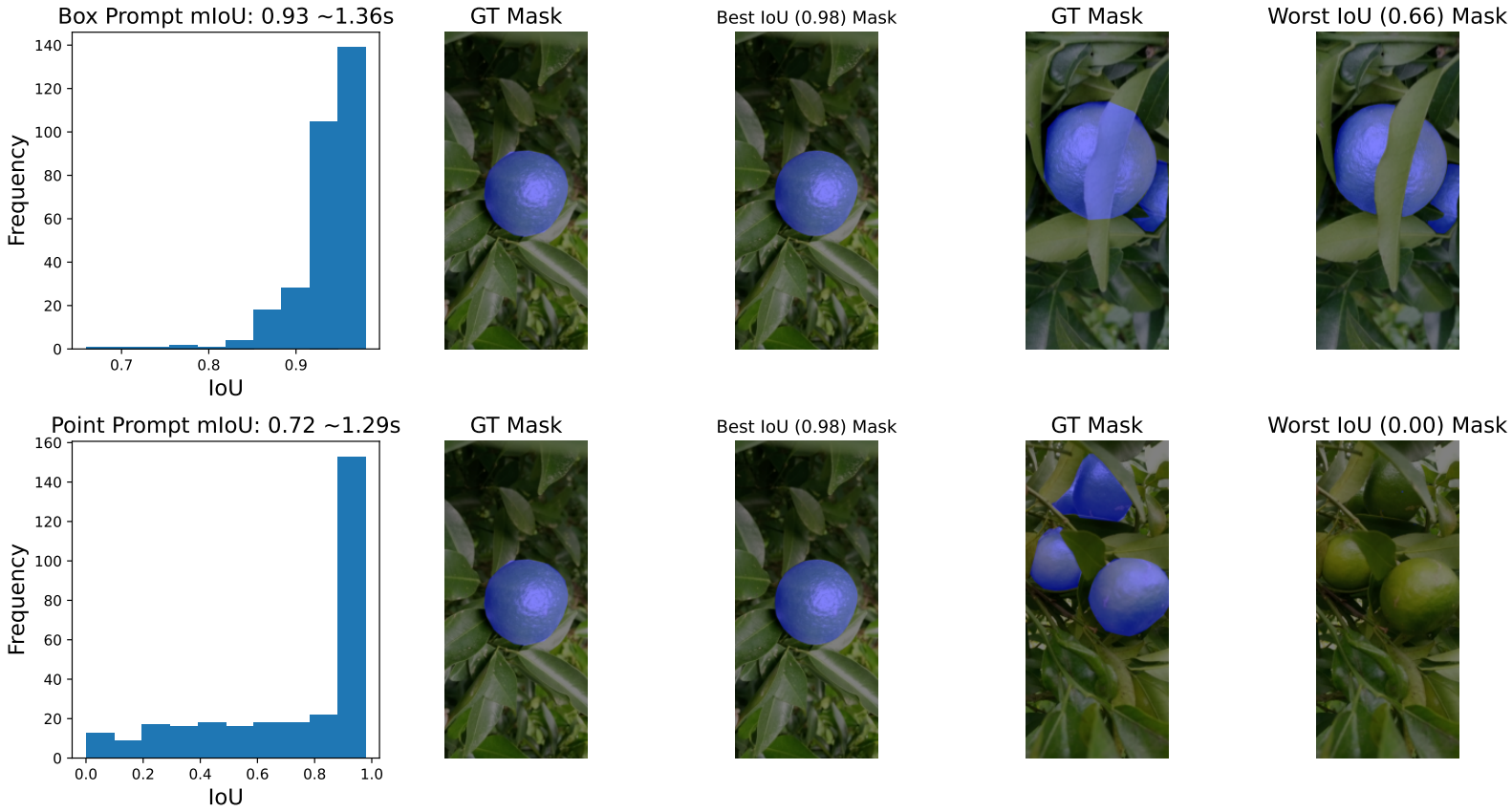


Figure: SAM2's Zero-Shot Ability on Test Orange Dataset

SAM2 is not showing significant improvements. Inference is made on Colab.

Results and Analysis

Training Curves

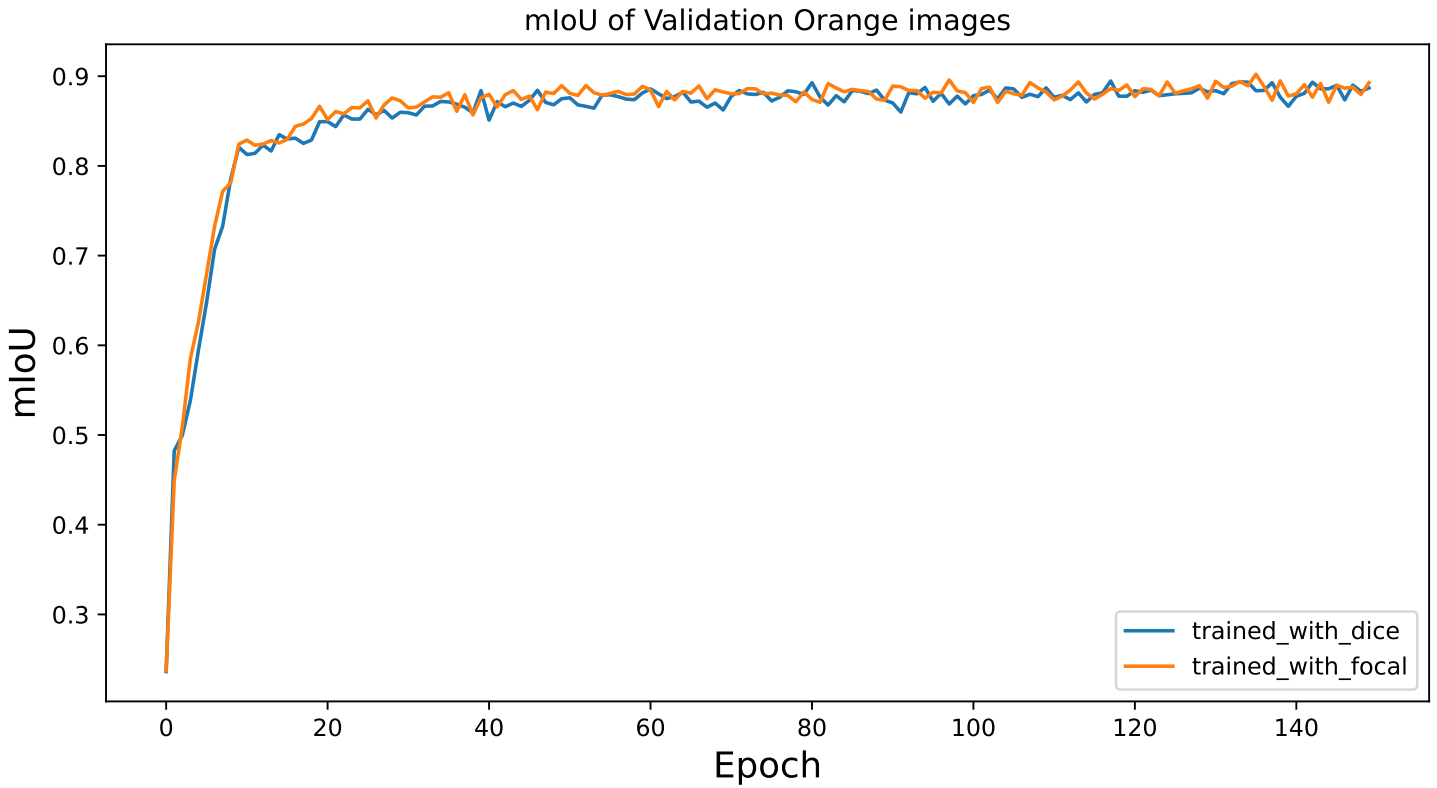


Figure: IoU of Validation Dataset using Unet

Trained for 150 epochs with ResNet18 encoder, 256 batch size, input size (448, 224), and Adam Optimizer with a LR of 1E-5.

Results and Analysis

U-Net's Performance

Figure: U-Net's Performance on the same dataset

Results and Analysis

Training Curves

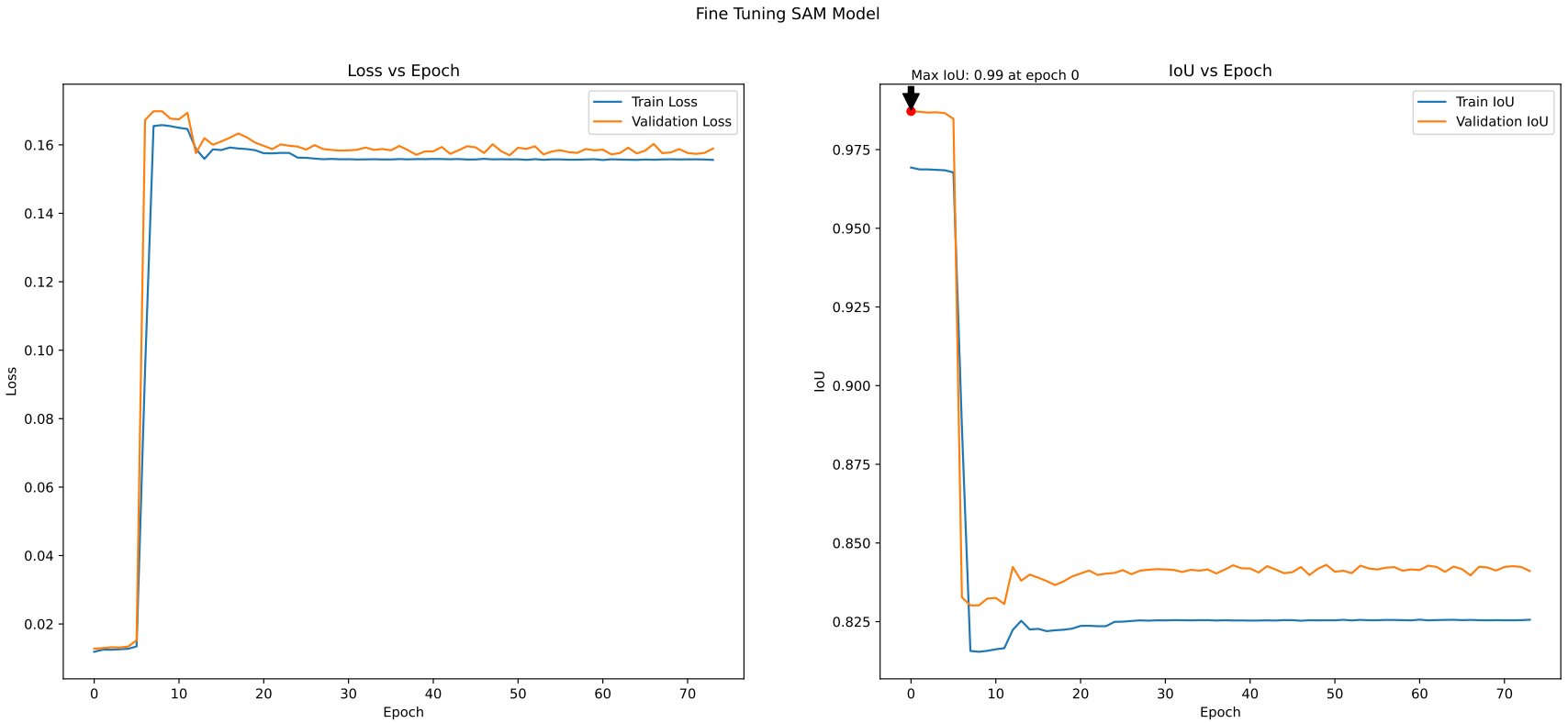


Figure: IoU of Validation Dataset using SAM

- Training config: 32 batch size, 70+ epochs, MSE loss, and Adam Optimizer.
- Trained for 10+ hours on Nvidia V100 on HPC.

Results and Analysis

Fine-tuned SAM's Results on Same Dataset

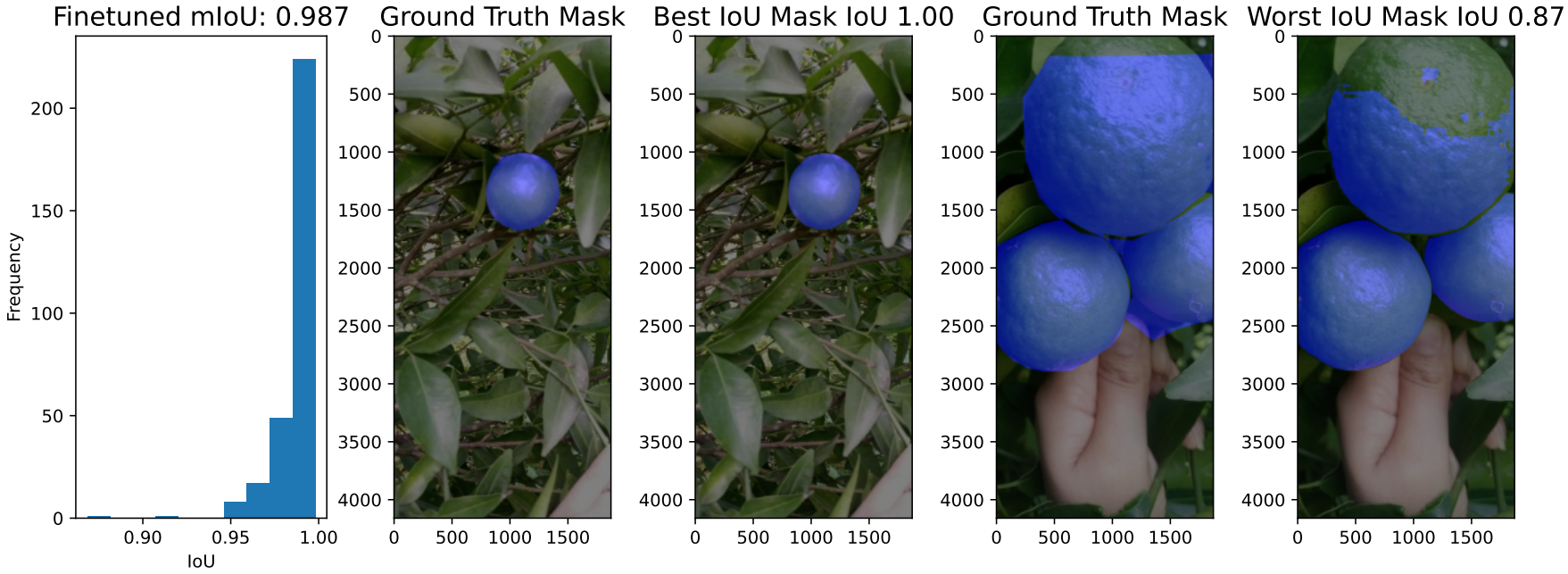


Figure: Fine-Tuned SAM on the same dataset

- Although Zeroshot predictions of SAM (with BBox) were already better than U-Net, fine-tuning improved it even more.

Discussion and Conclusion

SAM and Fine-tuning SAM

- + SAM is a powerful but still not feasible for edge devices.
- + Even without fine-tuning, it works better than standard U-Net for some tasks.
- - Fine-tuning is difficult without high-end resources.
- - Still needs all inputs to be resized as 1024 (applies ResizeLongestSide).

Evaluating Provided Code

github.com/facebookresearch/segment-anything

- + opensource
- + is easy to use and has instructions to run inferences
- + has 46k+ stars
- - does not guide how to reproduce results
- - does not provide training codes
- - does not provide text to mask inference
- - unittests and evaluation metrics are not provided

Finally

Thank you for your time.

References I

[ADF10] B. Alexe, T. Deselaers, and V. Ferrari. “What is an object?” In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2010, pp. 73–80. DOI: 10.1109/CVPR.2010.5540226.

[Bas+22] D. Bashkirova, M. Abdelfattah, Z. Zhu, J. Akl, F. Alladkani, P. Hu, V. Ablavsky, B. Calli, S. A. Bargal, and K. Saenko. “ZeroWaste dataset: Towards deformable object segmentation in cluttered scenes”. In: *CVPR*. 2022.

[Cai+19] J. C. Caicedo, A. Goodman, K. W. Karhohs, B. A. Cimini, J. Ackerman, M. Haghighi, C. Heng, T. Becker, M. Doan, C. McQuin, M. Rohban, S. Singh, and A. E. Carpenter. “Nucleus segmentation across imaging experiments: the 2018 data science bowl”. In: *Nature Methods* (2019).

[Can86] J. Canny. “A computational approach to edge detection”. In: *IEEE Transactions on pattern analysis and machine intelligence* PAMI-8.6 (1986), pp. 679–698.

[Che+22] J. Chen, Y. Xu, S. Lu, R. Liang, and L. Nan. “3D instance segmentation of MVS buildings”. In: *IEEE Transactions on Geoscience and Remote Sensing* (2022).

[Cia+21] L. Ciampi, C. Santiago, J. Costeira, C. Gennaro, and G. Amato. “Domain adaptation for traffic density estimation”. In: *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. 2021.

References II

[Cia+22] L. Ciampi, C. Santiago, J. Costeira, C. Gennaro, and G. Amato. “Night and day instance segmented park (NDISPark) dataset: a collection of images taken by day and by night for vehicle detection, segmentation and counting in parking areas”. In: *Zenodo* (2022).

[CNS22] N. Cohen, Y. Newman, and A. Shamir. “Semantic segmentation in art paintings”. In: *Computer Graphics Forum* (2022).

[Cor+16] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. “The Cityscapes dataset for semantic urban scene understanding”. In: *CVPR*. 2016.

[Dam+22] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, J. Ma, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray. “Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100”. In: *IJCV* (2022).

[Dar+22] A. Darkhalil, D. Shan, B. Zhu, J. Ma, A. Kar, R. Higgins, S. Fidler, D. Fouhey, and D. Damen. “EPIC-KITCHENS VISOR benchmark: Video segmentations and object relations”. In: *NeurIPS*. 2022.

[Dos+21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021.

References III

[FH04] P. F. Felzenszwalb and D. P. Huttenlocher. “Efficient graph-based image segmentation”. In: *International journal of computer vision* 59.2 (2004), pp. 167–181.

[For+22] J.-M. Fortin, O. Gamache, V. Grondin, F. Pomerleau, and P. Giguère. “Instance segmentation for autonomous log grasping in forestry operations”. In: *IROS*. 2022.

[FRR11] A. Fathi, X. Ren, and J. M. Rehg. “Learning to recognize objects in egocentric activities”. In: *CVPR*. 2011.

[GDG19] A. Gupta, P. Dollar, and R. Girshick. “LVIS: A dataset for large vocabulary instance segmentation”. In: *CVPR*. 2019.

[He+21] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. *Masked Autoencoders Are Scalable Vision Learners*. 2021.

[HFS20] J. Hong, M. Fulton, and J. Sattar. “TrashCan: A semantically-segmented dataset towards visual detection of marine debris”. In: *arXiv:2007.08097* (2020).

[HKS22] T. Haucke, H. S. Kühl, and V. Steinhage. “SOCRATES: Introducing depth in visual wildlife monitoring using stereo vision”. In: *Sensors* (2022).

[Kir+19] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár. *Panoptic Segmentation*. 2019.

[Kir+23] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick. *Segment Anything*. 2023.

References IV

[LH19] I. Loshchilov and F. Hutter. *Decoupled Weight Decay Regularization*. 2019.

[Li+22] Y. Li, Y. Chen, N. Wang, Z. Zhang, J. Sun, and J. Dai. “Exploring plain vision transformer backbones for object detection”. In: *arXiv preprint arXiv:2203.16527* (2022).

[Lin+14] T.-Y. Lin, M. M. Roth, P. Dollár, and J. Malik. “Microsoft COCO: Common Objects in Context”. In: *European Conference on Computer Vision (ECCV)* (2014), pp. 740–755. DOI: 10.1007/978-3-319-10602-1_48.

[Lin+15] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. *Microsoft COCO: Common Objects in Context*. 2015.

[Lin+18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. *Focal Loss for Dense Object Detection*. 2018.

[LYR15] Y. Li, Z. Ye, and J. M. Rehg. “Delving into egocentric actions”. In: *CVPR*. 2015.

[Min+16] M. Minervini, A. Fischbach, H. Scharr, and S. A. Tsaftaris. “Finely-grained annotated datasets for image-based plant phenotyping”. In: *Pattern Recognition Letters* (2016).

[PA24] D. Pokharel and R. Acharya. *Orange Infection Mask Dataset*. 2024. DOI: 10.34740/KAGGLE/DSV/7742783.

[PT22] M. Pugliatti and F. Topputo. *DOORS: Dataset fOr bOuldeRs Segmentation*. Zenodo. 2022.

[Pu+22] J. Pu, S. Zhu, W. Dong, C. Xu, and B. Zhou. “Edter: Edge detection with transformer”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 3748–3757.

References V

[Qi+22] J. Qi, Y. Gao, Y. Hu, X. Wang, X. Liu, X. Bai, S. Belongie, A. Yuille, P. Torr, and S. Bai. “Occluded video instance segmentation: A benchmark”. In: *ICCV*. 2022.

[Rad+21] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. *Learning Transferable Visual Models From Natural Language Supervision*. 2021.

[Rav+24] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer. “SAM 2: Segment Anything in Images and Videos”. In: *arXiv preprint arXiv:2408.00714* (2024).

[RFB15] O. Ronneberger, P. Fischer, and T. Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015.

[Rob+21] M. Roberts, J. Ramapuram, A. Ranjan, A. Kumar, M. A. Bautista, N. Paczan, R. Webb, and J. M. Susskind. “Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding”. In: *ICCV*. 2021.

[SD19] C. Snyder and M. Do. “STREETS: A novel camera network dataset for traffic flow”. In: *NeurIPS*. 2019.

[SPK21] K. Sofiiuk, I. A. Petrov, and A. Konushin. *Reviving Iterative Training with Mask Guidance for Interactive Segmentation*. 2021.

References VI

[Sud+17] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso. “Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations”. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer International Publishing, 2017, pp. 240–248. DOI: 10.1007/978-3-319-67558-9_28.

[Tan+20] M. Tancik, P. P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. T. Barron, and R. Ng. *Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains*. 2020.

[Tro+20] C. Trotter, G. Atkinson, M. Sharpe, K. Richardson, A. S. McGough, N. Wright, B. Burville, and P. Berggren. “NDD20: A large-scale few-shot dolphin dataset for coarse and fine-grained categorisation”. In: *arXiv:2005.13359* (2020).

[Vas+23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. *Attention Is All You Need*. 2023.

[Wan+21] B. Wang, L. Zhang, L. Wen, X. Liu, and Y. Wu. “Towards real-world prohibited item detection: A large-scale x-ray benchmark”. In: *CVPR*. 2021.

[XT15] S. Xie and Z. Tu. “Holistically-nested edge detection”. In: *Proceedings of the IEEE international conference on computer vision* (2015), pp. 1395–1403.

References VII

[Xu+16] N. Xu, B. Price, S. Cohen, J. Yang, and T. Huang. *Deep Interactive Object Selection*. 2016.

[Yan+21] L. Yang, Y. Z. Wei, Y. HE, W. Sun, Z. Huang, H. Huang, and H. Fan. “iShape: A first step towards irregular shape instance segmentation”. In: *arXiv:2109.15068* (2021).

[Yog+19] S. Yogamani, C. Hughes, J. Horgan, G. Sistu, P. Varley, D. O’Dea, M. Uricár, S. Milz, M. Simon, K. Amende, et al. “WoodScape: A multi-task, multi-camera fisheye dataset for autonomous driving”. In: *ICCV*. 2019.

[Zha+22] L. Zhang, S. Zhou, S. Stent, and J. Shi. “Fine-grained egocentric hand-object segmentation: Dataset, model, and applications”. In: *ECCV*. 2022.

[Zho+19] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba. “Semantic understanding of scenes through the ADE20K dataset”. In: *IJCV* (2019).