

The Missing Point in Vision Transformers for Universal Image Segmentation

Sajjad Shahabodini^{1*}, Mobina Mansoori^{1*}, Farnoush Bayatmakou¹, Jamshid Abouei²
Konstantinos N. Plataniotis³, Arash Mohammadi¹

¹I-SIP Lab @ Concordia University, ²Yazd University, ³University of Toronto

1. Upper Bound Analysis

As defined earlier, the upper bound results for a segmentation model represent the quality of the generated masks, while considering ideal mask classification accuracy. By analyzing these results, we can evaluate the model’s potential to produce precise segmentations regardless of classification errors. In the analysis of the upper bound results, we consider two distinct strategies for assigning labels to the generated masks: 1) leveraging mask proposals to determine the correct label by comparing them with ground truth masks, and 2) identifying the highest-value point within each generated mask and assigning the corresponding ground truth label based on that point. Presented in Table 1, the results show that the difference in accuracy between the two strategies is comparatively small. This finding highlights the efficacy of focusing on the most salient point within the mask, suggesting that a segmentation model can achieve robust performance even when label assignment relies on a single high-confidence region rather than the entire mask. Such observations underscore the feasibility of refining segmentation approaches by emphasizing key regions within masks while maintaining overall accuracy.

2. Mask-Prompting vs. Point-Prompting

While the ViT-P model classifies mask proposals based on input points, an alternative approach is to directly use mask proposals as inputs into a classification model. Prior works, such as MAFT [5], explored this approach by feeding the entire mask into a transformer model for labeling. In this section, we compare mask-based prompting with point-based prompting and explain why we favor point-based classification over using the complete mask proposal. To ensure a fair evaluation, both the ViT-P and MAFT [5] models were trained on the ADE20K [9] dataset using two different backbones, CLIP [8] and DinoV2 [7], with the Mask2Former [1] model serving as the mask generator.

As shown in Table 2, point-based classification, ViT-P, outperforms mask-based prompting, MAFT [5], across all

three segmentation tasks. By adding only a linear layer to the pre-trained transformer backbone, ViT-P preserves the original architecture. In contrast, the MAFT [5] approach integrates $12 - L$ additional randomly initialized attention layers into the L -layer backbone. This extra layer stack not only increases complexity but also limits the model’s adaptability when paired with standard vision transformer models such as CLIP [8] and DinoV2 [7] for dense prediction tasks, ultimately decreasing its performance.

In addition, mask-prompting models such as MAFT [5] suffer from mask quality during training. Typically, the number of generated masks exceeds the available ground truth masks, leading to the inclusion of additional low-quality mask proposals. Training on these generated masks reduces model accuracy, as overlapping masks introduce noise that negatively impacts learning efficiency. Conversely, training on high-quality ground truth masks creates an inconsistency between the training and inference, as the model is ultimately evaluated on imperfectly generated mask proposals during testing. These inaccuracies introduce inconsistencies that degrade classification performance. In contrast, the ViT-P model benefits from stable input points throughout both training and evaluation, ensuring more consistency. These findings highlight the limitations of mask-based prompting relative to point-based classification. Consequently, providing a strong justification for our decision to adopt a point-based approach in this study.

3. Limitations

While the proposed model demonstrates improved classification accuracy for generating mask proposals, a performance gap persists relative to the upper bound. This gap highlights the limitations of current mask classification strategies and indicates that further work is needed to enhance overall segmentation performance. Additionally, while ViT-P improves semantic segmentation performance, gains in instance and panoptic segmentation are marginal. Mask selection for individual objects is still a challenge for the model, highlighting the need for better instance selection methods.

Another limitation of the ViT-P model is its reliance on

*Equal contribution.

Table 1. Comparison of segmentation accuracy between point-based and mask-based label assignment strategies for OneFormer [4] on three benchmark datasets. The relatively small precision gap indicates that focusing on the most salient region within a mask is nearly as effective as evaluating the entire mask.

Model	Backbone	Classification Method	COCO mIoU	Cityscapes mIoU	ADE20K mIoU
OneFormer [4]	DiNAT-L [3]	Point Labeling	86.5	88.7	83.5
		Mask Labeling	87.3	89.7	84.7

Table 2. Comparison of point-based, ViT-P, and mask-based, MAFT [5], prompting on ADE20K [9] segmentation using CLIP [8] and DinoV2 [7] backbones.

Method	Classification Backbone	PQ	AP	mIoU
Mask2Former [1]	—	48.7	34.2	54.5
Mask2Former [1]+MAFT [5]	CLIP [8]	47.6	33.3	54.6
Mask2Former [1]+ViT-P		49.0	34.4	55.3
Mask2Former [1]+MAFT [5]	DinoV2 [7]	48.6	34.2	54.8
Mask2Former [1]+ViT-P		49.3	34.7	56.1

ensembling with the mask proposal model’s classification probabilities. Since mask generator models are trained on ground truth masks, mask proposal models tend to assign higher probability values to higher-quality masks, a nuance that is not present in point-based classifiers. Consequently, the ViT-P model achieves strong segmentation performance only when combined with the mask proposal’s probabilities. Eliminating this dependency would require more improvements in the point-based classification results, as the reported upper bound results are obtained without such ensembling.

4. Qualitative Results

We present segmentation examples on three benchmark datasets, ADE20K [9], COCO [6], and Cityscapes [2], illustrating ViT-P model’s performance, in Figure 1, 2, 3.

References

- [1] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 1, 2
- [2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2
- [3] Ali Hassani and Humphrey Shi. Dilated neighborhood attention transformer. *arXiv preprint arXiv:2209.15001*, 2022. 2
- [4] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2989–2998, 2023. 2
- [5] Siyu Jiao, Yunchao Wei, Yaowei Wang, Yao Zhao, and Humphrey Shi. Learning mask-aware clip representations for zero-shot segmentation. *Advances in Neural Information Processing Systems*, 36:35631–35653, 2023. 1, 2
- [6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision—ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 2
- [7] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 2
- [8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1, 2
- [9] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 1, 2

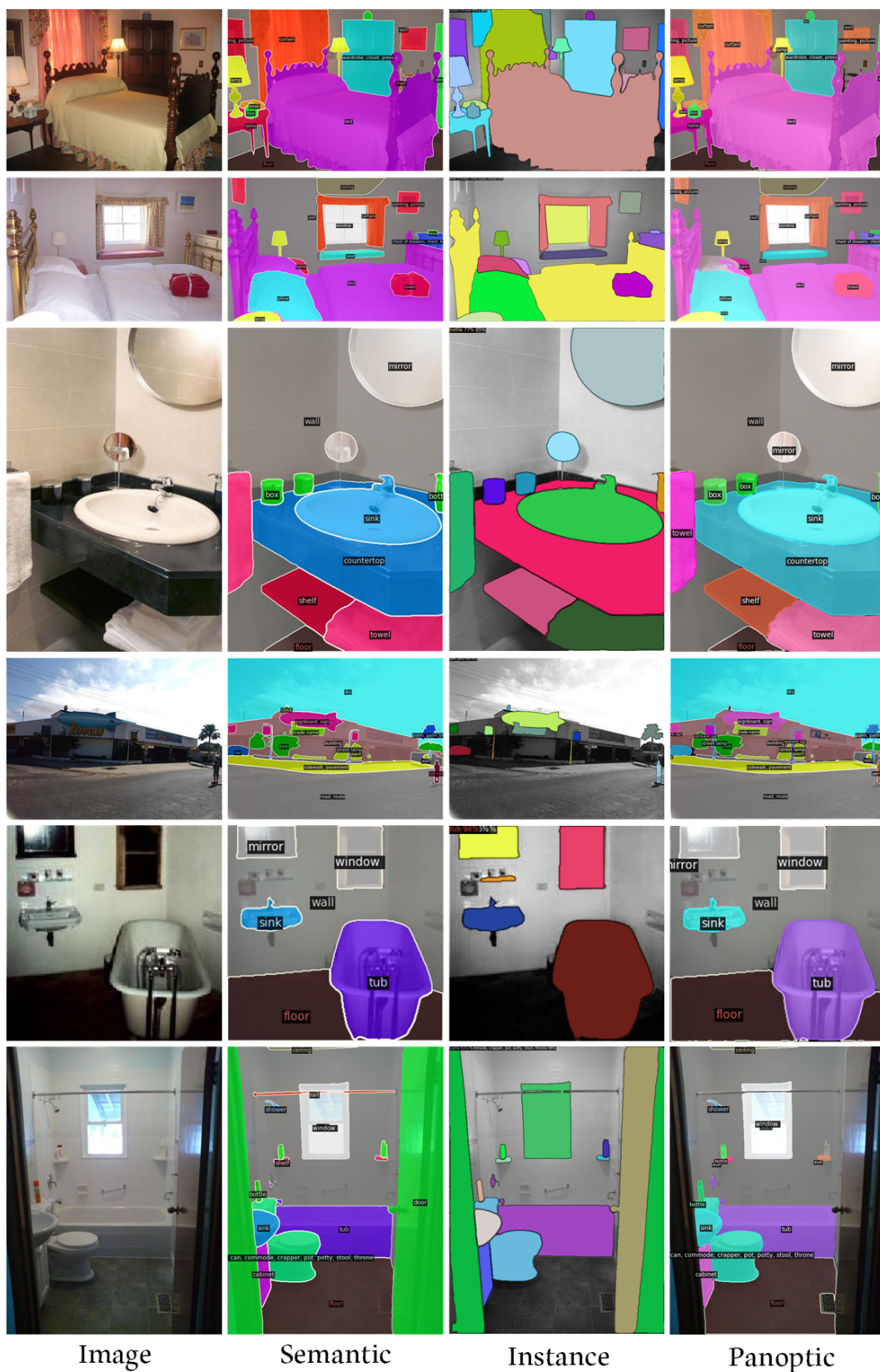


Figure 1. Segmentation examples on the ADE20K dataset.

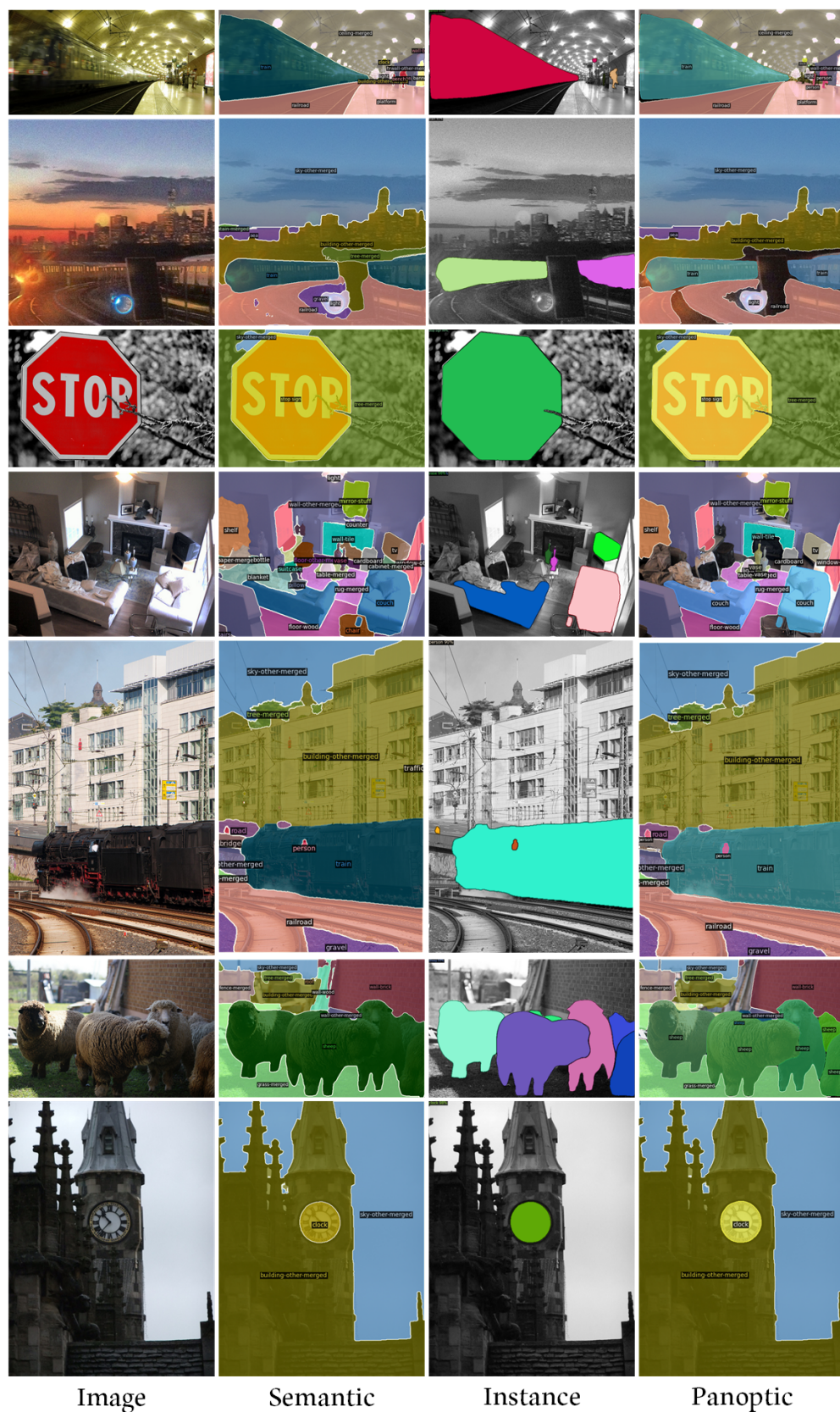


Figure 2. Segmentation examples on the COCO dataset.

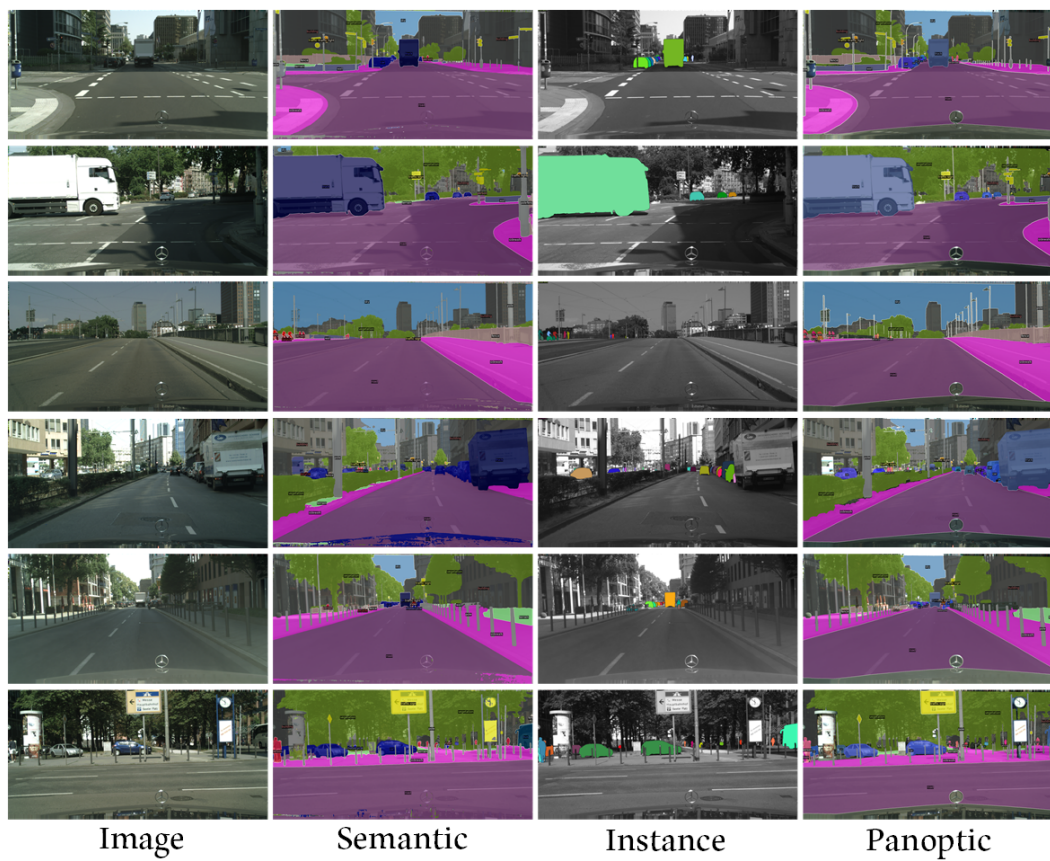


Figure 3. Segmentation examples on the Cityscapes dataset.