Can Segment Anything Model Improve Semantic Segmentation?

Maryam Qamar , Donghoon Kim, Muhammad Salman Ali ,
Chaoning Zhang, **Sung-Ho Bae**[*]
Kyung Hee University, Republic of Korea
{maryamqamar, dhkim2810, salmanali}@khu.ac.kr,
chaoningzhang1990@gmail.com, shbae@khu.ac.kr

Abstract

Recently, Segment Anything Model (SAM) has gained considerable attention in the field of computer vision establishing itself as a pioneering foundation model for segmentation. Notably, SAM excels in generating high-quality segmentation masks, yet it lacks in semantic labels. In contrast, conventional semantic segmentation models generate rather accurate semantic labels but often produce suboptimal segmentation masks. The notion of leveraging SAM's superior mask quality to enhance the performance of conventional semantic segmentation models appears intuitive. However, our preliminary experiments reveal that the integration of SAM with these models does not result in any discernible improvement. Specifically, when assessing the performance of SAM's integration into two baseline semantic segmentation models, DeepLab and OneFormer, we find no significant enhancements in the mean Intersection over Union (mIoU) on the Pascal VOC and ade20k datasets. Consequently, we conclude that, as it stands, the highly acclaimed foundational model is not the preferred solution for the semantic segmentation task. Instead, a more cautious and thoughtful approach is imperative to unlock any potential benefits in this context.

1 Introduction

Semantic segmentation is a fundamental computer vision task, which assigns a semantic category to each pixel in an image Everingham et al. [2015], Mottaghi et al. [2014]. This capability holds immense value across diverse industries, including medical diagnostics, autonomous driving, robotics etc Agrawal and Choudhary [2023], Cakir et al. [2022], Tzelepi and Tefas [2021]. In current era of deep learning, deep convolutional neural networks (DCNNs) have emerged as an effective approach for accomplishing semantic segmentation task. Their ability to learn complex visual features empowers them to achieve remarkable accuracy and efficiency in segmenting images, paving the way for numerous practical applications Cho et al. [2021], Kang et al. [2023]. However, when assessed qualitatively, as shown in Figure 1, many existing semantic segmentation models are observed to have deformed or irregular object boundaries not aligning well with the ground truth or overlapping to other close-by objects, resulting in performance degradation on complex datasets. Nevertheless, in spite of this apparent drawback, pertaining to their supervised training, the predicted category labels for object classes are more often accurate.

Recently introduced foundation segmentation model, segment anything (SAM) Kirillov et al. [2023], by Meta AI Research, is designed to excel in generalized segmentation tasks. This model is shown to have the ability to segment any object in a given image and remarkable zero-shot transfer capabilities. Regardless, while SAM is observed to segment objects with very refined borders it does not provide any class categories for the segmented objects.

[*] Corresponding Authors

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

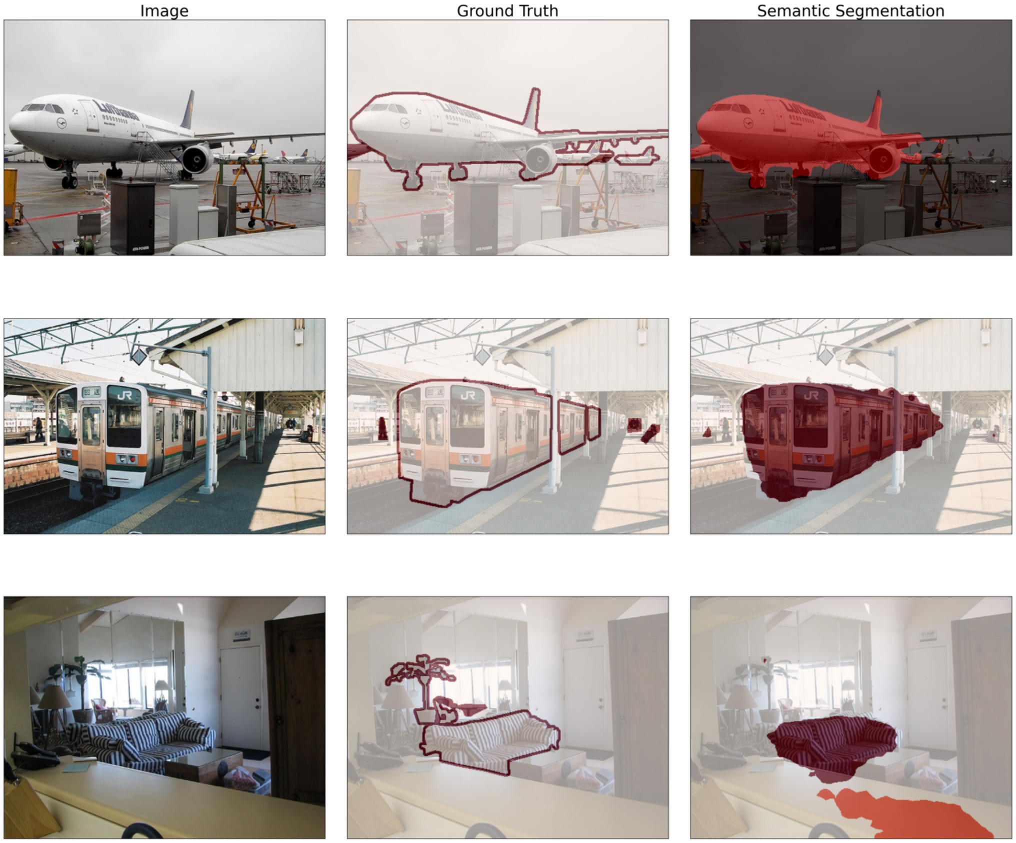

Figure 1: Example Outputs of Semantic Segmentation Featuring Noticeable Irregular Masks. From left to right: Original image, Ground truth, Deeplab output.

These observations regarding the apparent strengths and lacks in semantic segmentation models and SAM motivated us to combine the best of both in a low compute intensive way to empirically analyze if SAM can boost the semantic segmentation performance without further training or fine-tuning.

2 Related Work

Conventional Semantic Segmentation Models Early works based on fully convolutional networks (FCNs) Long et al. [2015] utilize an encoder-decoder structure. The design employs DCNNs across the entire image, transforming the last fully connected layers into convolutional layers. This modification allows the network to produce dense pixel-wise predictions for semantic segmentation.

However, there is potential loss of information concerning smaller-scale objects as the feature map undergoes downsampling through the DCNNs Chen et al. [2017]. Addressing this, DeepLab Chen et al. [2018] introduced atrous convolutions, preserving essential information for objects of varying scales while extracting dense features Giusti et al. [2013]. A more recent segmentation model, OneFormer Jain et al. [2023] adopts a transformer-based architecture that employs a task-conditioned joint training strategy, leading to a universal segmentation model capable of handling semantic, instance, and panoptic segmentation tasks. Remarkably, it achieves state-of-the-art performance across all three segmentation tasks.

Segment Anything Model (SAM) Segment anything Kirillov et al. [2023] presents an innovative and generalized segmentation task termed as promptable segmentation. The task involves generating a valid segmentation mask in response to any provided prompt. In other words, the proposed segmentation model SAM, aims to produce an accurate and meaningful segmentation result based on user-defined prompts, allowing for flexible and interactive segmentation capabilities. In detail, a prompt in SAM can be anything from a group of foreground or background points to a rough

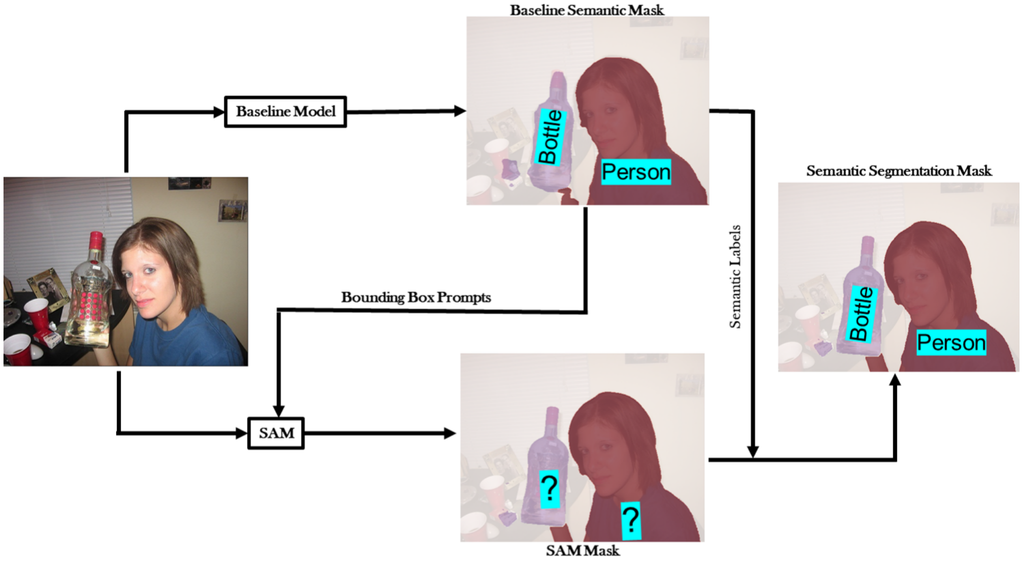

Figure 2: Proposed Framework: Bounding box prompts computed from baseline segmentation results are input to SAM with original images to generate final semantic segmentation

mask, approximate bounding box, or even free-form text. This versatility should enable researchers to provide input in different formats to guide the segmentation process. Thus it is anticipated that the capabilities of generalized segmentation and accommodation of different types of prompts should make this foundation model more adaptable towards specific segmentation tasks like semantic segmentation.

3 Proposed Methodology

In this paper, we ask a fundamental question, given SAM's acclaimed potential for image segmentation as a foundation segmentation model, can SAM boost the performance of the conventional semantic segmentation models? To answer this, we conduct a preliminary analysis. We utilize DeepLab and OneFormer as our baseline semantic segmentation models. And leveraging the promptable nature of SAM, we adopt a simple yet efficient approach, a conceptual overview is shown in Figure 2, to segment objects of interest in an image.

Specifically, segmentation masks are generated by giving a bounding box prompt to SAM, which was acquired via semantic masks of the baseline segmentation model, for each object class in a given image. Once we have a segmentation mask per object class, they are combined to create the final semantic mask. This new output mask, obtained by integrating SAM with a semantic segmentation baseline model, is expected to have refined segmentation masks attributing to the acclaimed quality and generalization power of SAM. In detail, we first obtain a semantic prediction from the baseline model and draw a bounding box on each object class. These bounding boxes are then passed as a prompt to the SAM model along with the original image to predict candidate segmentation masks for each corresponding object. Once these candidates masks are generated, most closely related mask for each particular object is selected based on the maximum IoU similarity between the baseline mask and the candidate masks. The employment of IoU similarity here is naive yet intuitive as the task of choosing the best masks from the candidate masks is intrinsically same as the actual segmentation task. Final semantic segmentation output is then computed as an aggregation of the selected masks for all the classes in a given image. The procedure is summarized in Algorithm 1.

4 Implementation Details

All experiments are conducted on a Nvidia RTX 3090 GPU, keeping hyperparameter settings as recomended in the baseline models. The datasets used for evaluation also correspond to the baselines, specifically Pascal VOC 2012 Everingham et al. for Deeplab and ade20k Zhou et al. [2017, 2019] for OneFormer. However, for the purpose of gaining initial insights regarding the potential of SAM for semantic segmentation task, we only conduct the experiments on a randomly selected subset of each dataset, specifically per dataset 100 images are utilized.

Algorithm 1 Boosting Semantic Segmentation with SAM

```
Input: Image
Output: Semantic Segmentation Mask
begin
BaselinePred $\leftarrow$ BaselineModel(image)
BoundingBoxes $\leftarrow$ drawBoundingBox(BaselinePred)
for each object class in BaselinePred do
    candidateMasks $\leftarrow$ SAM(image, BoundingBox)
    iouVector $\leftarrow$ calIOU(candidateMasks, BaselinePred)
    maxIOU $\leftarrow$ max(iouVector)
    selectedMask $\leftarrow$ selectMask(maxIOU)
end for
SemanticMask $\leftarrow$ aggregateMasks(selectedMask)
return SemanticMask
```

5 Evaluation Metrics

The most commonly used evaluation metric in semantic segmentation tasks is **mean Intersection over Union (mIoU)**. It measures the extent of overlap between the predicted segmentation mask and the ground truth mask Minaee et al. [2021]. IoU is calculated as:

$$IoU = \frac{1}{M+1} \sum_{i=0}^{M} \frac{|p_i \cap p_j|}{|p_i \cup p_j|} \tag{1}$$

where p_i and p_j are predicted and ground truth segmentation masks. IoU scores are then averaged across all classes M to get mean IoU, providing an overall measure of segmentation accuracy. **Mean Accuracy (mAcc.)** is used to assess the accuracy of pixel-level predictions. It measures the proportion of correctly classified pixels p_{ii} over the total number of pixels in an image averaged over classes M Minaee et al. [2021] as shown in equation 2:

$$mAcc = \frac{1}{M+1} \sum_{i=0}^{M} \frac{p_{ii}}{\sum_{j=0}^{M} p_{ij}} \tag{2}$$

6 Results and Analysis

Table 1: Semantic Segmentation Results

Method	**mIoU**	**mAcc.**
Deeplab	**75.75**	82.75
Deeplab+SAM	75.31	**84.21**
OneFormer	**40.99**	**65.09**
OneFormer+SAM	40.85	64.61

Table 1 shows the quantitative results, contrary to the expectations, integrating SAM fell short of boosting the performance of the two baseline methods in our experiments. Moreover, the performance is marginally dropped in comparison to the baseline. This failure can possibly be attributed to the existence of incomplete, multiple and overlapping objects in the prompt as shown in Figure 3. This raises the concern that SAM requires rather rigorous prompts to perform well for the semantic segmentation task. For further analysis, we qualitatively evaluate the segmentation masks of baseline (Deeplab) model versus the model boosted with SAM against the ground truth for randomly picked images from Pascal VOC 2012 in Figure 4. It seems evident that the boosted model outputs segmentation masks with very well defined boundaries in comparison to the baseline model's segmentation boundaries, for example, the train objects in the first example, the boat and cyclists in the second and third rows respectively, and the rider and the plane in the sixth and last examples respectively. Nonetheless, noticeable issues arise in the segmentation output when employing SAM. In the first example, noise becomes evident, while in the subsequent five examples, certain objects remain entirely unsegmented. The final two examples exhibit incomplete segmentation, with

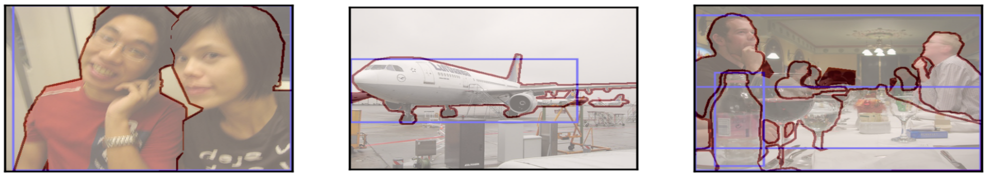

Figure 3: Failure Modes of SAM in Semantic Segmentation: Images from Pascal VOC 2012 overlaid with ground truth in red and box prompts acquired via baseline model (DeepLab) in blue.

portions of objects left out of the segmentation mask. In summary, both quantitative and qualitative analysis lead to debunking the illusion that simply utilizing a segmentation foundation model like SAM can help in boosting semantic segmentation performance. Admittedly, the current integration approach is simplistic, and a more refined method may be conducive to enhancing the performance of conventional semantic segmentation models. Nevertheless, when considered as a foundational model for segmentation, SAM's effectiveness in the context of semantic segmentation remains unproven, underscoring the need for further analysis.

7 Ablation Study

We further evaluated a variety of prompts and their combinations to asses their suitability for semantic segmentation task. These experiments are run using Deeplab as the baseline model. Points are computed as the centroids of the connected regions from the baseline segmentation output, mask prompt is the baseline output mask. Mean IoU values for diffent types of prompts is shown in Table 2. Single point prompt performance is very low and multiple points prompt while still lower than baseline is comparatively closer. This can be attributed to the presence of multiple objects in an image, SAM would presumably segment a single object when using a single point prompt degrading the performance. Bounding box prompts in combination with other prompts gives the best performance but this too can not surpass the baseline performance yet. A potential explanation for this behavior could be the presence of erroneous prompts, which might not accurately target the intended object. We intend to conduct a more thorough examination of these prompts and their performance characteristics, exploring ways to optimize them in order to harness SAM's full potential for semantic segmentation.

Table 2: Comparative Analysis of Prompt Types

Metric	mIoU	mAcc.
Baseline	**75.75**	82.75
Single Point	50.34	79.06
Multiple Point	64.30	74.09
Bbox	72.60	77.96
Points & Bbox	75.31	*82.96*
Points, Bbox & Mask	*75.50*	**84.27**

8 Discussion

This study serves as an initial exploration to assess the potential of SAM as a zero-shot foundational model in enhancing semantic segmentation performance of conventional models. The rationale behind this lies in the notable quality of SAM generated masks, characterized by refined borders. Given the baseline methods' near accurate class label predictions, SAM generated masks are anticipated to improve the overall accuracy of prediction in semantic segmentation. However, empirical findings and qualitative evidence indicate that SAM, at present, does not boost the performance without carefully engineered prompts. Thus, we conclude that regarding its role as a zero-shot foundational model for segmentation, SAM's efficacy in the realm of semantic segmentation remains yet unverified, emphasizing the necessity for additional insights.

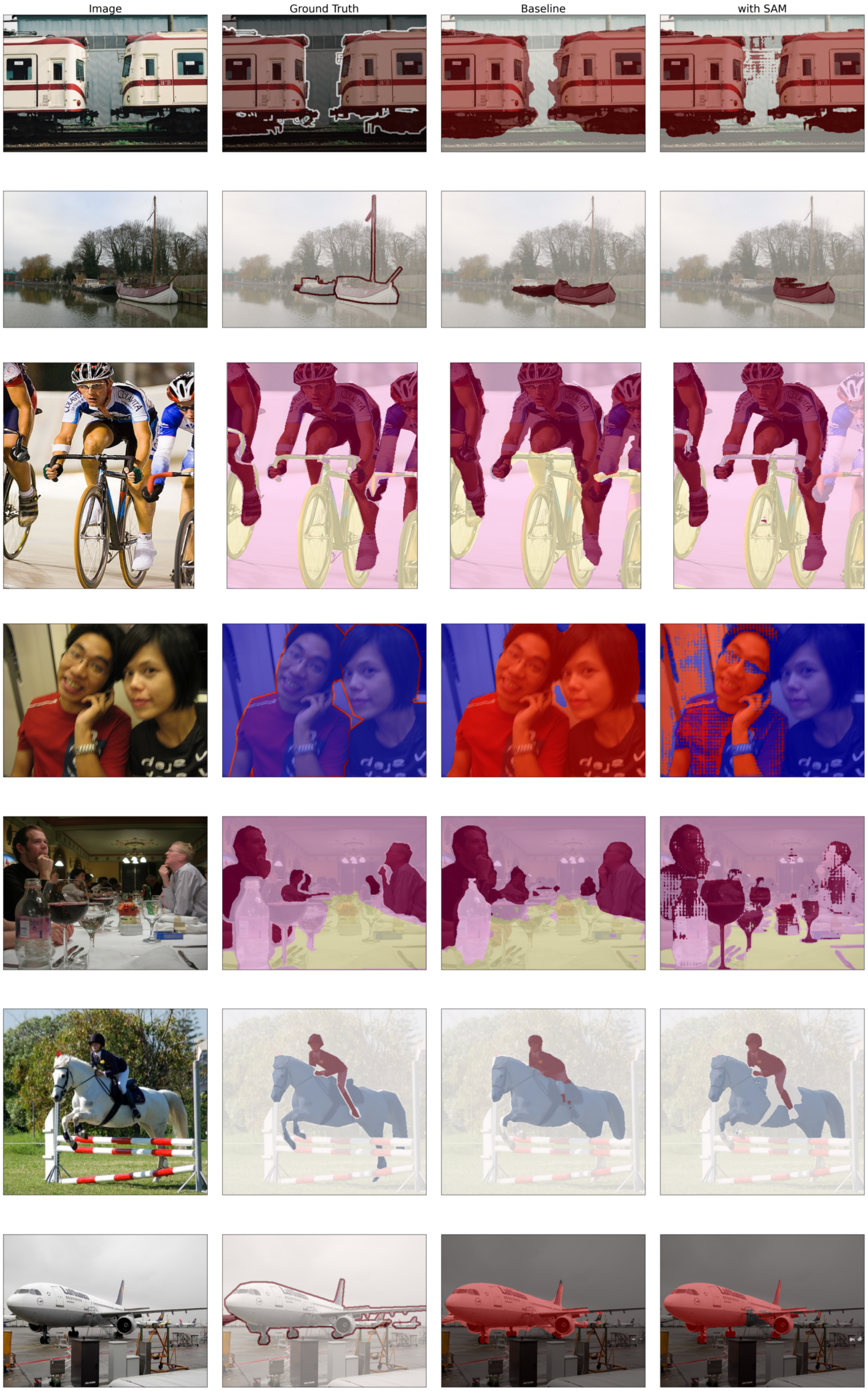

Figure 4: Qualitative Analysis of Baseline and Baseline+SAM: From left to right: Original image, Ground truth, Deeplab output, Deeplab+SAM output

Acknowledgments and Disclosure of Funding

This research was supported in part by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-RS-2023-00258649) supervised by the IITP (Institute for Information Communications Technology Planning Evaluation), and in part by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-RS-2023-00259004) supervised by the IITP (Institute for Information Communications Technology Planning Evaluation).

References

Tarun Agrawal and Prakash Choudhary. Covid-segnet: encoder–decoder-based architecture for covid-19 lesion segmentation in chest x-ray. *Multimedia Systems*, pages 1–14, 2023.

Senay Cakir, Marcel Gauß, Kai Häppeler, Yassine Ounajjar, Fabian Heinle, and Reiner Marchthaler. Semantic segmentation for autonomous driving: Model evaluation, dataset generation, perspective comparison, and real-time capability. *arXiv preprint arXiv:2207.12939*, 2022.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

Incheon Cho, Eunseop Shin, Muhammad Salman Ali, and Sung-Ho Bae. Dynamic structured pruning with novel filter importance and leaky masking based on convolution and batch normalization parameters. *IEEE Access*, 9:165005–165013, 2021.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111: 98–136, 2015.

Alessandro Giusti, Dan C Cireşan, Jonathan Masci, Luca M Gambardella, and Jürgen Schmidhuber. Fast image scanning with deep max-pooling convolutional neural networks. In *2013 IEEE International Conference on Image Processing*, pages 4034–4038. IEEE, 2013.

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.

Jung-Heum Kang, Muhammad Salman Ali, Hye-Won Jeong, Chang-Kyun Choi, Younhee Kim, Se Yoon Jeong, Sung-Ho Bae, and Hui Yong Kim. A super-resolution-based feature map compression for machine-oriented video coding. *IEEE Access*, 11:34198–34209, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542, 2021.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014.

Maria Tzelepi and Anastasios Tefas. Semantic scene segmentation for robotics applications. In *2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA)*, pages 1–4. IEEE, 2021.

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.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019.