Rebuttal for the ICML 2024 submission: Adaptive Hierarchical Certification for Segmentation using Randomized Smoothing

In the following document, we share tables, figures and images that address the comments of the reviewers **doKT**, **FNU9**, **EAKe**, and **BFDX**. Note that this is a supplementary document that we use to include figures which we reference in the rebuttal comment on OpenReview: https://openreview.net/forum?id=iOEReiiTitnoteId=iOEReiiTit.

1. AdaptiveCertify Performance on 4 datasets

1.1. Cityscapes

Table 1: Certified segmentation results for 200 images from the Cityscapes dataset. CIG and cCIG stand for per-pixel certified info. gain and class-average certified info. gain, respectively. $\%\oslash$ is the abstain rate, and $c\%\oslash$ is the class-average abstain rate. Our method AdaptiveCertify and SegCertify use $n_0 = 10, \alpha = 0.0001$, and their CIG is certified within a radius R at different noise levels σ, thresholds τ, and number of samples n. We follow the same setup for the following tables.

				Cityscapes				
		σ	R	CIG ↑	cCIG ↑	$\%\oslash$ ↓	$c\%\oslash$ ↓	mIoU ↑
	Uncertified HrNet	-	-	0.91	0.72	–	–	0.52
$n = 100$, $\tau = 0.75$	SegCertify	0.25	0.17	0.88	0.54	10	38	0.52
		0.33	0.22	0.75	0.40	22	53	0.39
		0.50	0.34	0.29	0.08	43	50	0.05
	AdaptiveCertify	0.25	0.17	**0.89** 1.1%	**0.57** 5.6%	**8** 20.0%	**32** 15.8%	–
		0.33	0.22	**0.77** 2.7%	**0.43** 7.5%	**18** 18.2%	**46** 13.2%	–
		0.50	0.34	**0.33** 13.8%	**0.10** 25.0%	**31** 27.9%	**36** 28.0%	–
$n = 500$, $\tau = 0.95$	SegCertify	0.25	0.41	0.85	0.50	13	45	0.52
		0.33	0.52	0.70	0.36	28	61	0.50
		0.50	0.82	0.26	0.07	53	62	0.05
	AdaptiveCertify	0.25	0.41	**0.86** 1.2%	**0.53** 6.0%	**11** 15.4%	**40** 11.1%	–
		0.33	0.52	**0.72** 2.9%	**0.38** 5.6%	**25** 10.7%	**56** 8.2%	–
		0.50	0.82	**0.30** 15.4%	**0.09** 28.6%	**42** 20.8%	**52** 16.1%	–

1.2. ACDC

Table 2: Certified segmentation results for 200 images from the ACDC dataset.

				ACDC				
		σ	R	CIG ↑	cCIG ↑	%⊘ ↓	c%⊘ ↓	mIoU ↑
	Uncertified HrNet	-	-	0.56	0.38	–	–	0.15
$n = 100$, $\tau = 0.75$	SEGCERTIFY	0.25	0.17	0.54	0.25	35	60	0.21
		0.33	0.22	0.44	0.18	44	67	0.15
		0.50	0.34	0.18	0.09	42	52	
	ADAPTIVECERTIFY	0.25	0.17	**0.56** 3.7%	**0.29** 16.0%	**29** 17.1%	**47** 21.7%	–
		0.33	0.22	**0.46** 4.5%	**0.21** 16.7%	**36** 18.2%	**57** 14.9%	–
		0.50	0.34	**0.21** 16.7%	**0.11** 22.2%	**32** 23.8%	**34** 34.6%	–
$n = 500$, $\tau = 0.95$	SEGCERTIFY	0.25	0.41	0.50	0.22	42	67	0.22
		0.33	0.52	0.39	0.15	53	75	0.15
		0.50	0.82	0.16	0.08	51	63	0.04
	ADAPTIVECERTIFY	0.25	0.41	**0.52** 4.0%	**0.26** 18.2%	**37** 11.9%	**57** 14.9%	–
		0.33	0.52	**0.41** 5.1%	**0.17** 13.3%	**46** 13.2%	**67** 10.7%	–
		0.50	0.82	**0.18** 12.5%	**0.09** 12.5%	**42** 17.6%	**47** 25.4%	–

1.3. COCO-Stuff-10K

Table 3: Certified segmentation results for 50 images from the COCO-Stuff-10k dataset.

				COCO-Stuff-10k				
		σ	R	CIG ↑	cCIG ↑	%⊘ ↓	c%⊘ ↓	mIoU ↑
	Uncertified HrNet	-	-	0.66	0.25	–	–	0.19
$n = 100$, $\tau = 0.75$	SEGCERTIFY	0.25	0.17	0.57	0.36	24	31	0.19
		0.33	0.22	0.51	0.31	27	37	0.17
		0.50	0.34	0.32	0.15	46	57	0.08
	ADAPTIVECERTIFY	0.25	0.17	**0.60** 5.3%	**0.39** 8.3%	**15** 37.5%	**22** 29.0%	–
		0.33	0.22	**0.53** 3.9%	**0.33** 6.5%	**18** 33.3%	**25** 32.4%	–
		0.50	0.34	**0.35** 9.4%	**0.18** 20.0%	**32** 30.4%	**40** 29.8%	–
$n = 500$, $\tau = 0.95$	SEGCERTIFY	0.25	0.41	0.51	0.31	36	46	0.19
		0.33	0.52	0.44	0.25	42	54	0.17
		0.50	0.82	0.26	0.10	62	74	0.07
	ADAPTIVECERTIFY	0.25	0.41	**0.54** 5.9%	**0.33** 6.5%	**28** 22.2%	**38** 17.4%	–
		0.33	0.52	**0.46** 4.5%	**0.27** 8.0%	**34** 19.0%	**45** 16.7%	–
		0.50	0.82	**0.29** 11.5%	**0.13** 30.0%	**50** 19.4%	**63** 14.9%	–

We show the certfied segmentation results of ADAPTIVECERTIFY against SEGCERTIFY on the Common Objects in COntext-Stuff (COCO-Stuff) (Caesar et al., 2018) dataset in Table 3. The COCO-Stuff dataset is a scene understanding dataset, that is an extension of the COCO dataset (Lin et al., 2014), which addresses the segmentation of pixels as either thing or stuff classes. It has 172 categories (80 things, 91 stuff and 1 unlabelled), from which we use the common evaluation mode of only using 171 categories while excluding the unlabelled class. We use the COCO-Stuff-10K v1.1 subset of the dataset, which contains a 9k and 1k train and validation splits. For the DAG hierarchy, we used the pre-defined hierarchy for things

and stuff officially provided by the dataset (Caesar et al., 2018). The model used is HrNetV2 (Wang et al., 2020; Yuan et al.), which we train on a Gaussian noise of $\sigma = 0.25$ and outline the details in Section 5. The threshold function T_{thresh} parameters used, which are found via a grid, are $(0, 0.3, 0.7)$.

1.4. PASCAL-Context

We show the certfied segmentation results of ADAPTIVECERTIFY against SEGCERTIFY on the PASCAL-Context dataset (Mottaghi et al., 2014). The PASCAL-Context dataset is a scene understanding dataset which contains 59 foreground and 1 unlabelled classes. We use the evaluation mode similar to (Fischer et al., 2021) where we only consider the 59 classes. We also use HrNetV2 (Wang et al., 2020; Yuan et al.) trained on a Gaussian noise of $\sigma = 0.25$ provided by (Fischer et al., 2021). We design and use a semantic hierarchy on top of the 59 classes in Figure 4. The threshold function T_{thresh} parameters, which are found via a grid, are $(0, 0.1, 0.4)$.

Table 4: Certified segmentation results for 100 images from the PASCAL-Context dataset.

				PASCAL-Context				
		σ	R	CIG ↑	cCIG ↑	%⊘ ↓	c%⊘ ↓	mIoU ↑
	Uncertified HrNet	-	-	0.55	0.20	–	–	0.18
$n = 100$, $\tau = 0.75$	SEGCERTIFY	0.25	0.17	0.56	0.26	21	29	0.20
		0.33	0.22	0.46	0.23	30	37	0.18
		0.50	0.34	0.13	0.05	39	45	0.02
	ADAPTIVECERTIFY	0.25	0.17	**0.57** 1.8%	**0.29** 11.5%	**16** 23.8%	**22** 24.1%	–
		0.33	0.22	**0.48** 4.3%	**0.25** 8.7%	**24** 20.0%	**30** 18.9%	–
		0.50	0.34	**0.14** 7.7%	**0.06** 20.0%	**34** 12.8%	**38** 15.6%	–
$n = 500$, $\tau = 0.95$	SEGCERTIFY	0.25	0.41	0.51	0.23	33	45	0.18
		0.33	0.52	0.39	0.19	45	55	0.13
		0.50	0.82	0.09	0.04	59	65	0.02
	ADAPTIVECERTIFY	0.25	0.41	**0.53** 3.9%	**0.25** 8.7%	**28** 15.2%	**38** 15.6%	–
		0.33	0.52	**0.41** 5.1%	**0.21** 10.5%	**40** 11.1%	**50** 9.1%	–
		0.50	0.82	**0.10** 11.1%	**0.04**	**55** 6.8%	**62** 4.6%	–

2. Fluctuations of unstable components

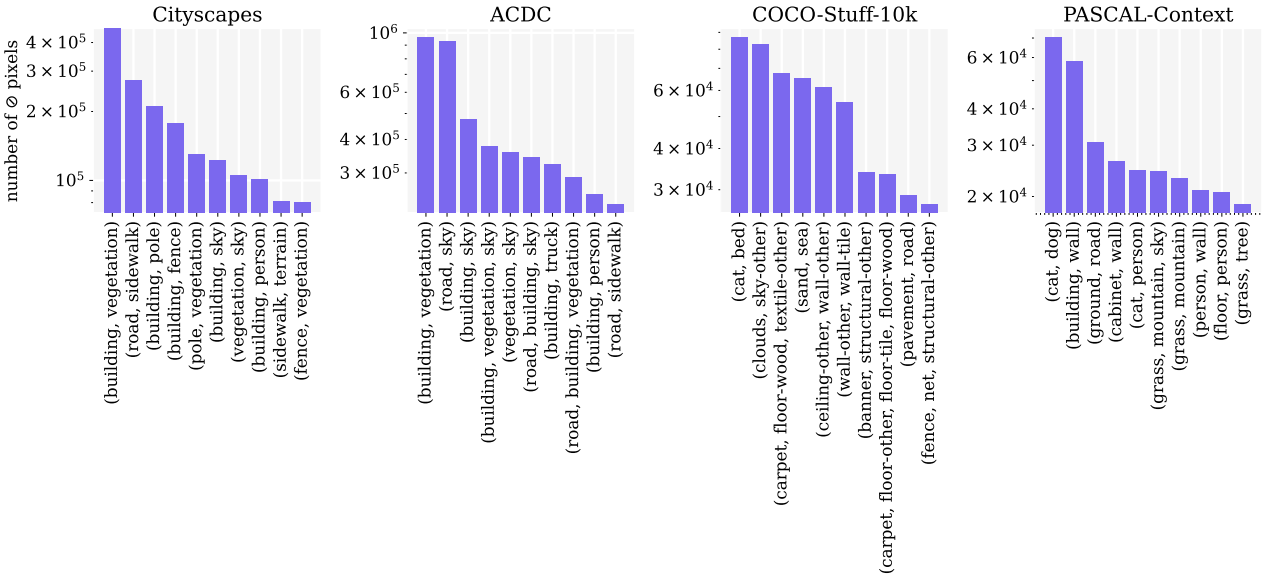

Figure 1: The frequency of the sets of classes the model fluctuates between in abstain pixels across 4 datasets: Cityscapes, ACDC, COCO-Stuff-10K and PASCAL-Context. The y-axis is a log scale. This is the result of running on 40 images per dataset, $n = 100$, $n_0 = 10$, $\sigma = 0.25$ and $\tau = 0.75$.

3. Certification rate and certified information gain (CIG) tradeoff

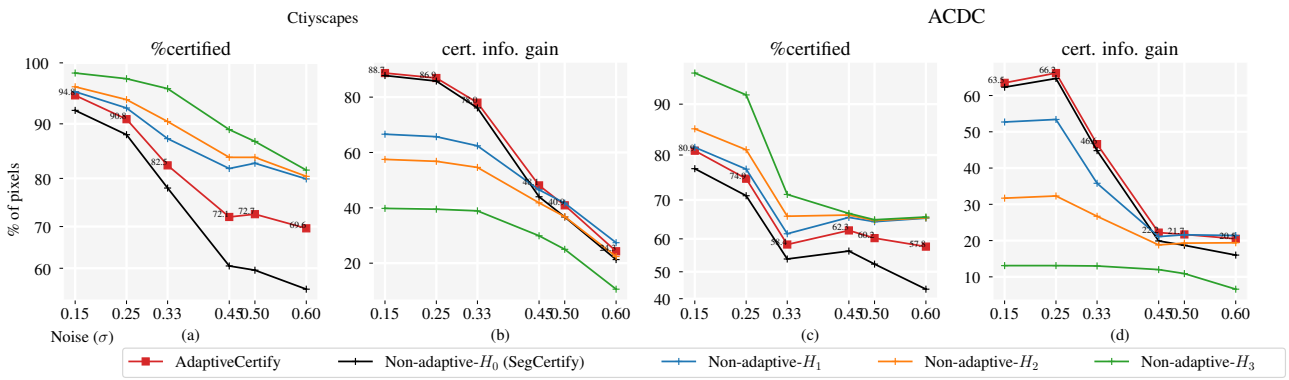

Figure 2: %certified (mean per-pixel certification rate) and cert. info. gain (mean per-pixel certified information gain) versus the number of samples n ($n_0 = 10, n = 100, \tau = 0.75$) on Cityscapes and ACDC.

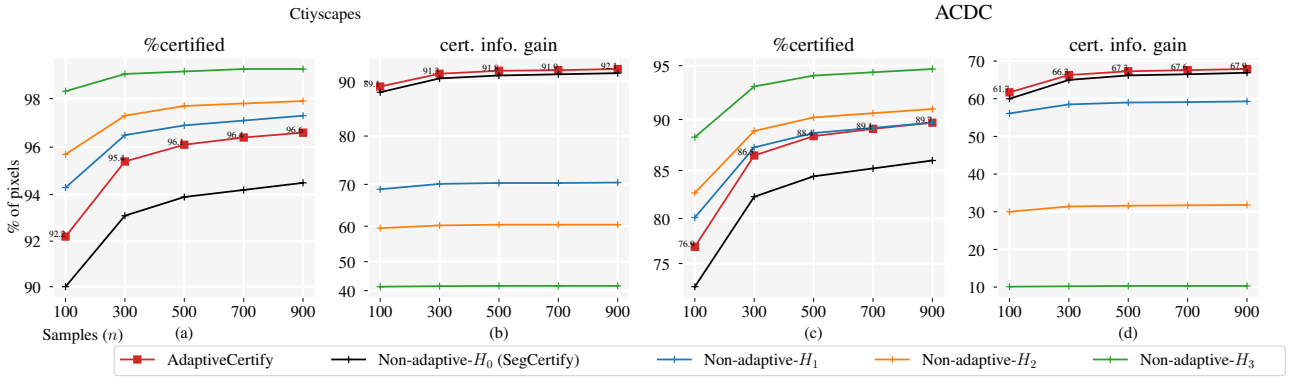

Figure 3: %certified (mean per-pixel certification rate) and cert. info. gain (mean per-pixel certified information gain) versus the noise levels σ ($n_0 = 10, n = 100, \tau = 0.75$) on Cityscapes and ACDC.

4. Pseudocode for HYPOTHESESTESTING

Algorithm 1 HYPOTHESESTESTING: algorithm to perform multiple hypotheses testing to certify or abstain from top vertices while bounding the Type I error rate by α

function HYPOTHESESTESTING (α, $\oslash$, (pv_1,...,pv_N), ($\hat{v}_1, \ldots, \hat{v}_N$))
  for i $\leftarrow$ 1, ..., N **do**
    if $pv_i > \frac{\alpha}{N}$ **then**
      $\hat{v}_i \leftarrow \oslash$
    end if
  end for
  return $\hat{v}_1, \ldots, \hat{v}_N$
end function

5. Experiment details

For the COCO-Stuff-10k dataset, we used the HrNetV2 model (Wang et al., 2020; Yuan et al.) with the HrNetV2-W48 Paddle Cls pre-trained backbone that follows the Object-Contextual Representations (OCR) approach, which is available in the official PyTorch implementation of the HrNetV2 paper (Paszke et al., 2019). We follow the same outlined training procedure, except by adding a Gaussian noise of $\sigma = 0.25$ in an alternating manner across the batches. We validate every epoch on both the clean validation split and the noisy one. During validation, we calculate the following metrics: the mean per-pixel accuracy, the mean accuracy, and the mean intersection over union (mIoU). The batch sizes used for training and validation were 12 and 1 respectively, per GPU. We only validate on non-resized images with a scale of 1, and also show the certification results on them.

For the Cityscapes, ACDC and PASCAL-Context, HrNetV2 was also used, with the HRNetV2-W48 backbone. We use the weights provided by (Fischer et al., 2021) in their official paper PyTorch implementation, which is the result of training the model on a Gaussian noise of $\sigma = 0.25$ following a similar training procedure to that of the PyTorch implementation HrNetV2.

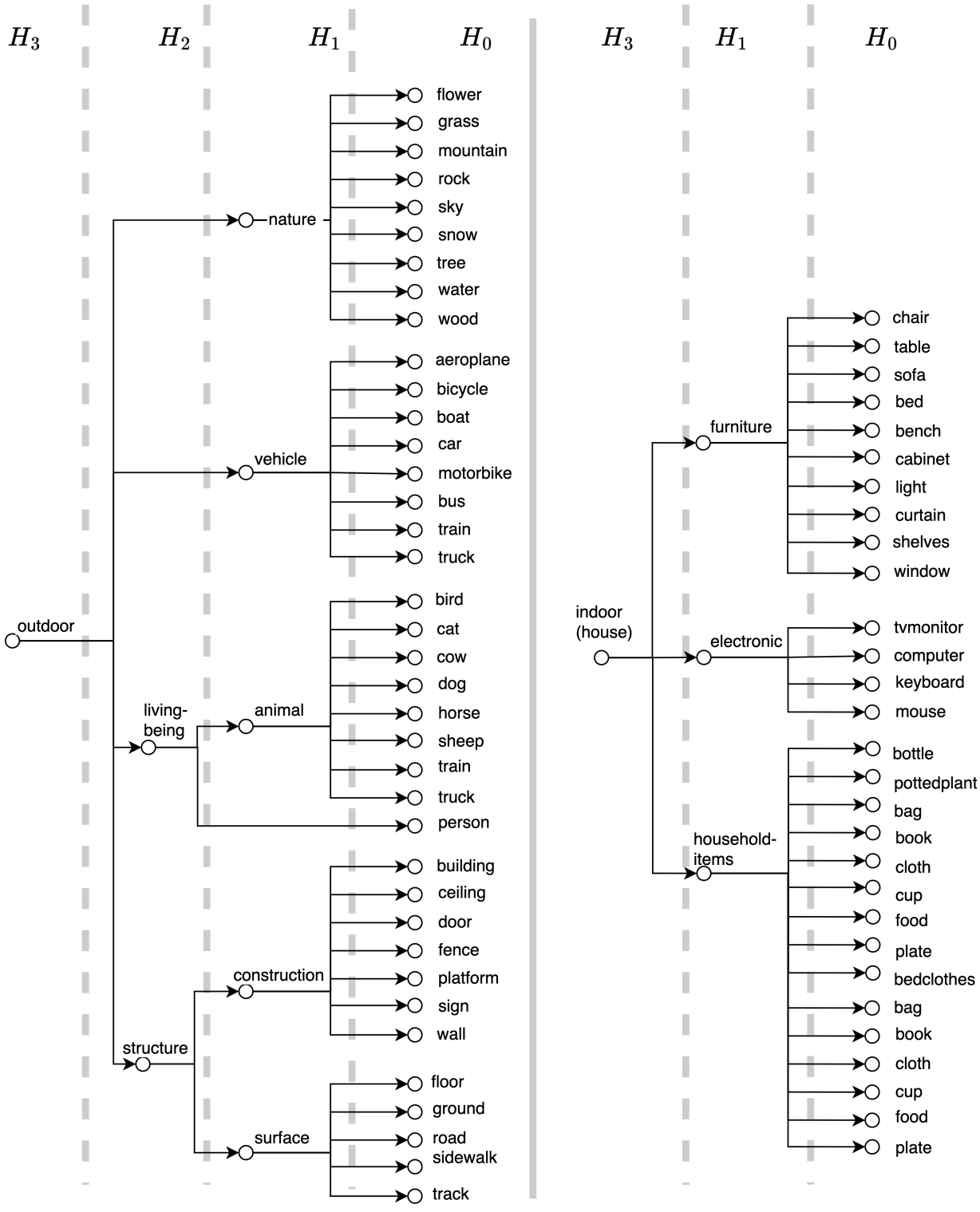

Figure 4: A DAG representing a semantic hierarchy on top of the 59 foregorund classes (located in level H_0) from PASCAL-Context dataset.

References

Caesar, H., Uijlings, J., and Ferrari, V. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1209–1218, 2018.

Fischer, M., Baader, M., and Vechev, M. Scalable certified segmentation via randomized smoothing. In *International Conference on Machine Learning (ICML)*, volume 139. PMLR, 2021.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Mottaghi, R., Chen, X., Liu, X., Cho, N.-G., Lee, S.-W., Fidler, S., Urtasun, R., and Yuille, A. 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*, pp. 891–898, 2014.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

Yuan, Y., Chen, X., Chen, X., and Wang, J. Segmentation transformer: Object-contextual representations for semantic segmentation. arxiv 2019. *arXiv preprint arXiv:1909.11065*.