Supplementary Material for Masking Strategies for Background Bias Removal in Computer Vision Models

Ananthu Aniraj[*1,3,4,5] Cassio F. Dantas[2,3,5] Dino Ienco[2,3,5] Diego Marcos[1,3,4,5]

[1]Inria [2]Inrae [3]University of Montpellier [4]LIRMM [5]UMR-Tetis

{ananthu.aniraj, diego.marcos}@inria.fr {cassio.fraga-dantas, dino.ienco}@inrae.fr

1. Binary Segmentation

1.1. Training Details

We fine-tuned a semantic segmentation model pretrained on the ADE20k [6] dataset on the FG-BG masks provided by the CUB dataset [5].

The training settings are exactly the same as the original Mask2Former paper [1]. We tested models with the ResNet50 and the Swin-Tiny [3] Backbones and trained both models for a total of 16000 epochs with the AdamW [4] optimizer

1.2. Evaluation Metrics

We used the Mean Dice Score [2] to evaluate the segmentation quality of the FG-BG segmentation models.

The Dice score is calculated using the equation given below.

$$Dice = \frac{2 \times |X \bigcap Y|}{|X| + |Y|} \quad (1)$$

Here X is the set of predicted pixels of a specific class from a model and Y is the pixels belonging to the ground truth. A higher dice score indicates higher segmentation quality.

1.3. Evaluation Results

Model	Backbone	CUB(%)		Waterbird(%)	
		BG	Bird	BG	Bird
Mask2Former	Swin-T	99.42	96.05	**98.74**	**91.84**
Mask2Former	ResNet50	99.43	**96.12**	98.72	91.81

Table 1. Evaluation Results - Binary Segmentation

The results of the evaluation are given in Table 1. From the table, we see that the model generalizes very well to both the in-distribution CUB and OOD Waterbirds test set.

We chose the model with the Swin-Tiny backbone as it performed better overall.

*Corresponding Author

References

[1] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar. Masked-attention Mask Transformer for Universal Image Segmentation. In *CVPR*, 2022. 1

[2] L. R. Dice. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):297–302, 1945. 1

[3] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9992–10002, 2021. 1

[4] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *7th International Conference on Learning Representations, ICLR 2019*, 2019. 1

[5] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 1

[6] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ADE20K dataset. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-Janua, pages 5122–5130, 2017. 1