Supplementary Materials

A. Training Setup

We train our models using the PyTorch framework [4]. We initialise our method with the Segment Anything Model (SAM) [2] from the `HuggingFace` library, which is a 94M parameter pre-trained image segmentation model. We optimise using the AdamW optimiser [3] for 10 epochs with an initial learning rate of 1e-5 and linear decay and a batch size of 2. We use an action space size of $K = 50$ and 300 MCTS rollouts for all our experiments. We set the hyperparameters for our model by comparing performance on the validation set. In order to promote robustness for our method, we opt to keep these hyperparameters the same across datasets. As a result, for SLIP-δ we set $\delta = 0.49$, for SLIP-p we set $\tau = 1.5$, and for SLIP-θ we set $\tau = 0.4$.

B. Dataset Preprocessing

To process the datasets, we split them into training, validation and test sets with a ratio of 80 / 10 / 10, except for ISIC, for which we use the existing train / test split, removing 10% of the training data for validation. As the images are not all of equal size, we reduce each image to 256 x 256 pixels to enable batchified training.

C. Metric Definitions

DSC is a region-based metric that considers the pixel-level harmonic mean between

$$\mathrm{DSC}(\hat{Y}, Y) = \frac{2|\hat{Y} \cap Y|}{|\hat{Y}| + |Y|} \tag{1}$$

NSD serves as a complementary metric to provide the consensus between the boundaries of the two masks, for a given tolerance t (set to 2, following [1]):

$$\mathrm{NSD}(\hat{Y}, Y) = \frac{|S_{\hat{Y}} \cap \mathcal{B}_Y^{(t)}| + |S_Y \cap \mathcal{B}_{\hat{Y}}^{(t)}|}{|S_{\hat{Y}}| + |S_Y|} \tag{2}$$

Where S represents the boundary of the mask, and $\mathcal{B}$ represents the regions within t pixels of the boundary [5].

References

[1] Yuhao Huang, Xin Yang, Lianli Liu, Hangyu Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongquan Chen, Chaoyu Chen, Haozhe Chi, Xindi Hu, Deng-Ping Fan, Fajin Dong, and Dong Ni. Segment anything model for medical images? *ArXiv*, abs/2304.14660, 2023. 1

[2] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. *ArXiv*, abs/2304.02643, 2023. 1

[3] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101, 2017. 1

[4] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 1

[5] Annika Reinke, Minu Dietlinde Tizabi, Carole Sudre, Matthias Eisenmann, Tim Rädsch, Michael Baumgartner, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Peter Bankhead, Arriel Benis, Manuel Jorge Cardoso, Veronika Cheplygina, Beth Cimini, Gary Collins, Keyvan Farahani, Ben Glocker, Patrick Godau, and Alican Noyan. Common limitations of image processing metrics: A picture story. 2022. 1