Variable Resolution Baseline Experiment

We compared our method to other networks with fixed rescaling factors φ. An alternative baseline involves training a regular model with $\varphi = 0.5$ and stochastically changing the resolution of the input images and segmentation labels during training. By exposing the model to multiple resolutions during training, we can then vary the input resolution at inference, providing an Quality-FLOPS trade-off.

Variable Resolution U-Net. We train a U-Net model equivalent to those in the paper with $\varphi = 0.5$. During training, at each training step the input is resized a factor $r \sim \mathcal{U}[0,1]$, and the network output is resized by a factor of $1/r$. At inference, we sweep the values of $r \in [0,1]$ to characterize the Pareto accuracy-efficiency curve.

Variable Resolution Hyper-UNet. We also consider a variable-resolution variant of our hypernetwork model where the input is rescaled by the a factor of φ and the prediction is rescaled by a factor of $1/\varphi$

We evaluate on the OASIS2d semantic segmentation task introduced in the paper, which features 24 brain structure labels. We evaluate using Dice score and average over the labels.

Figure 1 shows trade-off curves for the hypernetwork method, the set of fixed rescaling baselines, the variable-resolution baseline, and the variable-resolution hypernetwork model. We observe that both the variable-resolution baseline and hypernetwork model present substantially worse segmentation results to the methods evaluated in the paper. At the same time, the variable resolution methods explore a wider range of inference costs as they do not perform the outermost convolutional operations at full resolution. For the inference costs where the hypernetwork method is defined, it dominates all other approaches.

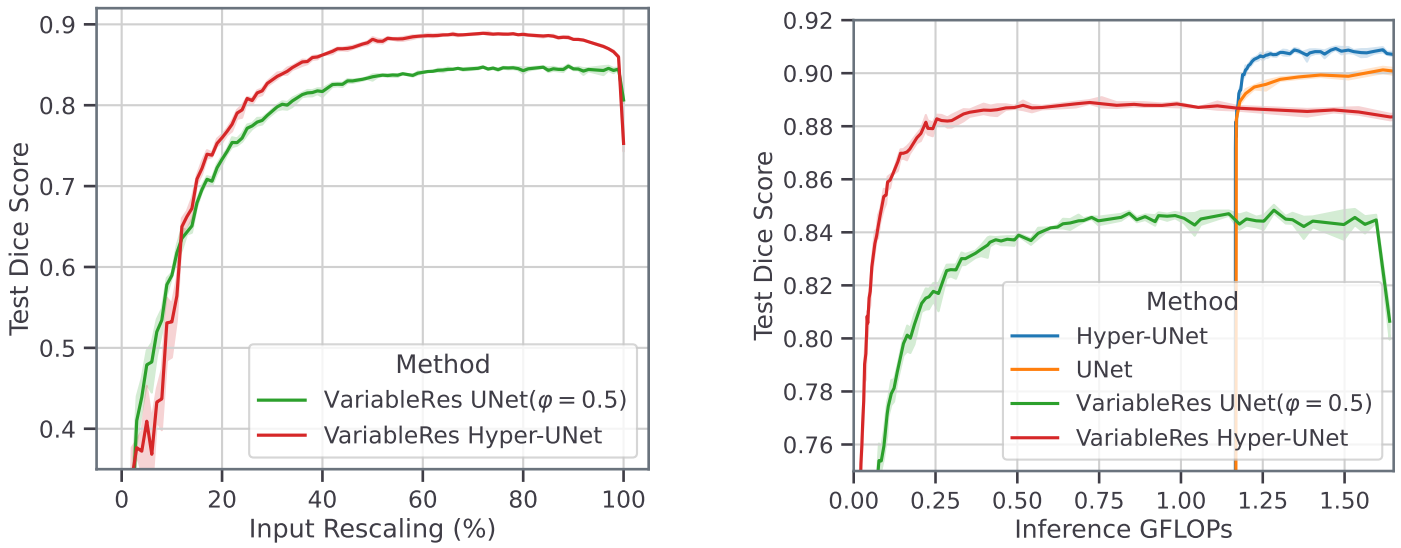

Figure 1: **Segmentation Architecture Ablation** (**Left**) Segmentation Dice coefficient for models trained with variable resolution inputs. (**Right**) Trade-off curves between test model accuracy measured in Dice score and inference computational cost measured in GFLOPs. Results include a variable-resolution UNet model with rescaling factor $\varphi = 0.5$ trained with variable resolutions (VariableRes UNet($\varphi = 0.5$)), as well as a hypernetwork model trained in a similar fashion (VariableRes Hyper-UNet). We also include the main hypernetwork method (Hyper-UNet) and the set of fixed rescaling baselines (UNet). Results are averaged across three random initializations, and the shaded regions indicate the standard deviation across them.