How do 3D image segmentation networks behave across the context versus foreground ratio trade-off?

Amith Kamath[1,*], Yannick Suter[1], Suhang You[1], Michael Müller[1], Jonas Willmann[2,3], Nicolaus Andratschke[2], Mauricio Reyes[1]

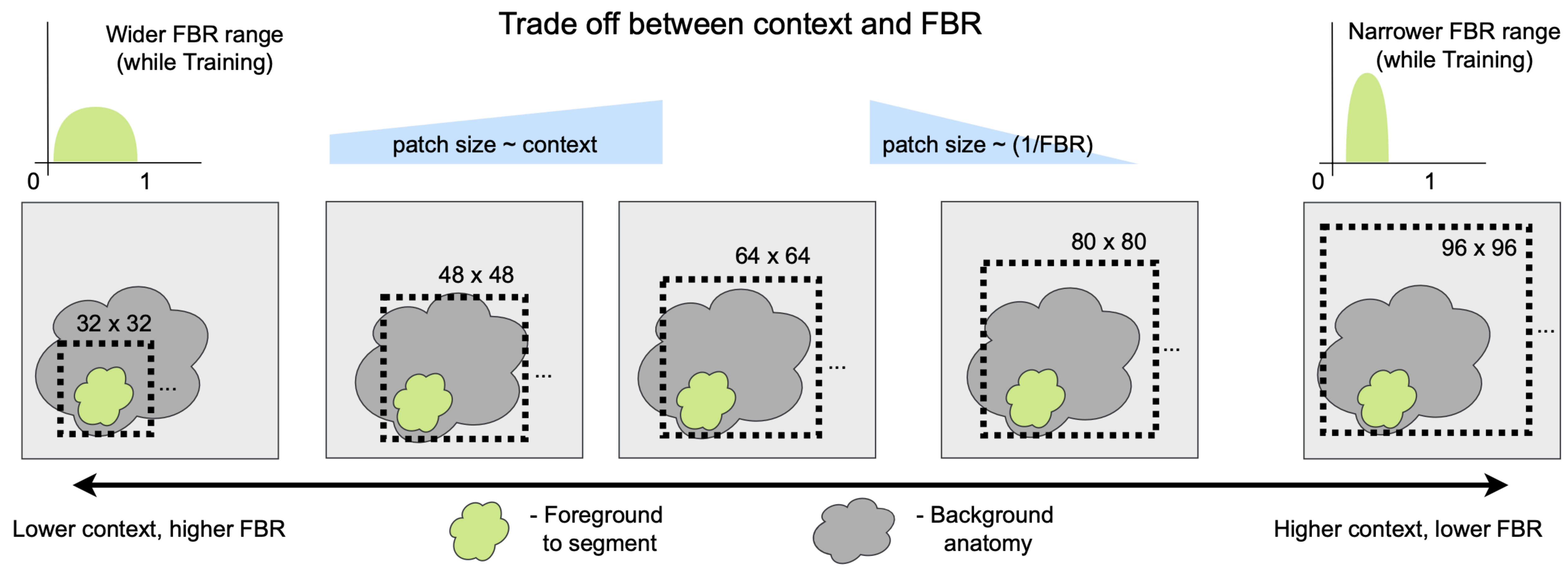
[1] ARTORG Center for Biomedical Engineering Research, University of Bern
[2] University Hospital Zurich, University of Zurich
[3] Center for Proton Therapy, Paul Scherrer Institute
* amith.kamath@unibe.ch

NEURAL INFORMATION PROCESSING SYSTEMS

Research Question:

There's a trade-off in image segmentation models between large patch sizes: higher context, but smaller foreground-to-background ratios (FBR) and small patch sizes (low context, higher FBR).

We aim to explore: How do Vanilla-Unet, UNETR, and Attention-Unet behave across patch-size choices?



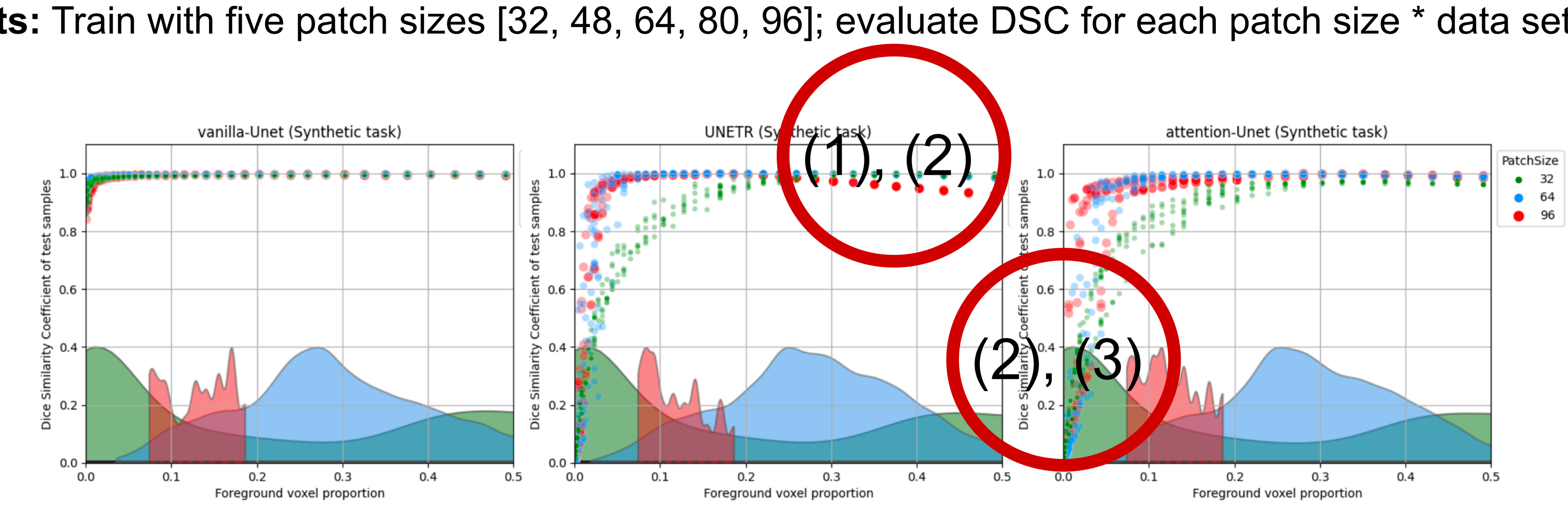
Experimental Setup:

Synthetic data: 100 volumes of size 96^3 voxels. Background: random noise, variance 0.8. Foreground: random white spheres with radius 25-30, random centers. 70 training, 30 test, 100 completely independent samples with out-of-training radii (5-48) for test.

Clinical data: Spleen data set from Medical Segmentation Decathlon. 26 training, 5 validation, and 10 test.

Models: Vanilla-Unet, UNETR, Attention-Unet with default settings using MONAI.

Experiments: Train with five patch sizes [32, 48, 64, 80, 96]; evaluate DSC for each patch size * data set * model.

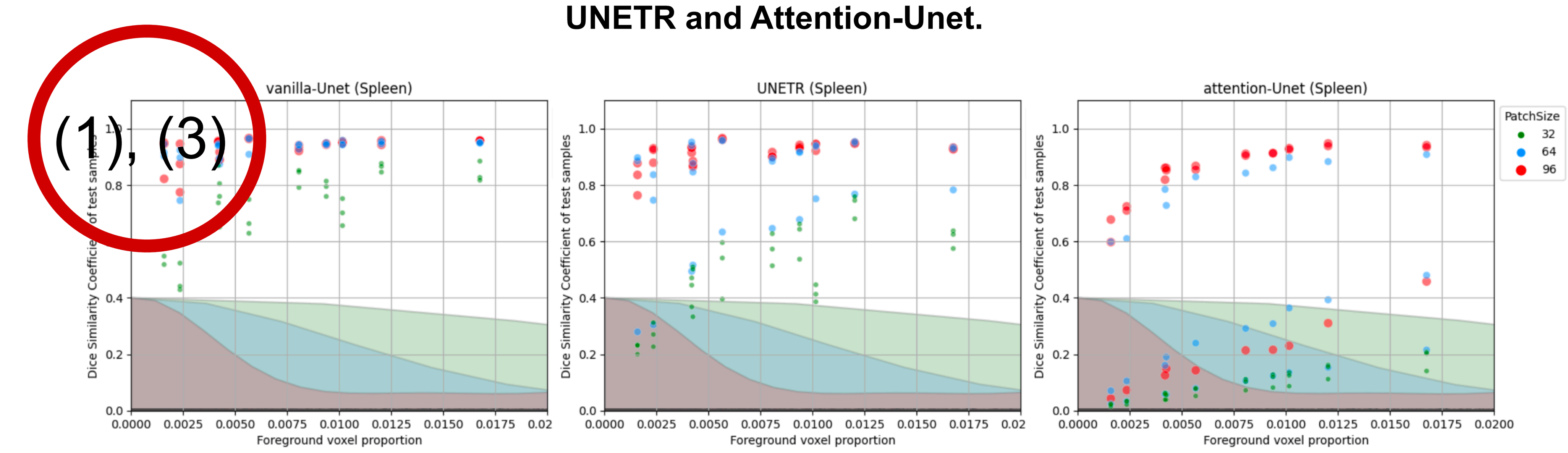


Dice Similarity Coefficient metrics for the synthetic task: using Vanilla-Unet (left), UNETR (middle), and Attention-Unet (right). **Distributions at the bottom indicate proportion of training samples with that FBR during training**. Only patch sizes 32, 64, 96 shown for clarity.

Results:

Network	PatchSize: 32	PatchSize: 48	PatchSize: 64	PatchSize: 80	PatchSize: 96
Unet (in-train)	**0.981** (0.024)	**0.986** (0.016)	**0.994** (0.001)	**0.994** (0.001)	0.993 (0.001)
drop outside	–	–	$\Delta = -0.0234$	$\Delta = -0.0139$	$\Delta = -0.0178$
UNETR (in-train)	0.677 (0.350)	0.643 (0.390)	0.957 (0.121)	0.991 (0.009)	**0.994** (0.003)
drop outside	–	–	$\Delta = -0.733$	$\Delta = \mathbf{-0.369}$	$\Delta = \mathbf{-0.275}$
Att-Unet (in-train)	0.632 (0.342)	0.663 (0.373)	0.948 (0.132)	0.992 (0.004)	0.970 (0.015)
drop outside	–	–	$\Delta = \mathbf{-0.836}$	$\Delta = -0.299$	$\Delta = -0.255$

Dice Similarity Coefficient metrics for the synthetic task: **note the drop in DSC outside the "training distribution" range for UNETR and Attention-Unet.**



Dice Similarity Coefficient metrics for the Spleen task: using Vanilla-Unet (left), UNETR (middle), and Attention-Unet (right). Distributions at the bottom indicate proportion of training samples with that FBR during training - **note that these cover all the test samples in this case.**

Network	PatchSize: 32	PatchSize: 48	PatchSize: 64	PatchSize: 80	PatchSize: 96
Unet	**0.721** (0.130)	**0.907** (0.045)	**0.928** (0.040)	**0.922** (0.047)	**0.932** (0.042)
UNETR	0.481 (0.158)	0.766 (0.178)	0.799 (0.186)	0.852 (0.116)	0.915 (0.042)
Att-Unet	0.086 (0.052)	0.102 (0.060)	0.384 (0.313)	0.582 (0.326)	0.634 (0.327)

Dice Similarity Coefficient metrics for the Spleen task: **note, Vanilla-Unet is the best performing across all patch sizes.**

Findings (see red circles for supporting details):

(1) **Larger patch sizes are preferred** across all three network architectures,

(2) **UNETR and Attention-Unet appear to be more sensitive** to patch size changes,

(3) Ensuring a **wide range in FBR during training is a prerequisite** for robustness.

References

See full list of references by scanning the QR code on the right.

MEDICAL IMAGE ANALYSIS

UNIVERSITÄT BERN | ARTORG CENTER BIOMEDICAL ENGINEERING RESEARCH

USZ Universitäts Spital Zürich

krebsliga schweiz
ligue suisse contre le cancer
lega svizzera contro il cancro