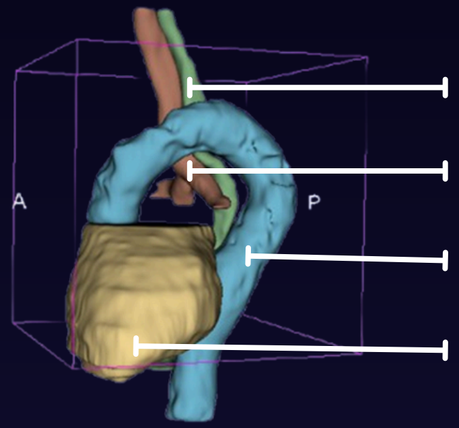
University of Amsterdam
AI for Medical Imaging - group 3
October 14, 2025

Improvements on CT segmentation

Emanuele Arcelli, Julia Blaauboer, Floor Laagland
Jasper van der Valk, Rick van der Veen, David Werkhoven

Clinical problem

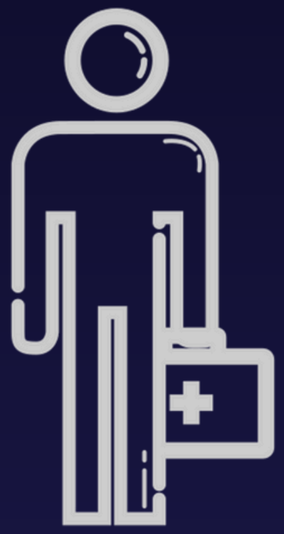


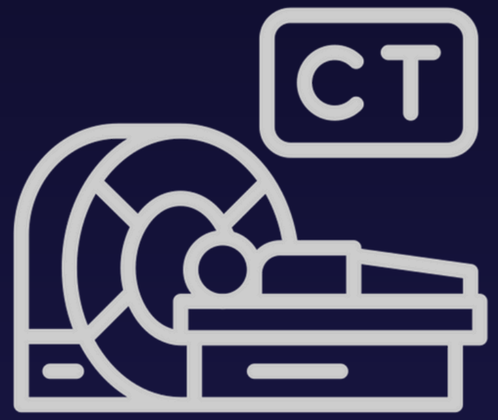
Esophagus

Trachea

Aorta

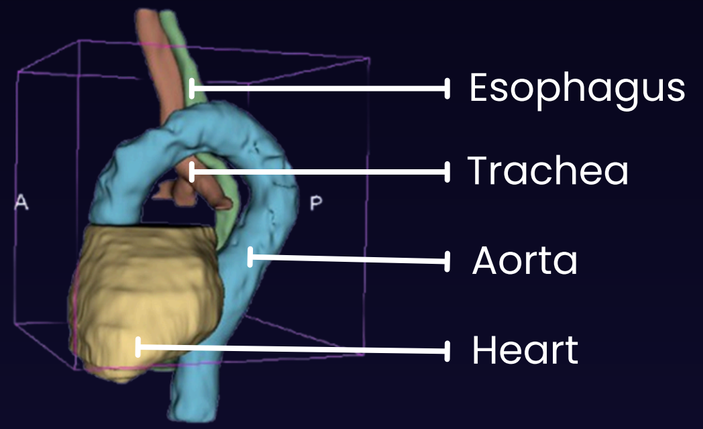
Heart

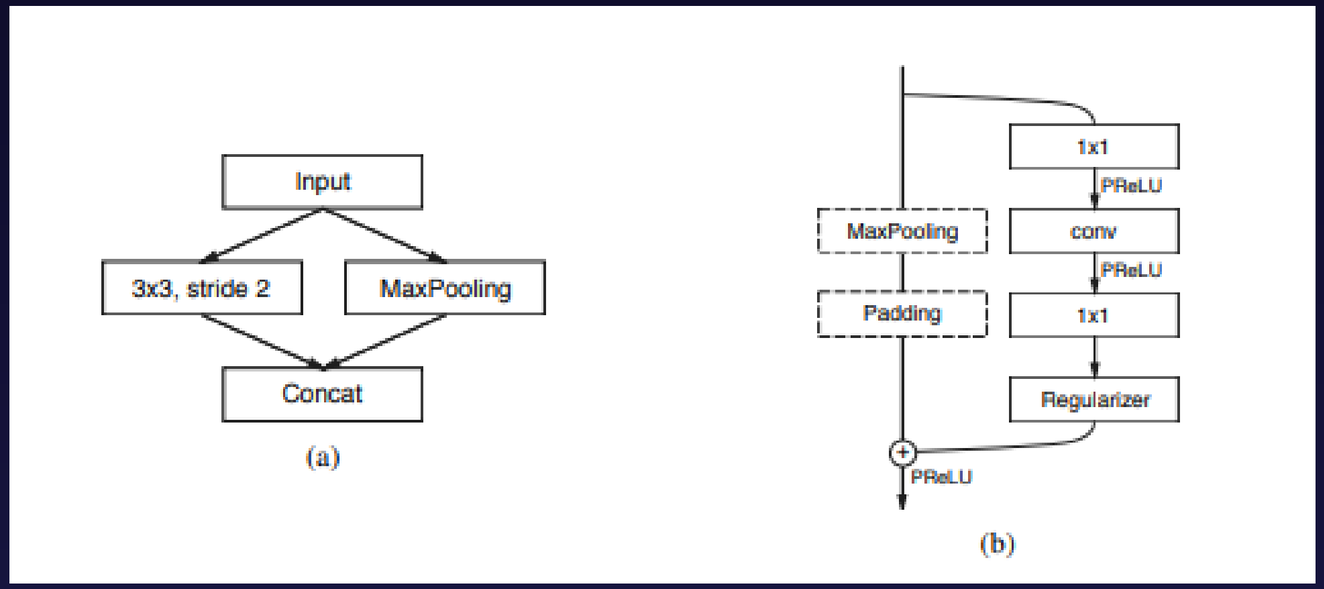




Icons made by IconJAM, SANB, Elzicon from www.flaticon.com
Ma, J., He, Y., Li, F., Han, L., You, C., & Wang, B. (2024). Segment anything in medical images. Nature Communications, 15(1). https://doi.org/10.1038/s41467-024-44824-z

Baseline model





	3D DICE	HD95
Esophagus	0.5737	10.3139
Heart	0.9172	9.1389
Trachea	0.7864	10.2712
Aorta	0.8408	8.9793
Baseline	0.7795	9.6758

Paszke, A., Chaurasia, A., Kim, S., & Culurciello, E. (2016). ENET: a Deep Neural Network architecture for Real-Time Semantic Segmentation. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1606.02147

Class imbalance

→

Voxel Distribution by Organ

Organ Class	Total Voxels	Percentage
background	1924610658	98.95
Esoshapous	880656	0.05
Heart	15121544	0.78
Trachee	674928	0.03
Aorta	3820694	0.20

Slice Distribution by Organ Class

Class	#Slices	% of slices
background	7420	100.00
Esoshapous	3920	52.83
Heart	1599	21.55
Trachea	1987	26.78
Aorta	3754	50.59
TOTAL	**7420**	**100.00**

Dice loss

Why?

Tries to predict the overall shape, rather than classifying pixels individually

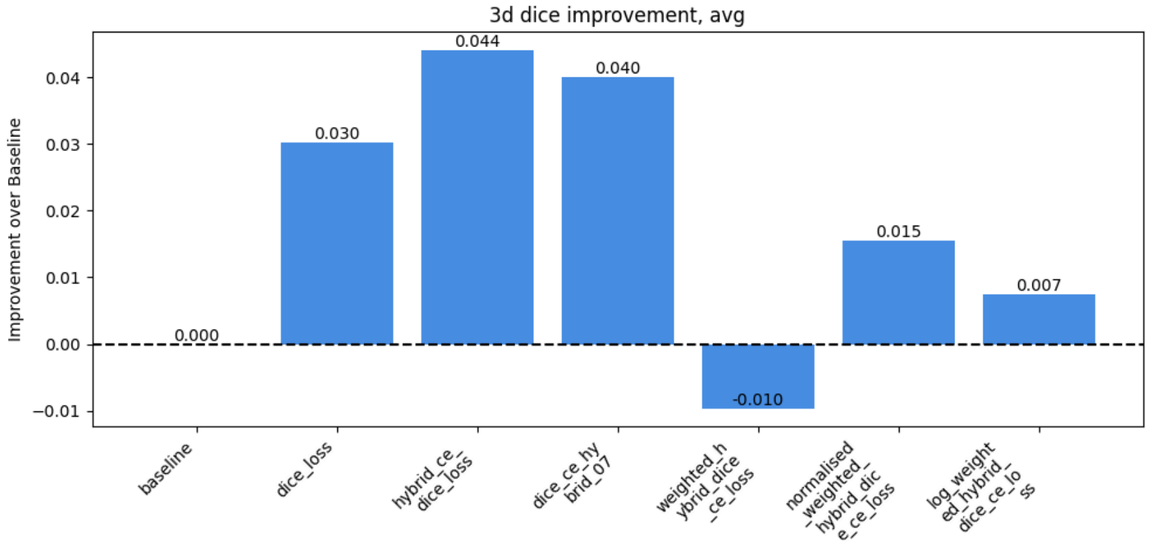
How?

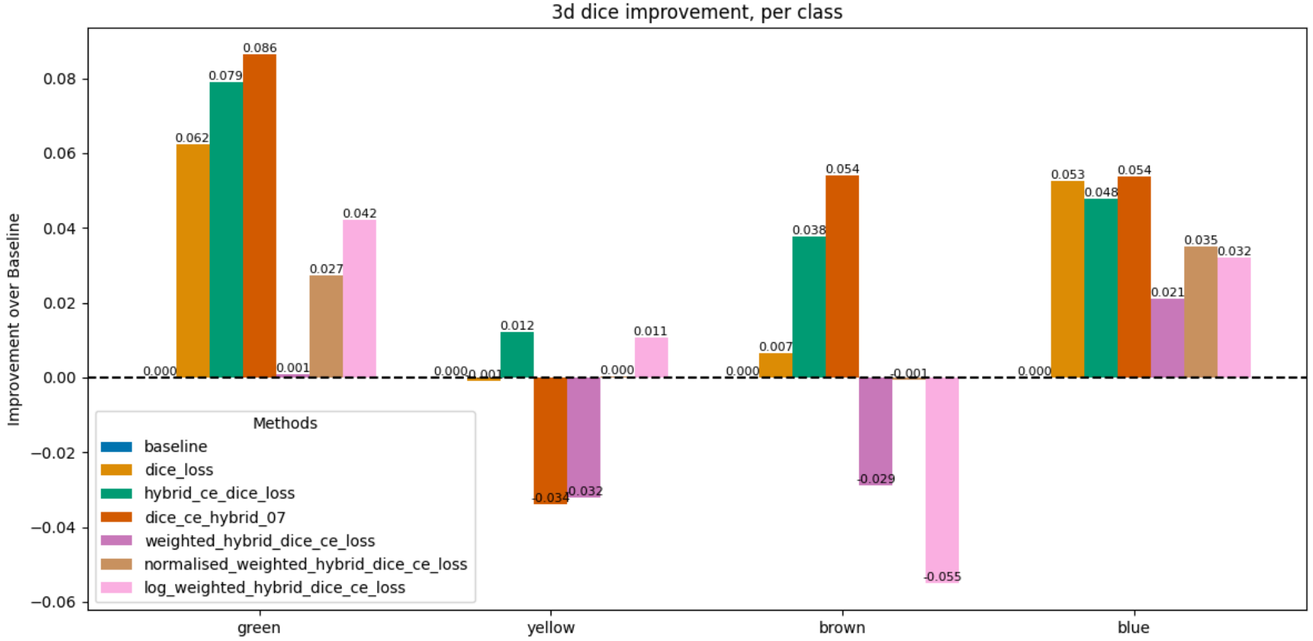
- By changing the cross entropy loss with a dice loss
- By using a hybrid loss that combines the cross entropy and dice loss

Impact?

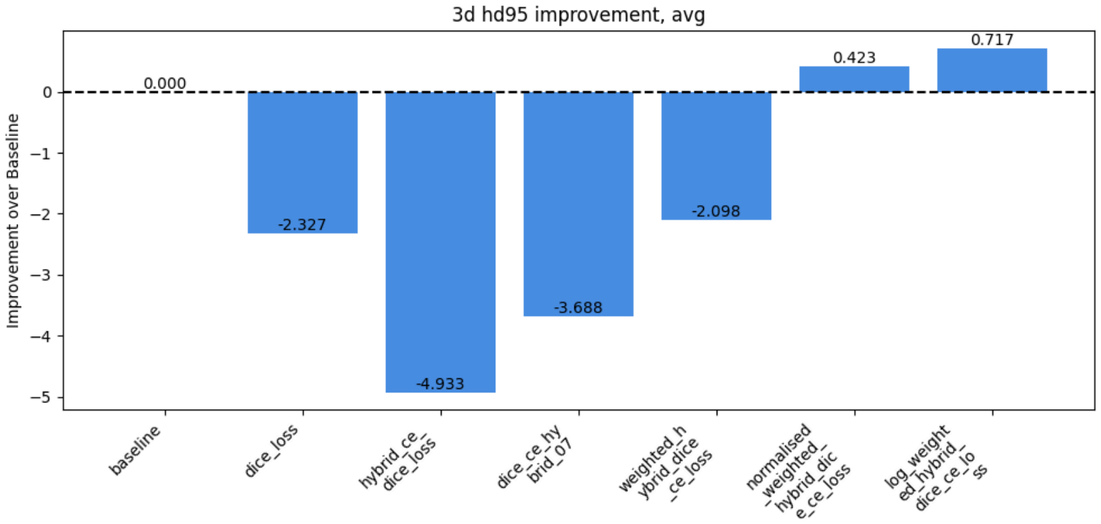
It improves performance on the rarer classes, but this comes at the trade-off of slightly lower accuracy for the more dominant ones.

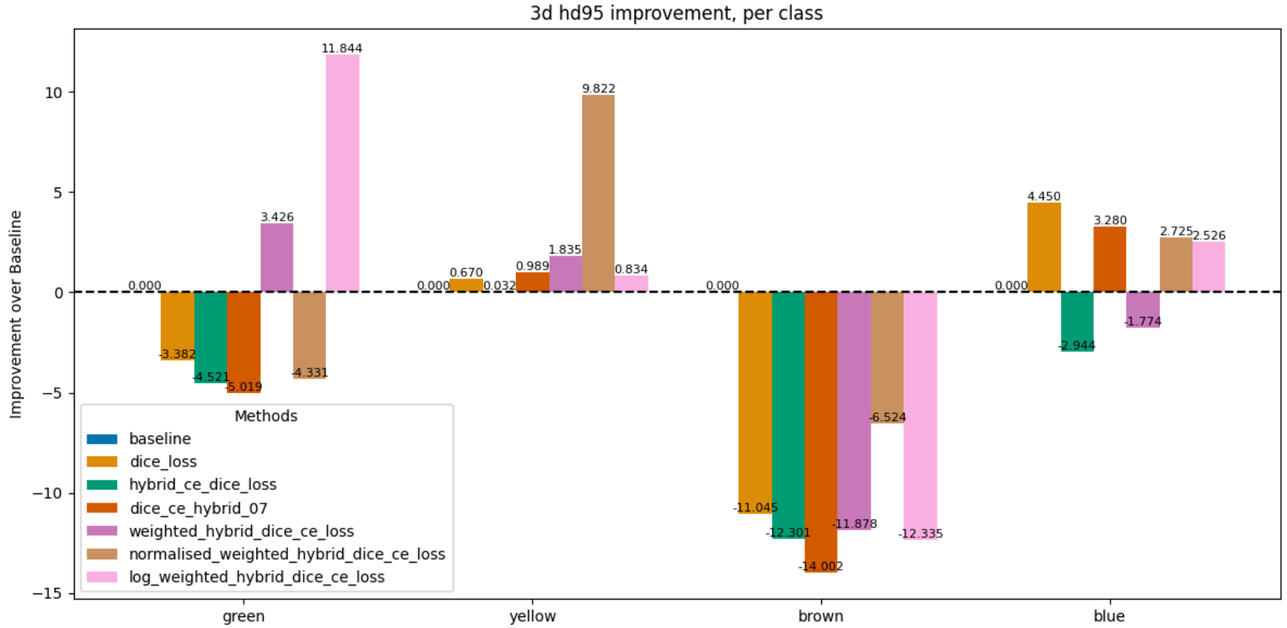
Dice loss


3d dice improvement, avg
Improvement over Baseline
0.04
0.03
0.02
0.01
0.00
−0.01
0.000
0.030
0.044
0.040
-0.010
0.015
0.007
baseline
dice_loss
hybrid_ce_dice_loss
dice_ce_hybrid_07
weighted_hybrid_dice_ce_loss
normalised_weighted_hybrid_dice_ce_loss
log_weighted_hybrid_dice_ce_loss


3d dice improvement, per class
Improvement over Baseline
0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
0.000
0.062
0.079
0.086
0.001
0.027
0.042
0.000
0.001
0.012
-0.034
0.032
0.000
0.011
0.000
0.007
0.038
0.054
-0.029
-0.001
-0.055
0.000
0.053
0.048
0.054
0.021
0.035
0.032
Methods
baseline
dice_loss
hybrid_ce_dice_loss
dice_ce_hybrid_07
weighted_hybrid_dice_ce_loss
normalised_weighted_hybrid_dice_ce_loss
log_weighted_hybrid_dice_ce_loss
green
yellow
brown
blue

Dice loss


3d hd95 improvement, avg
Improvement over Baseline
0.000
-2.327
-4.933
-3.688
-2.098
0.423
0.717
baseline
dice_loss
hybrid_ce_dice_loss
dice_ce_hybrid_07
weighted_hybrid_dice_ce_loss
normalised_weighted_hybrid_dice_ce_loss
log_weighted_hybrid_dice_ce_loss


3d hd95 improvement, per class
Improvement over Baseline
green
yellow
brown
blue
Methods
baseline
dice_loss
hybrid_ce_dice_loss
dice_ce_hybrid_07
weighted_hybrid_dice_ce_loss
normalised_weighted_hybrid_dice_ce_loss
log_weighted_hybrid_dice_ce_loss

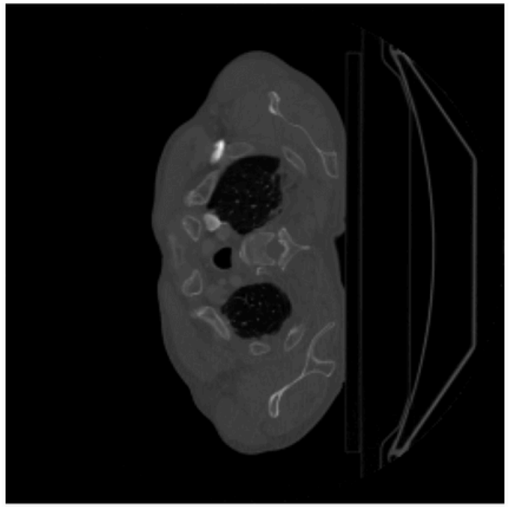
2.5D

Why?

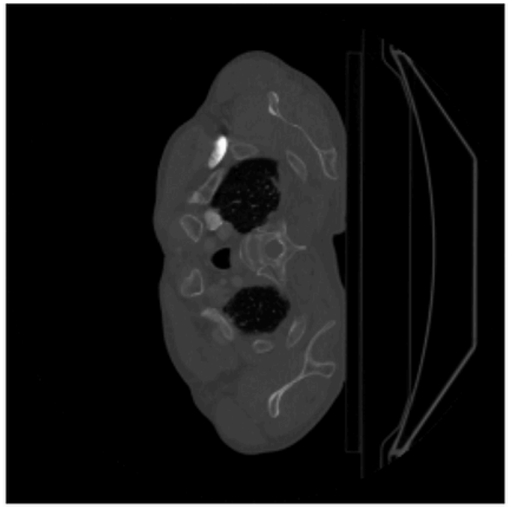
Some classes cover small areas (giving low importance for ce), but appear in many slices

→ use adjacent slices

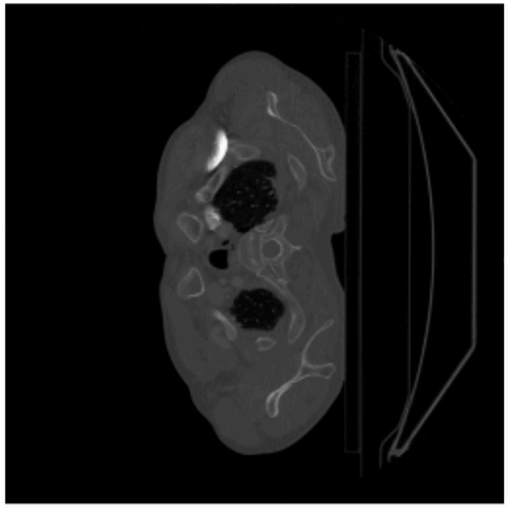
Slice n-1



Slice n



Slice n+1



Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. Proceedings of the Royal Society of London. Series B. Biological Sciences, 200(1140), 269-294.

2.5D

Why?

Some classes cover small areas (giving low importance for ce), but appear in many slices

→ use adjacent slices

How?

Stacking

Conv 3D

2.5D

Why?

Some classes cover small areas (giving low importance for ce), but appear in many slices

→ use adjacent slices

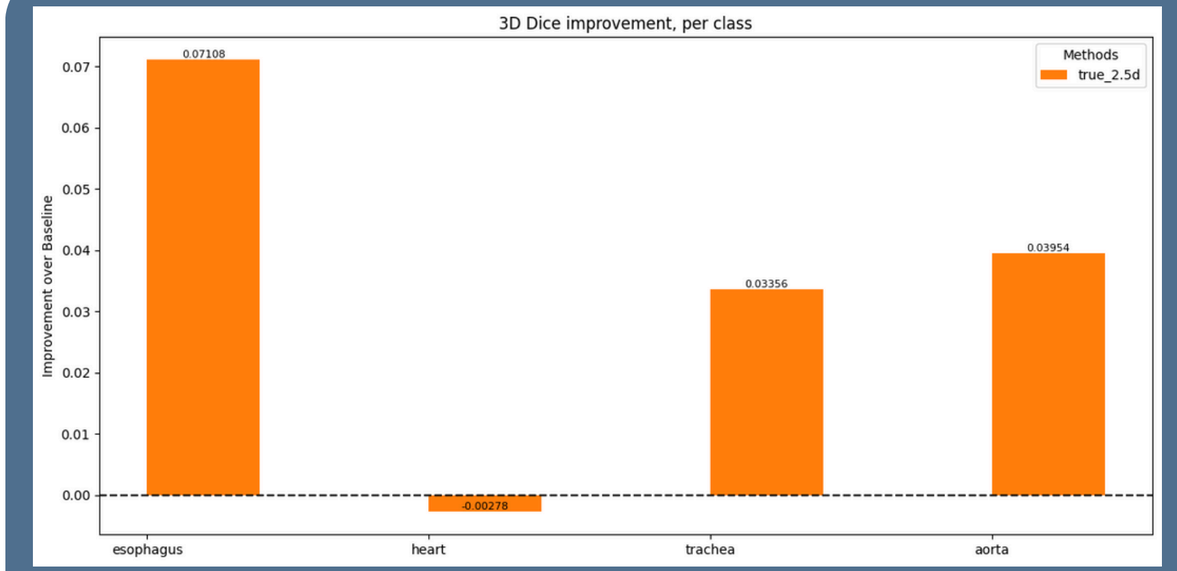
How?

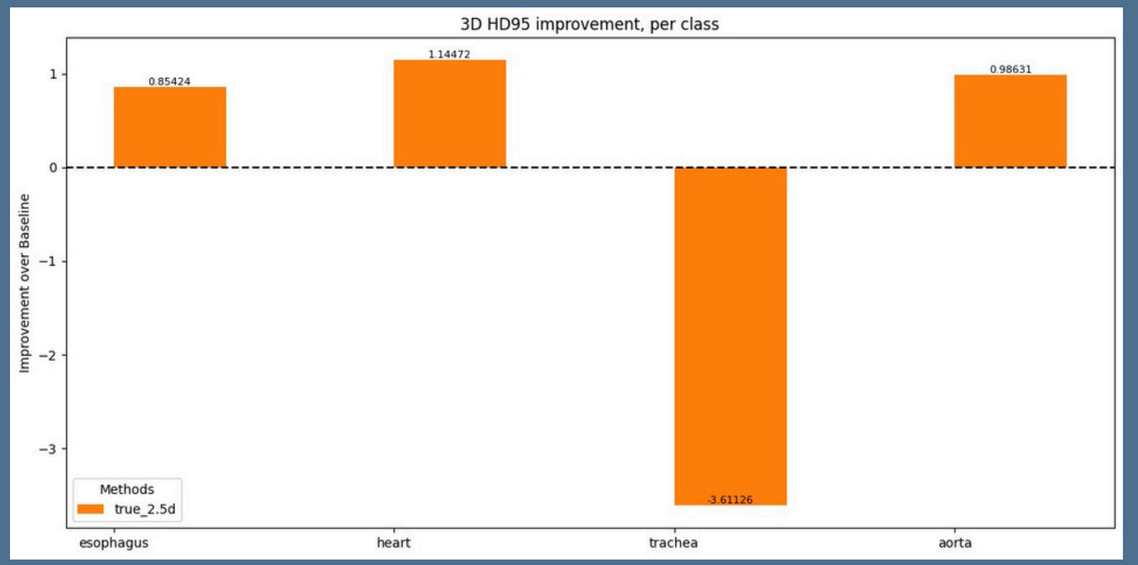
Stacking

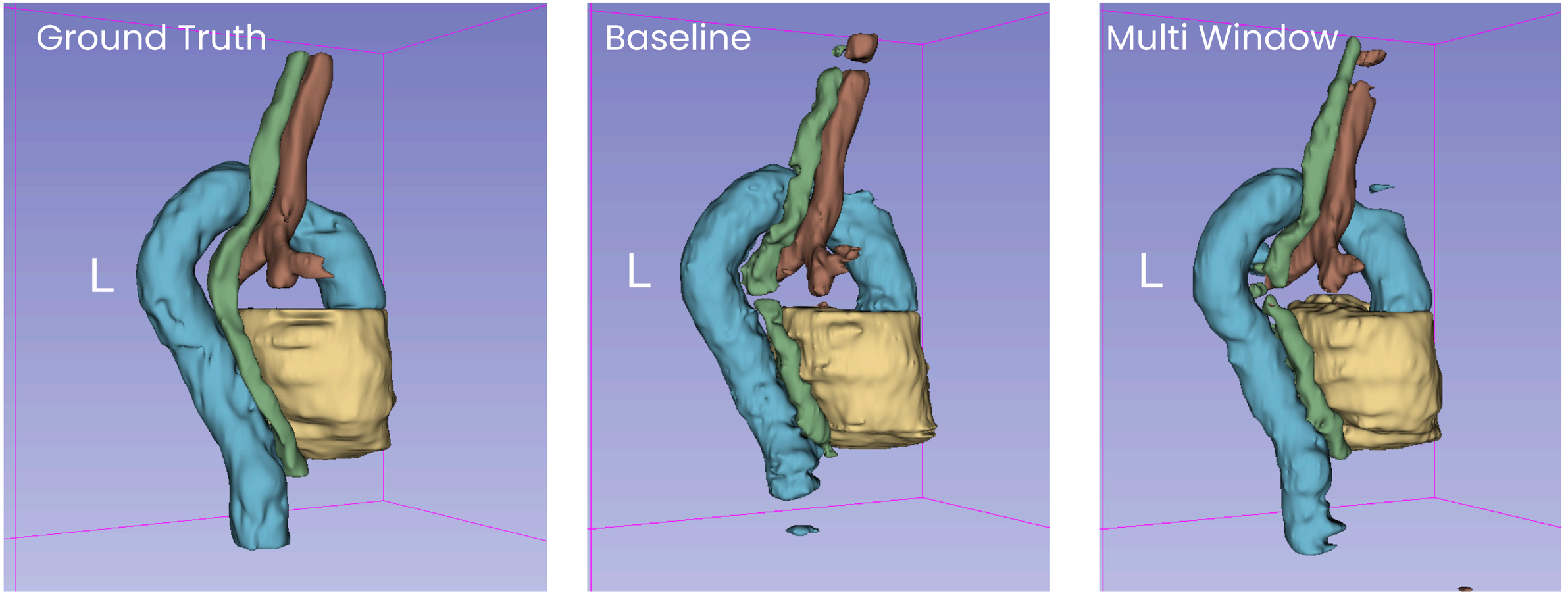
↓

Conv 3D

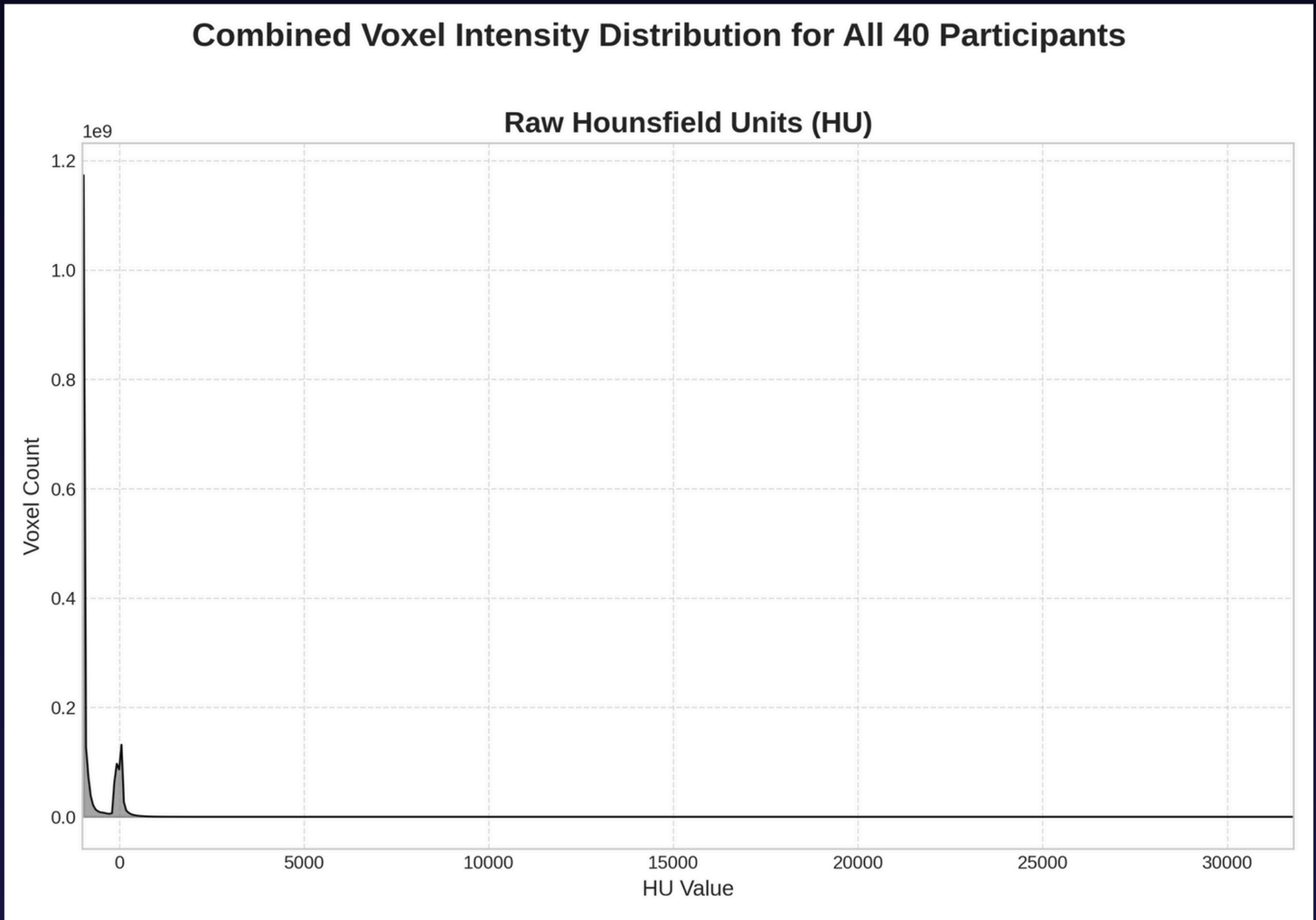
Impact



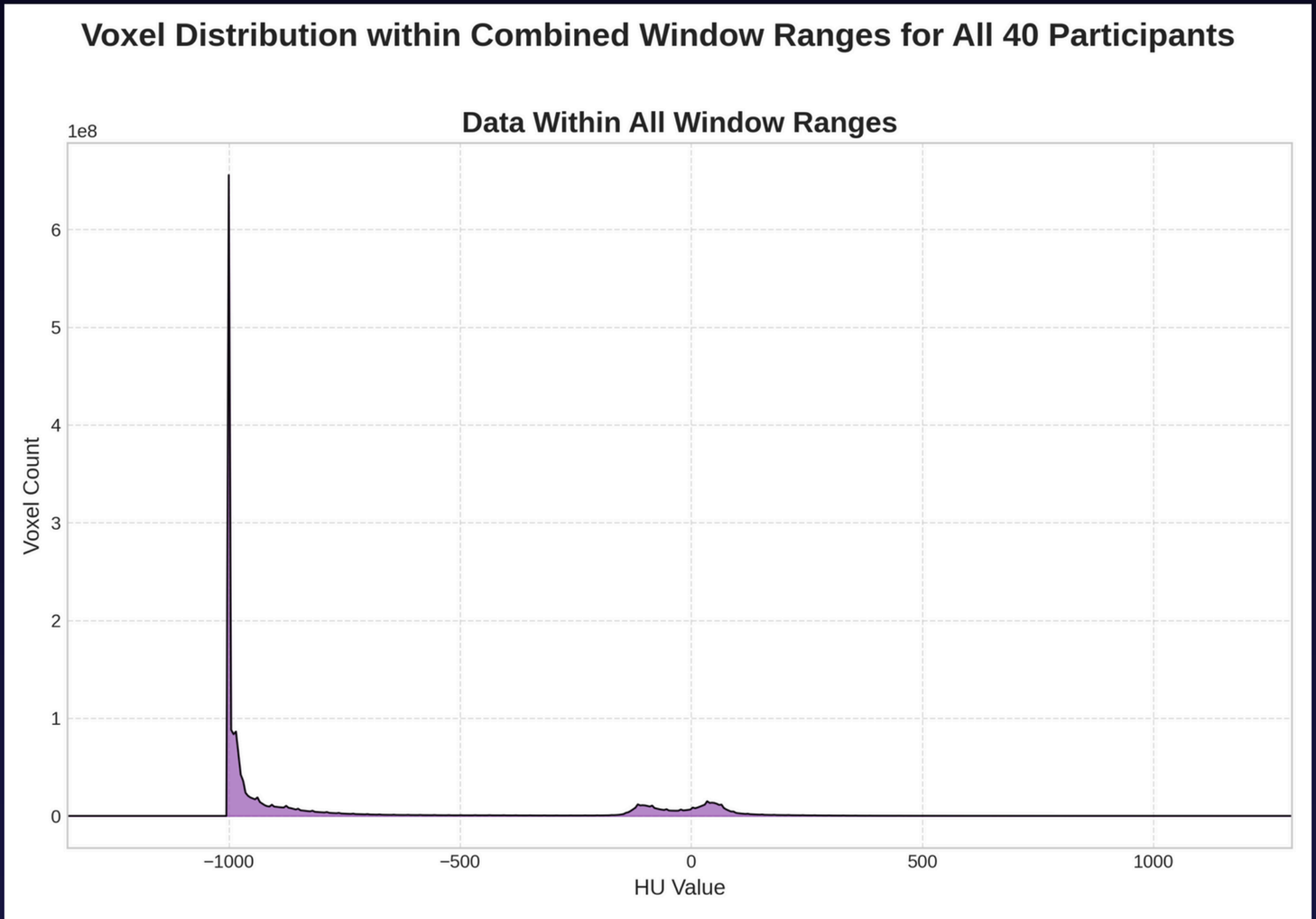



Ground Truth
L
Baseline
L
Multi Window
L

Range of the data

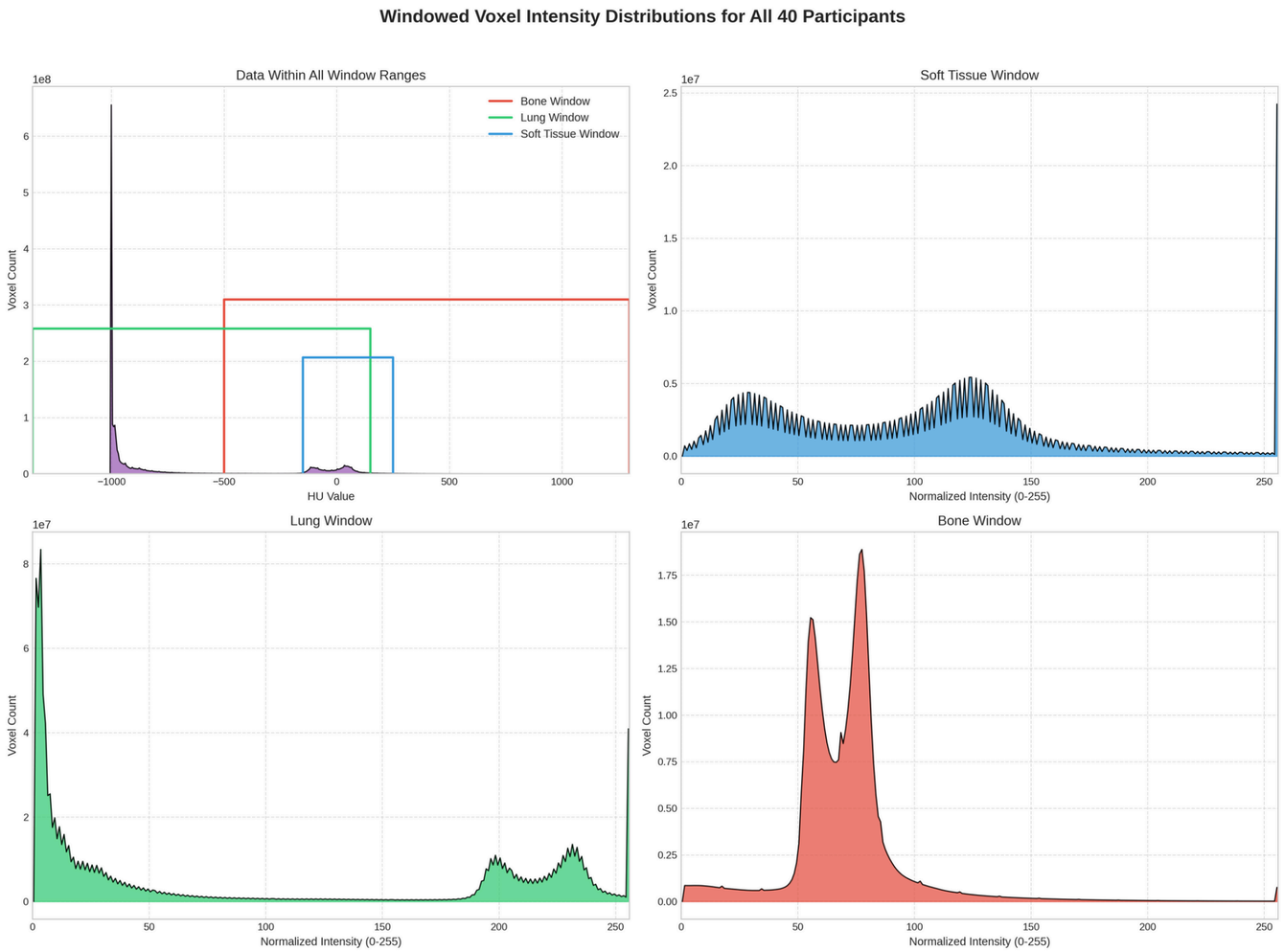


Range of the data



Let’s zoom in

Windowing

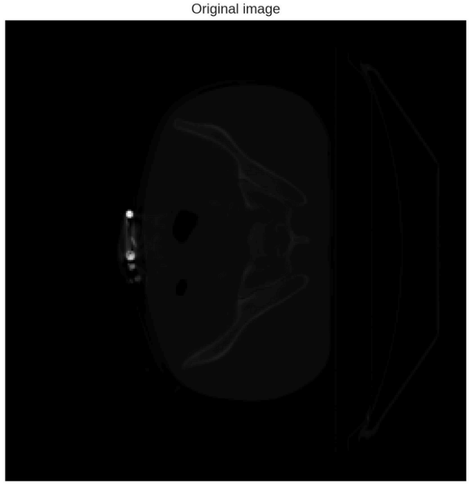


	Center/Range
Soft Tissue:	50/400
Lung:	-600/1500
Bone:	400/1800

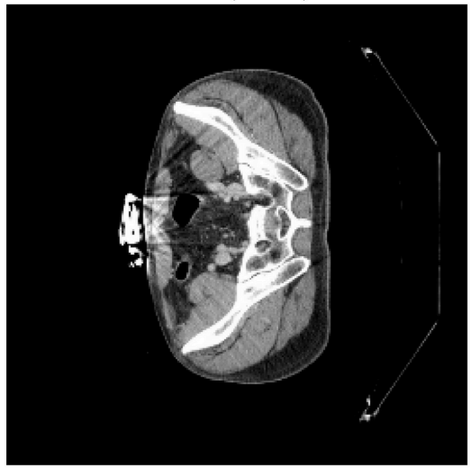
Ahn, H., Lee, K. W., Lee, K. H., Kim, J., Kim, K., Chung, J. H., & Lee, C. T. (2018). Effect of computed tomography window settings and reconstruction plane on 8th edition T-stage classification in patients with lung adenocarcinoma manifesting as a subsolid nodule. European journal of radiology, 98, 130-135.

Windowing

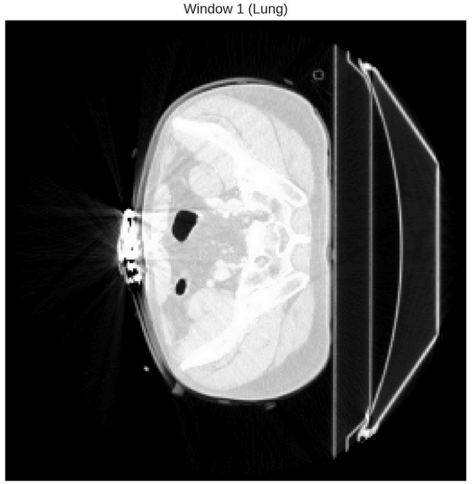
Original image



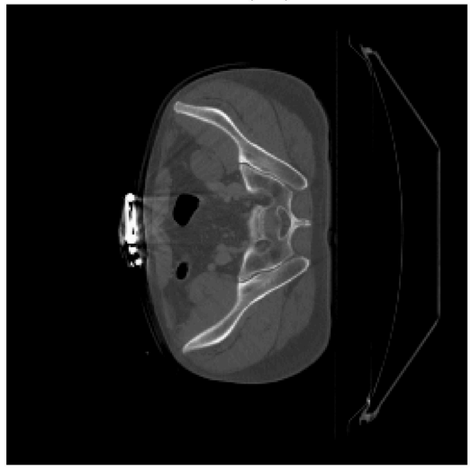
Window 0 (Soft Tissue)



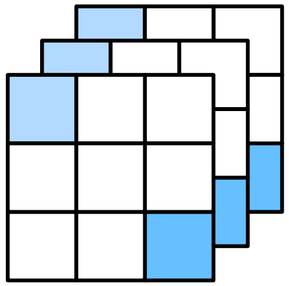
Window 1 (Lung)



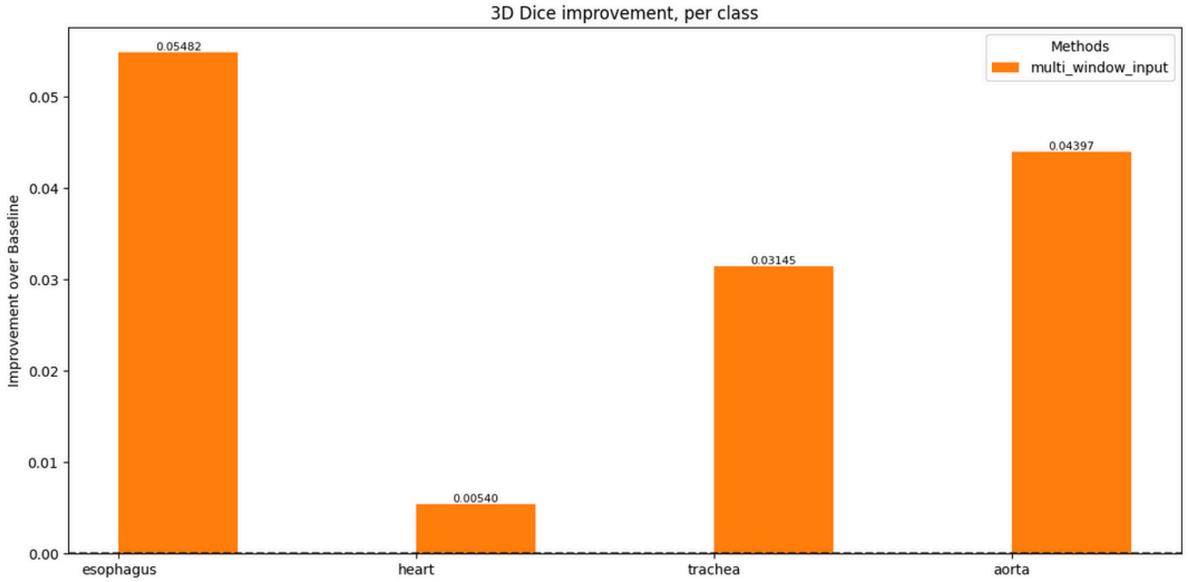
Window 2 (Bone)

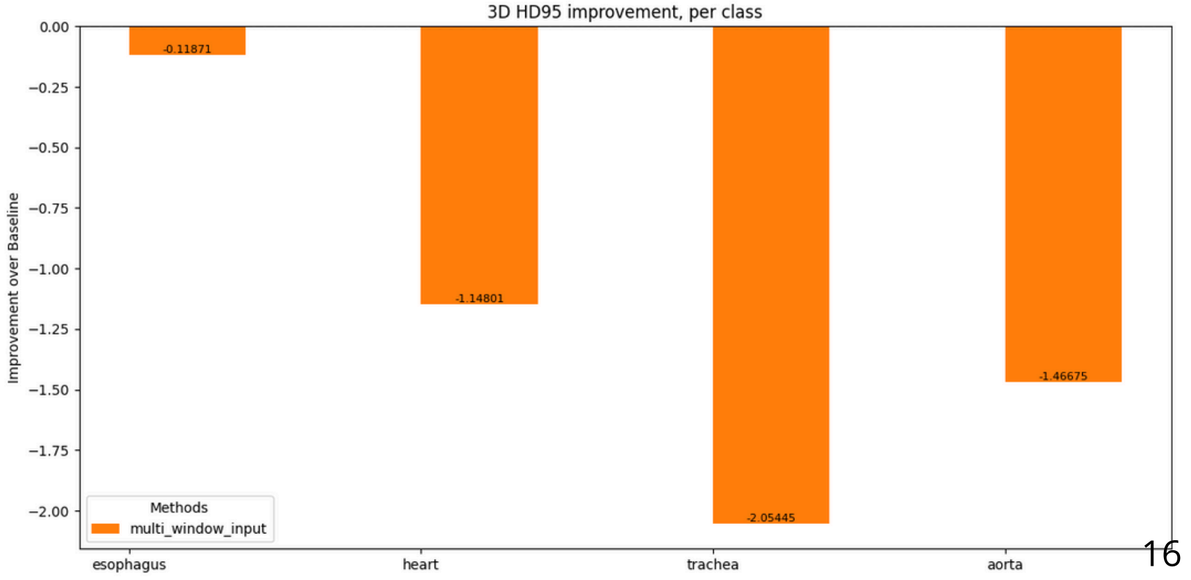


Stacking

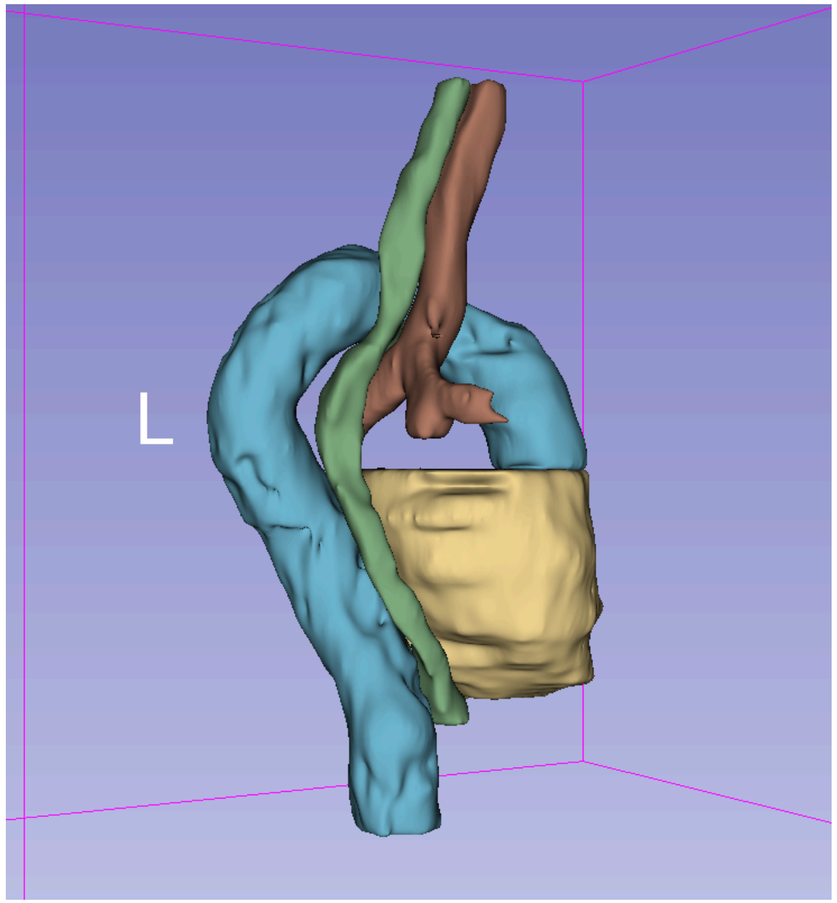


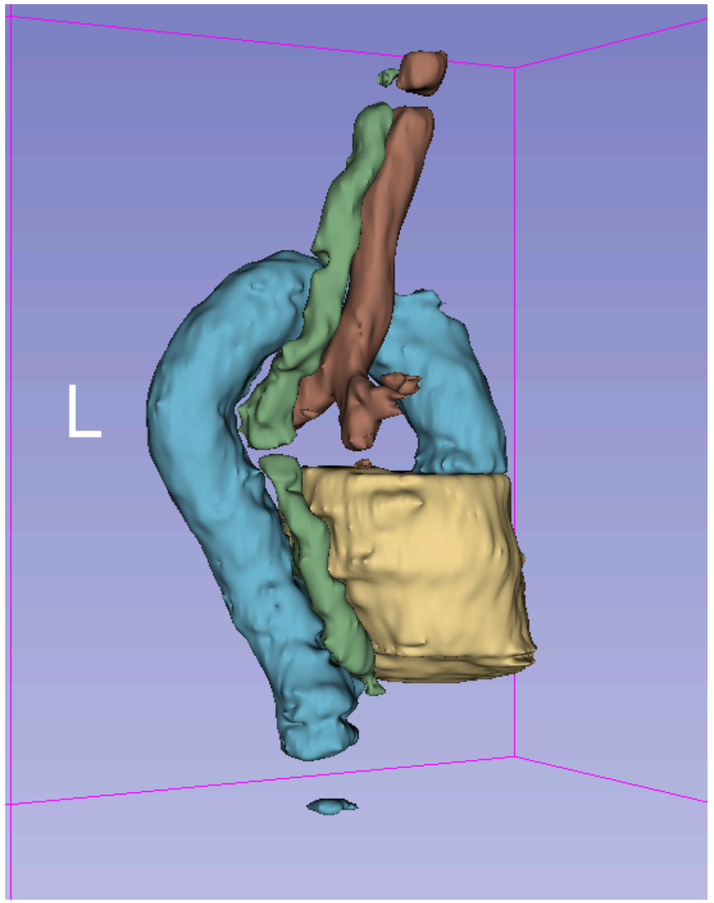
Conv 2D

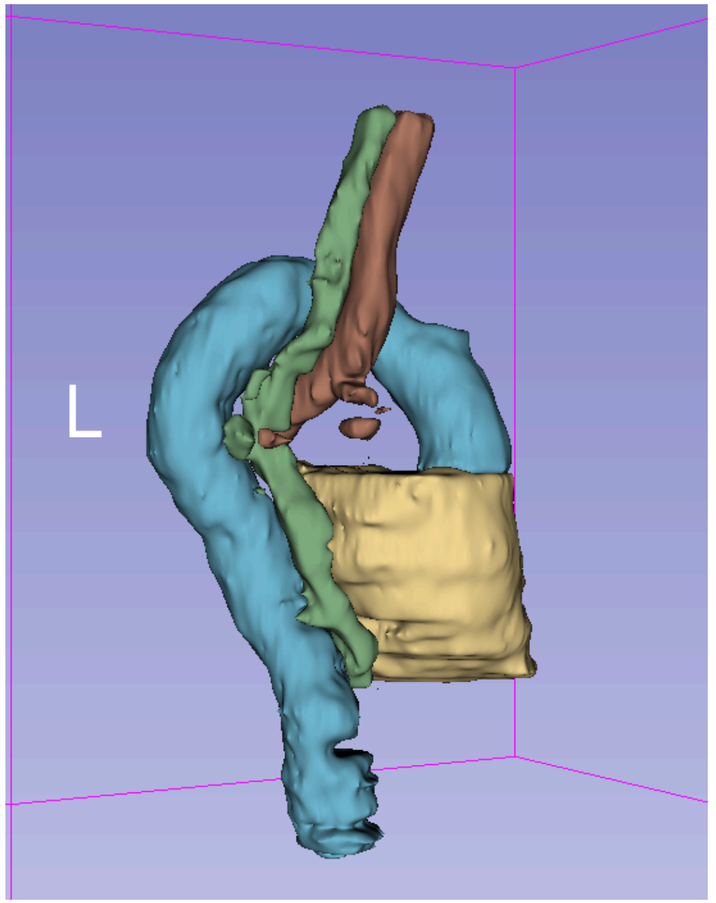




Windowing





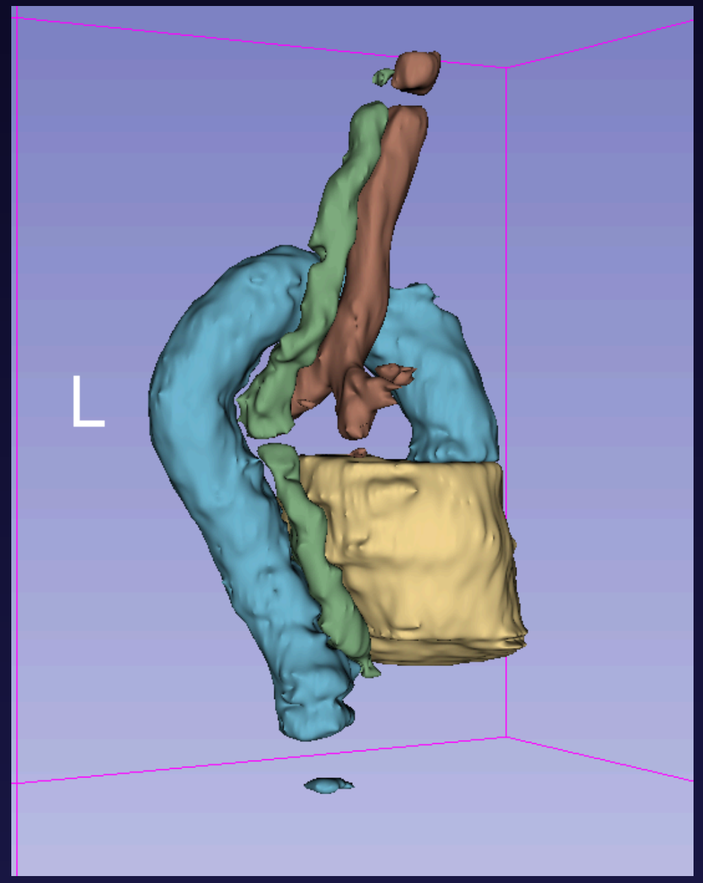


Noise in prediction

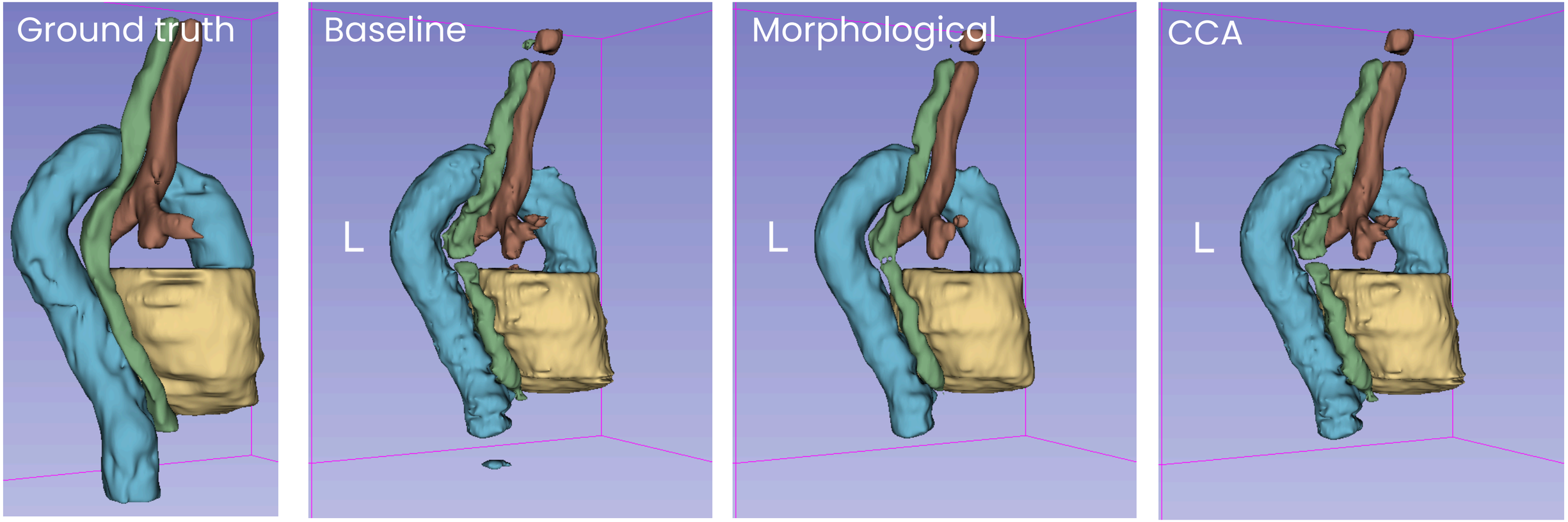
- Voxel noise
- Small islands

Ground Truth

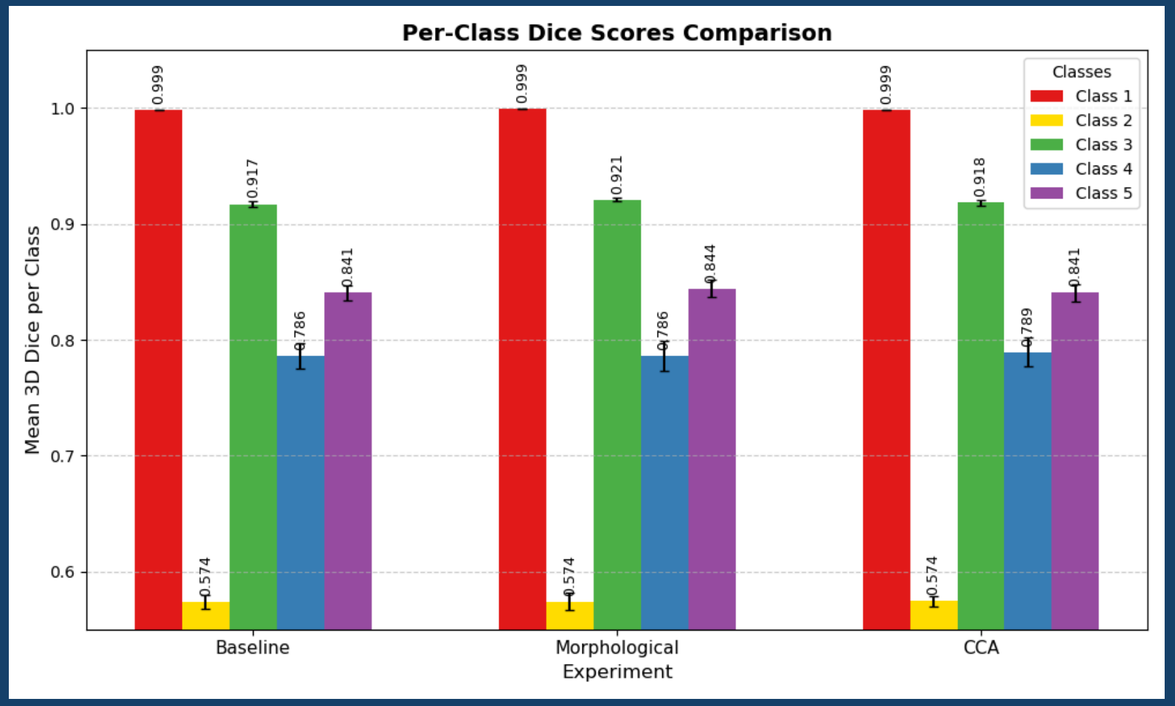
Baseline

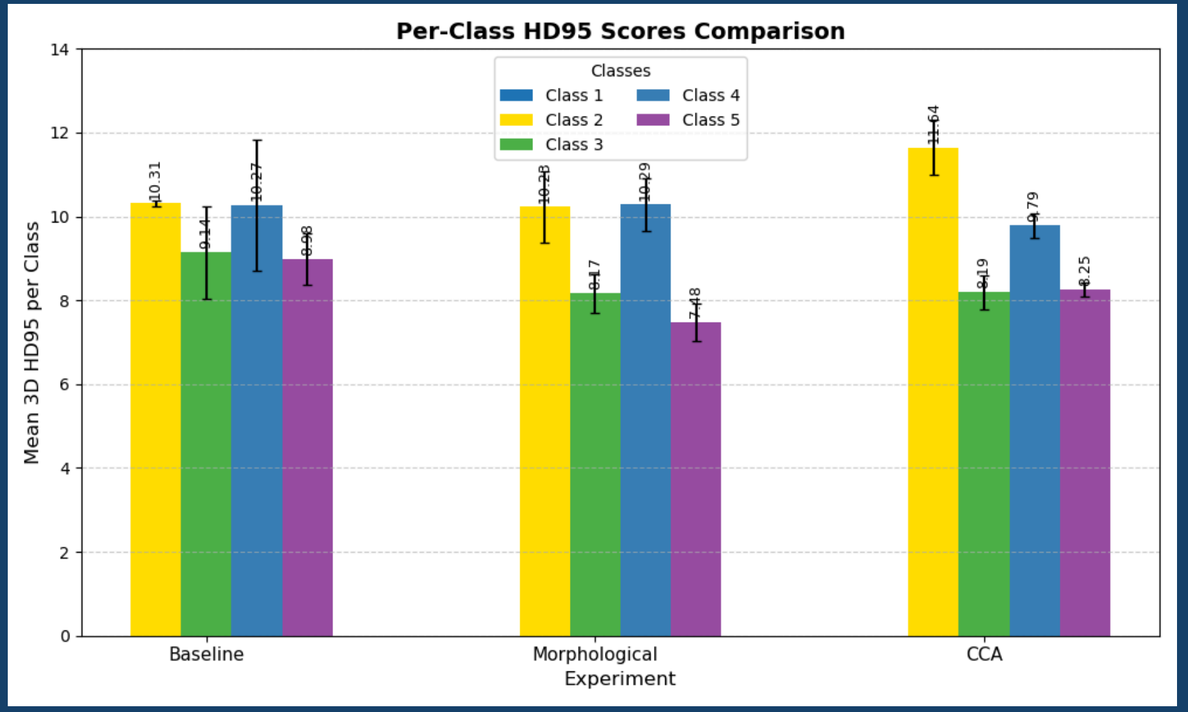


Morphological/CCA



Morphological/CCA analysis





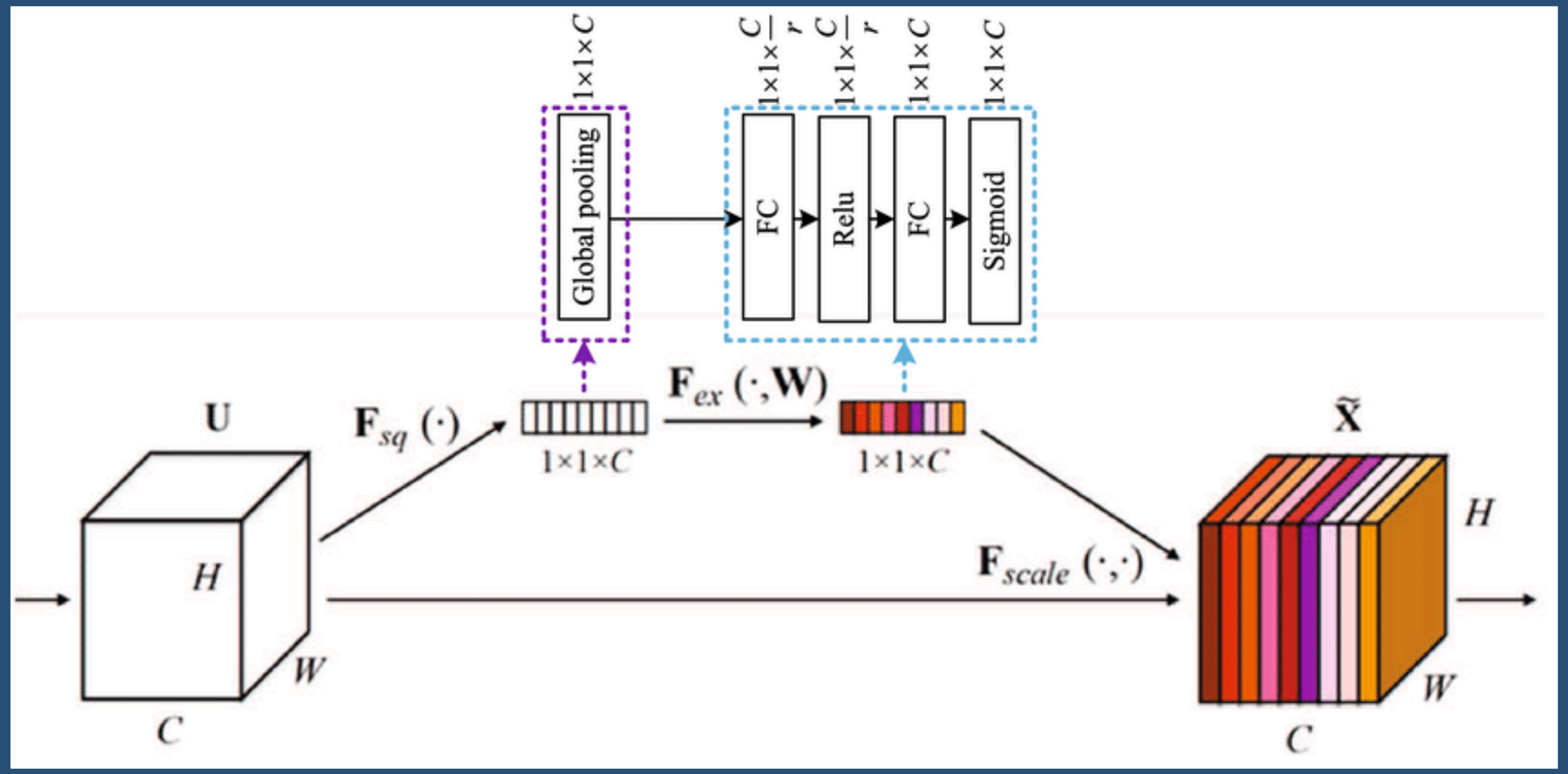
Squeeze and Excitation - why?

Reduce effect of noise

Emphasize important feature channels

Icons made by Freepik from www.flaticon.com

Squeeze and Excitation - how?



Zou, F., Xiao, W., Ji, W., He, K., Yang, Z., Song, J., Zhou, H., & Li, K. (2020). Arbitrary-oriented object detection via dense feature fusion and attention model for remote sensing super-resolution image. Neural Computing and Applications, 32(18), 14549–14562. https://doi.org/10.1007/s00521-020-04893-9

Squeeze and Excitation - impact?

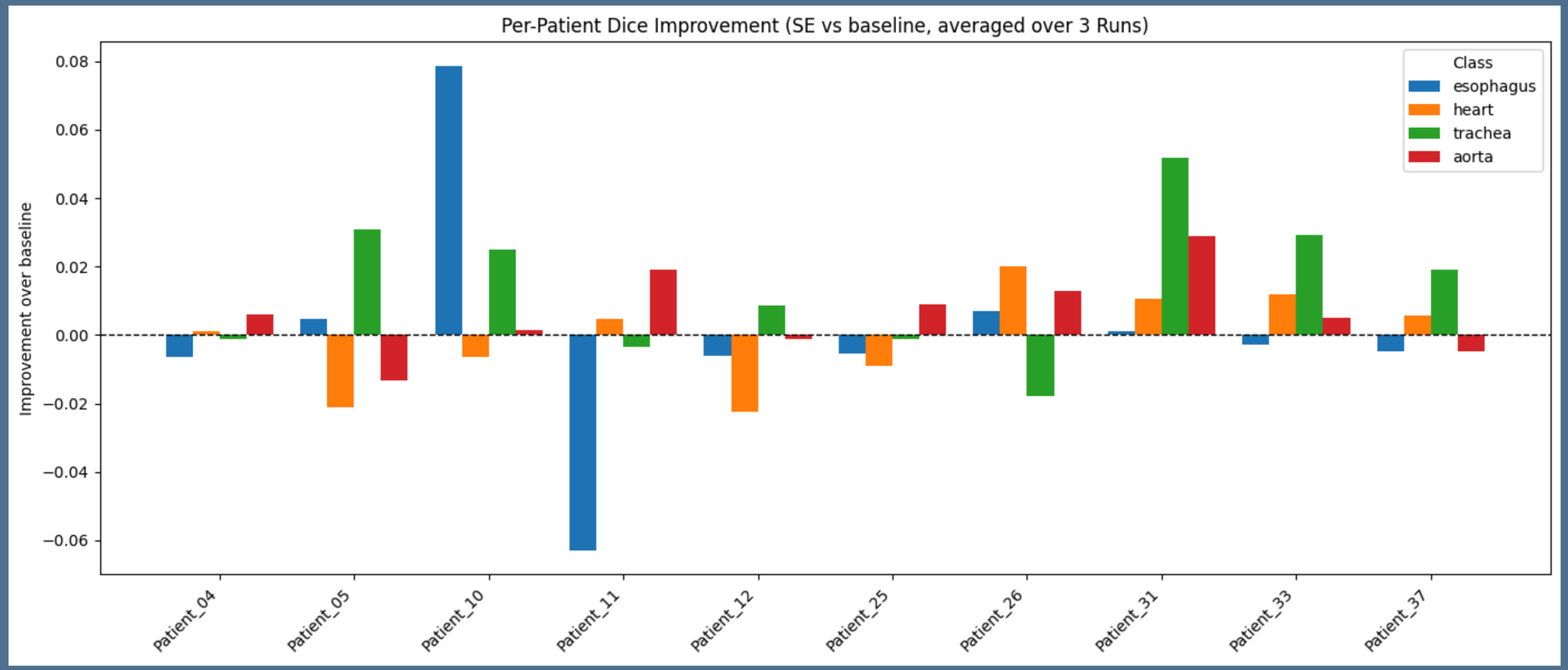
Overall, averaged across 3 runs

	3D DICE	HD95
ENet + SE	0.7845	10.2295
Improvement	+ 0.0050	+ 0.5536

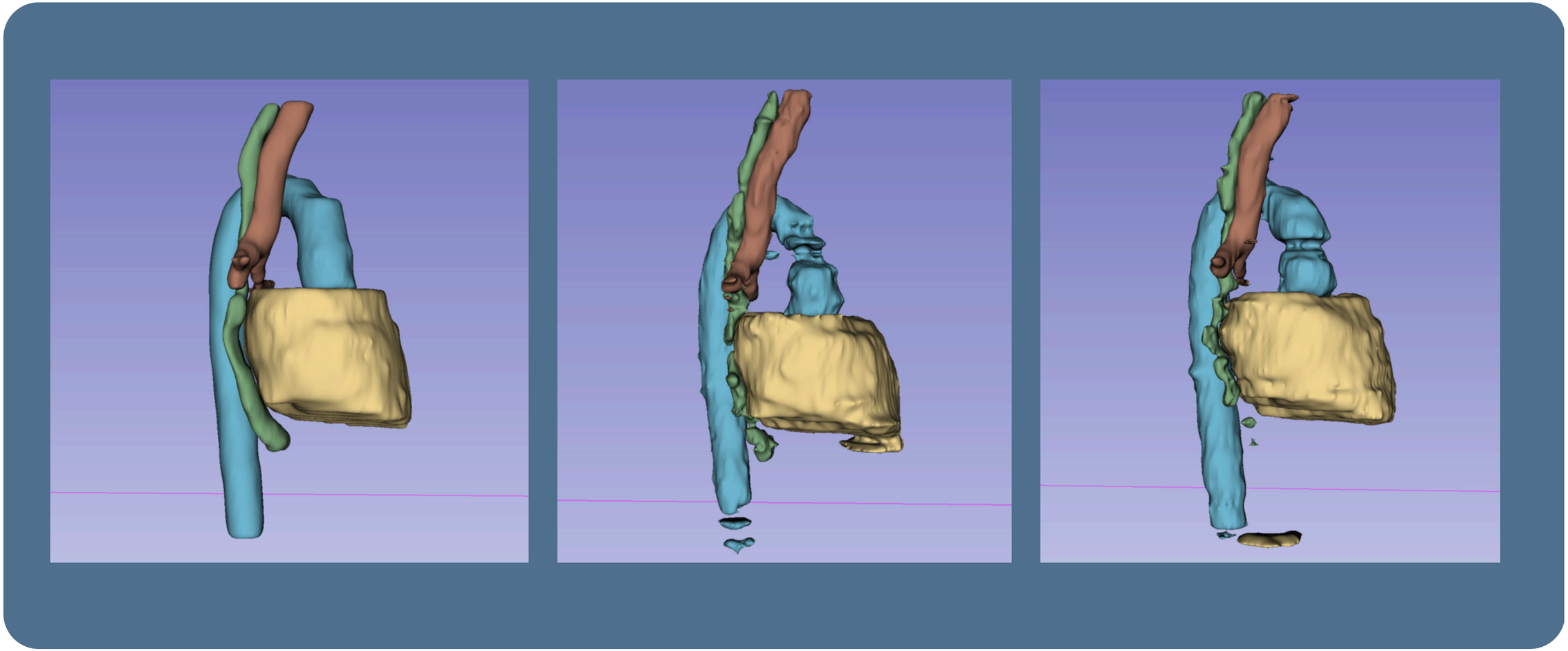
Per class, averaged across 3 runs

	3D DICE	HD95
Esophagus	+ 0.0003	+ 1.21
Heart	- 0.0006	+ 1.13
Trachea	+ 0.0141	- 0.66
Aorta	+ 0.0063	+ 0.54

Squeeze and Excitation - impact?



Squeeze and Excitation - impact?



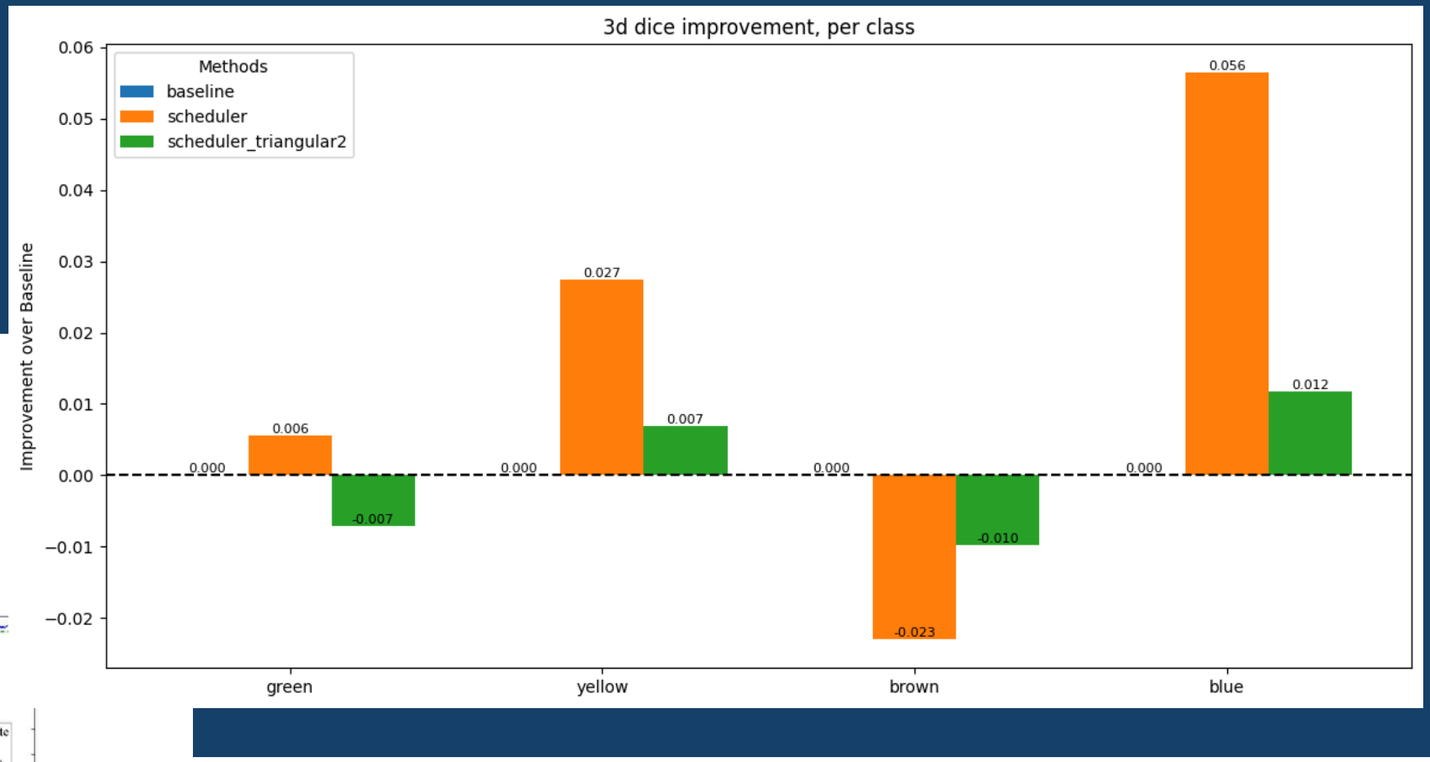
GT baseline SE

Standard optimization

Scheduler

CylicLR

- With amplitude scaling
- **Without amplitude scaling**



Cyclical Learning Rates for Training Neural Networks

Leslie N. Smith
U.S. Naval Research Laboratory, Code 5514
4555 Overlook Ave., SW., Washington, D.C. 20375
leslie.smith@nrl.navy.mil

Abstract

It is known that the learning rate is the most important hyper-parameter to tune for training deep neural networks. This paper describes a new method for setting the learning rate, named cyclical learning rates, which practically eliminates the need to experimentally find the best values and schedule for the global learning rates. Instead of monotonically decreasing the learning rate, this method lets the learning rate cyclically vary between reasonable boundary values. Training with cyclical learning rates instead of fixed values achieves improved classification accuracy without a need to tune and often in fewer iterations. This

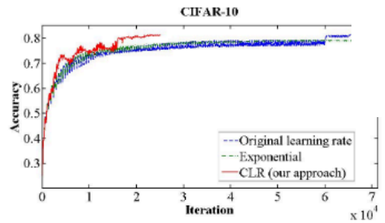
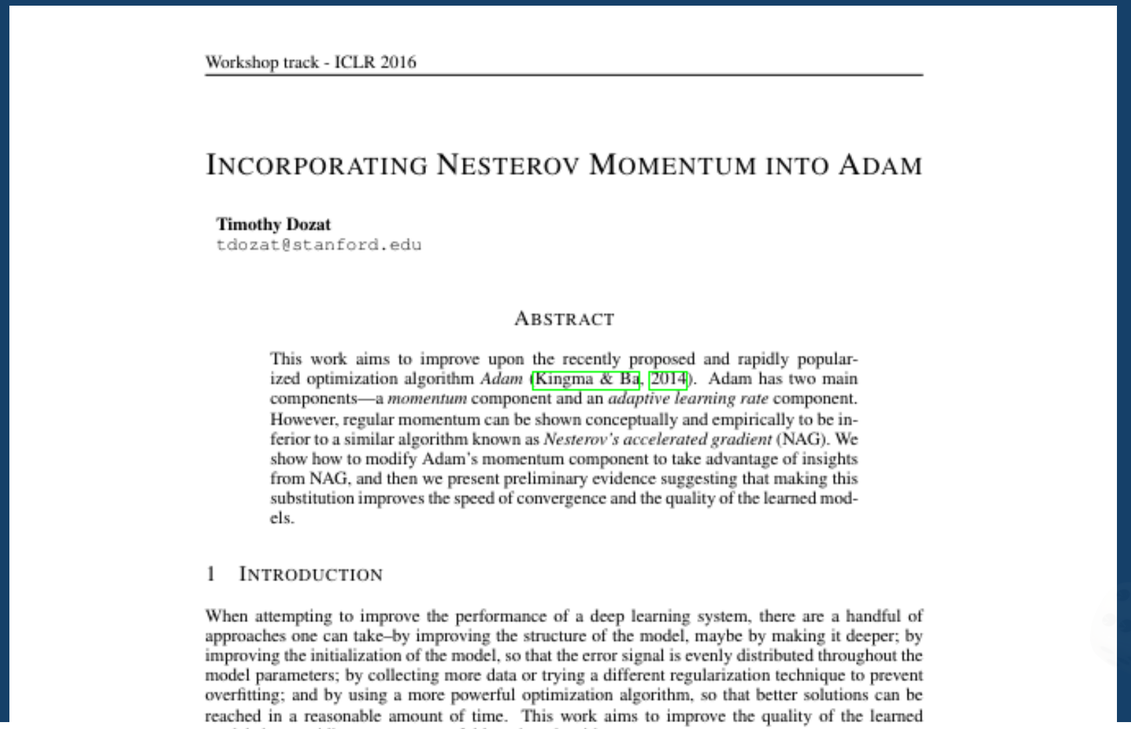


Figure 1. Classification accuracy while training CIFAR-10. The red curve shows the result of training with one of the new learning rate policies.

Smith L. N. (2017). Cyclical Learning Rates for Training Neural Networks, 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 2017, pp. 464-472

Optimizer

- Adam (baseline)
- AdamW
- NAG
- **NAdam**
- NAdamW



Workshop track - ICLR 2016

INCORPORATING NESTEROV MOMENTUM INTO ADAM

Timothy Dozat
tdozat@stanford.edu

ABSTRACT

This work aims to improve upon the recently proposed and rapidly popularized optimization algorithm *Adam* (Kingma & Ba, 2014). Adam has two main components—a *momentum* component and an *adaptive learning rate* component. However, regular momentum can be shown conceptually and empirically to be inferior to a similar algorithm known as *Nesterov's accelerated gradient* (NAG). We show how to modify Adam's momentum component to take advantage of insights from NAG, and then we present preliminary evidence suggesting that making this substitution improves the speed of convergence and the quality of the learned models.

1 INTRODUCTION

When attempting to improve the performance of a deep learning system, there are a handful of approaches one can take–by improving the structure of the model, maybe by making it deeper; by improving the initialization of the model, so that the error signal is evenly distributed throughout the model parameters; by collecting more data or trying a different regularization technique to prevent overfitting; and by using a more powerful optimization algorithm, so that better solutions can be reached in a reasonable amount of time. This work aims to improve the quality of the learned models by providing a more powerful learning algorithm.

Many popular learning algorithms for optimizing non-convex objectives use some variant of stochastic gradient descent (SGD); this work will consider a subset of such algorithms in its examination. Algorithm 1 presents SGD with the notation used in this paper–all following algorithms will add to or modify this basic template:

Algorithm 1 Stochastic Gradient Descent

Dozat, T. (2015). Incorporating Nesterov Momentum into Adam. International Conference on Learning Representations 2016

Final model vs baseline

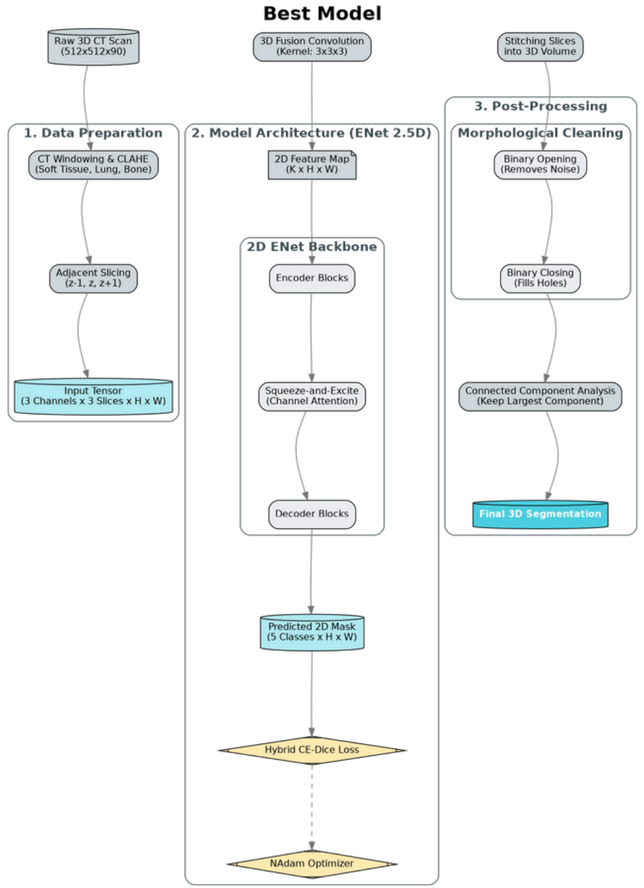
Methods used

?

Final 3D dice

- 2.5D
- NAdam optimizer
- Multi Window
- Squeeze and Excitation

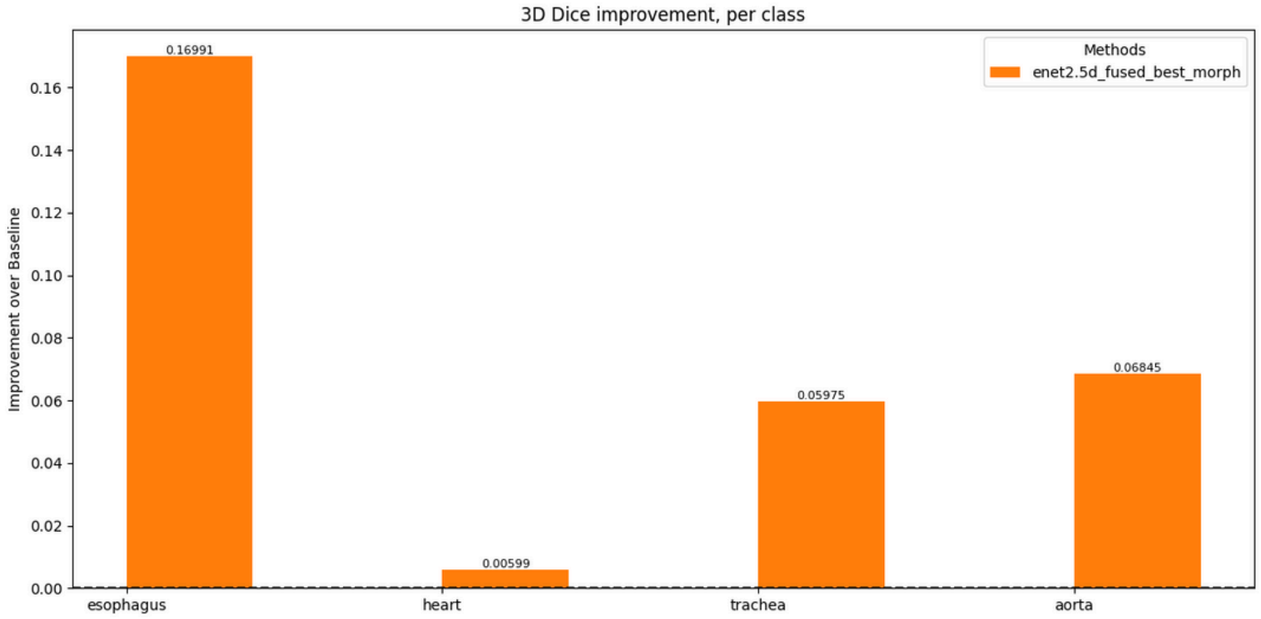
- Compound Cross Entropy and Dice loss
- Morphological/CCA postprocessing

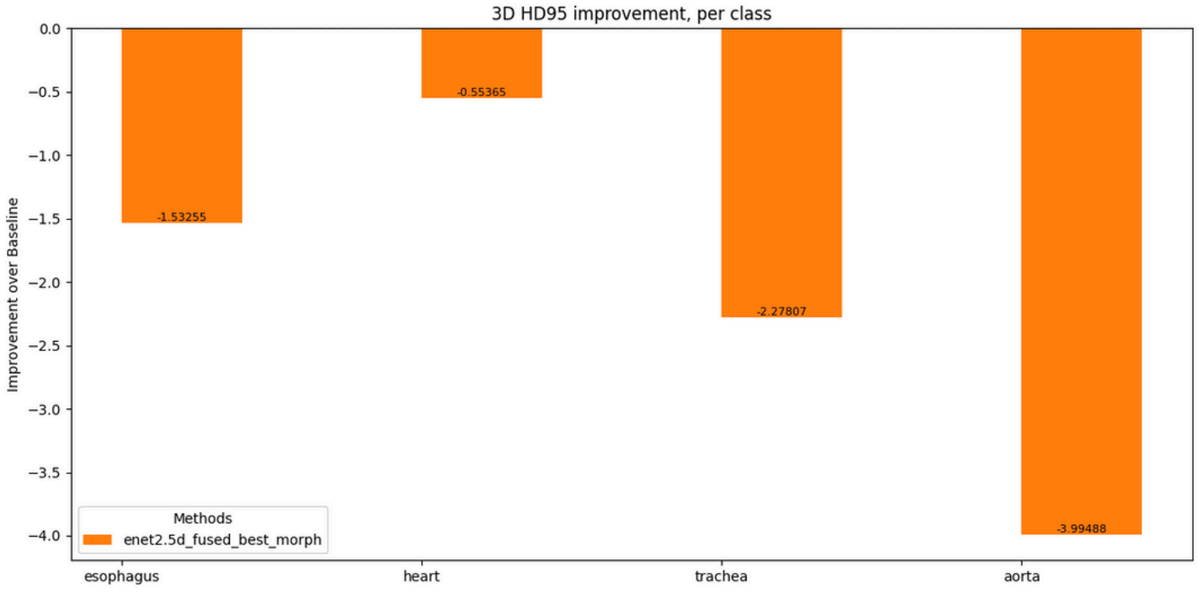


Baseline comparison

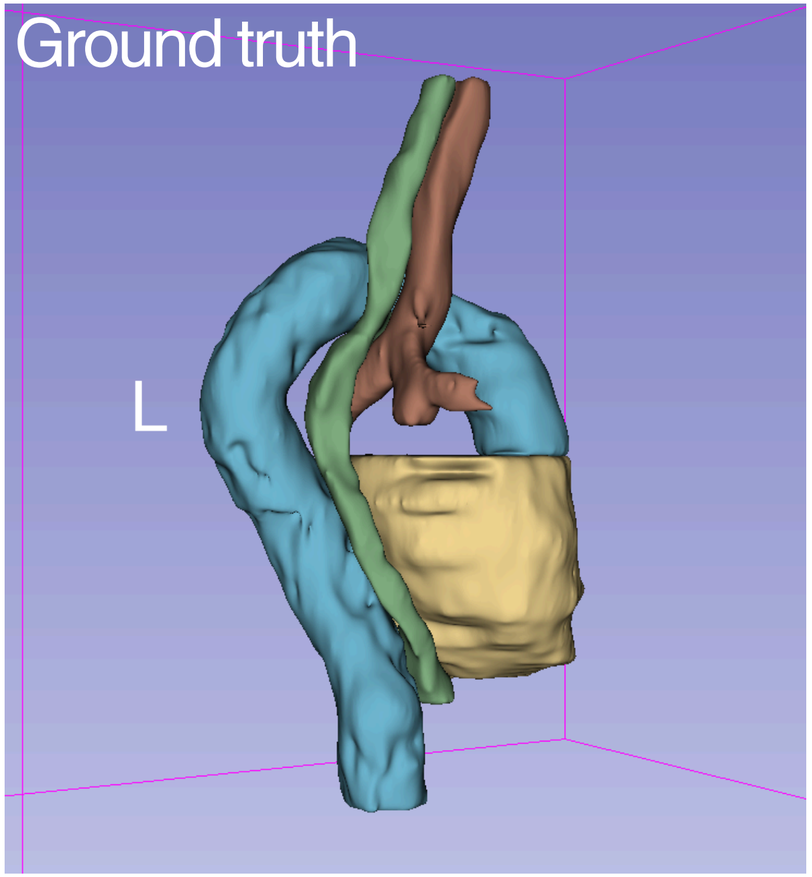
0.856
+0.076
Final 3D Dice

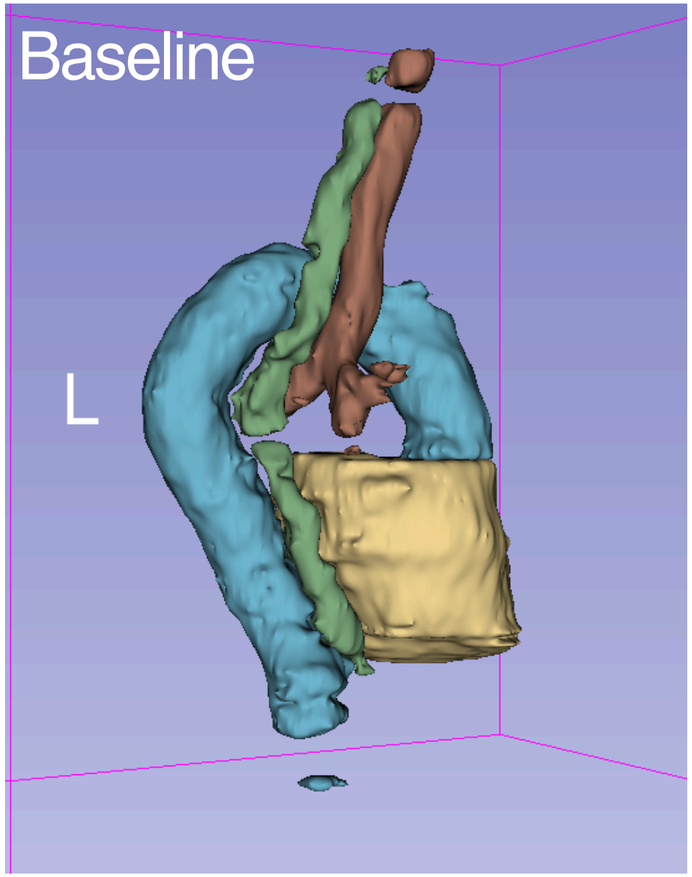
7.59
−2.08
Final 3D HD95

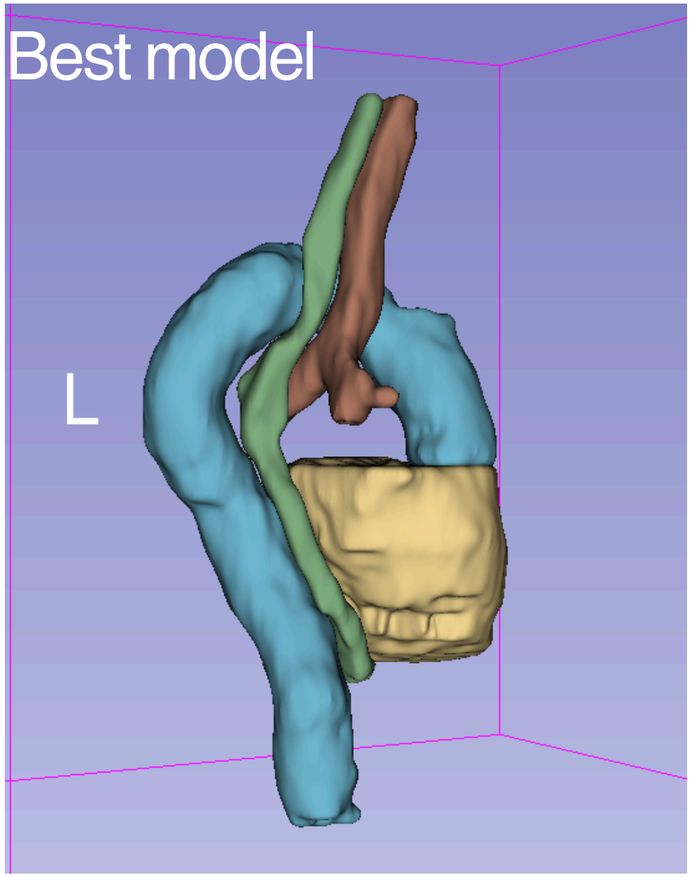




Comparison







Limitations

Limitations

The selection of the best epoch is based on 2D Dice, however the final metrics are all 3D. The model is not necessarily the best for the final task.

The severe class imbalance presents a core limitation, making it challenging to jointly optimize segmentation accuracy across both the rarest and the more dominant classes.

The convolution on the multi windows could be 1x1 convolutions to save computation and blend the information of the different windows

Thank you
Questions?