3D Self-Supervised Methods for Medical Imaging

Matthew Rubino, Caleb Schaefer, Aspen Smith, Shubhangi Thakur, Michael Wachsman

Cornell Bowers CIS Computer Science

Introduction & Background

Self-supervised learning (SSL) enables models to learn rich feature representations from unlabeled data by solving pretext tasks. It is especially useful in medical imaging, where obtaining labeled data is costly and time-consuming. This project recreates the methods from Taleb et al. [1], who extended 2D SSL techniques to 3D medical data. We reimplement and evaluate these methods on:

- **3D pancreas segmentation** using CT scans from the Medical Segmentation Decathlon [2].
- **2D diabetic retinopathy classification** using fundus images from the APTOS 2019 set [3].

We compare each SSL method's data efficiency by measuring downstream performance when fine-tuned with limited labeled data.

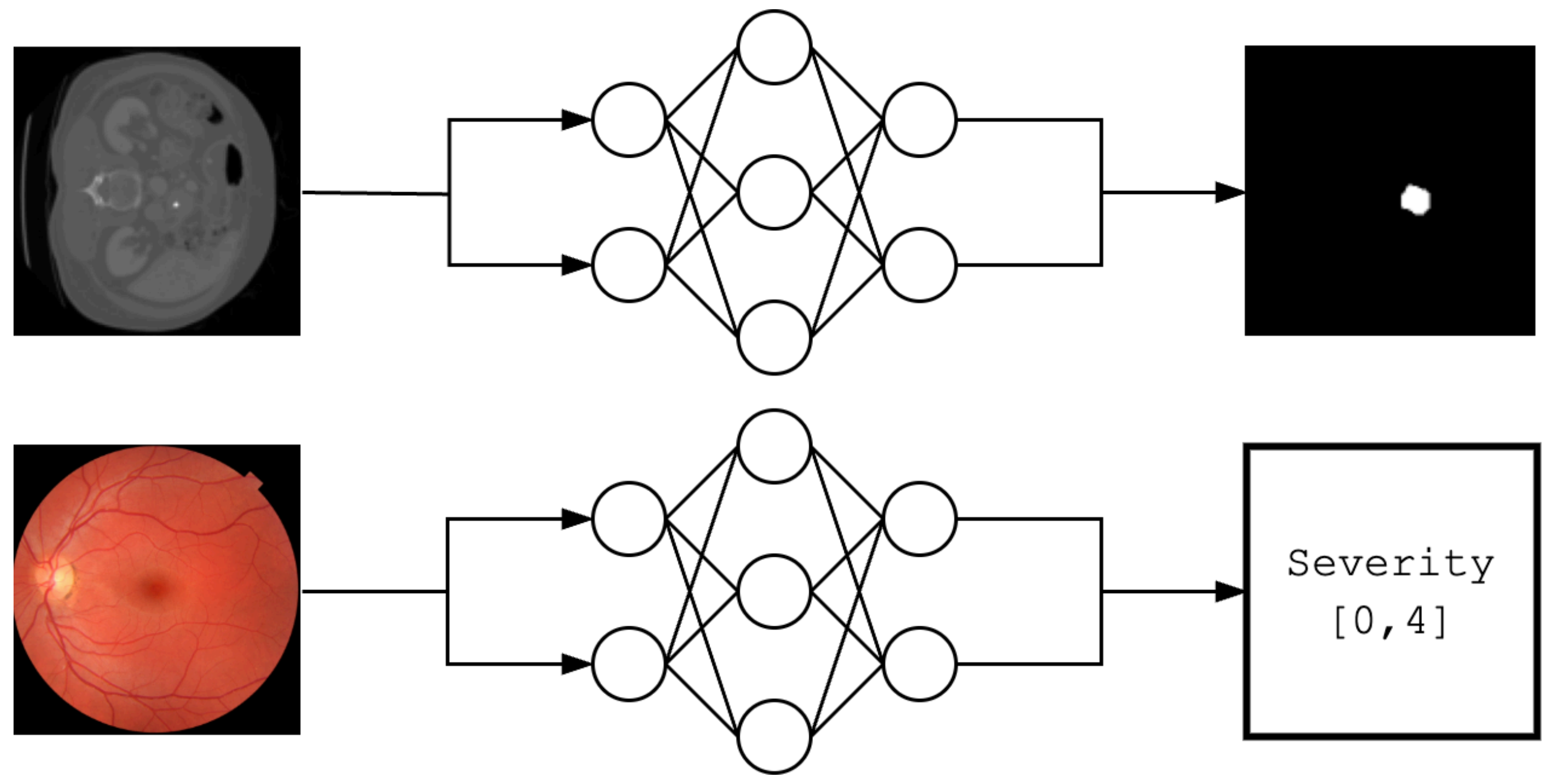


Figure 1 - Pancreas (t) and Fundus (b) Tasks

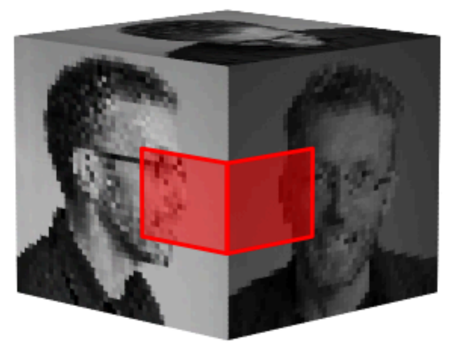
Pretext Tasks

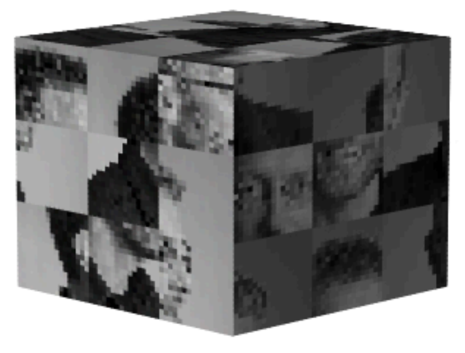
Task	Description	# Classes
Rotation	Predict which 3D rotation (e.g., 0°, 90°, 180°, 270°) was applied to the volume.	10
RPL	Predict the relative 3D offset between a pair of patches from the same volume.	26
Jigsaw	Predict the correct spatial arrangement of shuffled 3D patches.	100
Exemplar	Distinguish augmented views of the same volume from other volumes.	1024

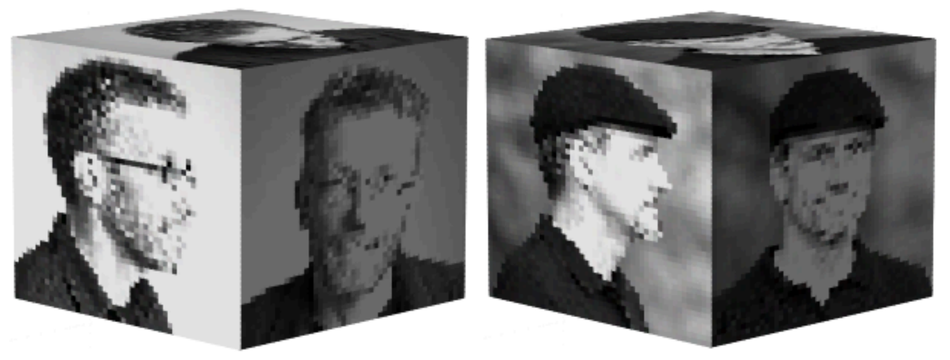
Table 1 - Pretext Task Descriptions

Input

Rotation


RPL


Jigsaw


Exemplar

Results

The graphs to the right show the effect of pretraining on finetuning efficiency. When finetuning on small amounts of data, pretraining has the greatest impact on downstream task performance.

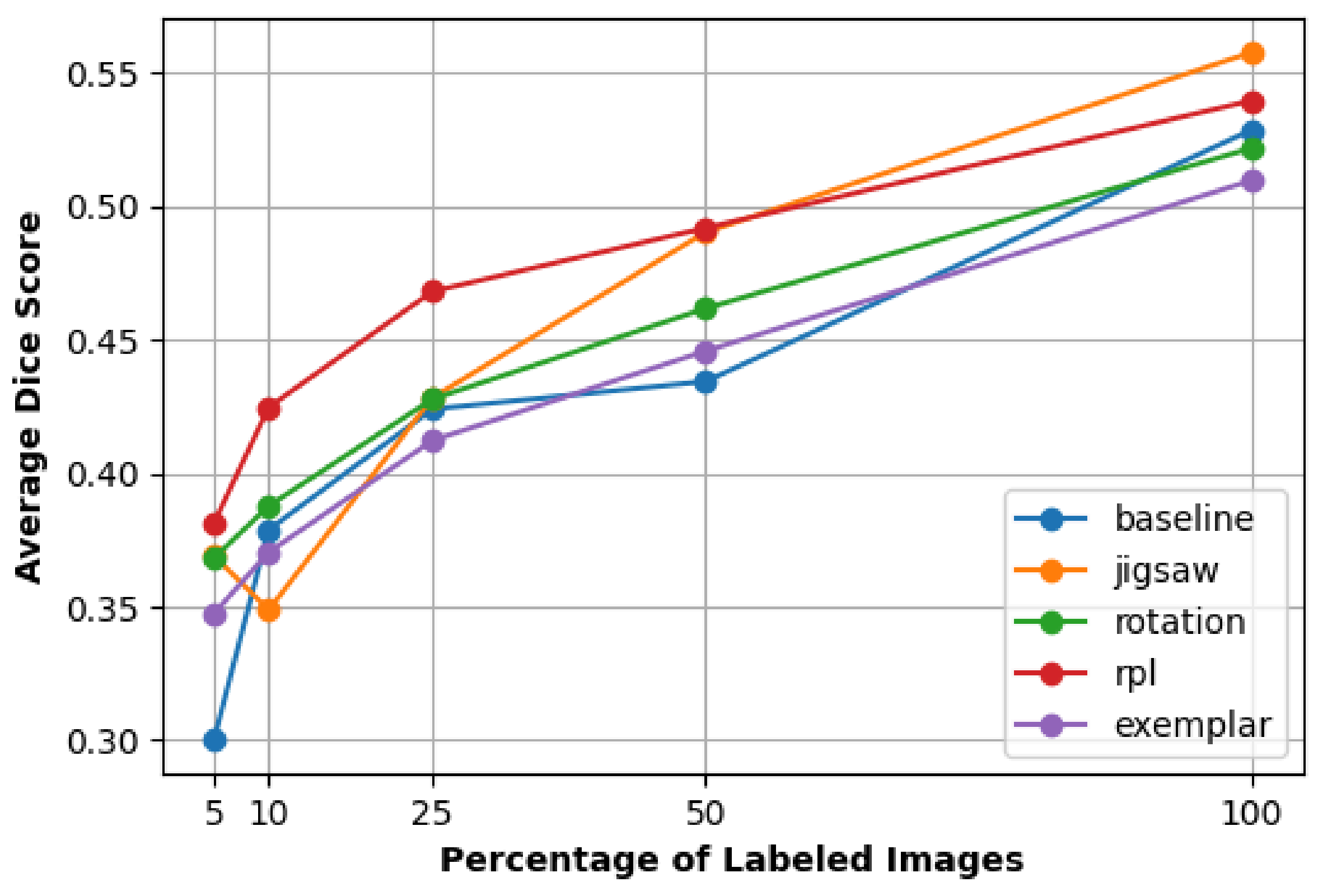


Figure 3 - Pancreas finetuning results (3D)

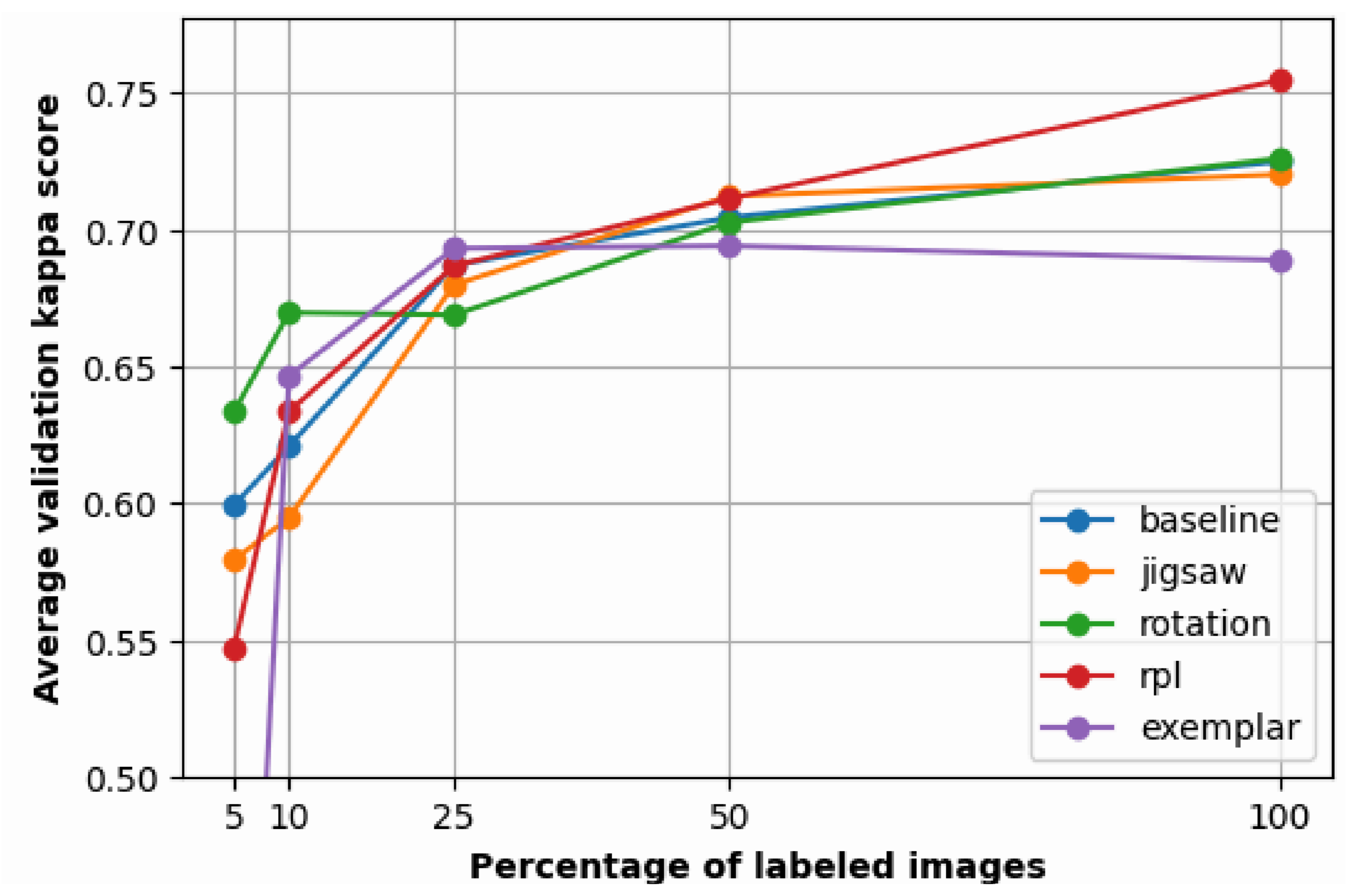


Figure 4 - Fundus finetuning results (2D)

We observe the same behavior as the authors with the following exceptions:

- **Lower 3D Dice scores**: Our 3D pancreas segmentation scores were consistently 5–10 points lower than those reported in the original paper. This is likely due to our use of a smaller UNet model due to limited VRAM, unlike the original multi-GPU setup.
- **Weaker 2D/3D Exemplar Performance**: We attribute this to our batch-level negative sampling, which likely offers less contrastive diversity than a memory bank-based strategy used in the original study.

Methodology

- **Architecture:** We adopt the original architecture choices, using a 3D UNet [4] for semantic segmentation and a DenseNet-121 [5] for 2D classification tasks. Using these, we implement the 3D SSL pretext tasks.
- **Finetuning Strategy:** During finetuning, a warmup routine is used for the encoder, while the classification head is trained at all epochs.
- **Evaluation Protocol:** Evaluation is done using dice (3D) and kappa (2D) scores using different amounts of data, helping measure downstream performance and data efficiency of each SSL method.

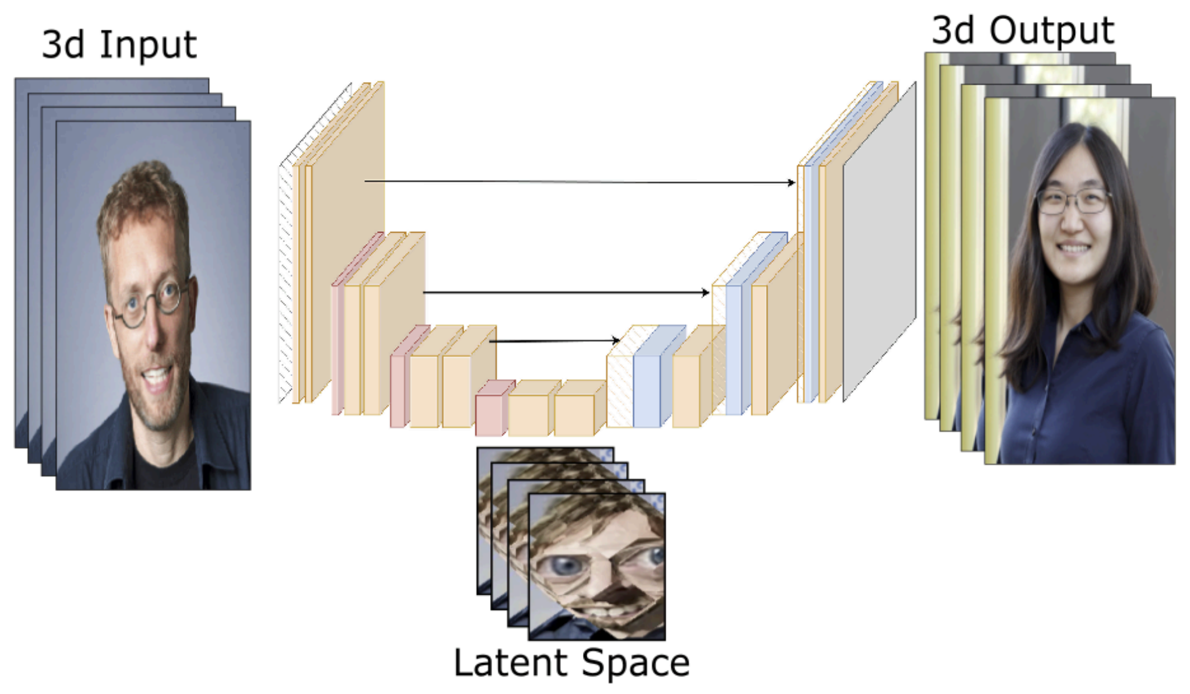


Figure 2 - 3D UNet Architecture

Conclusion & Future

- **Pretraining Boosts Efficiency:** Most self-supervised methods improved downstream performance compared to the baseline, highlighting the value of pretraining when labeled data is scarce.
- **Rotation and RPL Lead:** Rotation and RPL consistently outperformed other methods across both 2D and 3D tasks, demonstrating strong generalization for medical imaging pretext tasks.
- **Proposed Extension:** Apply self-supervised losses to intermediate encoder layers, rather than just the final one, to improve feature quality for decoding.

References

[1] A. Taleb et al., '3D self-supervised methods for medical imaging', in Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 2020.

[2] A. Simpson et al., 'A large annotated medical image dataset for the development and evaluation of segmentation algorithms', CoRR, vol. abs/1902.09063, 2019.

[3] M. Karthik and S. Dane, 'APTOS 2019 Blindness Detection', Kaggle, 2019.

[4] O. Ronneberger, P. Fischer, and T. Brox, 'U-Net: Convolutional Networks for Biomedical Image Segmentation', in Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2015, 2015, pp. 234–241.

[5] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, 'Densely Connected Convolutional Networks', in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261–2269.