We thank the reviewers for their supportive and insightful comments. We address their questions/concerns below.

Comments about novelty (*R1*: Limited technical contribution, *R3*: The novelty is modest):
As noted by *R2&R3*, existing results for self-supervised methods have mainly been obtained on ImageNet. We extend these methods to 3D medical imaging, where labels are expensive to obtain, by pretraining on a large unlabeled corpus (UK Biobank) or on images from the same dataset. Furthermore, as explained in *R2*-1, we propose extensions that work in 3D contexts, e.g. for CPC, which was not trivial. Moreover, generalizing a concept from lower- to higher-dimensions is common in the literature (see [1] and lines 95-126 in our paper), and can offer insights for novel applications.

R2: **1-** "The extensions from 2D to 3D seem relatively natural... pitfalls encountered when shifting from 2D to 3D." Extending CPC to 3D was not straightforward. In 1D, the future values are predicted based on history. In 2D, the prediction is performed row- and column-wise, i.e. solving many 1D tasks. In our experiments, similarly small contexts yielded poor results in 3D. Too large contexts (e.g. full surrounding of a patch) incurred prohibitive computations and memory use. The inverted-pyramid context was an optimal tradeoff. We will include a comparison of these variants.
2- In pancreas tumor and retinopathy experiments, "unsupervised data is created artificially by discarding labels.." Medical datasets are supervision-starved (lines 27-33), e.g. images may be collected as part of clinical routine, but much fewer (high-quality) labels are produced, due to annotation costs. However, we agree that a transfer learning setting is more significant, as it leverages additional data from different distributions. Hence, we pretrained on Retinopathy data from the UK Biobank (170K images), and fine-tuned on Kaggle data (5K images). Transfer learning yielded gains (in Qw-Kappa), in 24/25 settings (see table). We plan to include pancreas-tumor segmentation into this evaluation.

Model / (% of data)	CPC					RPL					Jigsaw					Rotation					Exemplar				
	5	10	25	50	100	5	10	25	50	100	5	10	25	50	100	5	10	25	50	100	5	10	25	50	100
Pretrained (UKB)	24	61	71	77	79	42	63	70	75	76	25	45	70	77	79	26	48	69	78	79	48	59	70	74	76
Baseline (Kaggle)	18	44	56	63	72	20	42	62	69	74	20	52	58	67	72	12	42	59	72	73	13	24	64	72	67

3- UK Biobank baseline pretrained on longitudinal segmentation labels, and transfer learning to BraTS (100% labels). The longitudinal labels are for fMRI. However, we added an experiment based on automatic labels from FSL-FAST, which include masks for three brain tissues. Our results in Tab.1 (paper) are comparable to this baseline (table below).

Model / BraTS Metrics	ET	WT	TC
Pretraining on FAST masks (UKB)	78.88	90.11	84.92

4- Discussion of the computational requirements (hardware used, flops spent, etc). We will add these to the final version.
5- For brain-tumor segmentation, our methods get near Isensee et al.'s. Discuss why their method is marginally better. Isensee et al. use more training data, a larger U-Net, and post-processing. Our 3D-RPL is comparable (lines 259-263).
6- Fig. 4, why Exemplar gets worse at 100%? Is this trend real or noise? If it is noise, then error bars are needed. We believe this drop at 100% of the data is caused by noise, and hence will add error bars to the final version.
7- How much does data augmentation matter, in particular for Exemplar. Recently, SimCLR shows big gains. Our findings are consistent with SimCLR, i.e. combined data augmentations in Exemplar improve learned representations. However, the types of augmentations may differ. An analysis about this will be added to the final version.
8- On ImageNet, Exemplar-based methods outperform others. Yet, in our experiments, Exemplar is not the best.. Exemplar-based methods can be affected by: training loss (contrastive vs. triplet), domain-specific tuning, negative sampling (batch vs. dataset), ... We discuss this in the final version. Also, implementing a 3D SimCLR is a future work.

R3: **1-** How to modify these methods by taking advantage of some specific prior knowledge in the medical domain. We aim to develop novel methods that utilize data-locality in 3D. Thank you for the suggestion.

R4: **1-** The motivation of five self-supervision approaches given that the SOTA is set by contrastive learning approaches. As accurately noted by *R2,R3*, all previous SOTA is set on ImageNet, and it is hard to generalize such results to different contexts (2D natural vs. 3D medical images). We plan to extend contrastive approaches to 3D contexts in the future.
2- Potential technical challenges, and a comparison to 2D+T methods on video inputs.
Please refer to *R2*-1 and *novelty comments* for a discussion about technical challenges. Moreover, as explained in lines (95-111), in contrast to 2D+T methods, our methods exploit the whole 3D context (including the depth dimension).
3- Solving a segmentation task with the self-learnt prediction embeddings. What about spatial/texture details? The applicability of our methods to several tasks is a benefit we had in mind. We agree that segmentation tasks require learning more details, however, our results in Fig.2&Fig.3 confirm that pretraining the encoder only is able to capture generic data representations, similar to other self-supervised methods [2]. This enforces the decoder network to capture these spatial and texture details during fine-tuning. We will add an analysis to the final version.
4- Some recent technical references are missing. SimCLR is ref. (24) in our paper. We will add the others, thanks.

References

[1] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
[2] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *ArXiv*, abs/1906.05849, 2019.