We thank the reviewers for their time and detailed comments. All reviewers appreciated the novelty of the method and the thoroughness in the experiments. We categorize the concerns of the reviewers and the corresponding responses into the following 3 groups: A. regarding the method, B. regarding experiments and evaluation, and C. miscellaneous.

A1. Contributions [R1]: The main concern of R1 is that the paper *just* proposed a new strategy for selecting +ve and -ve pairs for the contrastive loss. We emphasize that, as acknowledged by R3, R4, the paper has 3 main contributions: we (1) leverage domain knowledge to form appropriate +ve and -ve pairs, leading to clear gains over random augmentations as done in prior works [12], (2) propose a local contrastive loss useful for dense prediction tasks like segmentation and (3) show that pre-training is complementary to semi-supervised and data augmentation methods.

A2. Dependence on registration [R1, R4]: The method requires only rough alignment across volumes. This can be obtained with very basic registration, even using the transformation matrices located in the header files of medical images without an external registration algorithm. As a demonstration, in all the experiments presented in the article, we **did not** perform any registration and used volumetric images as they were distributed in the challenge datasets.

A3. Effect of multiple classes within a local region [R1]: L_l does not take any label information into account. So, even when a local region consists of several labels, its representation contains information about the entire local region. L_l seeks to make this representation consistent across various intensity transformations and simultaneously be different from other distant local regions within the image.

A4. Effect of domain-specific knowledge in local loss L_l [R1, R4]: L_l is a novel loss proposed by us, which improves performance as compared to only using the global loss L_g (as seen in Table 1, row 5 in the main article). We further propose and study the effect of two sampling strategies within L_l: (a) L^D, where local regions are matched across volumes (referred to as using domain knowledge), and (b) L^R, which does not assume such correspondences. Our experiments show that L^R performs better than L^D. We believe that this is not a drawback of the method, but instead, an indication that obtaining perfect in-plane alignment across volumes is difficult due to inter-subject variability (also pointed out by R4). We view L^R as a contribution of the proposed work.

A5. Effect of local region size ($K \times K$) [R1]: We ran this ablation experiment on the remaining datasets (with $d_l = 3$, and sampling strategies G^D, L^D). Results (Table 1) show that 3×3 works better for most settings, as seen with ACDC.

A6. Stage-wise v/s joint training [R4]: Results with joint training are shown in Table 2. We define the total loss: $L_{net} = L_g + \lambda_l * L_l$, where λ_l is a hyper-parameter to balance loss values. As per R4's idea, the encoder weights are updated with the net loss L_{net} that includes L_l, unlike our stage-wise training, where only L_g was used to update the encoder. We tried 4 values of λ_l on ACDC dataset for d_l=3. Results indicate that stage-wise training (where DSC is 0.725 for $|X_{tr}|$=1 and 0.789 for $|X_{tr}|$=2) performs better.

Dataset	$K \times K$	$\|X_{tr}\|$=1	$\|X_{tr}\|$=2
Prostate	1×1	0.554	0.614
	3×3	0.567	0.607
MMWHS	1×1	0.559	0.674
	3×3	0.574	0.681

Table 1: (A5) Effect of local region size.

A7. Relevance of the method for 3D CNNs [R4]: We agree that the proposed pre-training (L_g, G^D) may be informing the 2D CNN about the 3D structure of medical images. We believe that this is beneficial as compared to training 3D CNNs, where one faces memory issues as well as has more risk of overfitting due to a higher number of parameters.

λ_l	$\|X_{tr}\|$=1	$\|X_{tr}\|$=2
1	0.634	0.741
10	0.633	0.730
100	0.643	0.745
1000	0.644	0.739

Table 2: (A6) Joint training.

B1. Experimental setup: *(a) Data split [R3]*: The data split was chosen with the idea of keeping the number of volumes for pre-training (X_{pre}) and testing (X_{ts}) to be roughly around 50% of each dataset. For Prostate, although we have 48 volumes, labels were provided only for a subset of them, so the number of volumes for each set were adjusted accordingly. For ACDC, we ran the benchmark training with $|X_{tr}| = 78$ instead of 50 and obtained test DSC of 0.912, comparable to 0.908 obtained with $|X_{tr}| = 50$. We are happy to add these details in the revised supplementary. *(b) Validation set X_{vl} [R3]:* We use X_{vl} fixed to 2 3D volumes during fine-tuning to determine when to stop the training.*(c) Fine-tuning [R1]:* As mentioned in line 251, we experiment with 3 settings: $|X_{tr}|$ = 1, 2, and 8 3D volumes with X_{vl} fixed to 2 3D volumes.

B2. Comparison with [63] [R1]: We compare with [9], also based on data augmentation (like [63]), but more general in that it does not depend on a deformable registration step, which is difficult to achieve for anatomies other than the brain.

B3. Details of training time and convergence [R4]: On a Titan X GPU, training takes about: (a) 2 hours for L_g pre-training, (b) 4 hours for L_l pre-training, and (c) 2 hours for fine-tuning. Also, we found the pre-training convergence to be consistently stable. We will add these details in the revised supplementary and also make the code public.

C1. Clarity in notation [R4]: We really appreciate the detailed comments provided by R4. We will incorporate the suggested notational changes and required additional details in the revised version.

C2. Writing [R4]: We agree that the comment regarding batch size for pre-training with medical images is too strong. We will tone this down appropriately in the revised version.