Local Spatiotemporal Representation Learning for Longitudinally-consistent Neuroimage Analysis

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent self-supervised advances in medical computer vision exploit the global and local anatomical self-similarity for pretraining prior to downstream tasks such as segmentation. However, current methods assume i.i.d. image acquisition, which is invalid in clinical study designs where follow-up *longitudinal* scans track subject-specific temporal changes. Further, existing self-supervised methods for medically-relevant image-to-image architectures exploit only spatial or temporal self-similarity and do so via a loss applied only at a single image-scale, with naive *multi-scale spatiotemporal* extensions collapsing to degenerate solutions. To these ends, this paper makes two contributions: (1) It presents a *local* and *multi-scale* spatiotemporal representation learning method for image-to-image architectures trained on longitudinal images. It exploits the spatiotemporal self-similarity of learned multi-scale intra-subject image features for pretraining and develops several feature-wise regularizations that avoid degenerate representations; (2) During finetuning, it proposes a surprisingly simple self-supervised segmentation consistency regularization to exploit intra-subject correlation. Benchmarked across various segmentation tasks, the proposed framework outperforms both well-tuned randomly-initialized baselines and current self-supervised techniques designed for both i.i.d. and longitudinal datasets. These improvements are demonstrated across both longitudinal neurodegenerative adult MRI and developing infant brain MRI and yield both higher performance and longitudinal consistency.

1 Introduction

Tracking *subject-specific* anatomical trends over time is crucial to both clinical diagnostics and large-scale biomedical science. Such *longitudinal imaging* is especially relevant to analyzing neurological patterns of growth and degeneration via brain imaging in pediatric and elderly populations, respectively. As tracking individual structural changes requires precise and longitudinally-consistent segmentation methods with scarce annotated training volumes, we identify two major bottlenecks in existing self-supervised biomedical image analysis methods which we use to motivate our work.

Learning with few annotations. While modern imaging studies may scan hundreds to thousands of individuals, manually outlining volumetric structures of interest across multiple individuals for supervised segmentation training is prohibitively expensive. Therefore, current work focuses on leveraging large sets of unlabeled images to *pretrain* image-to-image architectures (e.g., the U-Net [48]), which can then be efficiently finetuned in the one or few-shot setting. These *self-supervised* methods may handcraft pre-text training objectives [10, 18, 45, 65] or may attempt to pretrain the base network to be equivariant to transformationsin order to preserve semantic meaning as in contrastive learning in medical [9, 63] and natural image vision [2, 24, 37, 60, 67, 68]. However, these methods typically leverage label supervision in sampling for their losses and impose a self-supervised loss

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

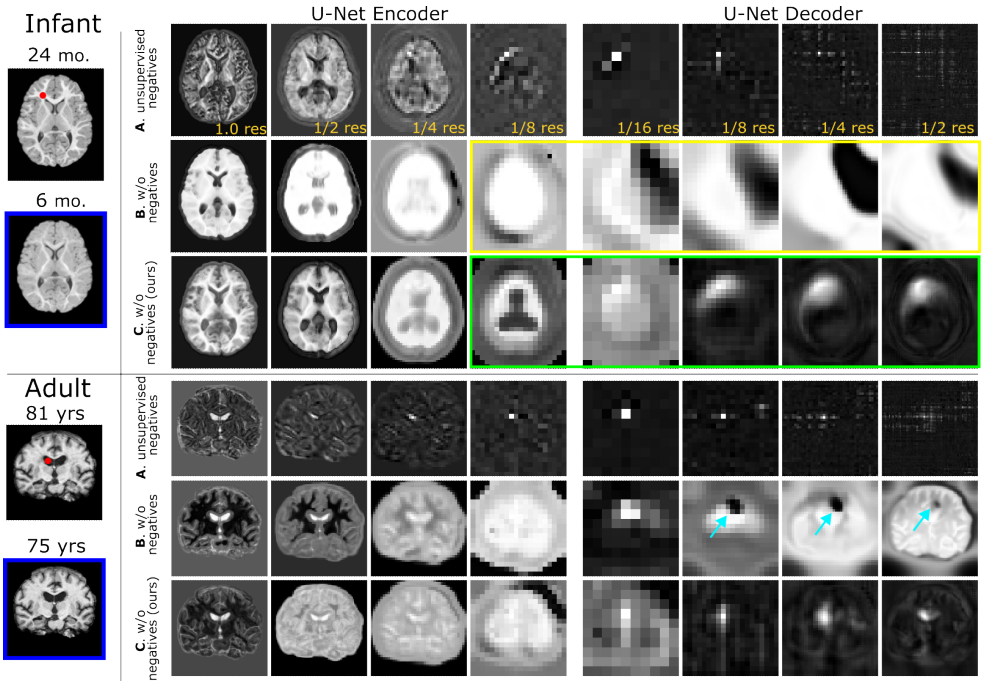

Figure 1: On pretraining an image-to-image network with *per-layer* spatiotemporal self-supervision, we visualize the **intra-subject multi-scale feature similarity** between a **query** channel-wise feature and all spatial positions within the **key** feature at a different age. **A:** Contrastive pretraining with unsupervised negatives [44] yields only positionally-dependent representations. **B:** Pretraining w/o negatives [11] by using corresponding intra-subject patch locations as positives leads to semantically-implausible representations with low-diversity (e.g., see yellow box) and artifacts (see arrows) in deeper layers. **C:** Our method attains both positionally and anatomically-relevant representations via proper regularization (e.g., see green box). Additional structures are visualized in Suppl. Figure 5.

only at a single feature scale (typically the encoder bottleneck or network output). Naive application of local unsupervised multi-scale contrastive losses [44] lead to degenerate representations (Fig. 1, row A) when applied to both the encoder and decoder of image-to-image architectures.

Violating i.i.d. assumptions. Further, most existing self-supervised frameworks assume i.i.d. data. Unfortunately, this assumption does not transfer to longitudinal studies where images of the same subject are highly correlated. Emerging longitudinal representation learning methods focus on imposing temporal-consistency into the encoder bottleneck [15, 43, 66], such that the encoder learns representations which are aware of the order of acquisition [15] and the overall trajectory [43]. These methods address image-level tasks such as disease classification or progression and age prediction. However, their extension to pixel-level applications with image-to-image architectures remains unclear.

Methods. Motivated by the above limitations, in this work, we claim that the spatiotemporal dependency of imaging data should be explicitly incorporated in self-supervised frameworks. We do so by exploiting the spatial and temporal self-similarity of local multiscale deep features in both encoder and decoder and further learn diverse intermediate representations by developing regularizations for self-supervised similarity objectives. Lastly, when finetuning with limited annotated data, we encourage predictions on unlabeled subject-wise images to be spatiotemporally consistent.

Contributions. This work makes the following contributions: (1) It presents a longitudinally-consistent spatiotemporal representation learning framework to learn from image time-series; (2) To impose multi-scale local self-supervision while avoiding degenerate solutions, it develops regularization terms on the variance, covariance, orthogonality of local features within the decoder; (3) To further self-supervise the fine-tuning stage, the proposed method encourages segmentations from adjacent timepoints on unlabeled data to be consistent; (4) Across three large-scale longitudinal one, few, and full-shot segmentation tasks on both elderly and pediatric populations, the developed framework yields improved segmentation performance and higher longitudinal segmentation consistency.

2 Related work

Self-supervision. Self-supervised learning (SSL) methods aim to learn hierarchical visual representations from unannotated data, which can then be transferred to low-annotation regimes. Early work focused on pretext tasks where a handcrafted loss is used to pretrain networks via orientation prediction [18], context restoration [10, 45], channel prediction [65], among others. Given their heuristic nature and suboptimal generalization, recent work instead focuses on data-driven SSL losses.

Contrastive learning. Constrastive learning [11, 21, 22, 23, 49, 55] (CL) typically transforms an input image and asks the embeddings of the input image and its transformation (the *positive* pair) to be close to one another and far apart from embeddings of other images (*negatives*) via a noise contrastive estimation [20, 42] (NCE) loss. While performant on image-level recognition tasks [12, 31], CL requires non-trivial modification to extend to segmentation tasks, as described further below.

Negative-free representation learning. In several applications, *true* negative samples may be difficult to construct [25]. For example, when learning on intra-domain internal image patches [44], non-local spatial positions may be semantically similar, but NCE objectives push their embeddings apart. This drawback may be mitigated via data-driven SSL methods which only use positive samples and avoid low-diversity (or *collapsed* [29]) embeddings solutions via predictor networks and custom backpropogation [13, 19] and careful regularization [6, 62]. However, as above, these methods operate on *global* image embeddings and require modification for pixel-level tasks.

Spatial self-supervision. Towards downstream segmentation, recent work [2, 24, 34, 37, 38, 60, 67, 68] encourages local single-scale features either within an image or across images to cluster according to their semantic labels by constructing positive pairs using ground-truth labels. To incorporate *local* and *multi-scale* spatial considerations into unpaired image translation, [44] imposed contrastive losses on randomly-sampled layer-wise encoder features by considering corresponding spatial indices as positives and all other locations as negatives. Our work builds on this by instead only considering temporal positives (described below) in the layerwise losses alongside custom regularization which avoids low-diversity decoder embeddings observed with naive application in Fig. 1.

Temporal self-supervision. SSL methods developed for video achieve high performance by exploiting temporally consistent transformation [5] and temporal pretext task [14, 28, 59]. However, longitudinal biomedical image time-series have sparser sampling (typically 2–5 timepoints/subject) and have greater spatial extents (3D volumes instead of 2D images), which leads to distinct modeling considerations. Emerging biomedical methods [15, 43, 66] enforce smooth trajectories for subject-wise images in the encoder latent space and deploy their methods on image-level downstream tasks such as disease classification and age regression. However, these methods focus on learning a global embedding, without a clear extension to pixel-level tasks such as segmentation.

Biomedical image segmentation. Major challenges specific to biomedical segmentation include: (1) large 3D volumes; (2) limited sample sizes and annotations; and (3) non-i.i.d. longitudinal acquisitions tracking temporal anatomical changes. To these ends, conventional approaches use a combination of intensity-based probabilistic models and registration-driven atlas-based models [1, 26, 39, 58]. In particular, longitudinal image analysis typically makes use of one or few longitudinal atlases [27, 32, 47, 51, 52], which motivates the one-shot segmentation setting benchmarked in this paper.

More recently, deep segmentation networks achieve strong performance [8, 7, 40, 41, 48] given enough training volumes. In the low-annotation setting, weakly supervised methods develop custom loss functions [30, 36], but may have drawbracks analogous to the handcrafted SSL losses described above. Fortunately, recent data-driven self and semi-supervised methods are well-suited to pixel-level prediction. For example, to pretrain an encoder for segmentation, [9, 63] develop application-specific positive and negative sampling for contrastive training, where 2D slices from similar locations in registered 3D volumes constitute positive pairs. While these methods have been successful in their applications, they are inherently slice-based methods and are outperformed by well-tuned randomly-initialized 3D baselines on our datasets (Tab. 1). Further, these self-supervised biomedical segmentation methods do not explicitly account for non i.i.d acquisitions. Lastly, to our knowledge, existing longitudinal deep learning work developed for biomedical segmentation is currently very specific to its target application [17, 35, 56] (for example, in tasks such as MS lesion change detection [56]) or require supervised pre-training on annotated cross-sectional datasets [61], whereas we develop a generic self-supervised spatiotemporal representation learning framework for non-i.i.d. data which can be applied to any downstream task in principle.

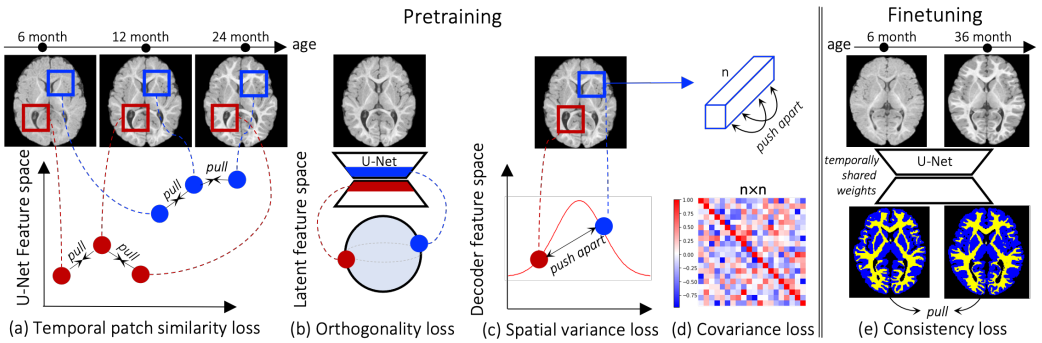

Figure 2: **Overview of proposed self-supervision.** Given nonlinearly-registered temporal images of a subject, **(a)** we assume that corresponding spatial locations in various network layers should have similar representations. As U-Net skip connections can cause degenerate decoder embeddings (see App E), we **(b)** encourage the decoder bottleneck to be orthogonal to encoder bottleneck and regularize the concatenated decoder features to have **(c)** high spatial variance and be **(d)** uncorrelated channel-wise. During fine-tuning, we **(e)** encourage temporal intra-subject network output consistency.

3 Methodology

Fig. 2 illustrates the proposed framework. The base U-Net architecture is pretrained end-to-end in a self-supervised manner with intra-subject spatiotemporal losses. On convergence, the pretrained network parameters serve as an initialization for training downstream local spatiotemporal pixel-level tasks (e.g., registration or segmentation). The main similarity loss is applied on multiscale local patches from different timepoints of the same subject, which attracts features in corresponding locations in separate timepoints together. The high-dimensional U-Net features corresponding to the boxes at the same locations in Fig. 2a should maintain high similarity despite varying appearance. The feature regularizers (Fig. 2b,c,d) avoid degenerate U-Net decoder embeddings in the patch similarity training and the output regularizer (Fig. 2e) encourages finetuning consistency on unannotated data.

Setup. The unlabeled dataset is a collection of N subjects, where each subject has at least two longitudinal image acquisitions available during pretraining. During every iteration, a pair of images $(\mathcal{X}_j^i, \mathcal{X}_{j+1}^i)$ are randomly sampled from subject $i \in \{1, 2, \ldots, N\}$ at distinct timepoints j and $j+1$, where $j \in \{1, 2, \ldots, T_i - 1\}$. T_i indicates the number of registered images from subject i, and $\mathcal{X} \in \mathcal{R}^{W \times H \times D \times C}$ is a 3D volume of spatial dimension $W \times H \times D$ and C channels. These channels are typically multi-modality acquisitions (e.g., T1w and T2w MRI from the same subject).

Spatiotemporal patchwise similarity loss. We aim to associate the local embeddings of $\mathcal{X}_j^i$ and $\mathcal{X}_{j+1}^i$ such that the representations are longitudinally-aware. To this end, a weight-sharing 3D U-Net G takes both images as input and produce a set of multi-scale CNN features $\{v\}_L$, where each element $v_{lij} = G^l(x_j^i) \in \mathcal{R}^{W_l \times H_l \times D_l \times C_l}$ indicates the output of the lth layer of interest. M feature vectors are then randomly sampled from the 3D spatial indices of the feature map, where each feature vector v_{lij}^m represents a local patch of the input image. These patch-wise activations from multiple layers of G form hierarchical representations of local regions of the input image.

To maximize the similarity of corresponding local features (e.g., blue boxes in Fig. 2a) without negative samples, we use a projector MLP f and a predictor head p and extend [13] to patchwise operation such that matching spatial indices have high agreement. The patchwise similarity loss between two representations is defined as follows: $\mathcal{L}(v_{lij}^m, v_{lik}^m) = \frac{1}{2}\mathcal{D}(p_1, z_2) + \frac{1}{2}\mathcal{D}(p_2, z_1)$ where $\mathcal{D}(p, z) = -\frac{p}{||p||_2} \cdot \frac{z}{||z||_2}$, $z_1, z_2 = f(v_{lij}^m), f(v_{lik}^m)$ and $p_1, p_2 = p(z1), p(z2)$. The total loss is an average of all sampled patches, across multilayer features: $\mathcal{L}_{sim} = \frac{1}{L}\sum_{l \in \{1,2,\ldots L\}} \frac{1}{M}\sum_{m \in \{1,2,\ldots M\}} \mathcal{L}(v_{lij}^m, v_{lik}^m)$.

Architectural challenges in multi-scale representation learning. $\mathcal{L}_{sim}$ applied to the hidden layers of a U-Net is found to maximize patchwise similarity of the encoder features (as expected) but lead to low-diversity and semantically-incoherent representations in the decoder layers as observed from the similarity maps in Fig. 1B and empirical observations in App E. We *speculate* that this is partially attributable to U-Net skip connections, which lead to a degenerate solution where the encoder learns

representations which are good enough to minimize the decoder losses and the decoder layers do not have to learn useful representations which transfer. To this end, we develop several regularization strategies such that the decoder layers obtain diverse and semantically-coherent embeddings.

Orthogonality. During model prototyping, we empirically observe that low-diversity embeddings first originate in the U-Net bottleneck (Fig 1B cols 4,5 and App. Fig. 14 rows 2,3) which are then upsampled hierarchically through the decoder. We therefore encourage decoupled bottleneck features between the encoder and decoder. Revisiting the U-Net skip-connection, the encoder features from the l-th layer $v_e \in \mathcal{R}^{W_l \times H_l \times D_l \times C_l}$ are concatenated with the upsampled decoder features $v_d \in \mathcal{R}^{W_l \times H_l \times D_l \times C_d}$, followed by a convolution to attain the same feature dimension as v_e. This yields $v_{\hat{d}} = \texttt{Conv}(\texttt{Concat}(v_e, v_d)) \in \mathcal{R}^{W_l \times H_l \times D_l \times C_l}$ and by regularizing for orthogonality between v_e and $v_{\hat{d}}$ in the projector embedding space, we implicitly encourage v_d to learn more meaningful representations instead of converging to degenerate solutions. Formally, the orthogonality loss is defined as $\mathcal{L}_O(z_e, z_{\hat{d}}) = \frac{1}{M} \sum_{m \in \{1,\dots M\}} \frac{z_e^m}{||z_e^m||_2} \cdot \frac{z_{\hat{d}}^m}{||v_{\hat{d}}^m||_2}$ where z_e^m and $z_{\hat{d}}^m$ are projected representations via f at sampling location m, from the matched encoder/decoder layers, respectively.

Variance and covariance. We further encourage spatial variation and channel-wise decorrelation of local decoder features to avoid degenerate representations.We extend the variance and covariance terms of [6] towards patchwise multi-layer operation and apply them to the decoder layers. A standard deviation loss encourages spatial variation above a threshold of $\eta = 1$ and is defined as $\mathcal{L}_S(z) = \frac{1}{k} \sum_{l \in \{L-k,\dots L\}} (\max(0, \eta - S(z_l))$, where $S(z_l) = \sqrt{Var(z_l) + \epsilon}$ is the standard deviation of M randomly-sampled projected features from the l-th layer and $\epsilon = 10^{-4}$. A covariance regularization decorrelates channelwise activations in the decoder to prevent low-diversity embeddings by minimizing $\mathcal{L}_C(z) = \frac{1}{k} \sum_{l \in \{L-k,\dots L\}} \frac{1}{n} \sum_{u \neq v} [C(z)]^2_{u,v}$, where $C(z) = \frac{1}{n-1} \sum_i^n (z_i - \overline{z})(z_i - \overline{z})^T$ is the covariance matrix of n-D representation z and (u, v) indicates its off-diagonal indices.

Reconstruction. To further pretrain the decoder, we investigate reconstruction losses. However, as high-resolution information is passed through skip connections, a reconstruction loss for a U-Net is near-trivially minimized and does not address degenerate solutions. Therefore, inspired by denoising autoencoders [57], we encourage network equivariance and invariance to geometric ($\mathcal{A}_g$) and intensity-based transformations ($\mathcal{A}_i$), respectively, of the input image x, by modifying the reconstruction objective to a denoising loss $\mathcal{L}_{rec} = \|G(\mathcal{A}_i(\mathcal{A}_g(x))) - \mathcal{A}_g(x)\|_2^2$.

Finetuning with longitudinal consistency regularization. In a longitudinal segmentation setup, if the finetuning data do not cover the overall age range, the network weights may perform well on the finetuned timepoints, but may perform poorly on out-of-distribution ages even when pretrained on unannotated data across all ages. We therefore develop a self-supervised longitudinal consistency regularization term applied at the network output during finetuning to increase the intra-subject agreement. Specific to segmentation, given registered image volumes, we formulate a segmentation consistency loss as $\mathcal{L}_{cs} = 1 - \texttt{Dice}(G(x_j^i), G(x_{j+1}^i))$ and this is minimized alongside the supervised segmentation term below during finetuning.

Total objective. During pretraining, our overall objective function is a weighted sum of the patch similarity loss, reconstruction loss, and three regularizations, and is defined as $\mathcal{L}_{PT} = \lambda \mathcal{L}_{sim} + \alpha \mathcal{L}_{rec} + \mu \mathcal{L}_S(z) + \gamma \mathcal{L}_C(z) + \beta \mathcal{L}_O(z)$. During finetuning, we use a combined dice and cross entropy loss on one-shot supervised training pairs as $\mathcal{L}_{sup} = (1 - \texttt{Dice}(G(x), y)) + \texttt{CE}(G(x), y)$ where y is the groundtruth segmentation label. Lastly, we additionally use the segmentation consistency loss $\mathcal{L}_{cs}$ for non i.i.d inputs, which forms our final finetuning loss $\mathcal{L}_{FT} = \mathcal{L}_{sup} + \mathcal{L}_{cs}$.

4 Experiments

Data and segmentation tasks. We conduct experiments on two de-identified longitudinal neuroimaging datasets, and specifically design three tasks to benchmark different extents of biomedical domain gaps between the finetuning and testing data. The main body of this work focuses on one-shot segmentation using one annotated subject (a common medical image analysis setting). Benchmarks on few-shot and fully-supervised segmentation tasks are provided in Appendix A. For both datasets, we perform a train/validation/test split on a subject-wise basis with 70%, 10% and 20% of the participants. The validation set is used for model and hyperparameter selection and results are reported on a held-out test set. ANTs [3, 4] is used to perform inter-subject affine alignment for all experiments, followed

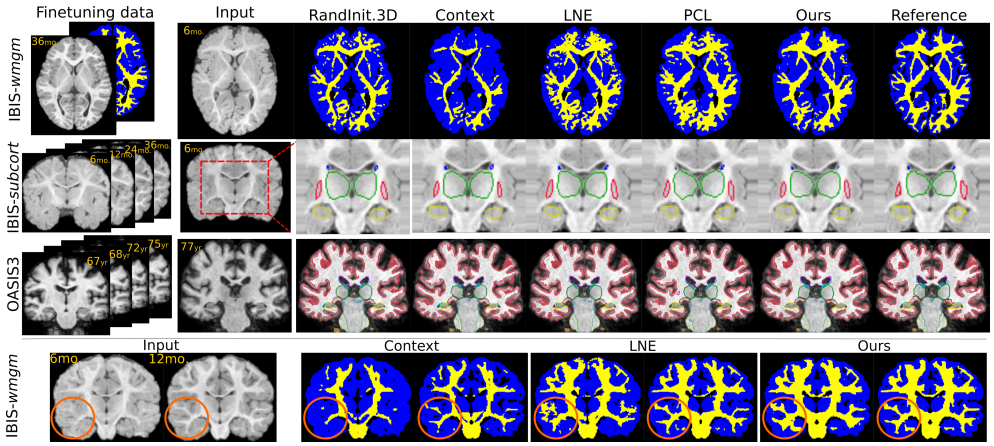

Figure 3: **One-shot segmentation**. **Top 3 rows:** Once pretrained on all unlabeled data, all benchmarked methods are finetuned on either a single annotated image (IBIS-*wmgm*) or a single annotated subject (IBIS-*subcort* and OASIS3). When deployed on other subjects at different ages, our method yields improved segmentation performance. **Bottom row:** When finetuned only on a 36 month-old image, our method generalizes to unseen timepoints by leveraging temporal anatomical consistency.

by intra-subject deformable registration to obtain accurate spatiotemporal correspondence for $\mathcal{L}_{sim}$ and $\mathcal{L}_{cs}$ calculation. All images are skull-stripped, bias-field corrected, and intensity-normalized.

OASIS3 [33] is a publicly-available dataset consisting of 1639 brain MRI scans of 992 longitudinally imaged subjects. Each subject has 1–5 temporal acquisitions over a $\sim$ 5-year long observation window, resulting in an aging cohort over the span 42 to 95 years which includes cognitively normal and mildly impaired individuals alongside subjects with Alzheimer's Disease. On OASIS3, we tackle whole-brain segmentation using the FreeSurfer label convention [16]. *Cross-sectional* FreeSurfer anatomical segmentation was done as part of the data release. Observing strong temporal inconsistency (see App. C.2), we further perform *longitudinal* FreeSurfer [47] to improve the temporal consistency of the reference segmentation. We exclude labels that have less than 100 voxels in all subjects, which results in 33 labels for segmentation training and evaluation. Finetuning for one-shot segmentation is performed on a single FreeSurfer-annotated subject with four timepoints.

IBIS is an infant brain imaging study, which longitudinally acquires 1272 structural T1w/T2w MRI from 552 infants across both controls and infants at a high-risk for Autism Spectrum Disorder (ASD) over a span of 3 to 36 months of age. We tackle two distinct tasks: subcortical segmentation (*IBIS-subcort*) and white/gray matter tissue segmentation (*IBIS-wmgm*). For *IBIS-subcort*, a multi-atlas method [54] cross-sectionally segments sub-cortical grey matter (relevant to ASD [49]) into 13 structures of interest, which are then followed by manual corrections. For one-shot benchmarking, finetuning is performed only on a single longitudinally-labeled subject. Further, we use the *IBIS-wmgm* setting to simulate a real-world use-case. Briefly, brain MRI segmentation into grey/white matter is straightforward at 24-36 months of age. However, $\sim$6-month-old brain segmentation remains elusive without manually labeled datasets for supervision due to white matter myelination leading to isointense appearance at that age [53] (e.g., Fig. 2a and App. D.2). We therefore investigate finetuning all benchmarked methods on a single 36 month old image which can be reliably segmented with [46] and then evaluate segmentation deployment with a strong domain shift on 6 month old isointense images and labels (whose ground truth labels are generated by supervised training [64]).

Baselines and Evaluation Strategies. We analyze segmentation performance and longitudinal consistency against well-tuned randomly initialized 2D/3D U-Nets (`RandInitUnet.2D/3D`) and various high-performing self-supervised pretraining methods. These include the pretext-task based `Context Restoration` [10], longitudinal representation learning based `LNE` [43], along with the contrastive learning based `GCL` [9] and `PCL` [63] methods which operate on image slices. We also repurpose `PatchNCE` [44] for segmentation to evaluate its generic representation learning capabilities. All methods are pretrained and finetuned with both geometric and intensity-based augmentation.

We quantify network performance via commonly used scores such as the Dice coefficient, IoU, and the 95-th percentile of the Hausdorff distance. More importantly, we also quantify the longitudinal

agreement between intra-subject non-linearly registered temporal segmentations via scores such as the spatiotemporal consistency of segmentation [35] $STCS = \frac{2|S_1 \cap S_2|}{|S_1|+|S_2|}$ where S_1 and S_2 are temporal segmentation predictions from non-linearly registered input images; and the absolute symmetrized percent change [47] $ASPC = 100\frac{|V_2 - V_1|}{0.5(V_1+V_2)}$, where V_1 and V_2 are the volume of a structure calculated from S_1 and S_2. We also report $STCS$ and $ASPC$ on the groundtruth segmentations as a reference.

Implementation details. We train a 3D U-Net [48] as the base image-to-image architecture with four levels of up/down sampling and repeated Conv-BN-ReLU blocks (all architectural details provided in App. B). The projector head consists of a 3-layer MLP with 2048 nodes per layer. Following [13], we apply batch normalization after each MLP, followed by ReLU activation, and l_2 normalization of the final activation. The predictor is a 3-layer bottlenecked MLP with widths of 2048-256-2048. We apply geometric (left-right flip, random affine warps) and intensity augmentations including random blur, noise, gamma contrast enhancement. Additional MRI-specific augmentations with random bias field and motion artifacts are also applied and followed by 128^3 random cropping. We use a batch size of 3 crops and an initial learning rate of 2×10^{-4} for both pretraining and finetuning. All networks are trained with the Adam optimizer ($\beta_1 = 0.9$ during pretraining and $\beta_1 = 0.5$ during finetuning, and $\beta_2 = 0.999$) on a single Nvidia RTX8000 GPU (45GB vRAM). The networks are pretrained for a maximum of $30,000$ steps and the best model based on validation performance is used for fine-tuning for another $35,000$ steps, alongside linear learning rate decay. All experiments are run on a fixed random seed due to limited computational budgets. Based on the ablation analysis in Tab. 2, we empirically choose $\lambda = 1, \alpha = 10, \gamma = 1e{-}3, \beta = 100$ for all datasets, and use $\mu = 10^{-2}$ for OASIS3, $\mu = 10^{-3}$ for IBIS. Further details on the configurations, implementation of our method and other baselines are provided in Appendix B.

Segmentation and longitudinal consistency results. Fig. 3 qualitatively demonstrates improved generalization using our method on unseen longitudinal data (row 1-3), especially on data displaying rapid intra-subject temporal developments (IBIS-*wmgm,subcort*). These improvements are consistent with the quantitative results presented in Fig. 4/Tab. 1 which indicate both improved segmentation performance and longitudinal consistency. In the strong domain shift setting of *IBIS-wmgm*, we see a near ten-point increase in median dice over most baselines. With moderate shifts in *IBIS-subcort*, we see appreciable increases in performance and consistency. We note that brain segmentation on adult brain MRI data from OASIS3 is a comparatively easier task as adult neuroimages do not significantly change appearance between imaging sessions. Therefore, several baselines are able to match (but not exceed) the segmentation performance of our method on OASIS3. However, all baselines are outperformed by ours on all datasets in terms of longitudinal-consistency which is essential to non-i.i.d. statistical analysis. In particular, both our pretraining (`Ours w/o` $\mathcal{L}_{cs}$) and finetuning (`Ours w/` $\mathcal{L}_{cs}$) methods show STCS and ASPC improvements over all of the compared settings. Fig. 3 (bottom row) shows an example of temporal predictions on two unseen timepoints from IBIS-wmgm. The predictions from Context Restoration [10] match only the input image intensity and lack anatomical and longitudinal consistency, LNE [43] introduces false positive predictions in temporal lobe (within the orange circle), and the proposed method yields a more spatiotemporally and anatomically consistent segmentation. While the main body of this work focuses on one-shot segmentation, we also observe gains in the **few-shot** and **fully-supervised** segmentation settings, with results provided in Appendix A and Suppl. Tabs. 3 and 4, respectively.

Qualitative self-supervised spatiotemporal similarity. In Fig. 1, we qualitatively examine the learned visual representations of the proposed method via intra-subject temporal self-similarity (**C**) and compare it to two of its variants which either use contrastive learning with unsupervised negatives (A) or negative-free representation learning (B). We calculate the per-layer multiscale feature self-similarity between the query and each key from the intra-subject feature maps at a different age (blue box). In row A, we see that assuming that all spatial indices not in correspondence constitute negative pairs leads to highly-positionally dependent representations in the decoder which carry low semantic meaning (e.g., in the adult data, the similarity to localize to the ventricles in the coronal view). By discarding all negatives in row B, we observe semantically-incoherent and low-diversity embeddings and artifacts in the decoder layers on both datasets. Finally, with careful regularization, our methods discards all negative pairs and attains semantically and positionally relevant representations.

Ablations. As the proposed method consists of several moving parts, an ablation analysis is conducted over different model configurations, hyperparameters, and loss functions, reported in Tab. 2 consisting of average dice coefficients. The combination of all proposed components yields optimal results.

Table 1: **One-shot segmentation benchmarking** of performance (median IoU & HD95) and longitudinal consistency (ASPC). Further scores alongside means and std. dev. are provided in Supplemental Table 7. **Few-shot** and **fully-supervised** results are provided in Suppl. Tabs. 3 and 4, respectively.

Method	IBIS-subcort			IBIS-wmgm			OASIS3		
	IoU ↑	HD95 ↓	ASPC↓	IoU↑	HD95↓	ASPC↓	IoU ↑	HD95 ↓	ASPC↓
GT	-	-	7.819	-	-	-	-	-	3.947
RandInitUnet.2D	0.707	4.485	7.127	0.510	3.274	11.506	0.687	2.545	10.189
RandInitUnet.3D	0.720	2.892	4.644	0.560	3.788	5.515	0.715	2.206	2.789
Context Restore [10]	0.711	4.403	7.831	0.444	8.273	29.235	0.717	3.323	5.577
LNE [43]	0.736	3.033	5.866	0.563	3.201	5.352	0.726	1.988	8.836
GCL [9]	0.718	3.203	5.514	0.550	4.112	8.472	0.695	2.264	4.622
PCL [63]	0.713	3.270	5.610	0.562	4.974	10.648	0.707	2.327	4.850
PatchNCE [44]	0.743	1.266	5.780	0.607	4.344	3.782	0.738	2.275	3.114
Ours w/o $\mathcal{L}_{cs}$	0.754	**1.145**	5.483	0.614	3.291	**2.462**	**0.739**	**1.940**	**2.729**
Ours w/ $\mathcal{L}_{cs}$	**0.757**	1.178	**4.475**	**0.676**	**3.237**	4.155	0.737	2.094	2.754

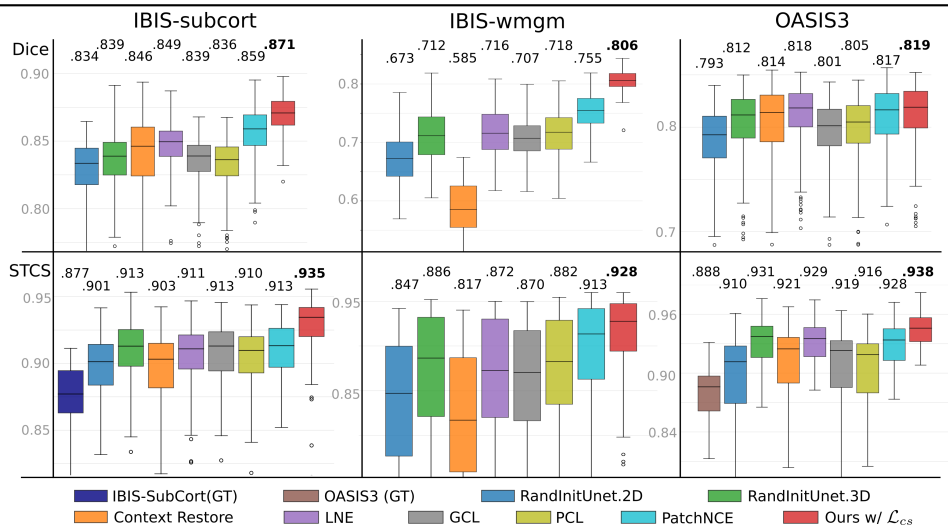

Figure 4: **One-shot segmentation benchmarking** quantifying performance with the Dice coefficient (**top**) and the spatiotemporal consistency of segmentation (**bottom**), visualizing the means and standard deviations alongside median values overlaid on the top of each subfigure (higher is better). **Few-shot** and **fully-supervised** results are provided in Suppl. Tabs. 3 and 4, respectively.

Row A starts with a base setting where only four encoder layers from the U-Net are selected for $\mathcal{L}_{sim}$ computation and a small MLP width of 256 is used for the projection and prediction heads. Here, IBIS-subcort and OASIS3 results are competitive with randomly initialized U-Net, as expected given the lack of auxiliary losses, data augmentation, and regularizations. However, the IBIS-wmgm experiment already shows a 2% improvement over random initialization, indicating benefits of using patchwise similarity losses for better out-of-distribution generalization even with suboptimal setups.

Row B: With larger projector and predictor networks, we observe improvements on two out of three datasets, which is consistent with trends observed on natural images [12, 13].

Rows C–F: On adding decoder layers to $\mathcal{L}_{sim}$ and introducing $\mathcal{L}_{rec}$ alongside data augmentation (without any regularization), we typically observe inconsistent dataset-specific trends which arise from unregularized representations (e.g., Fig. 1B). We speculate that a poorly trained decoder (due to a lack of regularization) may be equivalent to random initialization in this context. However, a combination of these components (Row F) leads to an appreciable increase in performance.

Rows G–K: When orthogonal regularization and/or covariance/variance regularization is used, we observe the best performance when they are applied together alongside augmentation and $\mathcal{L}_{Rec}$. In rows J and K, we observe that different hyperparameters are optimal for OASIS3 and IBIS (which already outperform all baseline methods in Tab. 1), which is intuitive as these are drastically different cohorts.

Table 2: **Ablation analysis** of our method over loss layers, projection+prediction layer widths (#MLP), loss functions ($\mathcal{L}_{Rec}$, $\mathcal{L}_{cs}$), use of augmentation, and hyperparameters (β, μ, γ). Mean dice is used for quantification on all datasets. *$\mu = 10^{-3}$ on IBIS-{*wmgm, subcort*} and $\mu = 10^{-2}$ on OASIS3.

Exp	Loss Layers	#MLP	$\mathcal{L}_{Rec}$	Aug.	β	μ	γ	$\mathcal{L}_{cs}$	IBIS-*subcort*	IBIS-*wmgm*	OASIS3
A	Enc	256							0.829(0.068)	0.733(0.062)	0.783(0.16)
B	Enc	2048							0.849(0.060)	0.732(0.073)	0.809(0.13)
C	Enc	2048	✓						0.859(0.058)	0.713(0.079)	0.811(0.13)
D	EncDec	2048	✓						0.860(0.058)	0.718(0.066)	0.810(0.13)
E	EncDec	2048		✓					0.858(0.060)	0.724(0.077)	0.809(0.13)
F	EncDec	2048	✓	✓					0.856(0.060)	0.739(0.067)	0.812(0.13)
G	EncDec	2048	✓		100				0.857(0.062)	0.728(0.074)	0.809(0.13)
H	EncDec	2048	✓			10^{-3}	10^{-3}		0.845(0.063)	0.739(0.074)	0.804(0.13)
I	EncDec	2048	✓		100	10^{-3}	10^{-3}		0.859(0.056)	0.735(0.061)	0.811(0.13)
J	EncDec	2048	✓	✓	100	10^{-3}	10^{-3}		0.863(0.057)	0.758(0.062)	0.808(0.14)
K	EncDec	2048	✓	✓	100	10^{-2}	10^{-3}		0.853(0.055)	0.745(0.058)	**0.813(0.13)**
L	EncDec	2048	✓	✓	100	*	10^{-3}	✓	**0.870(0.052)**	**0.806(0.030)**	0.810(0.13)

Row L: Finally, the overall proposed model is achieved when L_{cs} is added to the finetuning objective, which yields strong improvements for IBIS-{*wmgm,subcort*} and maintains OASIS3 performance.

5 Discussion

Limitations and future work. The presented work opens up many follow-up questions which will be tackled in future work: (1) Our proposed losses enable better training and performance, but require the tuning of several regularization weights and layer selections which may reveal dataset specific patterns (e.g., rows J, K in Tab. 2). The weights and layers selected here were chosen based on limited exploratory experiments on the validation sets due to computational budgets and future work will exhaustively search the hyperparameter space for optimal performance. (2) Our pretraining assumes accurate non-linear intra-subject registration, which may be non-trivial in edge cases like modalities with strong distortion and artifacts (e.g., eddy corruption in diffusion MRI). However, when studying pre and post-operative imaging (e.g., surgical excision of lesions), large topological changes break the assumptions of our model and will require the development of lesion-masked positive patch sampling methods. (3) We use a two-stage pre-training and finetuning approach and it is plausible that the proposed method can be reduced to a single-stage combined framework. (4) While this paper focused on the one-shot segmentation application, the pretraining framework is generic to any pixel-level task (e.g., registration) and its extension to such tasks will be explored. (5) While the proposed methods yield strong longitudinal segmentation consistency improvements across all datasets, we note that absolute segmentation performance gains in an elderly cohort (OASIS3) are modest in comparison to the rapidly developing infant dataset (IBIS) where higher gains are achieved. Future work will further investigate these performance differences between populations of differing temporal trends.

Potential negative impacts. The proposed method generically applies to medical image time-series and we do not anticipate negative impacts beyond those that currently exist for segmentation methods. However, while our tasks have impacts on understanding real-world disease mechanisms, we cannot claim any further insight into differences between subpopulations, as such analysis requires close collaboration with clinicians, neuroscientists, and biostatisticians, which is beyond the scope of this work.

Conclusions. This paper addressed several open questions regarding the self-supervised pretraining and finetuning of image-to-image architectures on longitudinal volumes using objective functions which exploit both intra-subject spatial and temporal self-similarity. It developed a local negative sample-free framework that trains multiple multi-scale hidden layers of image-to-image architectures that then enabled improved downstream segmentation performance, all while achieving semantically-meaningful representations via careful regularization of the decoder activations. During finetuning, it similarly developed a simple consistency-regularization objective which encourages longitudinal agreement between predictions on unlabeled data. When applied to large-scale neurodeveloping and neurodegerative longitudinal images, the proposed framework yielded improved segmentation performance and temporal consistency, both of which are crucial to statistical analyses of mechanisms of interest such as Alzheimer's Disease and Autism Spectrum Disorder. Our code is attached.

References

[1] Paul Aljabar, Rolf A Heckemann, Alexander Hammers, Joseph V Hajnal, and Daniel Rueckert. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. *Neuroimage*, 46(3):726–738, 2009.

[2] Iñigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, and Ana C. Murillo. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8219–8228, October 2021.

[3] Brian B Avants, Charles L Epstein, Murray Grossman, and James C Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1):26–41, 2008.

[4] Brian B Avants, Paul Yushkevich, John Pluta, David Minkoff, Marc Korczykowski, John Detre, and James C Gee. The optimal template effect in hippocampus studies of diseased populations. *Neuroimage*, 49(3):2457–2466, 2010.

[5] Yutong Bai, Haoqi Fan, Ishan Misra, Ganesh Venkatesh, Yongyi Lu, Yuyin Zhou, Qihang Yu, Vikas Chandra, and Alan Yuille. Can temporal information help with contrastive self-supervised learning? *arXiv preprint arXiv:2011.13046*, 2020.

[6] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

[7] Benjamin Billot, Magdamo Colin, Sean E Arnold, Sudeshna Das, Juan Iglesias, et al. Robust segmentation of brain mri in the wild with hierarchical cnns and no retraining. *arXiv preprint arXiv:2203.01969*, 2022.

[8] Benjamin Billot, Douglas N Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V Dalca, and Juan Eugenio Iglesias. Synthseg: Domain randomisation for segmentation of brain mri scans of any contrast and resolution. *arXiv preprint arXiv:2107.09559*, 2021.

[9] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in Neural Information Processing Systems*, 33:12546–12558, 2020.

[10] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical image analysis*, 58:101539, 2019.

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[12] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020.

[13] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758, June 2021.

[14] Hyeon Cho, Taehoon Kim, Hyung Jin Chang, and Wonjun Hwang. Self-supervised spatio-temporal representation learning using variable playback speed prediction. *arXiv preprint arXiv:2003.02692*, 8, 2020.

[15] Raphaël Couronné, Paul Vernhet, and Stanley Durrleman. Longitudinal self-supervision to disentangle inter-patient variability from disease progression. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 231–241. Springer, 2021.

[16] Bruce Fischl, David H Salat, Evelina Busa, Marilyn Albert, Megan Dieterich, Christian Haselgrove, Andre Van Der Kouwe, Ron Killiany, David Kennedy, Shuna Klaveness, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–355, 2002.

[17] Yang Gao, Jeff M Phillips, Yan Zheng, Renqiang Min, P Thomas Fletcher, and Guido Gerig. Fully convolutional structured lstm networks for joint 4d medical image segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1104–1108. IEEE, 2018.

[18] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[19] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

[20] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[22] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR, 2020.

[23] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[24] Hanzhe Hu, Jinshi Cui, and Liwei Wang. Region-aware contrastive learning for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16291–16301, October 2021.

[25] Tri Huynh, Simon Kornblith, Matthew R Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2785–2795, 2022.

[26] Juan Eugenio Iglesias and Mert R Sabuncu. Multi-atlas segmentation of biomedical images: a survey. *Medical image analysis*, 24(1):205–219, 2015.

[27] Juan Eugenio Iglesias, Koen Van Leemput, Jean Augustinack, Ricardo Insausti, Bruce Fischl, Martin Reuter, Alzheimer's Disease Neuroimaging Initiative, et al. Bayesian longitudinal segmentation of hippocampal substructures in brain mri using subject-specific atlases. *Neuroimage*, 141:542–555, 2016.

[28] Simon Jenni and Hailin Jin. Time-equivariant contrastive video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9970–9980, October 2021.

[29] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2022.

[30] Hoel Kervadec, Jose Dolz, Meng Tang, Eric Granger, Yuri Boykov, and Ismail Ben Ayed. Constrained-cnn losses for weakly supervised segmentation. *Medical image analysis*, 54:88–99, 2019.

[31] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

[32] Sun Hyung Kim, Vladimir S Fonov, Cheryl Dietrich, Clement Vachet, Heather C Hazlett, Rachel G Smith, Michael M Graves, Joseph Piven, John H Gilmore, Stephen R Dager, et al. Adaptive prior probability and spatial temporal intensity change estimation for segmentation of the one-year-old human brain. *Journal of neuroscience methods*, 212(1):43–55, 2013.

[33] Pamela J LaMontagne, Tammie LS Benzinger, John C Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G Vlassenko, et al. Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *MedRxiv*, 2019.

[34] Sangwoo Lee, Yejin Lee, Geongyu Lee, and Sangheum Hwang. Supervised contrastive embedding for medical image segmentation. *IEEE Access*, 9:138403–138414, 2021.

[35] Bo Li, Wiro J Niessen, Stefan Klein, Marius de Groot, M Arfan Ikram, Meike W Vernooij, and Esther E Bron. Longitudinal diffusion mri analysis using segis-net: a single-step deep-learning framework for simultaneous segmentation and registration. *NeuroImage*, 235:118004, 2021.

[36] Shijie Li, Neel Dey, Katharina Bermond, Leon Von Der Emde, Christine A Curcio, Thomas Ach, and Guido Gerig. Point-supervised segmentation of microscopy images and volumes via objectness regularization. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1558–1562. IEEE, 2021.

[37] Shikun Liu, Shuaifeng Zhi, Edward Johns, and Andrew J Davison. Bootstrapping semantic segmentation with regional contrast. *arXiv preprint arXiv:2104.04465*, 2021.

[38] Shikun Liu, Shuaifeng Zhi, Edward Johns, and Andrew J Davison. Bootstrapping semantic segmentation with regional contrast. In *International Conference on Learning Representations*, 2022.

[39] Jyrki MP Lötjönen, Robin Wolz, Juha R Koikkalainen, Lennart Thurfjell, Gunhild Waldemar, Hilkka Soininen, Daniel Rueckert, Alzheimer's Disease Neuroimaging Initiative, et al. Fast and robust multi-atlas segmentation of brain magnetic resonance images. *Neuroimage*, 49(3):2352–2365, 2010.

[40] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.

[41] Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *International MICCAI Brainlesion Workshop*, pages 311–320. Springer, 2018.

[42] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[43] Jiahong Ouyang, Qingyu Zhao, Ehsan Adeli, Edith V Sullivan, Adolf Pfefferbaum, Greg Zaharchuk, and Kilian M Pohl. Self-supervised longitudinal neighbourhood embedding. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 80–89. Springer, 2021.

[44] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345. Springer, 2020.

[45] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[46] Oula Puonti, Juan Eugenio Iglesias, and Koen Van Leemput. Fast and sequence-adaptive whole-brain segmentation using parametric bayesian modeling. *NeuroImage*, 143:235–249, 2016.

[47] Martin Reuter, Nicholas J Schmansky, H Diana Rosas, and Bruce Fischl. Within-subject template estimation for unbiased longitudinal image analysis. *Neuroimage*, 61(4):1402–1418, 2012.

[48] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[49] Kendrick Shen, Robbie Jones, Ananya Kumar, Sang Michael Xie, Jeff Z HaoChen, Tengyu Ma, and Percy Liang. Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation. *arXiv preprint arXiv:2204.00570*, 2022.

[50] Mark D Shen, Meghan R Swanson, Jason J Wolff, Jed T Elison, Jessica B Girault, Sun Hyung Kim, Rachel G Smith, Michael M Graves, Leigh Anne H Weisenfeld, Lisa Flake, et al. Subcortical brain development in autism and fragile x syndrome: evidence for dynamic, age-and disorder-specific trajectories in infancy. *American Journal of Psychiatry*, pages appi–ajp, 2022.

[51] Feng Shi, Yong Fan, Songyuan Tang, John H Gilmore, Weili Lin, and Dinggang Shen. Neonatal brain image segmentation in longitudinal mri studies. *Neuroimage*, 49(1):391–400, 2010.

[52] Feng Shi, Pew-Thian Yap, Guorong Wu, Hongjun Jia, John H Gilmore, Weili Lin, and Dinggang Shen. Infant brain atlases from neonates to 1-and 2-year-olds. *PloS one*, 6(4):e18746, 2011.

[53] Yue Sun, Kun Gao, Zhengwang Wu, Guannan Li, Xiaopeng Zong, Zhihao Lei, Ying Wei, Jun Ma, Xiaoping Yang, Xue Feng, et al. Multi-site infant brain segmentation algorithms: The iseg-2019 challenge. *IEEE Transactions on Medical Imaging*, 40(5):1363–1376, 2021.

[54] Meghan R Swanson, Mark D Shen, Jason J Wolff, Jed T Elison, Robert W Emerson, Martin A Styner, Heather C Hazlett, Kinh Truong, Linda R Watson, Sarah Paterson, et al. Subcortical brain and behavior phenotypes differentiate infants with autism versus language delay. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 2(8):664–672, 2017.

[55] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020.

[56] Minh-Son To, Ian G Sarno, Chee Chong, Mark Jenkinson, and Gustavo Carneiro. Self-supervised lesion change detection and localisation in longitudinal multiple sclerosis brain imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 670–680. Springer, 2021.

[57] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[58] Hongzhi Wang, Jung W Suh, Sandhitsu R Das, John B Pluta, Caryne Craige, and Paul A Yushkevich. Multi-atlas segmentation with joint label fusion. *IEEE transactions on pattern analysis and machine intelligence*, 35(3):611–623, 2012.

[59] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *European conference on computer vision*, pages 504–521. Springer, 2020.

[60] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7303–7313, October 2021.

[61] Jie Wei, Feng Shi, Zhiming Cui, Yongsheng Pan, Yong Xia, and Dinggang Shen. Consistent segmentation of longitudinal brain mr images with spatio-temporal constrained networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 89–98. Springer, 2021.

[62] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12310–12320. PMLR, 18–24 Jul 2021.

[63] Dewen Zeng, Yawen Wu, Xinrong Hu, Xiaowei Xu, Haiyun Yuan, Meiping Huang, Jian Zhuang, Jingtong Hu, and Yiyu Shi. Positional contrastive learning for volumetric medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 221–230. Springer, 2021.

[64] Guodong Zeng and Guoyan Zheng. Multi-stream 3d fcn with multi-scale deep supervision for multi-modality isointense infant brain mr image segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 136–140. IEEE, 2018.

[65] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1058–1067, 2017.

[66] Qingyu Zhao, Ehsan Adeli, and Kilian M Pohl. Longitudinal correlation analysis for decoding multi-modal brain development. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 400–409. Springer, 2021.

[67] Xiangyun Zhao, Raviteja Vemulapalli, Philip Andrew Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. Contrastive learning for label efficient semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10623–10633, October 2021.

[68] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7273–7282, October 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See abstract and Sec.1
   (b) Did you describe the limitations of your work? [Yes] See Sec. 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Sec. 5.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A] No new theoretical results claimed.
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Code: code with instructions is included with the supplementary materials. Data: public data links and procedures are described in the supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See 'Implementation details' in Sec. 4 and App. B.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Fig. 4, Tab. 1, Tab. 2
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See 'Implementation details' in Sec. 4
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See 'Baselines and Evaluation Strategies' in Sec. 4
   (b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Our code is included in the supplementary material

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] See [33] and [50] for how imaging consent was acquired for OASIS3 and IBIS, respectively.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See 'Data and preprocessing' in Sec. 4. Skull-stripping deidentifies human faces as only the brain is available.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes] See supplementary materials.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] See supplementary materials.

A Additional Results and Experiments

Self-supervised similarity maps from anatomically-relevant key points are visualized in Figure 5. Using only self-supervision, our model learns semantically and positionally-aware representations.

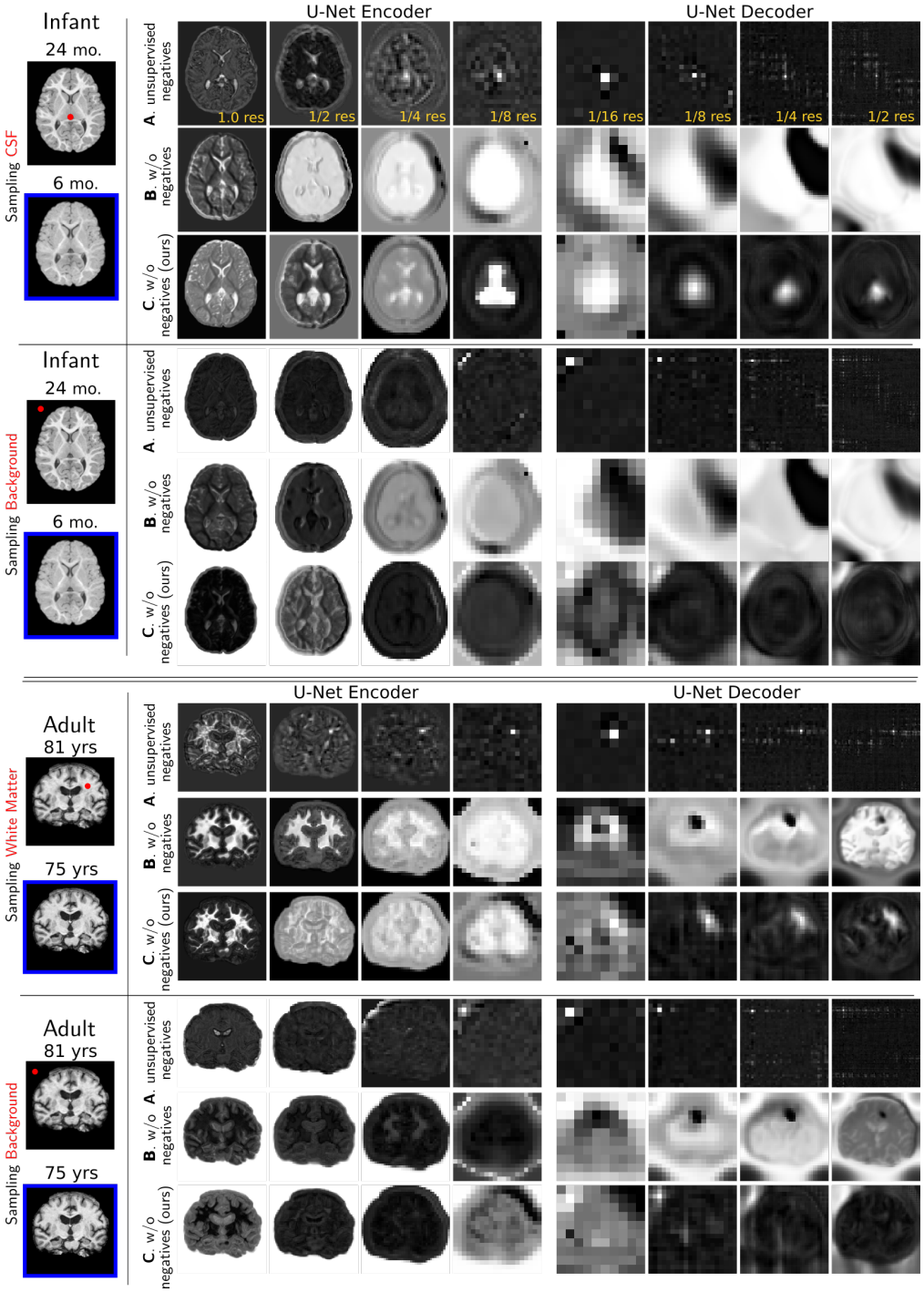

Figure 5: **Self-supervised spatiotemporal multi-scale similarity learning** visualized by sampling a spatial query point at given time point and computing its featurewise similarity to all locations in the key image acquired at a different age. The similarity-maps obtained from both datasets reveal semantically-relevant anatomical representation within the U-Net and projector/predictor networks. This figure follows the same nomenclature (A–C) as Fig. 1 with analysis in Sec. 4 (main text).

Label-wise longitudinal consistency. To observe structure-specific performance, we report label-wise longitudinal consistency scores (as measured by STCS [35]) for OASIS3 and IBIS-*subcort* on the test set for the best performing models (as measured by Figure 4 of the main text) in Figure 6. IBIS-*wmgm* was not included as it only has two anatomical structures.

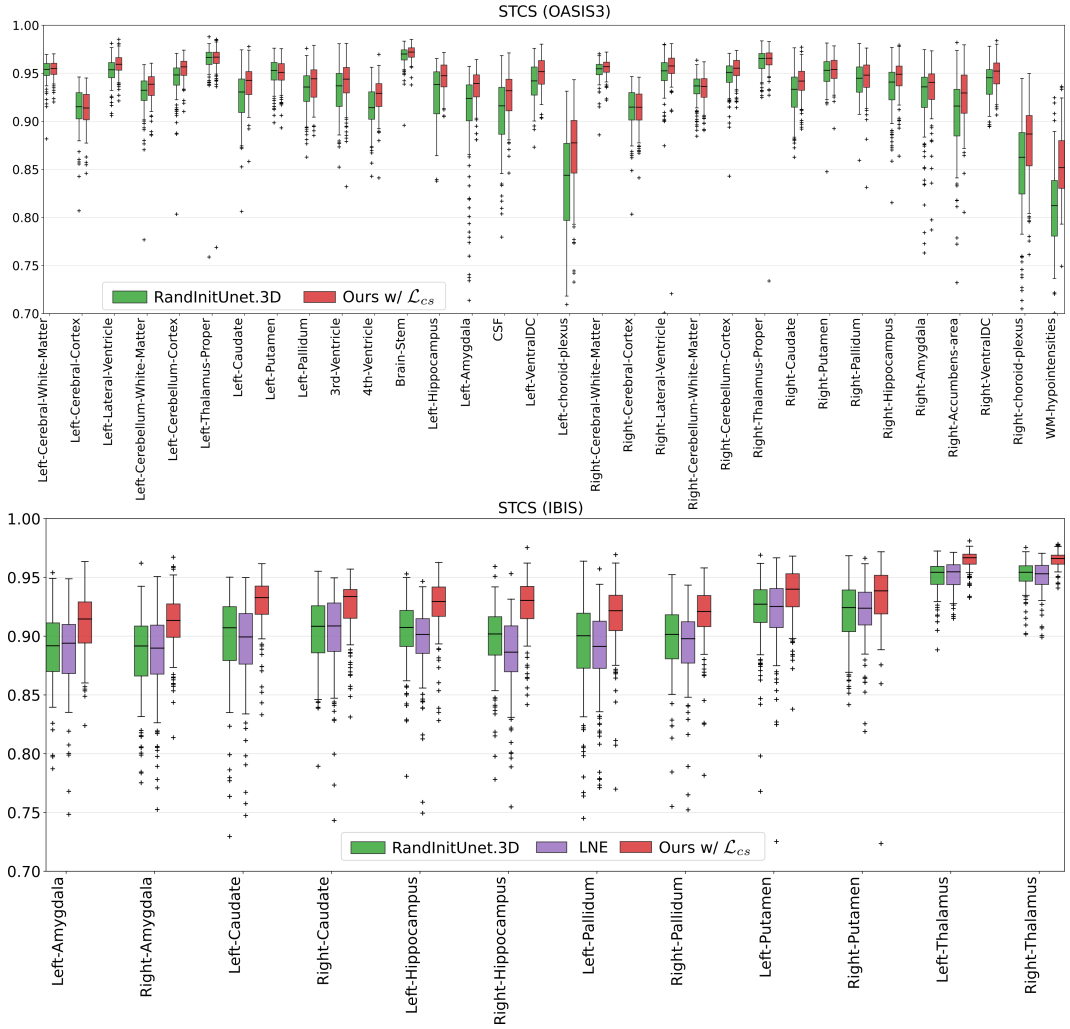

Figure 6: **Per-label longitudinal consistency comparison between the best performing models** as measured by STCS (spatiotemporal consistency of segmentations [35], higher is better) in the one-shot segmentation setting. For space considerations, only the best-performing two and three models are visualized on test set for the OASIS3 and IBIS-*subcort* tasks, respectively.

Few-shot and fully-supervised segmentation. While we focus on one-shot segmentation in our main paper (corresponding to the common real-world setting of single atlas-based segmentation), the proposed framework has benefits beyond one-shot segmentation. We present results obtained by using 10% and 100% of all the labeled training sets in Tables 3 and 4, respectively. As having more supervised data may necessitate changing our hyperparameters, we explore reducing the weight of the consistency-regularization from 1.0 to 0.1 for both experiments. In the 10% setting, we find that `Ours w/` $\mathcal{L}_{cs}$ with a weight of 0.1 obtains optimal performance and consistency. In the 100% setting, we find that including $\mathcal{L}_{cs}$ actually degrades performance and that optimal results are achieved with just the proposed pretraining (`Ours w/o` $\mathcal{L}_{cs}$). We hypothesize that these trends arise from imperfect deformable registration of the intra-subject images as they were warped considering only their intensity and not semantic structure (see App Section C.1). Therefore, having the imperfect self-supervision of $\mathcal{L}_{cs}$ in the low-annotation regime increases performance. However, hundreds of annotated samples remove the need for $\mathcal{L}_{cs}$, while maintaining the segmentation performance benefits of our pretraining framework over existing work.

Table 3: ***Few*-shot segmentation** using **10**% of the annotated training set for finetuning. This table quantifies test set performance (Dice, IoU, HD95) and longitudinal consistency (ASPC [47], STCS[35]). Mdn: Median.

IBIS-subcort

Model	Mean(std) Dice	Mdn Dice	Mean(std) IoU	Mdn HD95	Mdn IoU	ASPC	STCS
RandInitUnet.2D	0.895(0.03)	0.891	0.812(0.06)	2.390	0.804	4.618	0.909
RandInitUnet.3D	0.909(0.03)	0.905	0.834(0.05)	1.819	0.827	5.531	0.911
Context Restore [10]	0.901(0.03)	0.898	0.821(0.05)	2.240	0.815	4.895	0.918
LNE [43]	0.914(0.03)	0.909	0.842(0.05)	1.845	0.834	5.807	0.911
GCL [9]	0.905(0.03)	0.900	0.827(0.05)	2.486	0.819	4.369	0.913
PCL [63]	0.904(0.03)	0.900	0.826(0.05)	2.881	0.819	4.479	0.911
PatchNCE [44]	0.919(0.03)	0.915	0.851(0.05)	1.004	0.844	5.871	0.908
Ours w/o $\mathcal{L}_{cs}$	0.920(0.03)	0.916	0.853(0.05)	**1.001**	0.846	5.521	0.908
Ours w/ $0.1\mathcal{L}_{cs}$	**0.923(0.03)**	**0.920**	**0.858(0.04)**	1.002	**0.853**	5.077	0.910
Ours w/ $1\mathcal{L}_{cs}$	0.917(0.03)	0.913	0.848(0.05)	1.002	0.840	**4.655**	**0.925**

OASIS3

Model	Mean(std) Dice	Mdn Dice	Mean(std) IoU	Mdn HD95	Mdn IoU	ASPC	STCS
RandInitUnet.2D	0.852(0.09)	0.878	0.752(0.13)	1.569	0.785	4.297	0.908
RandInitUnet.3D	0.867(0.09)	0.888	0.774(0.12)	1.409	0.800	2.575	0.926
Context Restore [10]	0.847(0.10)	0.873	0.745(0.13)	1.660	0.777	4.528	0.916
LNE [43]	0.872(0.09)	0.894	0.782(0.12)	1.426	**0.811**	2.747	0.923
GCL [9]	0.861(0.09)	0.884	0.765(0.12)	1.387	0.794	4.005	0.911
PCL [63]	0.862(0.09)	0.885	0.766(0.12)	1.328	0.796	3.896	0.910
PatchNCE [44]	0.868(0.09)	0.888	0.776(0.12)	1.347	0.800	2.666	0.927
Ours w/o $\mathcal{L}_{cs}$	0.872(0.08)	0.894	0.782(0.12)	1.372	0.810	2.647	0.923
Ours w/ $0.1\mathcal{L}_{cs}$	**0.873(0.08)**	**0.894**	**0.783(0.12)**	**1.335**	0.809	2.602	0.926
Ours w/ $1\mathcal{L}_{cs}$	0.868(0.09)	0.890	0.776(0.12)	1.341	0.804	**2.446**	**0.932**

Table 4: ***Fully*-supervised segmentation** using **100**% of the annotated training set for finetuning. This table quantifies test set performance (Dice, IoU, HD95) and longitudinal consistency (ASPC [47], STCS[35]). Mdn: Median.

IBIS-subcort

Model	Mean(std) Dice	Mdn Dice	Mean(std) IoU	Mdn HD95	Mdn IoU	ASPC	STCS
RandInitUnet.2D	0.902(0.03)	0.898	0.824(0.05)	2.634	0.815	4.308	0.917
RandInitUnet.3D	0.922(0.03)	0.917	0.856(0.04)	1.730	0.847	5.517	0.916
Context Restore [10]	0.909(0.03)	0.906	0.835(0.05)	1.011	0.828	4.125	0.922
LNE [43]	0.928(0.02)	0.924	0.867(0.04)	1.000	0.859	5.253	0.911
GCL [9]	0.914(0.03)	0.909	0.843(0.05)	1.011	0.834	4.191	0.917
PCL [63]	0.913(0.03)	0.909	0.842(0.05)	1.006	0.833	**4.122**	0.918
PatchNCE [44]	0.928(0.02)	0.924	0.867(0.04)	1.001	0.860	5.470	0.910
Ours w/o $\mathcal{L}_{cs}$	**0.933(0.02)**	**0.930**	**0.876(0.04)**	1.000	**0.870**	6.666	0.901
Ours w/ $0.1\mathcal{L}_{cs}$	0.930(0.02)	0.926	0.870(0.04)	1.000	0.863	5.298	0.910
Ours w/ $1\mathcal{L}_{cs}$	0.922(0.02)	0.917	0.856(0.04)	1.000	0.846	4.605	**0.926**

OASIS3

Model	Mean(std) Dice	Mdn Dice	Mean(std) IoU	Mdn HD95	Mdn IoU	ASPC	STCS
RandInitUnet.2D	0.844(0.10)	0.874	0.741(0.13)	1.632	0.777	4.438	0.914
RandInitUnet.3D	0.878(0.08)	0.898	0.791(0.11)	1.371	0.817	5.570	0.920
Context Restore [10]	0.864(0.09)	0.886	0.769(0.12)	1.415	0.797	4.202	0.927
LNE [43]	0.882(0.08)	0.904	0.796(0.11)	1.280	0.825	4.140	0.926
GCL [9]	0.864(0.09)	0.887	0.769(0.12)	1.424	0.799	3.590	0.926
PCL [63]	0.865(0.09)	0.890	0.771(0.12)	1.315	0.802	3.328	0.926
PatchNCE [44]	0.869(0.08)	0.889	0.777(0.11)	1.328	0.801	2.599	0.937
Ours w/o $\mathcal{L}_{cs}$	**0.885(0.08)**	**0.907**	**0.801(0.11)**	1.246	**0.831**	2.919	0.933
Ours w/ $0.1\mathcal{L}_{cs}$	0.882(0.08)	0.904	0.796(0.11)	**1.233**	0.825	2.647	0.934
Ours w/ $1\mathcal{L}_{cs}$	0.877(0.08)	0.898	0.789(0.11)	1.245	0.817	**2.368**	**0.939**

U-Net configuration. For consistency, we use the same base U-Net architecture for all baselines. Its configuration is modeled based on one-shot segmentation mean Dice validation results on OASIS3 presented in Table 5 and evaluated over several training crop sizes, normalization layers, and channel width multipliers. Importantly, all configurations were trained without any pretraining to obtain a baseline. We observe that a random crop window of 128^3 is optimal (rows A, B, C) and find that Instance Normalization (row D) instead of Batch Normalization degrades performance. Finally, we observe almost no performance degradation when reducing the model size via the channel width multiplier from 24 to 16 (row C vs E) and therefore use 16 for computational efficiency.

Table 5: **Base network tuning.** Randomly initialized UNet performance over various crop sizes, normalization types, and the # channels of the first convolution layer. E is the final configuration for the base network used in all baselines as it best trades-off memory usage and performance.

Exp	Crop size	Normalization	Channel width	mean(std) dice
A	$64 \times 64 \times 64$	batch norm	24	0.768 ($\pm$ 0.14)
B	$128 \times 128 \times 128$	batch norm	24	**0.795 ($\pm$ 0.14)**
C	$160 \times 160 \times 192$	batch norm	24	0.786 ($\pm$ 0.14)
D	$128 \times 128 \times 128$	instance norm	24	0.777 ($\pm$ 0.14)
E	$128 \times 128 \times 128$	batch norm	16	0.792 ($\pm$ 0.15)
F	$128 \times 128 \times 128$	batch norm	8	0.786 ($\pm$ 0.14)

Training a randomly-initialized U-Net with $\mathcal{L}_{cs}$. To investigate the standalone benefit of $\mathcal{L}_{cs}$ without any regularized pretraining, we apply $\mathcal{L}_{cs}$ to a randomly initialized 3D U-Net in the one-shot segmentation setting for IBIS-*subcort* and report its results in Fig. 7.

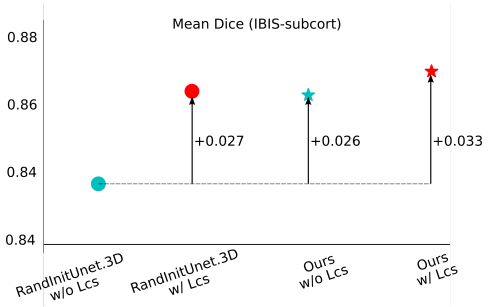

Figure 7: Stand-alone combination of the proposed $\mathcal{L}_{cs}$ consistency-regularization and one-shot segmentation training of a randomly-initialized U-Net. In the one-shot setting, $\mathcal{L}_{cs}$ improves mean dice by 0.027, which is comparable to the proposed model without $\mathcal{L}_{cs}$. Optimal results are obtained under combination with our regularized pretraining framework.

Using last decoder layer in loss. We investigate the impact of including the full-resolution feature (layer 21 in Table 9) in the `EncDec` loss configuration described in Section B.1. As shown in Figure 8, this notably degrades performance, indicating sensitivity to the exact layers used for $\mathcal{L}_{sim}$ calculation.

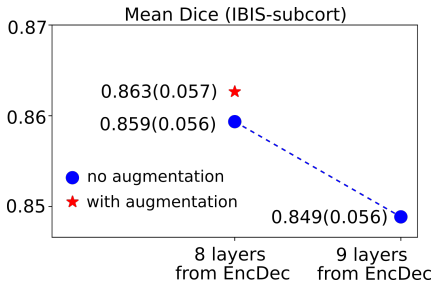

Figure 8: A comparison between using different number of layers during pretraining. We used '8 layers from EncDec' as our final layer selection as a trade-off between memory and performance.

Additional modeling decisions. In addition to the quantitative ablations over our modeling choices included in the main paper and appendix, we make a few modeling decisions detailed in Table 6 based on qualitative assessments of the representation quality using feature similarity maps.

Table 6: Additional modeling decisions. The final selections are bolded.

Parameter	Search space
# of layers in projector	2,**3**
non-linearity	ELU, **ReLU**, LeakyReLU
within subject registration	affine only, **affine and deformable**
feature selection for similarity loss	**conv output**, activation output

Additional one-shot quantification. As Table 1 and Figure 4 (main text) report median scores, Table 7 reports means and standard deviations for further interpretation. Similar to the medians, our method improves both mean performance and longitudinal consistency on IBIS-{*wmgm, subcort*}. On OASIS3, our Dice and IoU performance improves on all baselines except for LNE (with whom it is strongly competitive) and exceeds all baselines in terms of longitudinal consistency (ASPC, STCS).

Table 7: **One-shot segmentation performance** to complement results presented in Tab. 1 and Fig. 4 of the main text, using means instead of medians. *In OASIS3, we exclude the `Right Accumbens Area` and `4th Ventricle` labels when calculating mean HD95 due to numerical issues for some of the baselines (`RandInitUnet.2D`, `Context Restore`, and `PCL`).

IBIS-subcort

Model	Dice	IoU	HD95	ASPC	STCS
RandInitUnet.2D	0.828(0.07)	0.712(0.10)	4.784(6.32)	7.08(7.85)	0.893(0.03)
RandInitUnet.3D	0.836(0.06)	0.724(0.09)	3.603(5.27)	5.29(5.12)	0.908(0.02)
Context Restore [10]	0.838(0.07)	0.727(0.09)	5.196(9.27)	8.22(9.16)	0.893(0.02)
LNE [43]	0.847(0.06)	0.739(0.09)	3.478(5.33)	5.83(5.41)	0.905(0.02)
GCL [9]	0.835(0.07)	0.722(0.10)	3.640(4.16)	5.47(6.23)	0.905(0.03)
PCL [63]	0.833(0.07)	0.720(0.09)	3.528(3.56)	5.78(6.10)	0.902(0.03)
PatchNCE [44]	0.856(0.06)	0.753(0.09)	1.482(1.75)	5.93(5.79)	0.909(0.02)
Ours w/o $\mathcal{L}_{cs}$	0.863(0.06)	0.763(0.08)	1.510(2.62)	5.68(6.00)	0.909(0.02)
Ours w/ $\mathcal{L}_{cs}$	**0.870(0.05)**	**0.773(0.08)**	**1.408(1.24)**	**4.26(4.32)**	**0.929(0.02)**

IBIS-wmgm

Model	Dice	IoU	HD95	ASPC	STCS
RandInitUnet.2D	0.672(0.07)	0.510(0.08)	3.274(1.00)	11.51(11.79)	0.832(0.03)
RandInitUnet.3D	0.713(0.08)	0.560(0.09)	3.788(0.70)	5.52(5.48)	0.863(0.05)
Context Restore [10]	0.590(0.19)	0.444(0.19)	8.273(3.56)	29.24(26.76)	0.813(0.10)
LNE [43]	0.716(0.07)	0.563(0.09)	**3.201(0.98)**	5.35(6.34)	0.858(0.04)
GCL [9]	0.707(0.05)	0.550(0.06)	4.112(1.14)	8.47(8.85)	0.858(0.02)
PCL [63]	0.715(0.08)	0.562(0.09)	4.974(1.75)	10.65(13.29)	0.868(0.05)
PatchNCE [44]	0.753(0.05)	0.607(0.06)	4.344(0.66)	3.78(4.18)	0.889(0.03)
Ours w/o $\mathcal{L}_{cs}$	0.758(0.06)	0.614(0.08)	3.291(0.66)	**2.46(2.74)**	0.882(0.03)
Ours w/ $\mathcal{L}_{cs}$	**0.806(0.03)**	**0.676(0.04)**	3.237(0.47)	4.15(3.32)	**0.910(0.02)**

OASIS3

Model	Dice	IoU	HD95*	ASPC	STCS
RandInitUnet.2D	0.781(0.15)	0.661(0.17)	3.425(5.00)	10.27(28.84)	0.873(0.05)
RandInitUnet.3D	0.804(0.13)	0.689(0.16)	3.424(5.85)	3.53(6.35)	0.923(0.05)
Context Restore [10]	0.801(0.14)	0.687(0.16)	4.841(10.00)	6.88(15.21)	0.899(0.04)
LNE [43]	**0.813(0.13)**	0.700(0.15)	**3.171(5.50)**	8.81(31.94)	0.898(0.04)
GCL [9]	0.792(0.14)	0.674(0.16)	3.330(5.25)	5.30(11.12)	0.903(0.05)
PCL [63]	0.796(0.14)	0.680(0.17)	3.339(5.41)	5.69(13.40)	0.899(0.05)
PatchNCE [44]	0.809(0.13)	0.697(0.16)	3.298(5.93)	3.49(6.70)	0.920(0.04)
Ours w/o $\mathcal{L}_{cs}$	**0.813(0.13)**	**0.702(0.15)**	3.289(5.68)	3.40(6.11)	0.923(0.04)
Ours w/ $\mathcal{L}_{cs}$	0.810(0.13)	0.697(0.16)	3.733(6.41)	**2.76(4.96)**	**0.934(0.03)**

Additional IBIS-*wmgm* segmentation results. For one-shot IBIS-*wmgm* segmentation, we train on a single 36 month old T1w/T2w MR image and quantitatively evaluate on 6 month old MR images (for additional motivation see Appendix Section D.2). In Figure 9, we visualize predictions using additional baselines and our framework on a held-out test subject. As the infant brain matures over time (Appendix Section D.2), the rows are arranged in order of the most to the least amount of biological domain-shift w.r.t. the 36 month-old training image. Further, as the 12, 24, and 36 month old held-out images do not have ground truth segmentations available, we are restricted to qualitative evaluations of segmentation quality.

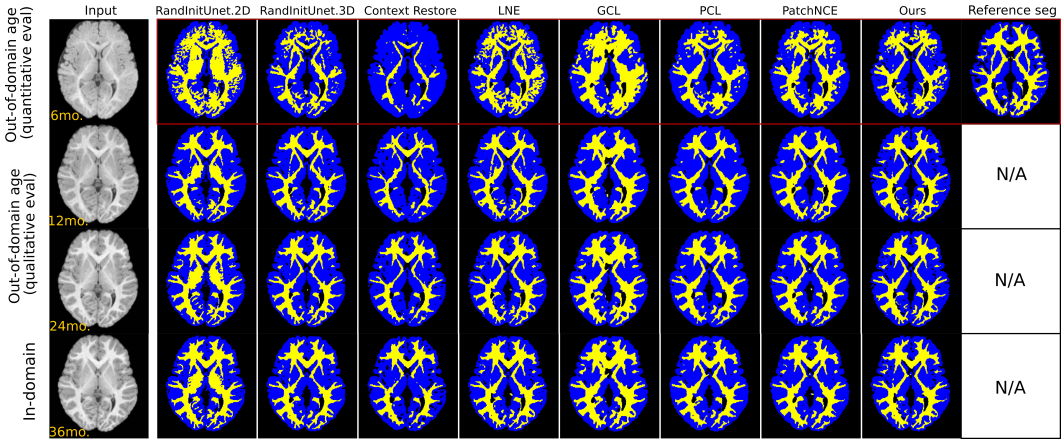

Figure 9: Infant brain segmentation on a held-out subject under the strong domain shift induced by training on a single 36 month-old image and testing on the remaining timepoints. Under the most extreme domain shift of row 1, we see that all baselines yield highly-inaccurate predictions, while our method yields higher segmentation quality.

B Additional Implementation Details

B.1 Additional Training Details

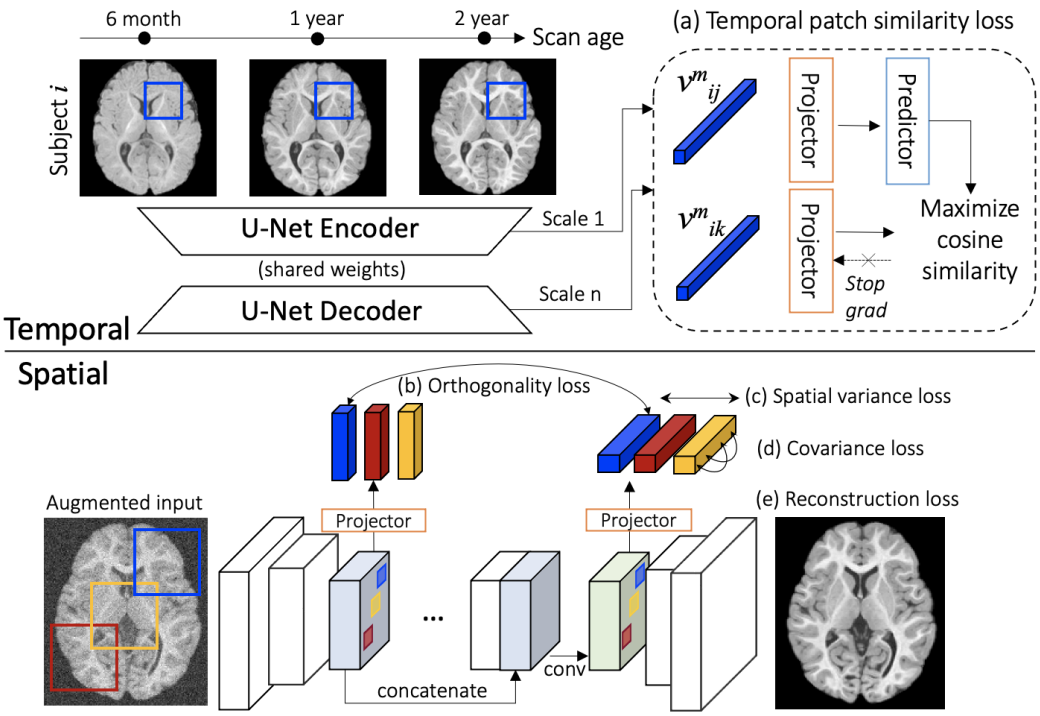

Figure 10: **Pre-training Overview.** A visual illustration of the proposed losses (a, b, c, d, e) for spatiotemporal representation learning with a U-Net. An overview of (a)–(d) is in main text Fig.1.

Figure 10 illustrates an in-depth overview of the proposed pretraining and representation learning framework. In the proposed loss configuration (`EncDec`), Pre-activation features from layers $\{1, 3, 5, 7, 9, 12, 15, 18\}$ in Table 5 are sampled spatially to extract channel-wise vectors which are fed into the projector and predictor MLPs for $\mathcal{L}_{sim}$ calculation. Patch features from layer $\{8, 12\}$ in the bottleneck (processed by the projector) are used for the orthogonality regularizer $\mathcal{L}_O$ and layers $\{12, 15, 18\}$ in the decoder are used for the $\mathcal{L}_S$ and $\mathcal{L}_C$ regularizers. In the ablations in Table 2 (main text), the `Enc` loss setting refers to using features from layers $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$ for $\mathcal{L}_{sim}$.

Each iteration of pretraining loads three corresponding crops from two intra-subject images each for $\mathcal{L}_{sim}$ calculation. Higher batch sizes could have been used for the proposed loss functions but were not for consistency with one of our baselines (`PatchNCE` [44]) which is highly memory-intensive due to the need for a large number of negative samples alongside 3D computation.

During finetuning, the segmentation loss ($\mathcal{L}_{sup}$) is applied to the original unwarped images and the consistency loss ($\mathcal{L}_{cs}$) is applied to the nonlinearly aligned images. When calculating $\mathcal{L}_{sup}$ and $\mathcal{L}_{cs}$, we perform individual forward/backward passes for each loss due to memory limitations as gradient accumulation over two forward passes was found to decrease performance on a validation set.

B.2 Architectures

The MLP architectures for the **projector** and **predictor** networks used for pretraining and representation learning are given in Table 8. The **U-Net** architecture used for both pretraining and finetuning is given in Table 9. During pretraining, an additional convolutional layer is used following layer 23 to reconstruct/denoise the input. During finetuning, a channel-wise softmax layer is attached following layer 23 for segmentation.

Table 8: Projector and predictor MLP architectures.

Projector (f)

id	Layer	Output size
0	FC(2048), BN, ReLU	(num patches $\times$ batch size) $\times$ 2048
1	FC(2048), BN, ReLU	(num patches $\times$ batch size) $\times$ 2048
2	FC(2048), BN	(num patches $\times$ batch size) $\times$ 2048

Predictor (p)

id	Layer	Output size
0	FC(256), BN, ReLU	(num patches $\times$ batch size) $\times$ 256
1	FC(2048), BN, ReLU	(num patches $\times$ batch size) $\times$ 2048

Table 9: U-Net architecture. All convolutional layers use 3×3 kernels. BN: Batch Normalization (using default PyTorch momentum). The batch size dimension is denoted bs and nc is the starting channel width multiplier of the model. We choose $nc = 16$ consistently throughout all models based on Table5. n indicates the number of output channels and is set to the number of labels.

id	Layer	Output size
0	Conv3D(nc), BN, ReLU	bs, w, h, d, nc
1	Conv3D(nc), BN, ReLU	bs, w, h, d, nc
2	Conv3D(nc), BN, ReLU	bs, w, h, d, nc
3	MaxPool(2), Conv3D(2nc), BN, ReLU	bs, w/2, h/2, d/2, 2nc
4	Conv3D(2nc), BN, ReLU	bs, w/2, h/2, d/2, 2nc
5	MaxPool(2), Conv3D(4nc), BN, ReLU	bs, w/4, h/4, d/4, 4nc
6	Conv3D(4nc), BN, ReLU	bs, w/4, h/4, d/4, 4nc
7	MaxPool(2), Conv3D(8nc), BN, ReLU	bs, w/8, h/8, d/8, 8nc
8	Conv3D(8nc), BN, ReLU	bs, w/8, h/8, d/8, 8nc
9	MaxPool(2), Conv3D(16nc), BN, ReLU	bs, w/16, h/16, d/16, 16nc
10	Conv3D(16nc), BN, ReLU	bs, w/16, h/16, d/16, 16nc
11	Upsample(2), Concatenate with 8	bs, w/8, h/8, d/8, 24nc
12	Conv3D(16nc), BN, ReLU	bs, w/8, h/8, d/8, 8nc
13	Conv3D(16nc), BN, ReLU	bs, w/8, h/8, d/8, 8nc
14	Upsample(2), Concatenate with 6	bs, w/4, h/4, d/4, 12nc
15	Conv3D(4nc), BN, ReLU	bs, w/4, h/4, d/4, 4nc
16	Conv3D(4nc), BN, ReLU	bs, w/4, h/4, d/4, 4nc
17	Upsample(2), Concatenate with 4	bs, w/2, h/2, d/2, 6nc
18	Conv3D(2nc), BN, ReLU	bs, w/2, h/2, d/2, 2nc
19	Conv3D(2nc), BN, ReLU	bs, w/2, h/2, d/2, 2nc
20	Upsample(2), Concatenate with 2	bs, w, h, d, 3nc
21	Conv3D(nc), BN, ReLU	bs, w, h, d, nc
22	Conv3D(nc), BN, ReLU	bs, w, h, d, nc
23	Conv3D(n)	bs, w, h, d, n

B.3 Baseline Reimplementation Details

LNE. LNE [43] is a self-supervised global representation learning method that models longitudinal effects in an autoencoder latent space and is repurposed here for longitudinal segmentation. During pretraining, we require the same base U-Net architecture for consistency with other baselines. Therefore, to extract a 1024D global representation as in the original paper, we add a global pooling layer on the output of layer 13 in Table 9, followed by a single 1024D fully-connected layer. With respect to their hyperparameters, due to the larger base network size, we reduce the batch size from 64 to 16 images due to memory considerations. We maintain the other hyperparameters of LNE including downsampling the images to 64^3, using $N = 5$ neighborhoods, and weighing auxiliary losses as $\lambda_{dir} = 1$ and $\lambda_{rec} = 2$ (see original work for definitions [43]).

PatchNCE. PatchNCE [44] proposes a framework for unsupervised multi-scale patchwise contrastive learning applied to the problem of unpaired image translation and is extended to serve as a baseline in this work as it shares a partially similar motivation. To adapt it to our generic longitudinal representation learning setting, we impose the same modeling assumption as our method (spatial indices in correspondence across an image time-series are positive samples and all other spatial indices are negatives) and modify both its data sampling and loss calculation. First, we perform forward passes on t corresponding intra-subject crops. At each selected layer, N feature vectors are randomly sampled for the longitudinal InfoNCE [31] loss: $\mathcal{L} = \sum_i \frac{1}{|P(i)|} \sum_{p \in P(i)} -\log \frac{e^{z_i \cdot z_p/\tau}}{\sum_{s \in S(i)} e^{z_i \cdot z_s/\tau}}$, where for any query feature index i, $P(i)$ is the set of indices of all positives (from the same subject at the same spatial index) and $S(i)$ is the set of all query indices (with size $(t \times N) - 1$, where t is the number of timepoints in the batch and N is the number of patches). We choose $t = 3$ timepoints and $N = 768$ patches in our experiments due to memory limits. An ablation study of PatchNCE over the layers used for the loss function and the temperature is given in Table 10. For prototyping efficiency, we perform baseline tuning using a three-layer MLP of size 256 and use validation Dice as the model selection criterion to choose F as our final configuration. For fair comparison with our model, we increase the MLP width in the `PathNCE` model to 2048 in all other experiments reported, although this does not significantly alter results.

Context Restoration. [10] proposes context restoration as a pretext task for self-supervised representation learning which consists of randomly and repeatedly swapping image patches and training the network to restore the original image. We extend Algorithm 1 in [10] to use 3D inputs and targets and tune it with a varying number of swapping iterations (#swaps) and swapped patch sizes. Downstream validation dice scores are provided in Table 11 and configuration D is chosen as the final model.

PCL, GCL, and 2D U-Net. PCL and GCL [9, 63] are 2D slice-based contrastive learning methods designed for 3D volume segmentation. We follow the optimal configurations (e.g. temperature, number of partitions, etc.) reported in the original papers. As these methods use a 2D U-Net trained on 2D slices, we additionally benchmark against a randomly initialized 2D U-Net to obtain a baseline for 2D methods. We use a batch size of 128 for all 2D baselines. Within each batch, we compare two different data sampling strategies: (1) randomly sampling 2D slices across all 3D volumes; (2) sampling intra-volume slices within each batch (i.e. treat one of the dimensions from a $128 \times 128 \times 128$ crop as the batch dimension for 2D networks). We train a randomly initialized 2D U-Net for the one-shot segmentation task, and found that (2) significantly increases the validation mean Dice over (1) on OASIS3 from $0.708(0.193)$ to $0.780(0.137)$ under the same architecture and optimization set up. We therefore use batch sampling strategy (2) for all 2D model benchmarks (PCL, GCL, RandInitUnet.2D).

Table 10: Parameter search (on IBIS-subcort) of patchNCE pretraining over layers included for multiscale patchNCE loss, and temperature. F is used for the final comparison.

Exp	layers	temperature	Mean(std) dice
A	Enc	0.07	0.834 ($\pm$ 0.060)
B	Enc	0.1	0.834 ($\pm$ 0.062)
C	Enc	0.2	0.823 ($\pm$ 0.067)
D	Enc	0.5	0.830 ($\pm$ 0.064)
E	EncDec	0.07	0.854 ($\pm$ 0.056)
F	EncDec	0.1	**0.856 ($\pm$ 0.054)**
G	EncDec	0.2	0.846 ($\pm$ 0.058)
H	EncDec	0.5	0.850 ($\pm$ 0.054)

Table 11: Context restoration parameter search over the size of local patches and the number of patch swap iterations on OASIS3. Validation mean(std) dice is used for selecting the configuration D.

Exp	patch size	# swaps	Mean(std) Dice
A	16^3	10	0.778 ($\pm$ 0.141)
B	24^3	10	0.779 ($\pm$ 0.142)
C	16^3	30	0.775 ($\pm$ 0.138)
D	16^3	50	**0.782 ($\pm$ 0.135)**
E	16^3	100	0.778 ($\pm$ 0.160)

C Additional data preparation details

C.1 Registration

C.1.1 Affine registration

To warp all images to a common space, we warp all images affinely to a constructed template [4] (with an affine transformation model) using `ANTs` [1] with the following command:

```
antsMultivariateTemplateConstruction2.sh \
-d 3 \
-o OUTPUT_FOLDER/T \
-i 1 -g 0.2 -j 128 -c 2 -r 1 -n 0 -m MI -l 1 \
-t Affine INPUT_FOLDER/*t1.nii.gz'
```

Once affinely aligned on a dataset-wide level, we proceed with longitudinal intra-subject deformable alignment for the calculation of $\mathcal{L}_{sim}$ and $\mathcal{L}_{cs}$, as described below.

C.1.2 Nonlinear Deformable Registration

Proper spatial correspondence of positive samples for $\mathcal{L}_{sim}$ and label maps for $\mathcal{L}_{cs}$ (the similarity and consistency losses, respectively) requires nonlinear/deformable registration of all intra-subject images to a common reference. To this end, we employ the `ANTs` SyN algorithm from [3] to register within-subject images to a single timepoint in a series of acquisitions with the following command:

```
fix = INPUT_FOLDER/{subj}_{trg_tp}_t1.nii.gz
for src_tp in tps:
    moving = INPUT_FOLDER/{subj}_{src_tp}_t1.nii.gz
    antsRegistration \
    --verbose 1 \
    --dimensionality 3 \
    --float 1 \
    --output [OUTPUT_FOLDER/{subj}_{src_tp}_to_{trg_tp}_t1_, \
              OUTPUT_FOLDER/{subj}_{src_tp}_to_{trg_tp}_t1_Warped.nii.gz, \
              OUTPUT_FOLDER/{subj}_{src_tp}_to_{trg_tp}_t1_InvWarped.nii.gz] \
    --transform SyN[0.15, 9, 0.2] \
    --metric CC[{fix}, {moving}, 1, 2, Random, 0.4] \
    --convergence [250x125x50, 1e-5, 10] \
    --shrink-factors 4x2x1 \
    --smoothing-sigmas 2x1x0vox \
    --interpolation Linear
```

where `fix` is the intra-subject image from a selected timepoint `trg_tp` to which all other timepoints `src_tp` are registered.

C.2 Additional preprocessing and label generation

With respect to OASIS3, the publicly-available dataset arrives preprocessed with cross-sectional FreeSurfer processing (intensity and geometric normalization and segmentation). As cross-sectional FreeSurfer does not account for longitudinal effects [47], FreeSurfer V6.0 is used for additional longitudinal processing on top of the publicly-released cross-sectional FreeSurfer segmentations. A comparison between longitudinal and cross-sectional FreeSurfer analysis based on training set is given in Fig. 11, where we observe notable improvements to longitudinal segmentation consistency (as quantified by STCS [35]) over all structures.

For IBIS-{*subcort, wmgm*}, all T1w/T2w MR images are preprocessed following standard procedures described in [50] of gradient distortion correction, bias-field correction, within-subject and within time-point multi-modality registration, and brain extraction. For IBIS-*subcort*, ground truth label generation follows the segmentation procedures outlined in [50]. For IBIS-*wmgm*, we use

[1] `https://github.com/ANTsX/ANTs`

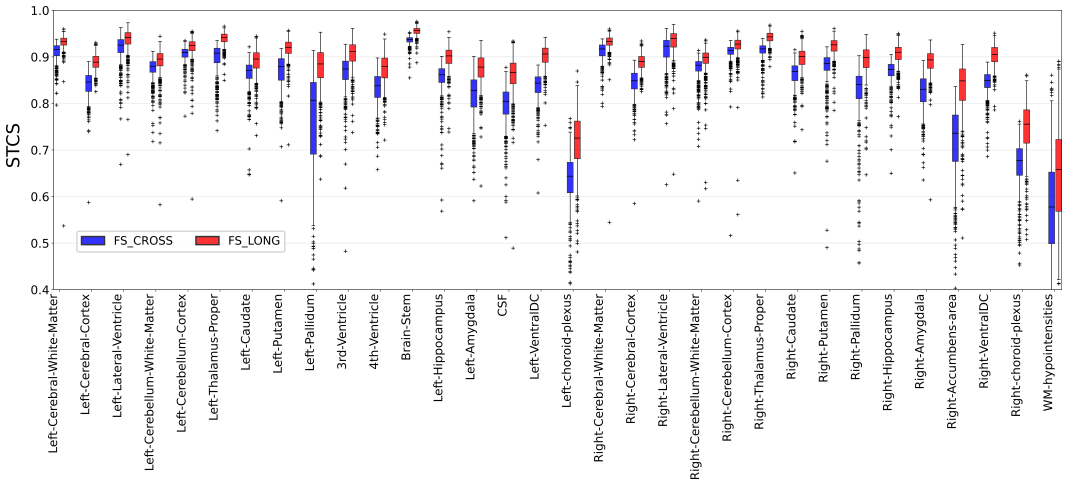

Figure 11: Longitudinal FreeSurfer processing (for anatomical segmentation used as ground-truth targets) leads to substantial improvements over the publicly-released cross-sectional FreeSurfer processing of OASIS3 in longitudinal-consistency metrics such as STCS [35] (higher is better). See section C.2 for further motivation.

segmentations generated by a fully-supervised network trained on an external dataset, see Appendix Section D.2 for more details.

Once preprocessed and aligned with methods described in Section C.1, the images are cropped to a common field-of-view ($128 \times 192 \times 160$ for IBIS, $160 \times 192 \times 160$ for OASIS3, corresponding to smaller brain volumes for infants vs. adults) for compatibility with common multi-resolution neural network architectures.

C.3 Data splitting with repeat acquisitions

We perform a 70-10-20 train, validation, and test split on subject-wise basis from a total of N subjects and M images ($M > N$ due to image time-series acquisitions), with the entire procedure illustrated in Figure 12. This level of the data split hierarchy (N_cross subjects and M_cross images) is used for segmentation training and evaluation. For longitudinal pretraining and longitudinal consistency evaluation (ASPC, STCS in Table 1 and Figure 4 of the main text), we further filter N_long subjects with at least 2 acquisitions per subject to obtain M_long images.

Finally, IBIS-*subcort* and IBIS-*wmgm* are segmentation tasks which share the same unlabeled pretraining data. Importantly, we note that as IBIS-*wmgm* is evaluated and tested only on 6-month-old MR images (representing the strongest domain shift, see Appendix Section D.2 for its motivation), its segmentation validation and test sets contain fewer images than IBIS-*subcort*. Table 12 provides the exact sample sizes for each task.

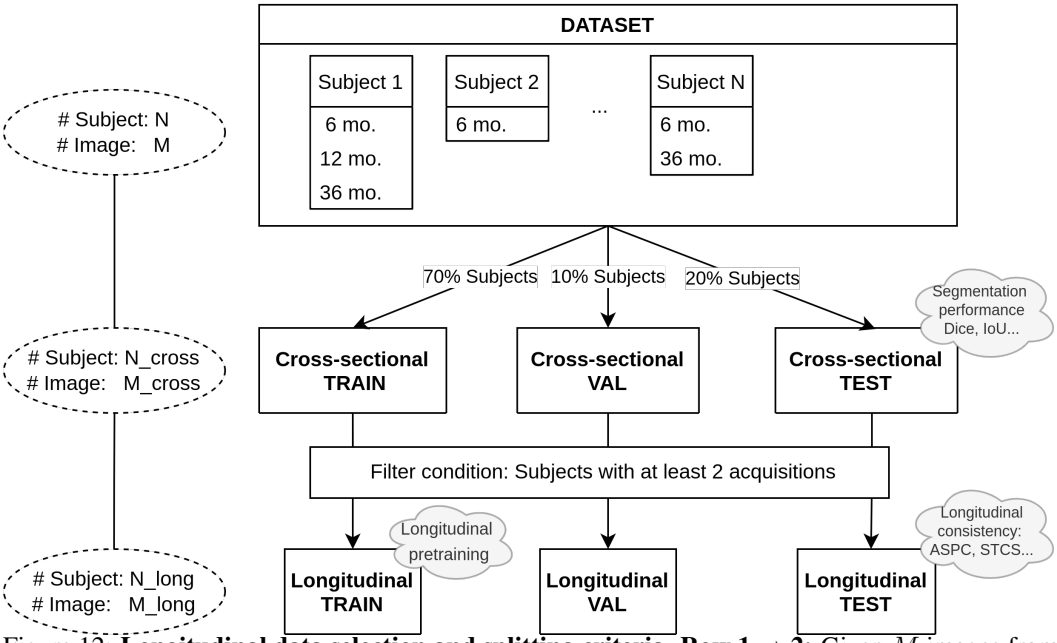

Figure 12: **Longitudinal data selection and splitting criteria. Row 1 → 2:** Given M images from N subjects, we apply a 70-10-20 train/val/test *subject-wise* split to avoid data leakage. This level of data split hierarchy is compatible with the calculation of segmentation performance scores which are agnostic to repeat acquisitions such as the Dice coefficient and mean Intersection-over-Union. **Row 2 → 3:** For longitudinal pretraining and calculation of longitudinal consistency scores (such as ASPC and STCS), we further filter the data splits to only contain subjects with two or more acquisitions. Actual dataset-specific numbers reported in Table 12.

Table 12: **Dataset-specific sample sizes of images (N) and subjects (M)**. Table best interpreted in conjunction with Fig. 12. Abbreviations include *_long*: longitudinal and *_cross*: cross-sectional.

IBIS

	IBIS-subcort		IBIS-wmgm	
Split	(N_cross, M_cross)	(N_long, M_long)	(N_cross, M_cross)	(N_long, M_long)
Total	(552, 1272)	(455, 1175)	(552, 1272)	(552, 1272)
Train	(386, 887)	(313, 814)	(386, 887)	(313, 814)
Validation	(55, 133)	(50, 128)	(55, 42)	(50, 128)
Test	(111, 252)	(92, 233)	(111, 80)	(92, 233)

OASIS3

Split	(N_cross, M_cross)	(N_long, M_long)
Total	(992, 1639)	(422, 1069)
Train	(694, 1147)	(293, 746)
Validation	(98, 166)	(48, 116)
Test	(200, 326)	(81, 207)

D Miscellaneous Details

D.1 Data availability and IRB information

IBIS. The IBIS/Autism MRI data is available through NIH NDA[2]. The Infant Brain Imaging Study (IBIS) Network is a National Institutes of Health–funded Autism Center of Excellence project and consists of a consortium of eight universities in the United States and Canada. Parents provided informed consent and the institutional review board at each site approved the research protocol.

OASIS-3. OASIS-3 [33] is publicly available through the OASIS webpage[3]. Ethical approval was obtained by the relevant ethics committees and informed consent was obtained from all participants following procedures set by the IRB at the Washington University School of Medicine.

D.2 IBIS-wmgm segmentation task details

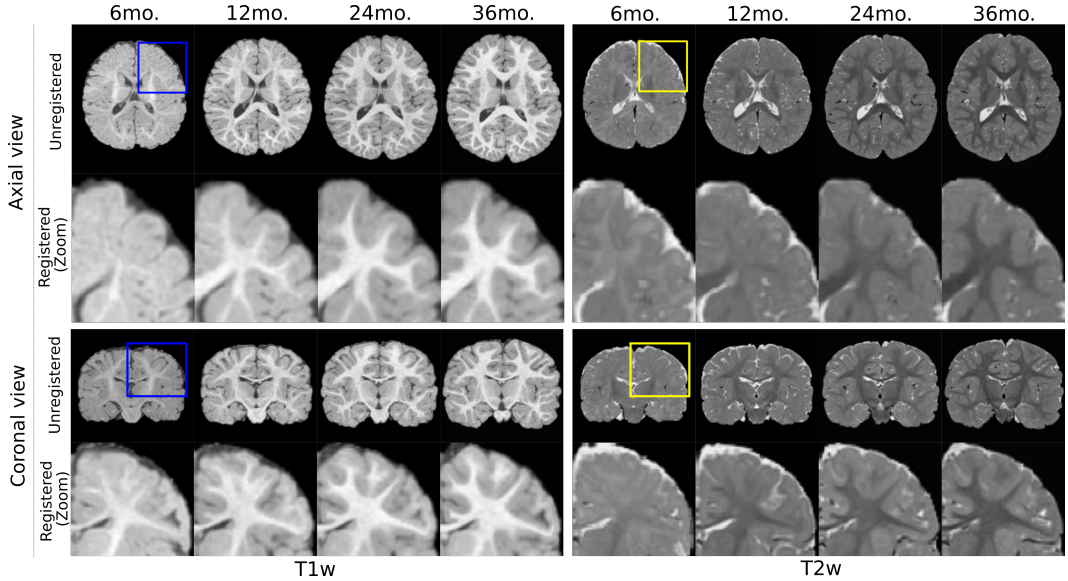

Figure 13: A showcase of **brain appearance maturation** through the early lifespan. Left: T1w; right: T2w images of developing infant brains from axial/coronal view.

Infant white matter maturation. Neurodevelopment in infants and toddlers is a highly-complex process of macro and micro-structural changes (for example, growth and myelination, respectively). Relevant to the scope of this paper, real-world tissue segmentation (into grey and white matter) of infant brains at approximately 6 months of age (post-birth) is highly difficult due to ongoing white matter myelination giving the brain an “isointense" (flat in intensity) appearance which ambiguates anatomical edges. Fig. 13 visualizes this phenomenon.

Currently, existing datasets for supervised segmentation of isointense infant brains [53] construct their ground truth labels by first nonlinearly registering algorithmic segmentations of an older timepoint from a given subject image time-series (which is straightforward to segment with existing tools) and then manually correcting the warped labels with the expertise of neuroradiologists.

In our one-shot segmentation setting, we take a partially analogous approach to segmentation of isointense infant brains. We first algorithmically segment [46] a single arbitrarily-selected 36 month old T1w/T2w MR image and use this reliable segmentation to train all segmentation baselines and the proposed method. These methods which have only been trained on a 36 month old brain are then quantitatively evaluated on isointense 6 month old brain tissue segmentation in this paper. The target labels for the 6 month old evaluation set are generated algorithmically with supervised segmentation networks following [64] which have been trained on a separate dataset [53].

[2] `https://nda.nih.gov/edit_collection.html?id=19`
[3] `https://www.oasis-brains.org/`

E Need for regularization

In this section, we qualitatively and quantitatively illustrate that proper regularization of patch-wise *negative-free* multiscale representation learning avoids low-diversity or degenerate decoder representations which hamper downstream segmentation performance. We primarily compare the complete proposed model (*'Ours w/ regularization'*) against an ablation (*'Our ablation w/o regularization'*) that only uses the patch-wise similarity loss and does not use the denoising, variance, covariance, and orthogonality regularization (which corresponds to ablation E from Table 2).

Figure 14 visualizes activations from layers {6,8,10,13,16} of Table 9 (left) and the sorted singular values of the spatial covariance matrices of multi-layer feature projections from the corresponding layer-wise projector output (right) trained with/without regularization. Given a flattened mid-axial feature projection $F \in \mathbb{R}^{wh \times c}$ (where $c = 2048$ and wh is the vectorized spatial dimensionality), we calculate the spatial covariance as $C = \frac{1}{c}\sum_{i=1}^{c}(z_i - \bar{z})(z_i - \bar{z})^T$, where $z_i \in \mathbb{R}^{wh \times 1}$ is a spatial feature vector and $\bar{z} = \frac{1}{c}\sum_{i=1}^{c} z_i$. The lower the rank of C, the lower the spatial variability of representations.

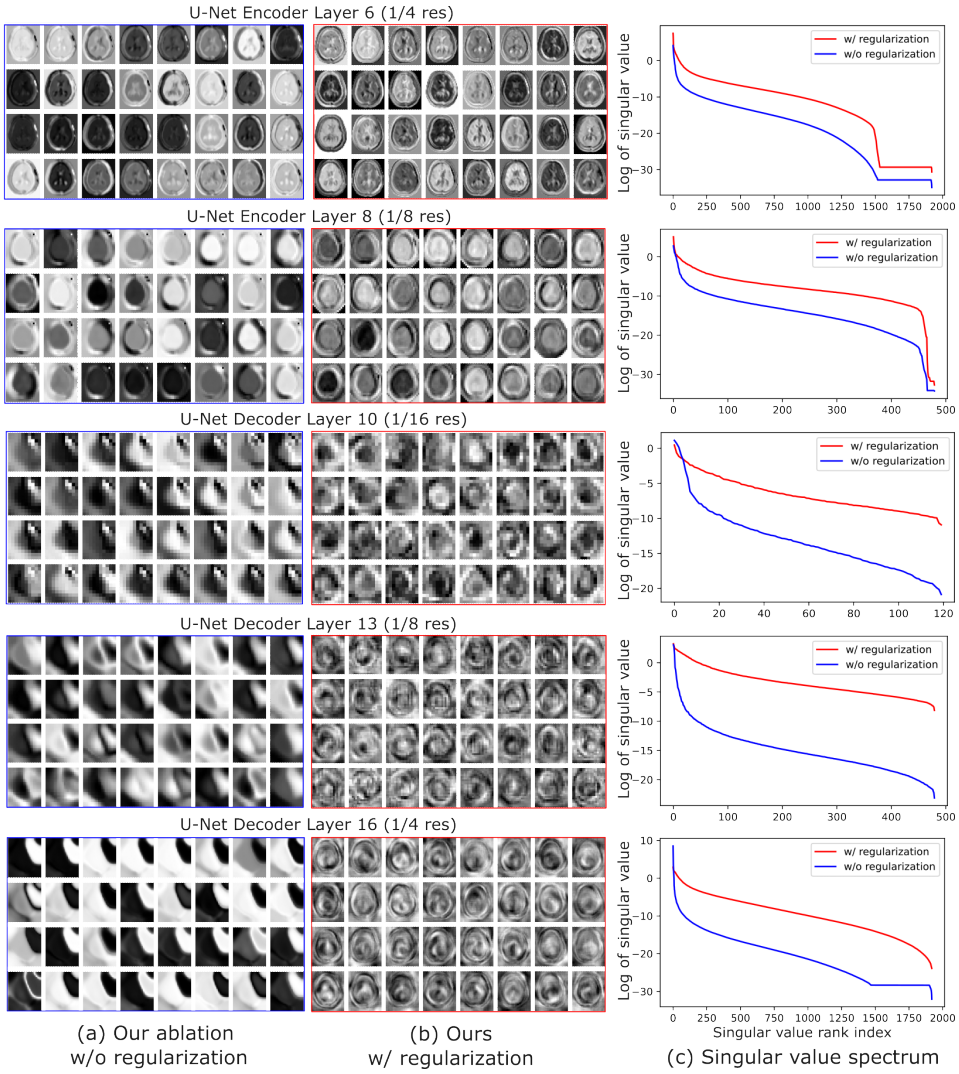

Figure 14: An illustration of U-Net activations (left) and the log-scale singular values of the spatial covariance matrix C (right), both trained with (red) and without (blue) regularization. While encoder activations are comparable both with and without regularization, the decoder U-Net activations converge to degenerate spatial patterns without regularization (left) and their projections have much lower spatial variability as shown by the singular values of their covariance matrices (right). With regularization, decoder layers avoid degenerate solutions and their spectra indicate higher rank.

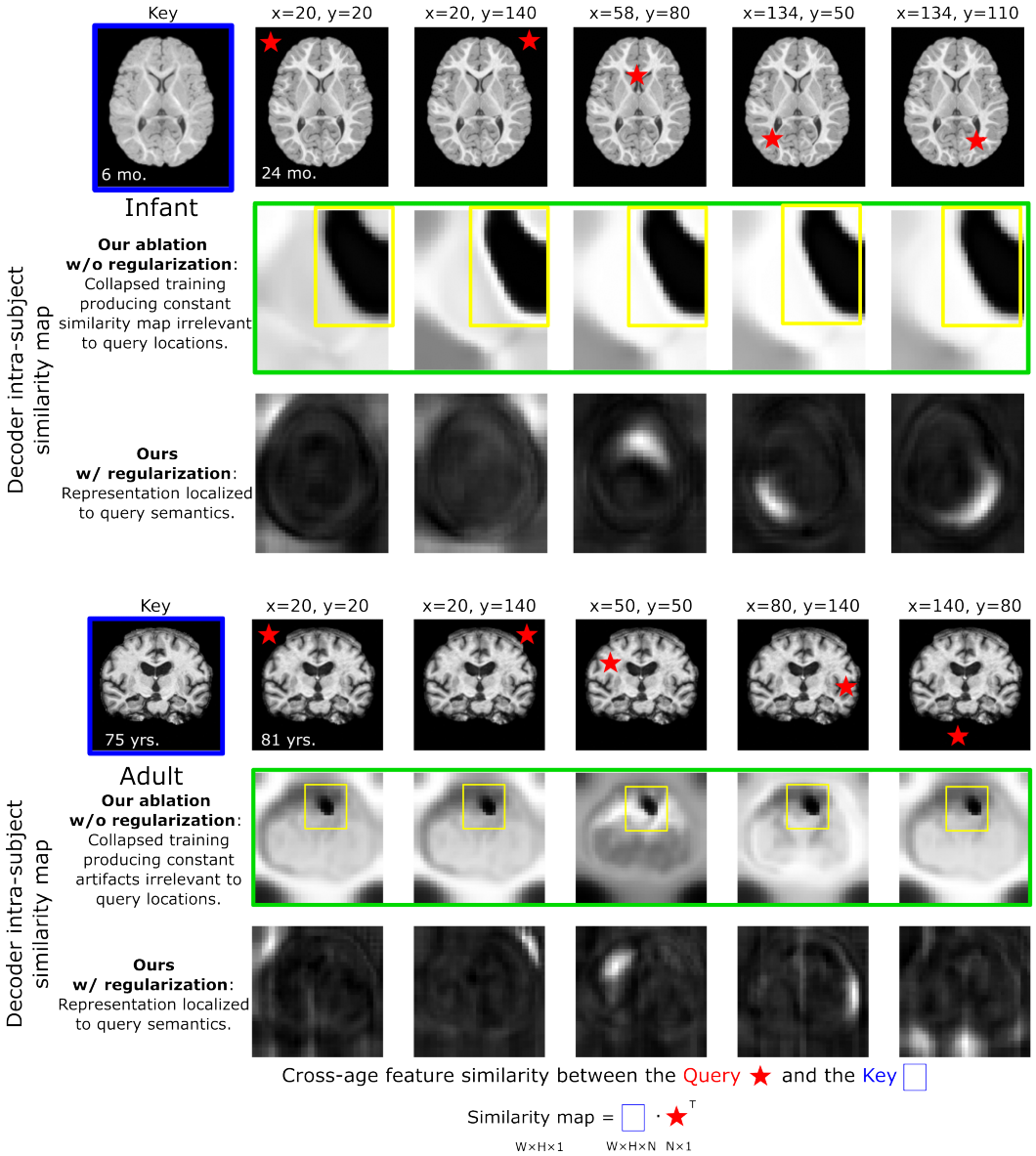

Figure 15: **Varying queries against a constant key.** To complement Figures 1 and 5, this figure qualitatively visualizes the intra-subject similarity maps obtained by querying the projections of decoder layer 16 (from Table 9) from several distinct regions (red stars, columns) and comparing them against the projection of the corresponding decoder activation from a key intra-subject temporal image (blue box, left). Without regularization (rows 2 and 5), the ablation yields intra-subject similarity values with very low spatial diversity across all queries, as indicated by similar appearance across all entries in the green boxes and the spatial artifacts in the yellow boxes. However, given proper regularization (rows 3 and 6), these values converge to semantically-similar regions, which indicate better representations for semantically-driven tasks as indicated by improved downstream segmentation performance.

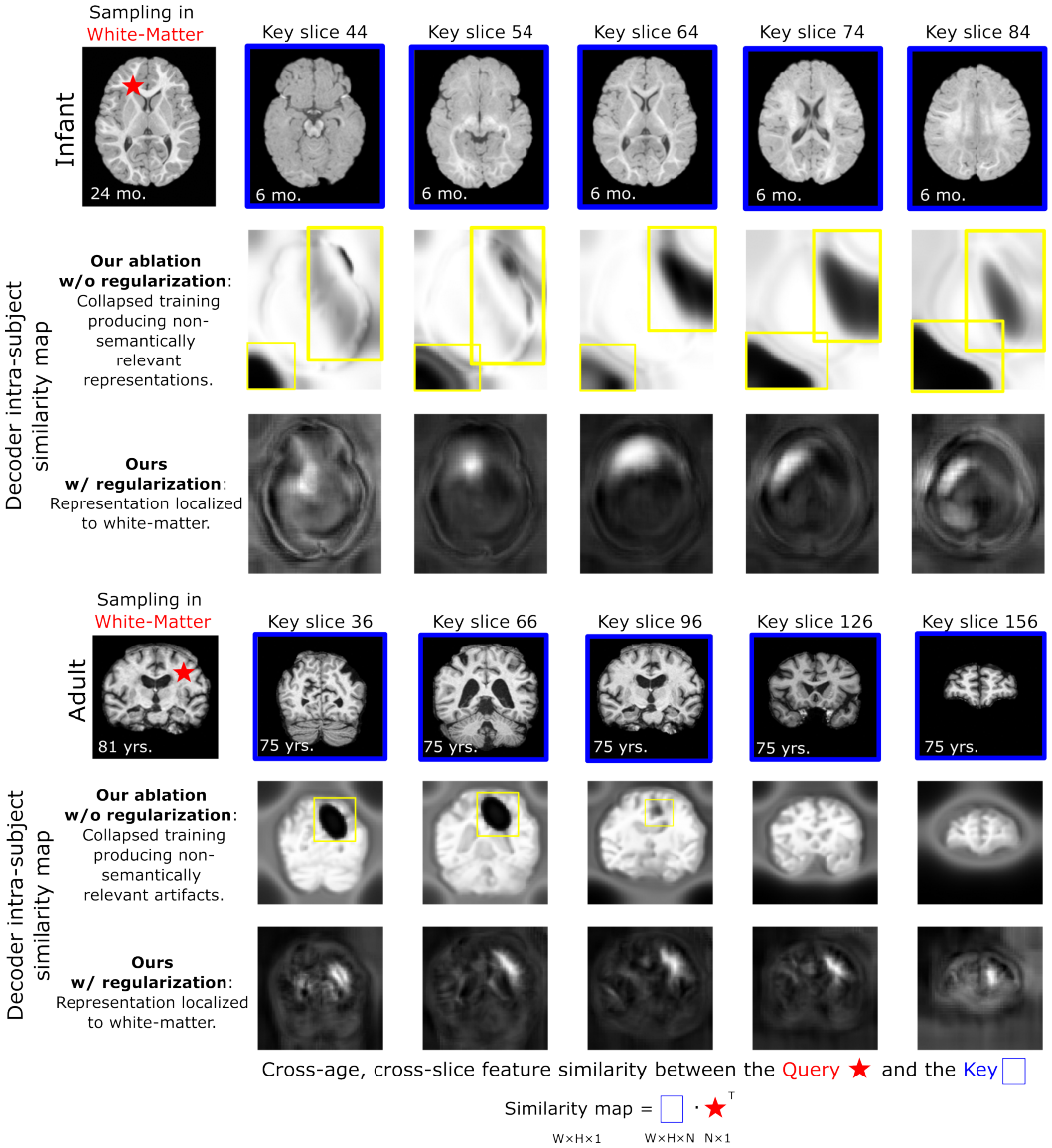

Figure 16: **Varying keys against a constant query.** To complement Figure 15, we now hold constant the query (red star, sampled from white matter) decoder feature projection from layer 19 (of Table 9) and visualize the intra-subject similarity values computed against several distinct key slice projections (blue boxes) from the corresponding decoder representations of an intra-subject temporal image. As in Figure 15, the ablation without regularization (rows 2 and 5) yields degenerate and semantically-unmeaningful similarity patterns (yellow boxes) with low spatial diversity and artifacts across 3D space. With regularization, self-similarity patterns are semantically-coherent in 3D.

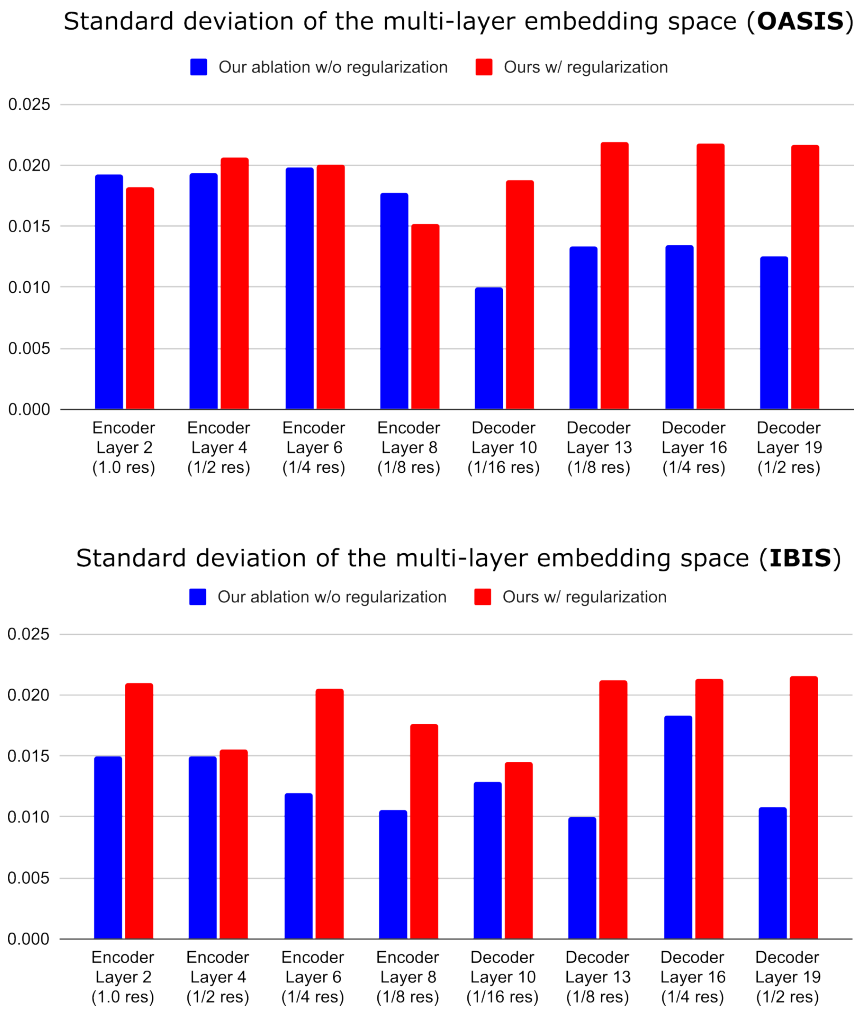

Figure 17: **Variability of local U-Net feature embeddings.** This figure visualizes the standard deviations of channel-wise projection vectors from multiple layers of a U-Net pretrained with (red) or without (blue) regularization. As previously suggested by Figures 14, 15, and 16, decoder projections have significantly lower spatial variability without regularization, indicating low-diversity representations. With regularization, spatial variability is increased which ultimately enables better transfer to downstream tasks such as segmentation.

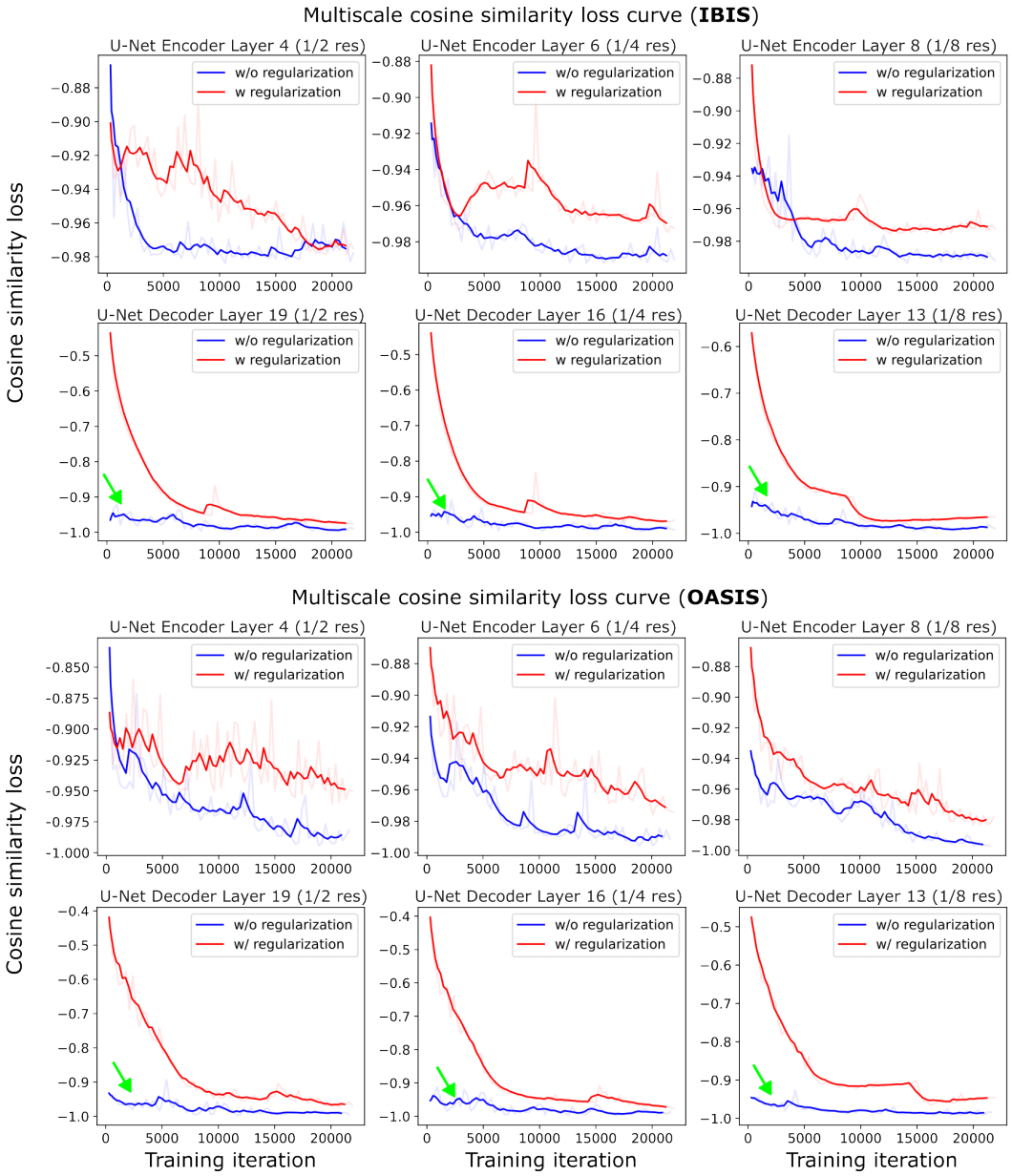

Figure 18: **Multi-layer patch-wise cosine similarity loss ($\mathcal{L}_{sim}$).** Here, we visualize the positive-only patch similarity loss against training steps for a selection of encoder and decoder layers (columns) across both datasets with (red) and without (blue) regularization. During initial training stages, we see that encoder layer losses (rows 1 and 3) gradually decrease in both regularized and unregularized settings. However, decoder layer losses (rows 2 and 4) show near-immediate convergence to a degenerate ideal solution (green arrows) without regularization and these unregularized representations do not transfer well to downstream segmentation as shown by the quantitative results in this paper. This phenomenon is consistent with previous literature on negative-free representation learning (see Figure 2a of [13]).