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 self-similarity of anatomical volumes for pretraining prior to downstream tasks such as segmentation. However, current methods assume i.i.d. image acquisition, which is invalid in several clinical studies 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 contributes two methods: (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 multi-scale intra-subject image features for pretraining and develops several feature-wise regularizations that avoid collapsed identity representations; (2) During finetuning, it proposes a surprisingly simple self-supervised consistency regularization to exploit intra-subject anatomical correlation. On finetuning, the proposed framework outperforms both well-tuned randomly-initialized baselines and current self-supervised techniques designed for both i.i.d. and longitudinal datasets in the one-shot segmentation setting. These improvements are demonstrated across both longitudinal neurodegenerative adult MRI and developing infant brain MRI and yield both higher performance and intra-subject 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 [46]), which can then be efficiently finetuned in the one or few-shot setting. These *self-supervised* methods may handcraft pre-text training objectives [10, 17, 43, 61] or may attempt to pretrain the base network to be equivariant to transformations which preserve semantic meaning as in contrastive learning in medical [9, 59] and natural image vision [2, 23, 35, 57, 63, 64]. 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.

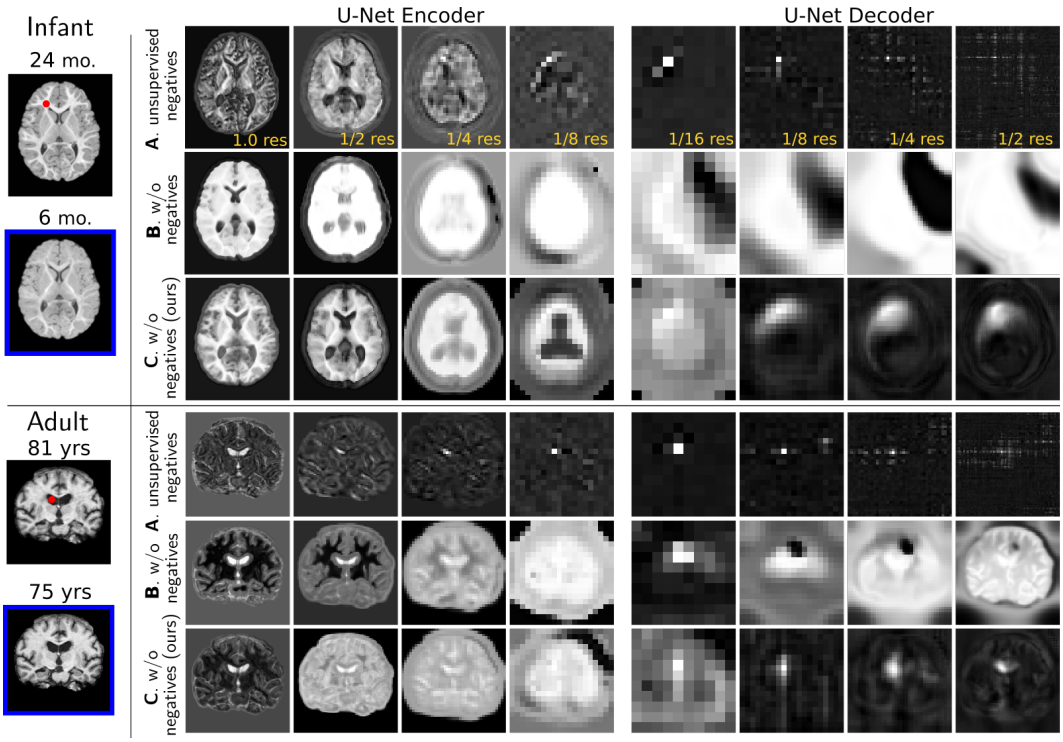

Figure 1: On pretraining an image-to-image network with *per-layer* multi-scale self-supervision, we visualize the **feature-wise self-similarity** between a **key** position and **query** its similarity to all positions at a different age. **A:** Contrastive pretraining with unsupervised negatives [42] yields only positionally-dependent representations. **B:** Pretraining w/o negatives [11] by using corresponding intra-subject patch locations as positives leads to mode collapse and artifacts in deeper layers. **C:** Our method attains both positionally and anatomically-relevant representations without collapse.

only at a single feature scale (typically the encoder bottleneck or network output). Naive application of local unsupervised multi-scale contrastive losses [42] lead to degenerate representations (Fig. 1) 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 and 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, 41, 62], such that the encoder learns representations which are aware of the order of acquisition [15] and the overall trajectory [41]. 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 do without mode collapse by developing regularizations for self-supervised similarity objectives. Further, 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 without mode collapse, it develops regularization terms on the variance, covariance, orthogonality of local decoder features within the decoder; (3) To further self-supervise the fine-tuning stage, the proposed method encourages segmentations from multiple timepoints on unlabeled data to be consistent; (4) Across three large-scale longitudinal one-shot segmentation tasks on both elderly and pediatric populations, the developed framework yields improved few-shot 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 [17], context restoration [10, 43], channel prediction [61], among others. Given their heuristic nature and suboptimal generalization, recent work instead focuses on data-driven SSL losses.

Contrastive learning. Constrastive learning [11, 20, 21, 22, 47, 53] (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 [19, 40] (NCE) loss. While performant on image-level recognition tasks [12, 29], 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 [24]. For example, when learning on intra-domain internal image patches [42], 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 identity solutions (*mode collapse*) via predictor networks and custom backpropogation [13, 18] and careful regularization [6, 58]. However, as above, these methods operate on *global* image embeddings and require modification for pixel-level tasks.

Spatial self-supervision. Towards downstream semantic segmentation, recent work [2, 23, 32, 35, 36, 57, 63, 64] 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, [42] 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 the decoder mode collapse 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, 27, 56]. However, longitudinal biomedical image time-series typically 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, 41, 62] 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, 25, 37, 55]. In particular, longitudinal image analysis typically makes use of one or few longitudinal atlases [26, 30, 45, 49, 50], which motivates the one-shot segmentation setting benchmarked in this paper.

More recently, deep segmentation networks achieve strong performance [8, 7, 38, 39, 46] given enough training volumes. In the low-annotation setting, weakly supervised methods develop custom loss functions [28, 34], 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, 59] 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 [16, 33], 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.

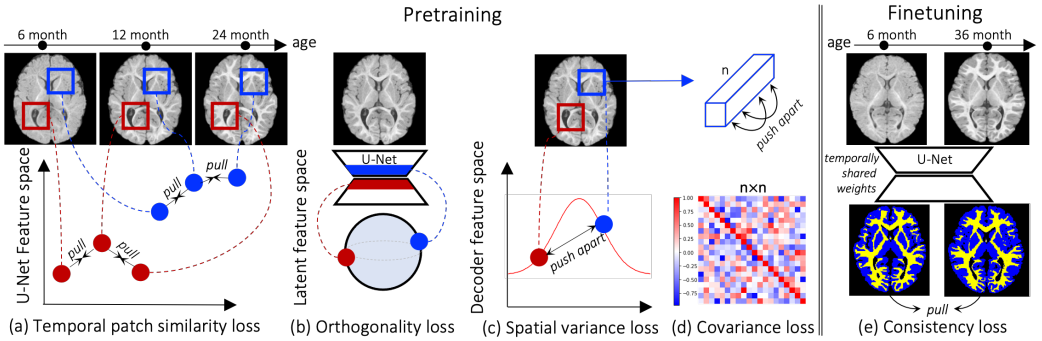

Figure 2: **Overview of proposed self-supervision.** Given nonlinearly-registered temporal images of a subject, **(a)** we assume that corresponding spatial locations in various network layers should have similar representations. As U-Net skip connections can cause decoder mode collapse, we **(b)** encourage 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 gives an overview of the proposed framework. The base U-Net architecture is pretrained end-to-end in a self-supervised manner with intra-subject temporal and spatial losses. On convergence, the pretrained network parameters serve as an initialization for training downstream local spatiotemporal pixel-level tasks (e.g., registration or segmentation). This work focuses on one-shot segmentation, where the base network will be finetuned on one annotated subject. 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 in Fig. 2a should maintain high similarity despite varying appearance. The feature regularizers (Fig. 2b,c,d) avoid U-Net decoder mode collapse 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 and the patch-wise activations from the 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. 2) 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 mode collapse in the decoder layers as observed from the learned feature similarity maps (Fig. 1 B). We speculate that this is partially due to the skip connections between the encoder and decoder,

which lead to a degenerate solution where the encoder learns representations which are good enough to minimize the decoder losses and therefore the decoder layers do not have to learn meaningful representations. To this end, we develop several regularization strategies (described below) such that the decoder layers avoid collapse and all layers of the U-Net are adequately pretrained.

Orthogonality. During model prototyping, we empirically observe that decoder mode collapse first originates in the bottleneck *latent-space* of the U-Net. We therefore encourage decoupled latent 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 alleviate mode collapse. We do this by extending the variance and covariance terms of [6] towards patchwise multi-scale operation and apply these terms 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, $\epsilon = 10^{-4}$ is added for numerical stability. A covariance regularization decorrelates channelwise activations in the decoder to prevent collapsed embeddings by minimizing $\mathcal{L}_C(z) = \frac{1}{k} \sum_{l \in \{L-k,\dots L\}} \frac{1}{n} \sum_{u \neq v} [C(z)]_{u,v}^2$, 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 collapse. Therefore, inspired by denoising autoencoders [54], 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 an 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. 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 the results are reported on a held-out test set. ANTs [3, 4] is used to perform inter-subject affine alignment for all methods and experiments, followed by intra-subject nonlinear deformable registration to obtain accurate

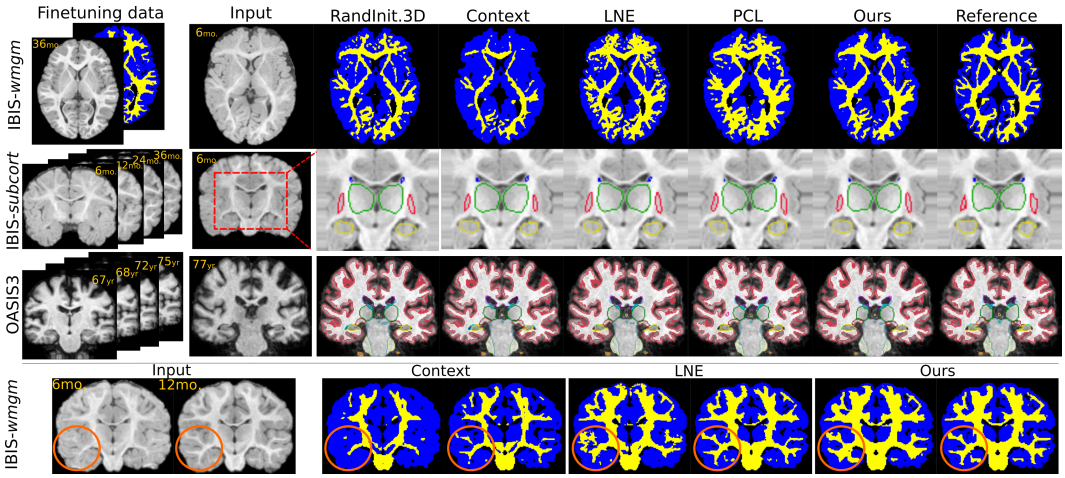

Figure 3: **One-shot segmentation**. **Top 3 rows:** Once pretrained on all unlabeled data, all benchmarked methods are finetuned on either a single annotated image (IBIS-*wmgm*) or a single annotated subject (IBIS-*subcort* and OASIS3). When deployed on other subjects at different ages, our method yields 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.

spatiotemporal correspondence for $\mathcal{L}_{sim}$ and $\mathcal{L}_{cs}$ calculation. All images are skull-stripped, bias-field corrected, and intensity-normalized with further details given in the appendices.

OASIS3 [31] is a publicly-available dataset consisting of 1639 MRI scans of 992 longitudinally imaged subjects (each with 1–5 temporal acquisitions over a $\sim$ 5-year long observation window), resulting in an aging cohort over the span 42 to 95 years, including cognitively normal and mildly impaired individuals alongside subjects with with early-stage Alzheimer's Disease. *Cross-sectional* FreeSurfer anatomical segmentation was done as part of the data release. Observing strong temporal inconsistency (see appendices), we further perform *longitudinal* FreeSurfer [45] 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 is performed on a single FreeSurfer-annotated subject with four timepoints with the remainder used for evaluation.

IBIS is an ongoing 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 [52] cross-sectionally segments sub-cortical grey matter (relevant to ASD [47]) into 13 structures of interest, which are then followed by manual corrections. Similar to OASIS3 above, 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, but 6-month-old segmentation remains elusive without manually labeled datasets for supervised training due to myelination trends leading to isointense appearance at those ages [51] (examples in Fig. 2a). We therefore investigate finetuning all benchmarked methods on a single 36 month old image which can be reliably segmented with [44] 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 [60]).

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 including the pretext-task based `Context Restoration` [10], longitudinal representation learning based `LNE` [41], along with contrastive learning based `GCL` [9] and `PCL` [59] methods which operate on image slices. We also repurpose `PatchNCE` [42] to evaluate its generic representation learning capabilities. All methods are both 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 [33] $STCS = \frac{2|S_1 \cap S_2|}{|S_1|+|S_2|}$ where S_1 and S_2 are temporal segmentation predictions from non-linearly registered input images; and the absolute symmetrized percent change [45] $ASPC = 100\frac{|V_2 - V_1|}{0.5(V_1+V_2)}$, 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 [46] as the base image-to-image architecture with four levels of up/down sampling and repeated Conv-BN-ReLU blocks, with full architectural details provided in the supplementary material. 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. For both pretraining and finetuning, we use a batch size of 3 crops, with an initial learning rate of $2e-4$. All networks are trained with the Adam optimizer ($\beta_1 = 0.9$ during pretraining and $\beta_1 = 0.5$ during finetuning, and $\beta_2 = 0.99$) 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 the appendices.

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, where the predictions from Context Restoration [10] match only the input image intensity and thus lack anatomical and longitudinal consistency; the predictions from LNE [41] introduce false positive predictions in temporal lobe (within the orange circle) and our method yields a more spatiotemporally and anatomically consistent segmentation.

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 key and each query 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 mode collapse and artifacts in the decoder feature maps 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. provided in the appendices.

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 [41]	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 [59]	0.713	3.270	5.610	0.562	4.974	10.648	0.707	2.327	4.850
PatchNCE [42]	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

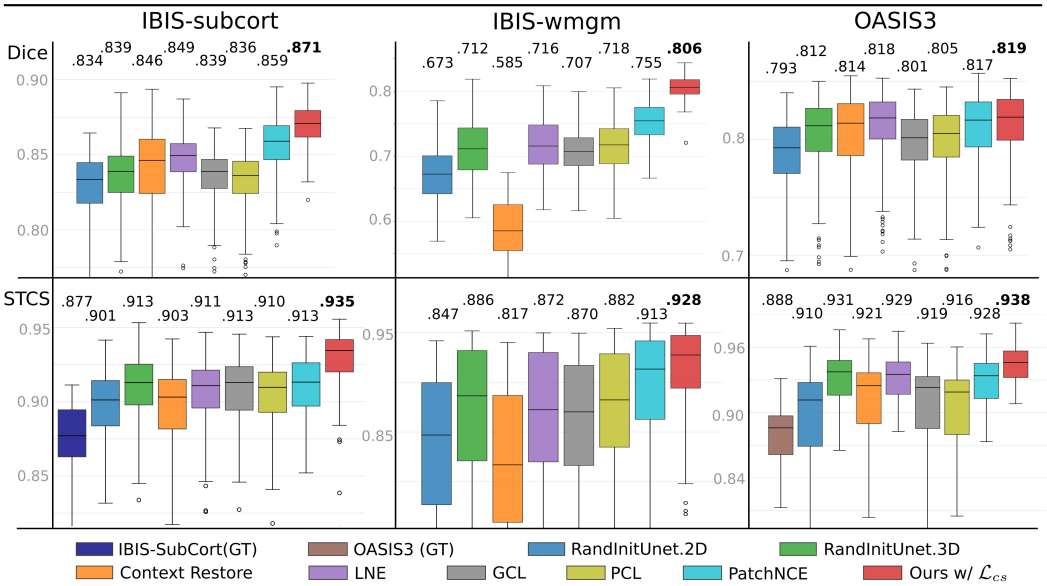

Figure 4: **One-shot segmentation benchmarking** quantifying performance with the Dice coefficient (**top**) and the spatiotemporal consistency of segmentation (**bottom**), visualizing the means and standard deviations alongside median values overlaid on the top of each subfigure (higher is better).

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 mode collapsed representations (e.g., Fig. 1 B). We speculate that mode collapse (due to a lack of regularization) may be equivalent to random initialization. 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 medical 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 avoiding mode collapse via careful regularization of the decoder activations. During finetuning, it similarly developed a simple consistency-regularization objective which encourages longitudinal agreement between predictions on unannotated data. When applied to large-scale neurodeveloping and neurodegerative longitudinal image sets, 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 (OASIS3) and Autism Spectrum Disorder (IBIS). Our source 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] 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.

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

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

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

[33] 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.

[34] 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.

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

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

[37] 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.

[38] 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.

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

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

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

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

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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: OASIS3 is a publicly available dataset which can be accessed with grant. However, IBIS is an on-going research study that has not been publicly released as of now.
   (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 supplementary.
   (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 supplementary
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] See [31] and [48] 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.