A. Method configuration

Across all baseline methods, we utilize a common set of hyperparameters. For all baseline methods we utilize the same pre-training dataset with the same image preprocessing. Moreover we use the same amount of pre-training steps (250k) as for our S3D-B method and the same fine-tuning scheme, as highlighted in Tab. 10. Aside from this, we employ the SGD optimizer with LR 1e-2 with a PolyLR schedule, momentum 0.99 and weight decay 3e-5 across all pre-training experiments, as they showed to be highly robust and reliable in the supervised medical image segmentation setting using CNNs [20]. Moreover, we denote that all these methods have their backbones replaced with a ResEnc U-Net to minimize confounding effects of different architectures.

A.1. Models Genesis

Models genesis [48] pre-text task is centered around restoring original patches from transformed versions. The transformed version is achieved by applying four different transformations in various combinations, with the following transformations: a composition of four separate pre-training schemes: (i) **Non-linear intensity transformation:** Alters the intensity distribution while preserving the anatomy, focusing on learning the appearance of organs. (ii) **Out--painting:** Removes part of the image and requires the model to extrapolate from the remaining image, forcing it to learn the global structure of the organs. (iii) **In-painting:** Masks a part of the image, and the model learns to restore the missing parts, focusing on local continuity and context. After having transformed the original image through these augmentations, the model is trained to recover the original image through a convolutional encoder-decoder architecture. This approach consolidates different tasks (appearance, texture, and context learning) into one unified image restoration task, making the model more robust and generalizable.

Model specific Hyperparameters: The entire set of hyperparameters of Models Genesis are contained within the data-augmentation. This allows us to transfer this transformation pipeline, as provided in the official repository without any changes to the hyperparameters.

A.2. VolumeFusion

Volume Fusion [43] is a pseudo-segmentation task using two sub-volumes from different 3D scans, which are fused together based on random voxel-level fusion coefficients. The fused image is treated as input, and the model predicts the fusion category of each voxel, mimicking a segmentation task. Pretraining is optimized using a combination of Dice loss and cross-entropy loss.

Method specific parameters: Volume Fusion has unique parameters defining the size ranges of the rectangles used for fusing together images. In our experiments we utilize a rectangle size range between [8, 100] sampled uniformly for each axis. This represent the 62.5% of our input patch size, and identical percentage as in the original paper. Moreover the amount of rectangles sampled is an important parameter. Like in the original paper we sample $M \sim \mathcal{U}(10, 40)$ different rectangles, iteratively. Lastly, the number of categories was chosen to be 5, as in the original paper (this represent $K = 4$).

A.3. VoCo

The 'Volume Contrastive Learning Framework' (VoCo) [46] is designed to enhance self-supervised learning for 3D medical image analysis by leveraging the consistent contextual positions of anatomical structures. The method involves generating base crops from different regions of 3D images and using these as class assignments. The framework then contrasts random sub-volume crops against these base crops, predicting their contextual positions using a contrastive learning approach. The authors utilize a SwinUNETR model architecture, employing the AdamW optimizer with a cosine learning rate schedule for 100,000 pre-training steps. The specific hyperparameters include cropping non-overlapping volumes with a size of 64x64x64, and generating 4x4 base crops during the position prediction task. This represents an input patch size of 384×384×96 which is rescaled and resized to fit exactly 4x4 64x64x64 crops.
Since our chosen patch size 160x160x160 is incompatible with the 64 cube length, we adjusted our patch size for VoCo to 192x192x64. This accommodates a 3x3 grid of 64x64x64. Unfortunately the 4x4 grid led to exceeding the memory limit hence a reduction was necessary. Moreover we increased the target crop size from 4 originally to 5 and increased the batch size from 6 (default in our other experiments) to 12, to fully utilize the 40GB VRAM of an A100 node.

B. Longer Training schedule

MAEs are known to benefit from increasing the length of the training schedule, as shown in He et al. [16]. We evaluate if this effect transfers to 3D medical pre-training by increasing the training batch size by x8 to 48, the learning rate to $3e-2$, and the iteration steps by x5 to a total of $1.25M$ steps. We refer to this model as **S3D-Long** to denote the longer training schedule with more data seen. We denote that the architecture remains identical to the previous architecture, to isolate the effect of the length of the steps as well as the amount of samples seen. Results are presented in Tab. 12. It can be observed that this x32 actually leads to a decrease in overall model performance on our test datasets,

Table 5. **Publicly available checkpoint trained on the ABCD dataset:** To provide a public available checkpoint, we retrained our proposed model on the ABCD dataset, indicated by a *. It performs slightly worse than the network pre-trained on the private dataset.

SSL Method	No (Dyn.)	No (Fix.)	**S3D***	**S3D**
Dataset	**Dice Similarity Coefficient (DSC)**			
MS FLAIR (D1)	57.81	59.82	59.75	**60.35**
Brain Mets (D2)	63.66	56.53	64.20	**65.24**
Hippocampus (D3)	89.18	89.24	89.45	**89.60**
Atlas22 (D4)	63.28	65.52	66.61	**66.95**
CrossModa (D5)	**85.64**	83.44	83.61	84.08
Cosmos22 (D6)	60.28	78.17	**80.01**	80.00
ISLES22 (D7)	77.94	79.44	78.94	**79.70**
Hanseg (D8)	59.00	61.85	61.27	**62.11**
HNTS-MRG24 (D9)	66.73	65.90	67.03	**68.62**
BRATS24 Africa (D10)	**93.07**	92.51	92.49	92.19
Avg. DSC	71.66	73.24	74.34	**74.88**
Avg. Rank	3.2	2.9	2.4	**1.5**

showing a 0.6% lower average DSC as well as a 0.6% lower Average NSD.

The observed performance degradation of the S3D-Long model, despite the longer training schedule and increased data exposure, suggests several possible factors at play. While MAEs have shown benefits from extended training schedules in general computer vision tasks, the same assumptions may not directly transfer to 3D medical image pre-training due to the unique nature of this domain. The findings highlight the importance of tailoring training strategies to the domain.

C. Additional results

Aside from the quantitative data on the development and test dataset, we provide the quantiative data of the ablation experiments here. The following additional results are provided: 1. Results when fine-tuning in a low-data regime are presented in Tab. 6. 2. Experiment on how to best transfer weights when transferring to a dataset with more than 1 input channel is provided in Tab. 8 3. Results on how the pre-training effects generalization is provided in Tab. 9. 4. Experiment results of investigating if one can reduce the fine-tuning steps are presented in Tab. 10

C.1. Public weights trained on the ABCD dataset

Due to patient privacy concerns and data ownership regulations, we are unable to share the original pre-trained weights. As an alternative, we retrained our best-performing model on the National Institute of Health's Adolescent Brain Cognitive Development (ABCD) dataset. This dataset comprises about 41k MRI scans with a 50-to-50 ratio of T1-weighted to T2-weighted scans. The results, shown in Tab. 5, indicate that the original model slightly outperforms the version pre-trained on the ABCD dataset. This discrepancy is likely attributable to the greater diversity and variation in the images within our private dataset, which enables a more robust feature representation.

Table 6. **Forty images with SSL are almost as good as all data from-scratch!** The pre-trained S3D model almost reaches the performance of the model trained from-scratch with only 40 training cases, with the exception of D4. Overall train/val/test dataset size was 38/10/12 for D1, 67/17/21 for D2, 166/42/52 for D3, 419/105/131 for D4, 134/34/42 for D5. Results in the table are reported on the validation set. *full: Uses all train samples of the dataset. * D1 has only 38 training cases for the train split.*

SSL Method	N Train	D1	D2	D3	D4	D5	Avg. D1-D5
Scratch	10	40.78	43.52	84.94	44.11	76.66	58.00
	20	44.46	59.46	86.75	46.33	78.67	63.13
	30	45.42	64.20	87.14	48.22	78.47	64.69
	40	49.37*	60.13	87.59	50.43	78.37	65.18
	full	49.37	69.13	88.78	60.74	81.33	69.87
S3D (ours)	10	43.48	48.44	84.12	41.51	77.70	59.05
	20	46.58	65.30	86.61	45.50	79.52	64.70
	30	48.12	68.41	86.77	51.62	78.88	66.76
	40	51.49*	72.91	87.46	53.05	80.82	69.15
	full	51.49	74.01	88.83	62.39	81.54	71.65

Table 7. **Pre-training length ablation:** Longer pre-training does not lead to improved performance. Interestingly, when exceeding 250k steps.

PT Iterations	D1	D2	D3	D4	D5	Avg. D1-D5	Train Time [h]
62.5k	49.49	70.79	88.82	62.95	81.27	70.67	28
125k	50.56	70.48	88.86	62.51	81.69	70.82	56
250k	51.02	74.07	88.91	62.81	81.50	71.66	112
500k	50.93	72.71	88.88	62.17	81.86	71.31	224
1M	50.45	71.55	88.92	62.78	81.82	71.10	448

C.2. Comparison to previous work on CT data

Although this study focuses on brain MRI images, we also evaluate our method on a CT downstream task. Table 11 presents the results of our approach on the BTCV multi-organ segmentation task [22]. For comparison, we incorporate a diverse set of results reported in [33], which were trained using an unspecified 80/20 data split. Additionally, we fine-tune the publicly available HySpark checkpoint [33] using the same five-fold cross-validation split as for our method. Remarkably, M3D outperforms all other methods, despite being pretrained exclusively on brain MRI data. Notably, none of the related approaches surpass our backbone trained on scratch. This highlights the critical role of leveraging state-of-the-art networks and advanced training frameworks, such as nnU-Net. [20, 21].

D. Distinction to AMAEs

Concurrently with this work, [25] introduced the AMAEs framework. Like our approach, it utilizes a dataset of ap-

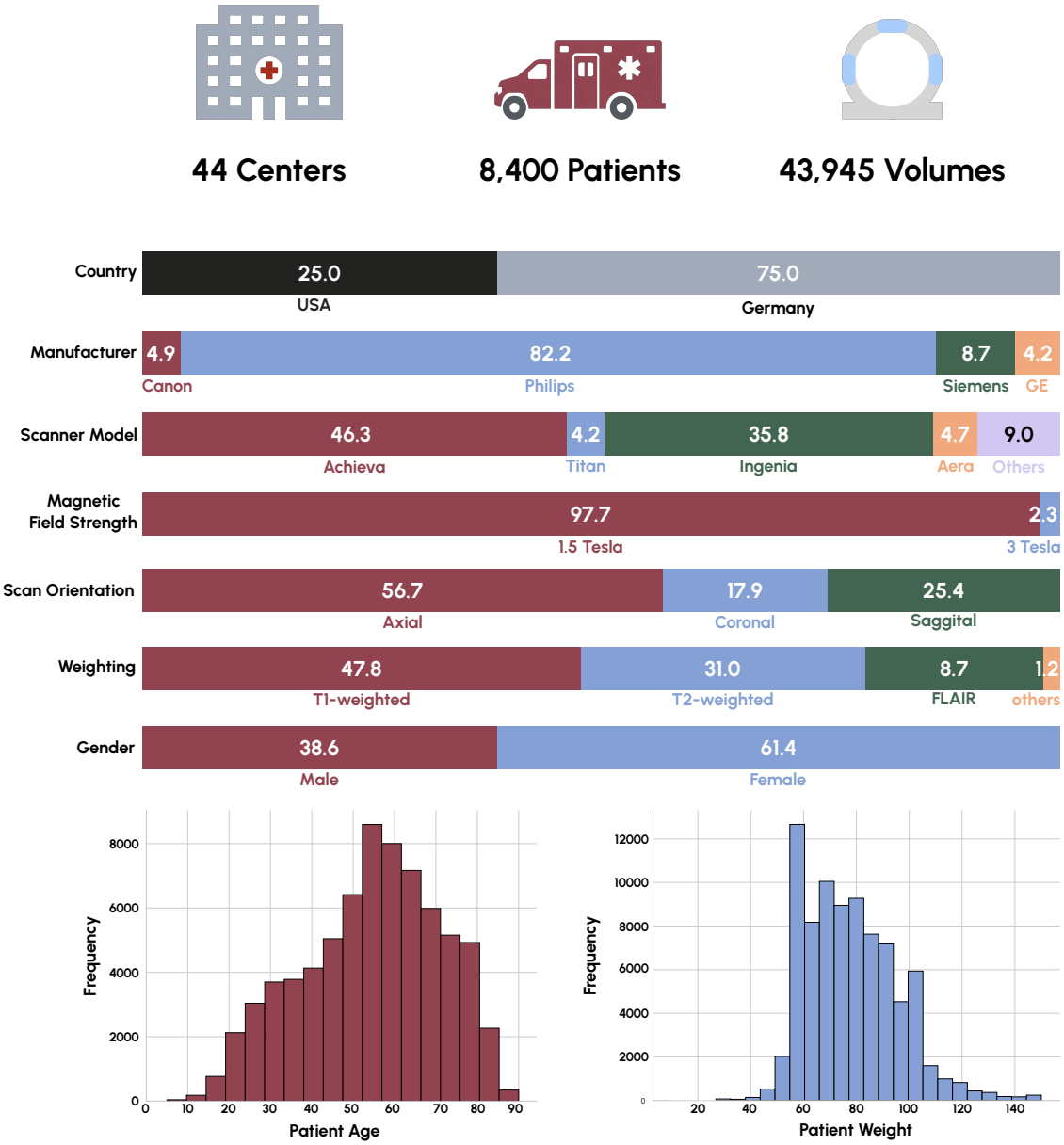

Figure 3. Distribution of our pre-training dataset. The dataset stems from 44 centers and includes 8400 Patients with a 60 to 40 female-to-male ratio. Most patients were imaged with a 1.5 Tesla Philips Achieva or Ingenia scanner. The most prevalent modalities are T1 and T2-weighted images with some additional FLAIR images present. While other modalities were in the dataset, these were not used as prevalence was deemed too low.

Table 8. Replicating the pre-trained stem weights and freezing them during the decoder warm-up phase yields the most stable and equally best results.

Initialization	Decoder Warm-Up	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4	Average	STD
Replication	Frozen	72.84	64.42	66.11	62.86	62.85	65.82	4.15
Replication	Unfrozen	72.68	63.07	65.60	66.02	61.08	65.69	4.39
Random	Frozen	74.38	60.89	65.10	67.55	61.31	**65.85**	5.51
Random	Unfrozen	72.20	63.16	62.25	66.71	61.47	65.16	4.42

Table 9. **Pre-training can improve generalization:** We investigate generalization to a new modality time-of-flight (ToF) MRI (top), and the generalization of a resulting method when translating it to a different clinic (bottom).

Experiment	Setting	No Dyn.	No Fixed	VoCo	VF	MG	S3D-B
Modality shift	TOF Angio. Aneurysms(D12)	**42.61**	22.76	22.32	31.21	34.60	28.72
In Distribution	Brain Mets (D2)	**72.81**	67.93	64.34	71.69	69.05	71.56
Patient shift	Brain Mets (D13)	64.08	61.61	56.78	63.95	64.22	**64.54**

Table 10. **Fine-tuning length:** When initializing from our pre-trained checkpoint, it is possible to achieve a large fraction of the final performance after less than 15% of the normal training time. Despite this a full training schedule reaches better performance. These experiments were conducted using S3D long on the validation splits.

FT Iterations	D1	D2	D3	D4	D5	Avg. D1-D5
25k	50.85	73.99	88.51	55.49	46.00	62.97
37.5k	51.69	74.03	88.85	60.22	81.68	71.29
50k	51.13	73.53	88.93	60.14	81.92	71.13
75k	51.41	72.80	89.08	63.14	81.83	71.65
150k	50.95	71.28	88.96	62.51	81.92	71.13
275k	53.10	71.24	89.14	63.55	82.53	71.91

proximately 40k 3D images and employs a state-of-the-art CNN architecture. However, in contrast to our more comprehensive evaluation, their assessment is based on three in-distribution and one out-of-distribution downstream dataset.

While evaluation on three in and one out-of-distribution datasets can suffice to draw some insights, their evaluation setup has additional limitations. First, they constrain themselves to a low-data regime and do not assess whether their performance surpasses a default nnU-Net baseline that is trained for 1000 epochs. Second, their fine-tuning strategy is fixed to a single-channel input to align with the pre-trained stem weights. While this would be okay for their method, they apply the same limitation to their baselines, which typically utilize all available modalities, which can lead to stronger performance. For instance, in their evaluation on the BraTS dataset, only one of the four available modalities was used, potentially limiting the effectiveness of the baseline models. These choices may impact result reliability and align with Pitfall 3—unreliable evaluation practices. To address such concerns, our work adopts a very comprehensive evaluation strategy to allow drawing reliable conclusions on the state-of-the-art in self-supervised learning for 3D medical image segmentation for the first time.

E. Acknowledgement of ABCD

Data used in the preparation of this article were obtained from the Adolescent Brain Cognitive Development^{SM} (ABCD) Study (https://abcdstudy.org), held in the NIMH Data Archive (NDA). This is a multisite, longitudinal study designed to recruit more than 10,000 children age 9-10 and follow them over 10 years into early adulthood. The ABCD Study® is supported by the National Institutes of Health and additional federal partners under award numbers U01DA041048, U01DA050989, U01DA051016, U01DA041022, U01DA051018, U01DA051037, U01DA050987, U01DA041174, U01DA041106, U01DA041117, U01DA041028, U01DA041134, U01DA050988, U01DA051039, U01DA041156, U01DA041025, U01DA041120, U01DA051038, U01DA041148, U01DA041093, U01DA041089, U24DA041123, U24DA041147. A full list of supporters is available at https://abcdstudy.org/federal-partners.html. A listing of participating sites and a complete listing of the study investigators can be found at https://abcdstudy.org/consortium_members/. ABCD consortium investigators designed and implemented the study and/or provided data but did not necessarily participate in the analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or ABCD consortium investigators.

Table 11. **S3D outperforms all related work on the BTCV dataset [22].** Despite being pretrained exclusively on brain MRI data, our network outperforms all related methods. Values above the line are sourced from Tang et al. [33], which used an unknown 80/20 split. To fine-tuned their published pre-trained weights using a 5-fold cross-validation. Below the line, all models, including ours, were trained on the same 5-fold cross-validation. Notably, even though all related work leveraged CT data for pretraining, none surpassed the performance of our backbone model trained from scratch, emphasizing the importance of Pitfall 2.

Pre-training Method	Network	Spl	Kid	Gall	Eso	Liv	Sto	Aor	IVC	Veins	Pan	AG	Avg
vox2vec [11]	3D UNet(FPN) [30]	91.40	**90.70**	59.50	72.70	**96.30**	83.20	91.30	83.90	69.20	73.90	65.20	79.50
SUP [14]	Swin UNETR [14]	84.20	86.70	58.40	70.40	94.40	76.00	87.70	82.10	67.00	69.80	61.00	75.80
MAE [16]	UNETR [15]	90.71	87.63	62.50	72.60	96.09	**94.73**	86.11	**90.36**	71.00	75.47	63.77	79.07
SimMIM [36]	Swin UNETR [14]	88.33	86.82	62.43	74.36	92.35	90.70	83.03	87.43	68.04	68.43	58.65	76.44
SparK [36]	MedNeXt [31]	90.92	87.66	62.43	74.36	95.03	84.85	86.04	80.63	68.83	76.57	61.43	79.21
HySparK [33]	MedNeXt+ViT [33]	90.67	88.32	68.18	74.20	95.03	87.46	90.17	84.50	70.04	78.36	66.75	80.67
No	MedNeXt+ViT [33]	90.35	87.46	63.18	74.49	95.09	86.00	89.29	83.22	71.85	79.48	62.59	80.27
HySparK [33]	MedNeXt+ViT [33]	90.94	86.99	63.43	74.39	95.12	87.15	88.92	83.48	72.77	79.66	64.84	80.67
No (Dyn.)	nnU-Net [20]	90.44	88.52	68.86	78.14	95.53	88.06	91.59	86.47	76.27	81.78	71.06	83.34
No (Fix.)	ResEncL (fixed) [21]	91.97	89.58	68.76	**79.18**	95.96	91.97	92.80	87.16	**77.29**	84.01	72.21	84.63
S3D	ResEncL (fixed) [21]	**92.00**	90.40	**70.77**	78.71	96.01	92.51	**92.83**	87.04	77.28	**84.79**	**72.58**	**84.99**

Table 12. **Longer training schedule degrades performance.** When training with a larger batch size, higher learning rate, and more train steps we observe a degradation in performance for DSC and NSD. Ranks are calculated only between the four methods presented in the table.

	Dice Similarity Coefficient			
Dataset	No Dyn.	No Fixed	S3D	S3D-Long
MS FLAIR (D1)	57.81	59.82	**60.35**	59.85
Brain Mets (D2)	63.66	56.53	**65.24**	64.81
Hippocampus (D3)	89.18	89.24	**89.60**	89.34
Atlas22 (D4)	63.28	65.52	**66.95**	64.58
CrossModa (D5)	**85.64**	83.44	84.08	84.02
Cosmos22 (D6)	60.28	78.17	80.00	**80.01**
ISLES22 (D7)	77.94	79.44	79.70	**79.89**
Hanseg (D8)	59.00	61.85	**62.11**	61.93
HNTS-MRG24 (D9)	66.73	65.90	**68.62**	67.94
BRATS24 Africa (D10)	**93.07**	92.51	92.19	92.90
T2 Aneurysms (D11)	46.76	41.97	**47.26**	44.15
Avg. DSC	69.40	70.40	**72.37**	71.77
Avg. Rank	3.09	3.27	**1.55**	2.09
	Normalized Surface Distance			
MS FLAIR (D1)	78.77	80.16	80.03	**80.40**
Brain Mets (D2)	80.72	76.72	**82.53**	82.32
Hippocampus (D3)	**99.46**	99.42	99.46	99.44
Atlas22 (D4)	70.52	73.77	**75.35**	73.45
CrossModa (D5)	**99.85**	99.76	99.81	99.80
Cosmos22 (D6)	72.60	96.47	**97.45**	96.75
ISLES22 (D7)	88.55	90.45	90.59	**90.72**
Hanseg (D8)	82.20	85.94	85.80	**86.20**
HNTS-MRG24 (D9)	71.83	71.26	**74.07**	73.17
BRATS24 Africa (D10)	95.66	95.36	95.06	**95.72**
T2 Aneurysms (D11)	**62.24**	55.56	61.18	57.07
Avg. NSD	82.04	84.08	**85.58**	85.00
Avg. NSD Rank	2.82	3.18	**2.00**	**2.00**