MAPSeg: Unified Unsupervised Domain Adaptation for Heterogeneous Medical Image Segmentation Based on 3D Masked Autoencoding and Pseudo-Labeling

Supplementary Material

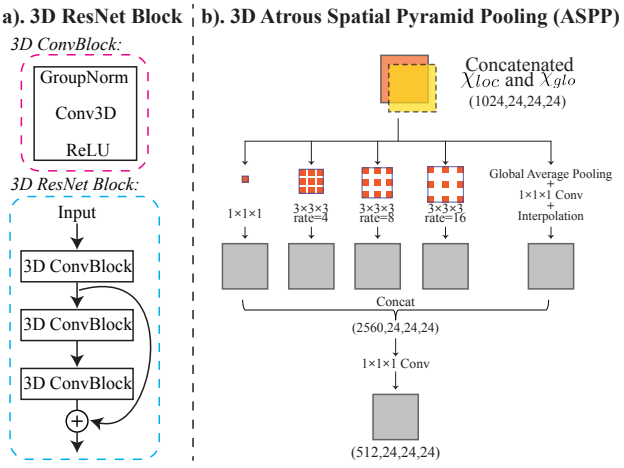
1. Appendix

1.1. Model Architecture

MAPSeg is implemented using PyTorch. Detailed configurations of model and training can be found below.

3D Multi-Scale Masked Autoencoder (MAE). We implement the 3D MAE using 3D ResNet Blocks [10, 24] instead of Vision Transformers, different from the previous study [11], due to the constraint of GPU memory. The encoder consists of eight 3D ResNet Blocks. The 3D ResNet Block is depicted in Suppl.Fig.1a. Following the previous study [11], we adopt an asymmetric design by employing a lightweight decoder (Suppl.Tab.1).

3D Global-Local Collaboration (GLC). The segmentation backbone (Suppl.Tab.1) consists of the pretrained encoder and a segmentation decoder that is adapted from DeepLabV3 [4]. In the decoding path, we take advantage of the Atrous Spatial Pyramid Pooling (ASPP), which employs dilated convolution at multiple scales and provides access to larger FOV (Suppl.Fig.1b). After feature extraction, the GLC module fuses the local and global features and forms a latent representation with a dimension of 1024, which is then fed into the ASPP layer. During training, each local sub-volume with size of $96 \times 96 \times 96$ is randomly sampled from global scan. During inference, the final output is formed by sliding window with stride of 80 across entire volumetric scan.



Suppl.Fig. 1. Illustrations of 3D ResNet Block and 3D Atrous Spatial Pyramid Pooling (ASPP) layer.

Encoder			
Layer Name	Input Size	Output Size	Architecture
enc_res1	(1,96,96,96)	(512,24,24,24)	[4×4×4, 512; 3×3×3, 512; 3×3×3, 512] ×1
enc_res2.x	(512,24,24,24)	(512,24,24,24)	[3×3×3, 512; 3×3×3, 512; 3×3×3, 512] ×7
MAE Decoder			
Layer name	Input size	Output size	Architecture
trans_conv1	(512,24,24,24)	(32,96,96,96)	4×4×4, 32, stride 4
dec_res1	(32,96,96,96)	(16,96,96,96)	[3×3×3, 16; 3×3×3, 16; 3×3×3, 16] ×1
final_recon	(16,96,96,96)	(1,96,96,96)	3×3×3, 1, stride 1
Segmentation Decoder			
Layer name	Input size	Output size	Architecture
ASPP	(1024,24,24,24)	(512,24,24,24)	Suppl. Fig.1b
trans_conv2	(512,24,24,24)	(64,96,96,96)	4×4×4, 64, stride 4
seg_head	(64,96,96,96)	(cls_num,96,96,96)	[3×3×3, 64; 1×1×1, cls_num] ×1

Suppl.Tab. 1. Architectures of different components of MAPSeg. Building blocks ([kernal size, output channels]) are shown in brackets, with the number of blocks stacked. Downsampling is performed by the first block of enc_res1 with a stride of 4.

1.2. Training Recipe

MAE Pretraining. For the MAE Pretraining, we follow the training configurations listed in Suppl.Tab.2. Each mini-batch contains a pair of randomly sampled local patch x and downsampled global scan X. The masking patch in Suppl.Tab.2 only applies to x and is always half-sized for X because of the larger FOV. For example, in the ablation study of masking patch size, a masking patch of 16 to x indicates a masking patch of 8 to X. We implement the augmentation using TorchIO [21]. During the MAE stage, we employ random 3D affine transformation, with isotropic scaling 75-150% and rotation [-40°, 40°].

Centralized UDA. For the centralized UDA on brain MRI segmentation tasks, detailed training configuration can be found in Suppl.Tab.3. Similarly, each mini-batch contains a pair of x and X from the source domain and another pair from the target domain (four $96 \times 96 \times 96$ patches). During warmup epochs, the model is only trained on source domain. We utilize *Score* to select the best model and the patience is set as 50 epochs. For the target domain, we design a similar random 3D affine transformation, with scaling 70-130% and rotation [-30°, 30°]. A stronger augmentation

config	value
masking patch	8×8×8
masking ratio	70%
optimizer	AdamW [19]
learning rate	$2e^{-4}$
weight decay	0.05
optim. momentum	$\beta_1, \beta_2 = 0.9, 0.95$
lr scheduler	cosine annealing [18] T_{max}=20, min_lr=$1e^{-6}$
total epochs	300
annealing epochs	last 100
batch size	4
iters/epoch	500
aug. prob.	0.35
augmentation	random affine

Suppl.Tab. 2. MAE Pretraining Configurations

config	value
masking patch	8×8×8
masking ratio	70%
optimizer	AdamW [19]
learning rate	$1e^{-4}$
weight decay	0.01
optim. momentum	$\beta_1, \beta_2 = 0.9, 0.999$
lr scheduler	cosine annealing warm restart [18] T_0=10, T_{mult}=2, min_lr=$1e^{-8}$
total epochs	100
warmup epochs	first 10
annealing epochs	all
early stop	50
batch size	1
iters/epoch	100
aug. prob.	0.35
source aug.	random affine random bias field random gamma trans.
target aug.	random affine

Suppl.Tab. 3. Centralized UDA configurations for brain MRI segmentation.

strategy is applied to the source domain, consisting of random affine (scaling 70-140% and rotation [-30°, 30°]), random bias field [22, 23], and random gamma transformation ($\gamma \in [e^{-0.4}, e^{0.4}]$). For the centralized UDA on public cardiac CT→MRI segmentation, we use the same configuration except for training epochs of 150 and warmup epochs of 50. For MRI→CT cardiac segmentation, we use a less aggressive augmentation strategy because MRI is noisier than CT. We set the scaling ratio to 85-115% and rotation to [-15°, 15°] for both source and target domains, and exclude random bias field and gamma transformation. The warmup epoch is set as 70.

Federated UDA. For the federated UDA tasks, we follow the procedure detailed in Algorithm 1. We initialize the encoder of the global model f_ϕ with the encoder pretrained on the large-scale data mentioned in Sec.4.3. We set the global FL round $R = 100$. We set both the server and client update steps to 1 epoch with batch size of 1. Training configuration inherits mostly from that of the centralized UDA, except a global cosine annealing learning rate schedule is adopted to decay the learning rate from $1e^{-4}$ to $1e^{-6}$ over the course of the FL rounds.

Test-Time UDA. For the test-time UDA tasks, we follow

Algorithm 1 Federated MAPSeg (Fed-MAPSeg)

Require: Source domain dataset $D_S = \{(x_s, y_s)\}$ and target domain datasets $D_T^k = \{(x_t^k)\}$ for each client k, pretrained global model f_ϕ, number of FL round R, number of server update steps T_s, number of client update steps T_t

for $r = 1, 2, \cdots, R$ **do**
  Initialize server EMA teacher model: $\theta \leftarrow \phi$
  for $t = 1, 2, \cdot, T_s$ **do**
    Sample patches (x_s, y_s) from D_S and generate downsampled global volume and masked inputs X_s, X_s^M, x_s^M
    Update f_ϕ on server by minimizing $\mathcal{L}_s$ (Eq.9)
    Update server EMA teacher model parameter θ with (Eq.3)
  end for
  Server broadcast θ to clients
  for each client k in parallel **do**
    $\phi_k \leftarrow \theta$, $\theta_k \leftarrow \theta$
    for $t = 1, 2, \cdots, T_t$ **do**
      Sample patches x_t^k from D_T^k and generate downsampled global volume and masked inputs X_t^k, $(X_t^k)^M$, $(x_s^k)^M$
      Generate pseudolabels for unmasked inputs x_t^k and X_t^k using the teacher model f_{θ_k}: $f_{\theta_k}(x_t^k)$ and $f_{\theta_k}(X_t^k)$
      Update f_{ϕ_k} by minimizing $\mathcal{L}_u$ (Eq.10)
      Update client EMA teacher model parameter with (Eq.3)
    end for
    Upload θ_k to server
  end for
  The server aggregates θ_k from clients:

$$\bar{\theta} \leftarrow \sum_k \frac{|D_T^k|}{\sum_k |D_T^k|} \theta_k$$

  Update server model parameters $\phi \leftarrow \bar{\theta}$
end for

the same configuration as listed in Suppl.Tab.3. The difference is that the model can only access source domain data (image and label) during warmup epochs and can only access target domain data (image only) after that, while centralized UDA has synchronous access to both source and target domain data throughout the whole training process.

1.3. Implementation of Comparing Methods

For other comparing methods in centralized UDA, we adapt their official implementations. For DAFormer, HRDA, and MIC, we modify the ground truth labels to make them denser, as we observe that the original sparse annotations

cause trouble for those methods. Specifically, we crop the scans to include only brain regions. In addition to having foreground classes of 7 subcortical regions (which account for approximately 2% of overall voxels), we assign another foreground class to the remaining brain regions. Therefore, there are 9 classes for DAFormer, HRDA, and MIC, 8 foreground and 1 background classes. This modification significantly improves the results. For the FL baselines FAT [20] and DualAdapt [25], since there is no public official implementation available, we implement both methods following the description in the original papers and finetune thoroughly. We use the same network backbone initialized with the same pretrained encoder and training configuration (FL rounds, global learning rate schedule, local update steps, batch size, etc.) as Fed-MAPSeg whenever possible.

1.4. Dataset Description

We include a diverse collection of 2,421 brain MRI scans from several international projects, each with its unique focus on infant brain development. From the Developing Human Connectome Project (dHCP) V1.0.2 data release[1] [8] in the UK, we incorporate 983 scans (426 T1-weighted, T1w), acquired shortly after birth. The Baby Connectome Project (BCP) [12] in the USA contributes 892 scans (519 T1w), featuring longitudinal data. Additionally, from the Environmental Influences on Child Health Outcomes (ECHO) project, also in the USA, we have 433 scans (218 T1w) from newborn infants. The 'Maternal Adversity, Inflammation, and Neurodevelopment' (Healthy Minds) project from Brazil, conducted at Hospital São Paulo - Federal University of São Paulo (UNIFESP), adds 103 T2-weighted (T2w) MRI scans, acquired shortly after birth and available in the National Institute of Mental Health Data Archive (collection ID 3811). Lastly, the Melbourne Children's Regional Infant Brain (M-CRIB) project [1] from Australia provides 10 additional T2w scans. All studies involved have received Institutional Review Board (IRB) approvals. MAPSeg takes normalized scans as inputs. During training, the intensity of each volumetric scan is clipped at a percentile randomly drawn from a uniform distribution $\mathcal{U}(99, 100)$, then normalized to 0-1. During inference, the intensity clip is fixed at 99.5%. The top 0.5% intensity is clipped as 1 to cope with outlier pixels (hyperintensities) that are usual in MRI.

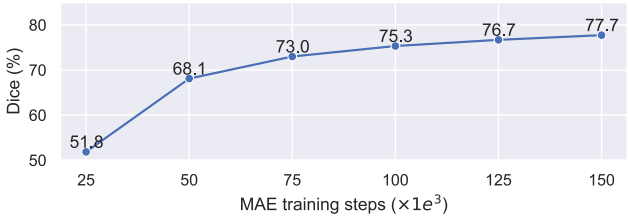
1.5. Results of MRI → CT cardiac segmentation

The performance of MAPSeg on the public cardiac MRI→CT segmentation is reported in Suppl.Tab.4. Similarly, we use the same dataset partition as previous studies. MAPSeg consistently outperforms other baseline methods, although the performance gap is smaller than CT→MRI.

[1] https://www.developingconnectome.org/data-release/data-release-user-guide/

Suppl.Tab. 4. Results of cardiact MRI→CT segmentation.

Cardiac CT → MRI segmentation					
Method	Dice(%) ↑				
	AA	LAC	LVC	MYO	Avg
PnP-AdaNet[7]	74.0	68.9	61.9	50.8	63.9
SIFA-V1[2]	81.1	76.4	75.7	58.7	73.0
SIFA-V2[3]	81.3	79.5	73.8	61.6	74.1
DAFormer[13]	85.5	88.2	74.5	60.2	77.1
MPSCL[17]	90.3	87.1	86.5	72.5	84.1
MA-UDA[14]	90.8	88.7	77.6	67.4	81.1
SE-ASA[9]	83.8	85.2	82.9	71.7	80.9
FSUDA-V1[15]	86.4	86.9	84.8	81.8	85.0
PUFT[6]	88.1	88.5	87.5	74.1	84.6
SDUDA[5]	87.9	88.1	88.4	78.7	85.8
FSUDA-V2[16]	88.2	**88.9**	85.2	**82.2**	86.1
MAPSeg (Ours)	**93.3**	87.3	**89.1**	78.9	**87.1**



Suppl.Fig. 2. Downstream cross-sequence centralized UDA performance vs. MAE pretraining iterations.

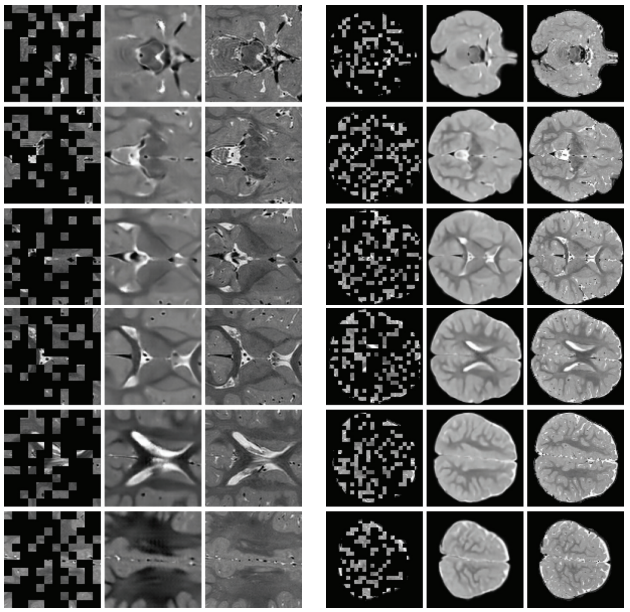
1.6. Additional Analysis

Influence of MAE Pretraining on UDA Results. We conduct an additional analysis to investigate the relationship between MAE training steps and downstream UDA performance. The experiments are conducted on cross-sequence brain MRI segmentation (Suppl.Fig.2). We observe significant improvement in UDA performance at the first 75,000 MAE training steps, which then gradually saturates. We choose 150,000 MAE training steps as the benefits of further training diminish.

Sensitivity to hyperparameters. We conduct additional experiments on cross-sequence brain MRI segmentation to investigate the sensitivity of MAPSeg to hyperparameters (Suppl.Tab.5). Specifically, we investigate the step size (α) of EMA update as well as weights of loss terms (γ and δ). When one parameter is varying, other parameters remain unchanged. We notice that the performance is relatively stable across a wide range of hyperparameters. Since we did not tune the hyperparameters extensively during development, the default parameters may not represent the optimal setting.

Suppl.Tab. 5. Influence of hyperparameters on results, bold indicates used parameters.

α	**0.999/0.9999**	0.99/0.999	0.99	0.999	0.9999
Dice (%)	77.73	74.00	74.26	74.74	78.06
γ	**0.05**	0.5	0.1	0.01	0.005
Dice (%)	77.73	77.22	77.97	77.98	77.99
δ	**0.025**	0.25	0.1	0.01	0.0025
Dice (%)	77.73	76.74	78.08	77.82	78.57



Suppl.Fig. 3. A randomly sampled T2w scan in cross-sequence task. MAE parameters is same as in Suppl.Tab.2

1.7. Visualization

MAE. Some visualizations of MAE results (axial slices) are provided in Suppl.Fig.3.

UDA Results. We provide qualitative comparisons of different methods on cross-sequence (X-Seq), cross-site (X-Site), and cross-age (X-Age) brain MRI segmentation tasks in Suppl.Fig.4. MAPSeg consistently provides accurate segmentation in different UDA settings. It is worth noting that, despite the second best performance in cross-sequence, DAR-UNet tends to oversegment on cross-site and cross-age tasks, partially because of translation errors. On cross-site and cross-age tasks, despite DAFormer, HRDA, and MIC generate reasonably good segmentation inside the subcortical regions, they exhibit extensive false positives outside the subcortical regions, leading to suboptimal overall Dice score.

References

[1] Bonnie Alexander, Andrea L Murray, Wai Yen Loh, Lillian G Matthews, Chris Adamson, Richard Beare, Jian Chen, Claire E Kelly, Sandra Rees, Simon K Warfield, et al. A new neonatal cortical and subcortical brain atlas: the melbourne children's regional infant brain (m-crib) atlas. *Neuroimage*, 147:841–851, 2017. 3

[2] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng-Ann Heng. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):865–872, 2019. 3

[3] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng Ann Heng. Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. *IEEE Transactions on Medical Imaging*, 39(7):2494–2505, 2020. 3

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1

[5] Zhiming Cui, Changjian Li, Zhixu Du, Nenglun Chen, Guodong Wei, Runnan Chen, Lei Yang, Dinggang Shen, and Wenping Wang. Structure-driven unsupervised domain adaptation for cross-modality cardiac segmentation. *IEEE Transactions on Medical Imaging*, 40(12):3604–3616, 2021. 3

[6] Shunjie Dong, Zixuan Pan, Yu Fu, Dongwei Xu, Kuangyu Shi, Qianqian Yang, Yiyu Shi, and Cheng Zhuo. Partial unbalanced feature transport for cross-modality cardiac image segmentation. *IEEE Transactions on Medical Imaging*, 42 (6):1758–1773, 2023. 3

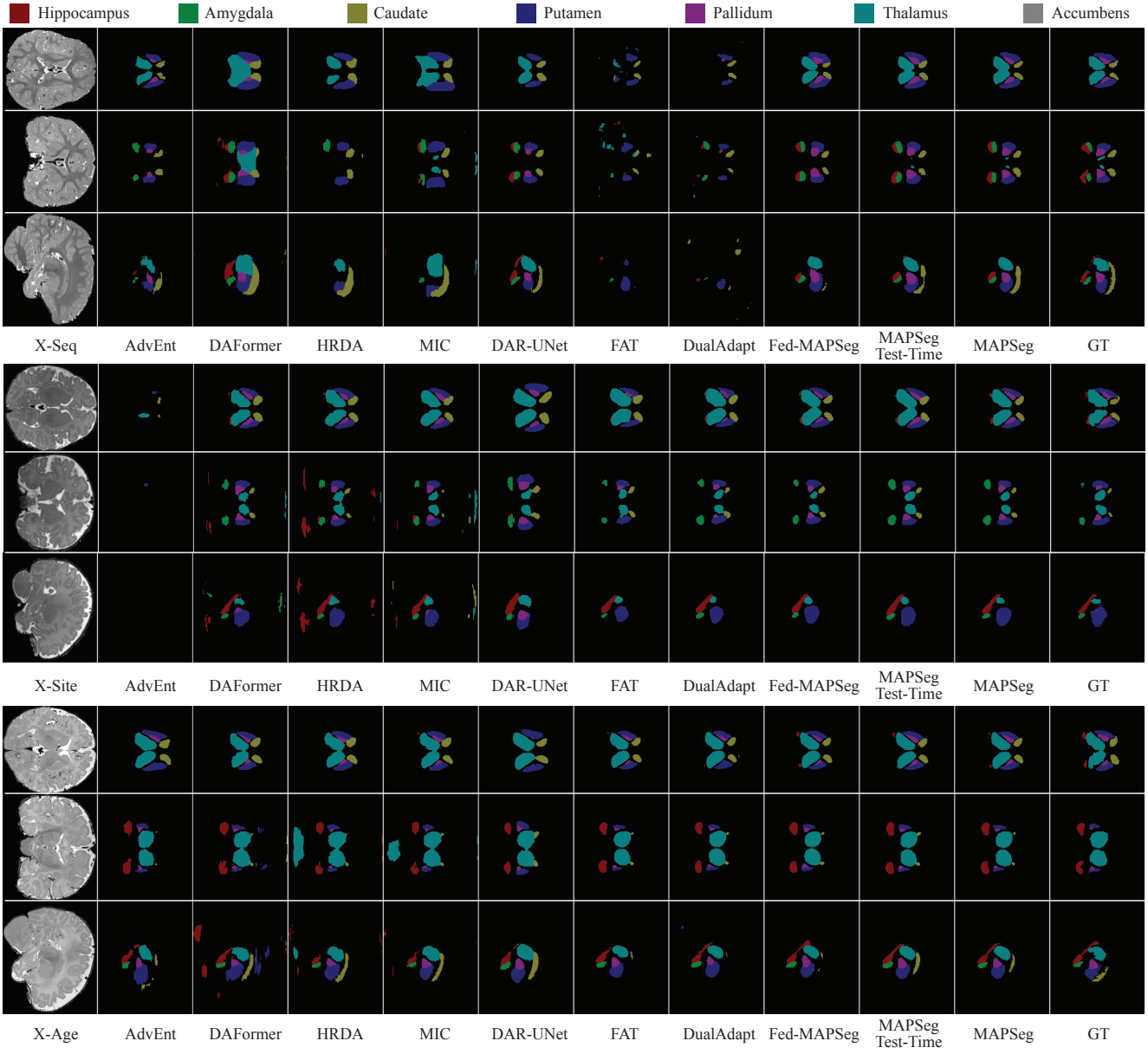
[7] Qi Dou, Cheng Ouyang, Cheng Chen, Hao Chen, Ben Glocker, Xiahai Zhuang, and Pheng-Ann Heng. Pnp-adanet: Plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation. *IEEE Access*, 7:99065–99076, 2019. 3

[8] A David Edwards, Daniel Rueckert, Stephen M Smith, Samy Abo Seada, Amir Alansary, Jennifer Almalbis, Joanna Allsop, Jesper Andersson, Tomoki Arichi, Sophie Arulkumaran, et al. The developing human connectome project neonatal data release. *Frontiers in neuroscience*, 16, 2022. 3

[9] Wei Feng, Lie Ju, Lin Wang, Kaimin Song, Xin Zhao, and Zongyuan Ge. Unsupervised domain adaptation for medical image segmentation by selective entropy constraints and adaptive semantic alignment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1):623–631, 2023. 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[11] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 1



Suppl.Fig. 4. Qualitative comparisons. Three rows (top to bottom) of each task represent axial plane, coronal plane, and sagittal plane, respectively.

[12] Brittany R Howell, Martin A Styner, Wei Gao, Pew-Thian Yap, Li Wang, Kristine Baluyot, Essa Yacoub, Geng Chen, Taylor Potts, Andrew Salzwedel, et al. The unc/umn baby connectome project (bcp): An overview of the study design and protocol development. *NeuroImage*, 185:891–905, 2019. 3

[13] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9924–9935, 2022. 3

[14] Wen Ji and Albert C. S. Chung. Unsupervised domain adaptation for medical image segmentation using transformer with meta attention. *IEEE Transactions on Medical Imaging*, pages 1–1, 2023. 3

[15] Shaolei Liu, Siqi Yin, Linhao Qu, and Manning Wang. Reducing domain gap in frequency and spatial domain for cross-modality domain adaptation on medical image segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):1719–1727, 2023. 3

[16] Shaolei Liu, Siqi Yin, Linhao Qu, Manning Wang, and Zhijian Song. A structure-aware framework of unsupervised cross-modality domain adaptation via frequency and spatial knowledge distillation. *IEEE Transactions on Medical Imag-*

ing, pages 1–1, 2023. 3

[17] Zhizhe Liu, Zhenfeng Zhu, Shuai Zheng, Yang Liu, Jiayu Zhou, and Yao Zhao. Margin preserving self-paced contrastive learning towards domain adaptation for medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 26(2):638–647, 2022. 3

[18] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. 2

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 2

[20] Erum Mushtaq, Yavuz Faruk Bakman, Jie Ding, and Salman Avestimehr. Federated alternate training (fat): Leveraging unannotated data silos in federated segmentation for medical imaging. In *20th IEEE International Symposium on Biomedical Imaging, ISBI 2023, Cartagena, Colombia, April 18-21, 2023*, pages 1–5. IEEE, 2023. 3

[21] Fernando Pérez-García, Rachel Sparks, and Sébastien Ourselin. Torchio: A python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Computer Methods and Programs in Biomedicine*, 208:106236, 2021. 1

[22] Carole H. Sudre, M. Jorge Cardoso, and Sebastien Ourselin. Longitudinal segmentation of age-related white matter hyperintensities. *Medical Image Analysis*, 38:50–64, 2017. 2

[23] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens. Automated model-based tissue classification of mr images of the brain. *IEEE Transactions on Medical Imaging*, 18(10): 897–908, 1999. 2

[24] Adrian Wolny, Lorenzo Cerrone, Athul Vijayan, Rachele Tofanelli, Amaya Vilches Barro, Marion Louveaux, Christian Wenzl, Sören Strauss, David Wilson-Sánchez, Rena Lymbouridou, et al. Accurate and versatile 3d segmentation of plant tissues at cellular resolution. *Elife*, 9:e57613, 2020. 1

[25] Chun-Han Yao, Boqing Gong, Hang Qi, Yin Cui, Yukun Zhu, and Ming-Hsuan Yang. Federated multi-target domain adaptation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1081–1090, 2022. 3