GS-EMA: INTEGRATING GRADIENT SURGERY EXPONENTIAL MOVING AVERAGE WITH BOUNDARY-AWARE CONTRASTIVE LEARNING FOR ENHANCED DOMAIN GENERALIZATION IN ANEURYSM SEGMENTATION

*Fengming Lin[1] *, Yan Xia[1] *, Michael MacRaild[1], Yash Deo[1], Haoran Dou[1], Qiongyao Liu[1], Nina Cheng[1], Nishant Ravikumar[1] †, Alejandro F. Frangi[1 2] †*

[1] University of Leeds, [2] University of Manchester https://github.com/fmlinks/domain

ABSTRACT

The automated segmentation of cerebral aneurysms is pivotal for accurate diagnosis and treatment planning. Confronted with significant domain shifts and class imbalance in 3D Rotational Angiography (3DRA) data from various medical institutions, the task becomes challenging. These shifts include differences in image appearance, intensity distribution, resolution, and aneurysm size, all of which complicate the segmentation process. To tackle these issues, we propose a novel domain generalization strategy that employs gradient surgery exponential moving average (GS-EMA) optimization technique coupled with boundary-aware contrastive learning (BACL). Our approach is distinct in its ability to adapt to new, unseen domains by learning domain-invariant features, thereby improving the robustness and accuracy of aneurysm segmentation across diverse clinical datasets. The results demonstrate that our proposed approach can extract more domain-invariant features, minimizing over-segmentation and capturing more complete aneurysm structures.

***Index Terms*—** Domain Generalization, Gradient Surgery, Contrastive Learning

1. INTRODUCTION

The accurate segmentation of cerebral aneurysms is vital for diagnosing and treating patients effectively. This process is not just about detecting aneurysms early; it involves precise measurements of their size and shape, which are critical for formulating treatment plans [1, 2]. However, the variability in imaging data quality (see Fig.1) from different medical centers presents a significant challenge, complicating the segmentation process.

This variability necessitates a domain generalization (DG) approach, where a model trained on data from multiple sources can adapt to new, unseen domains. The diversity of multi-source data makes DG a daunting challenge in medical imaging, pushing the need for models that generalize well across different medical centers and data types.

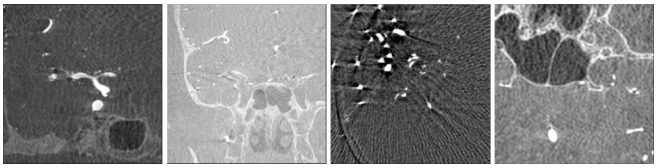

Fig. 1. Illustration of the variability in imaging data quality from different medical centers.

Unlike traditional DG approaches such as domain alignment [3], data augmentation [4], ensemble learning [5], self-supervised learning [6], disentangled representation learning [7], and others, our method takes a different approach. We enhance domain generalization by leveraging gradient surgery exponential moving average (GS-EMA), offering an innovative solution to address DG challenges.

In deep learning, EMA is a frequently used technique for parameter averaging in models, aimed at enhancing the generalization performance and stability of the model. In a teacher-student [8] network setup, the teacher network undergoes a process of parameter smoothing, driven by the student network. However, initially, there are no specific conditions set for this transfer. Consequently, all parameters learned by the student network, whether they are domain-invariant or domain-specific, are updated into the teacher network at some rate. This approach poses a challenge as it fails to distinguish between domain-invariant and domain-specific parameters. To address this issue, we introduce the concept of gradient surgery.

Deep neural networks are trained using gradient descent, where gradients guide the optimization process across the landscape defined by the loss function and training data. The gradient surgery framework [9, 10] aims to resolve conflicts arising in multi-task learning. The conflicting gradients are typically averaged to obtain a final gradient for parameter updates. GSMorph [11] propose alternative methods like

* Contribute equally to this work

† Joint last authors

† AFF is supported by the Royal Academy of Engineering INSILEX Chair (CiET1919/19), UKRI Frontier Research Guarantee INSILICO (EP/Y030494/1), and EC Sixth Framework Programme @neurIST (FP6-2004-IST-4-027703).

normal vector projection to derive the ultimate gradient for parameter updates. Instead of devising a new projection method as suggested by others, we approach the problem by analyzing the relationships between gradients to determine whether EMA parameter updates should occur.

Additionally, there is a class imbalance problem in 3D data segmentation due to the small proportion of aneurysms. After multiple downsampling steps, these small features are more likely to be overlooked in the latent space. To tackle this, we introduce the concept of boundary-awareness to traditional contrastive learning [12].

Contributions: Our study introduces innovative techniques that enhance model adaptability. We integrate gradient surgery with EMA updates, strengthening the ability of model to learn domain-invariant features. This novel approach promises to elevate the performance of DG tasks in medical imaging, ensuring that our model can generalize effectively to new datasets and medical centers. Additionally, we pioneer the use of boundary-aware contrastive learning, enabling our model to discern small target features especially for cerebral aneurysms.

2. METHODS

Fig. 2 depicts our neural network architecture dedicated for domain generalization tasks. It initiates with 3D source images, which undergoes image transformation to produce target images. Once both the source and target images are obtained, they are separately fed into the encoders of the student and teacher networks.

After acquiring the latent space features, a boundary-aware contrastive learning loss is computed. The central notion here is to amalgamate the same instance subjected to diverse transformations, while distancing different instances, aiming to grasp instance-aware representations. This contrastive learning differs from transformation predictions, as it strives to attain transformation-invariant representations. The latent space features are then decoded to yield predictions, which are supervised using ground truth.

Within the student network, the green arrow signifies fully supervised learning for the source images, and the yellow arrow represents the same for the target images. By analyzing the gradient relationship between these two losses, a novel GS-EMA strategy is devised to update the parameters of teacher network. If the gradient angle between the losses is less than 90 degrees, it indicates that the network has learned domain-invariant features, prompting an EMA update. Conversely, if the gradient angle exceeds 90 degrees, no EMA update is performed, as this suggests the network has grasped domain-specific features, which is not conducive to domain generalization tasks. Ultimately, after several updates, a teacher network enriched with more domain-invariant features is achieved, readying it for domain generalization tasks.

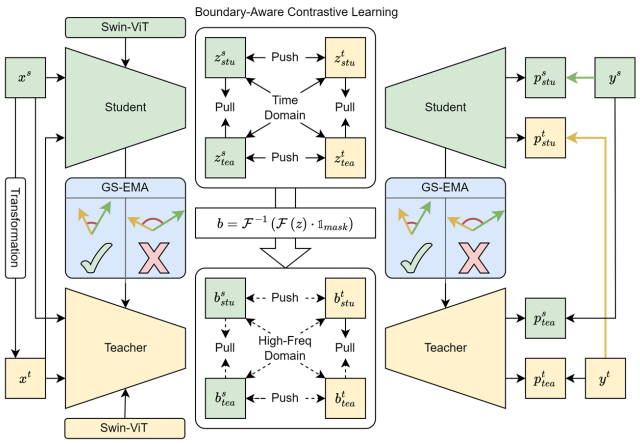

Fig. 2. Schematic of the proposed model.

2.1. Problem Definition and Data Transformation

Let $\mathcal{X}$ be the input (image) space and $\mathcal{Y}$ be the segmentation (label) space, a domain is defined as a joint distribution P_{XY} on $\mathcal{X} \times \mathcal{Y}$. In the context of DG, we have access to K similar but distinct source domains $\left\{\left(x_s^k, y_s^k\right)\right\}_{k=1}^K$, each associated with a joint distribution P_{XY}^k. Note that $P_{XY}^i \neq P_{XY}^j$ with $i \neq j$ and $i, j \in \{1, ..., K\}$. The goal of DG is to learn a predictive model using only source domain data such that the prediction error on an unseen target domain is minimized.

To enhance the model adaptability to previously unseen data domains, we use data transformations to simulate the distribution of the target domain data. The simulated target data is represented as $\{(x_k^t, y_k^t)\}_{k=1}^K$. The process of data transformation encompasses several key steps, including geometric transformations, intensity alterations, noise injection and smoothing, histogram shifting, as well as bias field correction. These operations collectively aim to generate diverse target data, empowering the model with enhanced generalization capabilities to adapt to various data sources and target domains.

2.2. Gradient Surgery Exponential Moving Average

In a teacher-student network setup, when the student network is tasked with learning from data originating from different domains, we calculate distinct losses for each domain in Eq. 2. This allows us to obtain gradient information specific to each domain. Our fundamental hypothesis is that when the angle between gradients from different domains is less than 90 degrees, it suggests that the student network has effectively learned how to extract domain-invariant features. In such cases, we employ EMA to update the parameters of student network, subsequently transferring these parameters to the teacher network. This transfer is performed to better capture universal features.

However, when the angle between gradients from differ-

ent domains exceeds 90 degrees, it indicates that the student network is primarily focused on learning domain-specific features. In such scenarios, we abstain from utilizing EMA for parameter updates and refrain from transmitting these parameters to the teacher network. This strategic approach ensures that the student network can efficiently discriminate between features originating from different domains, enabling it to adapt effectively to the challenges of multi-task learning.

$$\mathcal{L}_{DCE}(p, y) = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{2\,|p \cap y|}{|p| + |y|} - y \log p\right) \quad (1)$$

$$\mathcal{L}_{stu}^{src} = \mathcal{L}_{DCE}(p_{stu}^{s}, y^{s}),\ \mathcal{L}_{stu}^{trg} = \mathcal{L}_{DCE}(p_{stu}^{t}, y^{t}) \quad (2)$$

Algorithm 1: Gradient Surgery Exponential Moving Average

Data: Student network parameters θ_{stu}; Teacher network parameters θ_{tea}; Loss on source data in student network $\mathcal{L}_{src}$; Loss on target data in student network $\mathcal{L}_{trg}$; EMA decay coefficient α.

Result: Decide whether updated θ_{tea} with EMA from θ_{stu}.

for *each mini-batch* **do**
- $\nabla\mathcal{L}_{stu}^{src} \rightarrow g_{src}$;
- $\nabla\mathcal{L}_{stu}^{trg} \rightarrow g_{trg}$;
- **if** $\langle g_{src}, g_{trg}\rangle \leq 0$ **then**
  - $\theta'_{tea} = \theta_{tea} \cdot \alpha + (1-\alpha)\cdot\theta_{stu}$;
- **else if** $\langle g_{src}, g_{trg}\rangle > 0$ **then**
  - $\theta'_{tea} = \theta_{tea}$;
- Update θ_{tea} with θ'_{tea};
- Update θ_{stu} as needed;

2.3. Boundary-Aware Contrastive Learning

In our study, we tackle the challenge of uneven distribution of classes in the segmentation of aneurysms by proposing a unique contrastive learning approach that operates within a teacher-student network configuration. This method enhances the distinction between matching (positive) and non-matching (negative) sample pairs by employing a Fourier transformation strategy, which is particularly adept at isolating high-frequency elements that delineate boundaries. Transitioning from volume-based to boundary-based analysis ensures that the presence of small aneurysms is not disproportionately low compared to larger vessels.

Both the student and teacher branches receive two distinct sets of data: the original data from the source domain, represented as x^s, and the corresponding transformed data x^t. Consequently, the latent feature representations from the student network are symbolized as z_{stu}^{s} and z_{stu}^{t}, while those from the teacher network are signified as z_{tea}^{s} and z_{tea}^{t}.

Advancing further, we harness the power of Fourier transformation paired with a high-frequency filter in Eq. 3 to extract features that are cognizant of the boundaries within the data. These extracted features from both student and teacher networks are represented as b_{stu}^{s}, b_{stu}^{t}, b_{tea}^{s}, and b_{tea}^{t} respectively. Our primary objective within this feature space is to cultivate instance-specific representations that are closely aligned when the same instance is encoded differently, while simultaneously ensuring a clear demarcation between distinct instances, irrespective of the encoder used.

To clarify the relationships within our contrastive learning framework, we delineate the instances processed through different encoders as positive pairs when they originate from the same instance. This includes pairs like z_{stu}^{s} with z_{tea}^{s}, and z_{stu}^{t} with z_{tea}^{t}. In contrast, negative pairs consist of different instances that have been encoded either by the same or by different encoders, such as z_{stu}^{s} with z_{stu}^{t}, and z_{tea}^{s} with z_{tea}^{t}, as well as cross-encoder pairs like z_{stu}^{s} with z_{tea}^{t}, and z_{tea}^{s} with z_{stu}^{t}. These delineations form the basis of our contrastive learning process.

Moving forward, we apply a Fourier transformation to the volume features to construct an amplitude map, which is crucial for identifying the salient high-frequency components that highlight boundaries in Eq.3. A specialized square mask is then utilized to isolate this high-frequency information in Eq.3. Here, the Fourier transform and its inverse are denoted by $\mathcal{F}$ and $\mathcal{F}^{-1}$ respectively. The mask $\mathbb{1}_{mask}$, with value zero in its center and one at the periphery, has the same shape as z.

For boundary features, positive pairs are formed by analogous instances across the student and teacher networks, such as b_{stu}^{s} with b_{tea}^{s}, and b_{stu}^{t} with $b_{tea}^{s\rightarrow t}$. Conversely, negative pairs are created by combining features from distinct instances, which may be within the same network or across both, exemplified by pairs such as b_{stu}^{s} with b_{stu}^{t}, and b_{tea}^{s} with b_{tea}^{t}, as well as inter-network pairs like b_{stu}^{s} with b_{tea}^{t}, and b_{tea}^{s} with b_{stu}^{t}.

$$b = \mathcal{F}^{-1}\left(\mathcal{F}\left(z\right)\cdot\mathbb{1}_{mask}\right) \quad (3)$$

$$h(u, v) = \frac{u^T v}{\|u\|_2\|v\|_2} \quad (4)$$

$$\mathcal{L}_c = -\log\frac{\sum_{i=1}^{N_p} e^{\left(h\left(u_i^+, v_i^+\right)\right)}}{\sum_{i=1}^{N_p} e^{\left(h\left(u_i^+, v_i^+\right)\right)} + \sum_{j=1}^{N_n} e^{\left(h\left(u_j^+, v_j^-\right)\right)}} \quad (5)$$

To quantify the similarity of these pairs, we compute the cosine similarity for each within both the time and frequency domains in Eq.4 and Eq.5. The similarity for positive pairs is expressed as $h(u_i^+, v_i^+)$ where i spans all positive pair indices, and the similarity for negative pairs is articulated as $h(u_j^+, v_j^-)$ where j represents the indices of all negative pairs. Here, N_p stands for the count of positive pairings, and N_n corresponds to the count of negative pairings.

The contrastive learning loss for these high-frequency boundary pairs is then calculated using the same equation as

for the volumetric pairs. By summing up the volumetric contrastive learning loss with the boundary contrastive learning loss, we derive a comprehensive boundary-aware contrastive learning loss. This loss function is designed to finely tune our model to discriminate between the nuanced features of aneurysms, enhancing its segmentation performance.

2.4. Overall framework and training objective

The loss function consists of two parts. $\mathcal{L}_{DCE}$ fully supervises the four outputs of the teacher and student networks. BACL includes volume contrast $\mathcal{L}_c^z$ and boundary contrast $\mathcal{L}_c^b$. The ratio of λ_1 to λ_2 is set at 0.25:0.5.

$$\mathcal{L} = \lambda_1 \cdot (\mathcal{L}_{stu}^{src} + \mathcal{L}_{stu}^{trg} + \mathcal{L}_{tea}^{src} + \mathcal{L}_{tea}^{trg}) + \lambda_2 \cdot (\mathcal{L}_c^z + \mathcal{L}_c^b) \quad (6)$$

3. EXPERIMENTS

3.1. Experimental Setting

Dataset: We tested our method with 3DRA images from 223 patients from the @neurIST dataset [13]. These images were collected from four distinct medical institutions, each employing varied scanning equipment and imaging protocols. Consequently, this dataset exhibits a broad various in both visual characteristics and resolution. The data diversity can evaluate the robustness and adaptability of our proposed GS-EMA method.

Implementation details: Our study was conducted on a NVIDIA RTX 3090 GPU. We utilized the Swin-UNet [14] architecture for both the student and teacher networks in our framework. The training was set to 100 epochs. To determine whether to apply EMA updates, we experimented with setting the EMA coefficient α to either 0.9999 or 0.9. We started with an initial learning rate of 0.001 and adjusted it downwards by multiplying by 0.1 after every ten epochs. The code will be publicly available soon.

3.2. Quantitative Results

	DSC (%) ↑	Sen (%) ↑	Jac (%) ↑	VS (%) ↑
nnUNet [15]	59.61	57.51	47.38	70.91
VASeg [16]	60.28	54.47	49.82	67.91
FedDG [4]	64.50	64.31	54.26	74.73
CMDG [17]	65.01	64.10	54.11	73.38
Ours	**71.89**	**70.88**	**62.36**	**80.00**
no EMA	61.52	55.64	50.91	69.01
EMA	64.71	62.86	54.03	72.64
GS-EMA	**68.49**	**72.79**	**58.40**	**76.63**
BACL-V	68.49	72.79	58.40	76.63
BACL-B	70.62	**75.14**	60.54	78.22
BACL	**71.89**	70.88	**62.36**	**80.00**

Table 1. Quantitative results including compare with SOTAs and ablation studies. Critical metrics includes the Dice similarity coefficient (DSC), Sensitivity (Sen), Jaccard index (Jac) and Volume similarity (VS).

Table.1 includes comparison with state-of-the-art (SOTA) methods and two ablation studies. Our model outperforms traditional segmentation approaches like nnUNet [15] and aneurysm-focused VASeg [16], as well as domain-generalizing methods for medical image segmentation, including CMDG [17] and FedDG [4]. The ablation study highlights that our GS-EMA algorithm, regulating EMA updates with gradient relation, surpasses both regular and non-EMA methods in segmenting aneurysms. It also indicates superior results for BACL when integrating volume (BACL-V) and boundary (BACL-B) learning, compared to using either alone.

3.3. Visual Inspection

Fig. 3 offers a visual comparison of aneurysm segmentation between our method and the SOTAs. It is evident from the comparison that our approach is less prone to over-segmentation while also being able to segment aneurysms more completely.

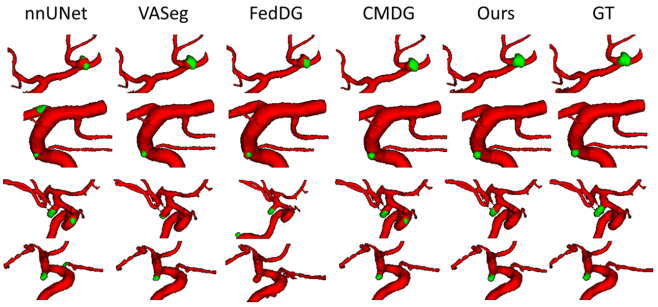

Fig. 3. Comparative visualization of SOTAs and ours method on aneurysm segmentation.

Fig. 4 displays a t-SNE comparison of latent features using EMA and GS-EMA. The larger overlap achieved by GS-EMA indicates a stronger capability of the model to extract domain-invariant features.

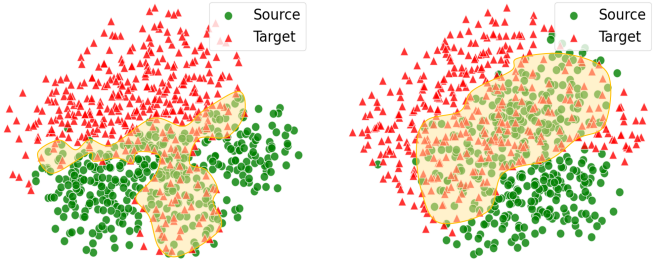

Fig. 4. The t-SNE visualization of latent features from EMA (left) and GS-EMA (right).

4. CONCLUSION

In summary, our study introduces an effective GS-EMA algorithm and a boundary-aware contrastive learning technique

for aneurysm segmentation. These methods outperform existing approaches by minimizing over-segmentations and capturing more complete aneurysm structures. For future work, we plan to apply our GS-EMA technique to a wider array of medical imaging datasets for further validation and enhancement.

5. REFERENCES

[1] Qiongyao Liu, Ali Sarrami-Foroushani, Yongxing Wang, Michael MacRaild, Christopher Kelly, Fengming Lin, Yan Xia, Shuang Song, Nishant Ravikumar, Tufail Patankar, Z Taylor, Toni Lassila, and Alejandro F. Frang, "Hemodynamics of thrombus formation in intracranial aneurysms: an in-silico observational study," *APL Bioengineering*, 2023.

[2] Francesco Pappalardo, Giulia Russo, Flora Musuamba Tshinanu, and Marco Viceconti, "In silico clinical trials: concepts and early adoptions," *Briefings in bioinformatics*, vol. 20, no. 5, pp. 1699–1708, 2019.

[3] Wang Lu, Jindong Wang, Haoliang Li, Yiqiang Chen, and Xing Xie, "Domain-invariant feature exploration for domain generalization," *Transactions on Machine Learning Research*, 2022.

[4] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng, "Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1013–1023.

[5] Shujun Wang, Lequan Yu, Kang Li, Xin Yang, Chi-Wing Fu, and Pheng-Ann Heng, "Dofe: Domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 4237–4248, 2020.

[6] Rayan Krishnan, Pranav Rajpurkar, and Eric J Topol, "Self-supervised learning in medicine and healthcare," *Nature Biomedical Engineering*, vol. 6, no. 12, pp. 1346–1352, 2022.

[7] Hanlin Zhang, Yi-Fan Zhang, Weiyang Liu, Adrian Weller, Bernhard Schölkopf, and Eric P Xing, "Towards principled disentanglement for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8024–8034.

[8] Fengming Lin, Yan Xia, Nishant Ravikumar, Qiongyao Liu, Michael MacRaild, and Alejandro F Frangi, "Adaptive semi-supervised segmentation of brain vessels with ambiguous labels," *arXiv preprint arXiv:2308.03613*, 2023.

[9] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn, "Gradient surgery for multi-task learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5824–5836, 2020.

[10] Lucas Mansilla, Rodrigo Echeveste, Diego H Milone, and Enzo Ferrante, "Domain generalization via gradient surgery," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6630–6638.

[11] Haoran Dou, Ning Bi, Luyi Han, Yuhao Huang, Ritse Mann, Xin Yang, Dong Ni, Nishant Ravikumar, Alejandro F Frangi, and Yunzhi Huang, "Gsmorph: Gradient surgery for cine-mri cardiac deformable registration," in *MICCAI 2023*, Cham, 2023, pp. 613–622, Springer Nature Switzerland.

[12] Chen Yang, Xiaoqing Guo, Zhen Chen, and Yixuan Yuan, "Source free domain adaptation for medical image segmentation with fourier style mining," *Medical Image Analysis*, vol. 79, pp. 102457, 2022.

[13] Siegfried Benkner, Antonio Arbona, Guntram Berti, Alessandro Chiarini, Robert Dunlop, Gerhard Engelbrecht, Alejandro F Frangi, Christoph M Friedrich, Susanne Hanser, Peer Hasselmeyer, et al., "@ neurist: infrastructure for advanced disease management through integration of heterogeneous data, computing, and complex processing services," 2010, vol. 14, pp. 1365–1377, IEEE.

[14] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*. Springer, 2022, pp. 205–218.

[15] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[16] Fengming Lin, Yan Xia, Shuang Song, Nishant Ravikumar, and Alejandro F Frangi, "High-throughput 3dra segmentation of brain vasculature and aneurysms using deep learning," *Computer Methods and Programs in Biomedicine*, vol. 230, pp. 107355, 2023.

[17] Cheng Ouyang, Chen Chen, Surui Li, Zeju Li, Chen Qin, Wenjia Bai, and Daniel Rueckert, "Causality-inspired single-source domain generalization for medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 4, pp. 1095–1106, 2022.