Safe Generalization: Mitigating Catastrophic Forgetting in Single-Source Multi-Organ Segmentation via Collaborative Causal Learning

Yucheng Song[1], Ruoxi Yu[1], Feng Shu[2], Cheng Huang[3], Haokang Ding[1], and Zhifang Liao[1] *

[1] School of Computer Science and Engineering, Central South University, Changsha, China
[2] Department of Clinical Laboratory, Hunan Key Laboratory of Oncotarget Gene, Hunan Cancer Hospital and The Affiliated Cancer Hospital of Xiangya School of Medicine, Central South University, Changsha, China
[3] Artificial Intelligence Research Institute, iFLYTEK Co., Ltd., Hefei, Anhui, China

Abstract. Single-Source Domain Generalization (SDG) in medical multi-organ image segmentation confronts severe challenges of distribution shifts caused by varying imaging protocols. However, we identify a critical yet overlooked phenomenon in current methodologies: Catastrophic Forgetting in SDG (CF-SDG). Specifically, existing generalization strategies often lack explicit protection for source domain features, leading to performance degradation on the source domain while improving generalization capabilities. To address this, we propose the Collaborative Causal Learning Network (CCL-Net), a unified framework that simultaneously achieves robust generalization and source knowledge preservation. From a causal perspective, we construct a dual-path causal intervention mechanism with source domain knowledge constraints. First, we design an Orthogonal Structure Disentanglement (OSD) module to learn structural mediator representations, establishing a pure structural pathway by blocking the intrusion of non-causal information via orthogonalization strategies. Second, we introduce an Anatomy-Guided Local Causal Intervention (AG-LCI) module to physically sever the spurious correlations between anatomical semantics and local appearances through mask-guided counterfactual generation. Furthermore, to mitigate CF-SDG, a novel Collaborative Learning Constraint mechanism is designed to anchor source domain memory via dynamic local-global consistency regularization. Extensive experiments on multi-modality and multi-organ datasets demonstrate that CCL-Net not only achieves state-of-the-art generalization performance on unseen domains but also maintains a significantly lower forgetting rate on the source domain.

Keywords: Single-Source Domain Generalization · Medical Image Segmentation · Causal Intervention · Catastrophic Forgetting

* Corresponding author. E-mail: zfliao@csu.edu.cn

1 Introduction

In recent years, deep learning approaches have made remarkable strides in medical image segmentation [21]. However, the performance of these models often suffers from severe degradation when deployed on datasets acquired across disparate clinical sites or with varying scanning protocols [20]. To address this challenge, Domain Generalization (DG) techniques have been extensively investigated [15,29]. Nevertheless, in real-world clinical scenarios, constrained by data privacy regulations and high annotation costs, labeled data is often accessible from only a single source during the training phase. This renders Single-Source Domain Generalization (SDG) a more challenging yet practically relevant research problem [1].

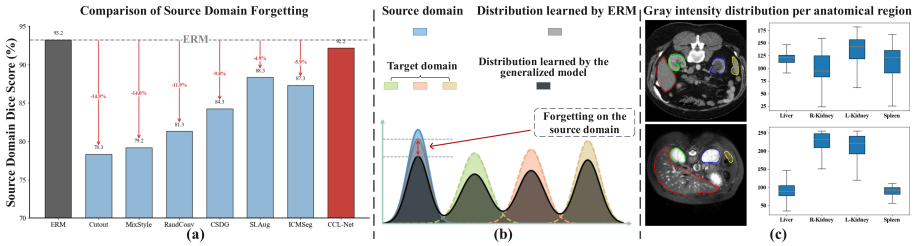

Fig. 1: Conceptual overview of our research motivation. (a) When existing models are trained to pursue cross-domain stability, their segmentation accuracy on the source domain is significantly lower than that of the baseline model (ERM) trained solely on source domain data. (b) The domain-generalized model can fit the distribution of other data well, but its ability to fit the distribution of source domain data declines. (c) Different anatomical structures in different imaging modalities exhibit distinct tissue contrasts.

While existing SDG approaches have mitigated distribution shifts to a certain degree by employing strategies such as style transfer [9, 28], feature alignment [2, 10], or cross-domain augmentation [13, 22], they predominantly oversimplify the SDG task as merely mapping source features into the target domain space. Consequently, they often overlook domain-invariant knowledge that is intrinsically linked to the underlying anatomical structures. Crucial semantic information is frequently attenuated or even obliterated during aggressive augmentation processes, precipitating a decline in performance on the source domain, as illustrated in Fig. 1(a). We term this phenomenon "**Catastrophic Forgetting in Single-Source Domain Generalization (CF-SDG).**" This indicates that in their attempt to eliminate domain shifts, these models inadvertently compromise their representational capability regarding the original distribution, as evidenced in Fig. 1(b). Such performance degradation not only undermines the reliability of the model in its original deployment scenarios but also poses potential risks to clinical applicability.

Based on the above observations, we posit that the fundamental cause of CF-SDG lies in the model's failure to capture genuine causal features, specifically, anatomical structures. Unlike strategies that oversimplify domain shifts as mere global style discrepancies, we focus on a more prevalent phenomenon in multi-organ segmentation: "Confounding Coupling", as illustrated in Fig. 1(c). Due to variations in organ density distribution and inter-tissue contrast during imaging, the underlying anatomical semantics (D) simultaneously dictate both structural morphology (S) and tissue appearance (A). This induces a statistical dependency $S \leftarrow D \rightarrow A$ within the source domain. Consequently, during single-source training, segmentation networks are prone to exploiting the appearance shortcut provided by A to predict the label Y. However, this dependency is inherently unstable and leads to performance degradation when imaging parameters shift across different clinical sites.

Based on the above motivations, this paper proposes a novel Collaborative Causal Learning Network (CCL-Net). Specifically, CCL-Net performs collaborative learning through two related tasks: source domain image reconstruction and generalized segmentation. On this basis, we design a dual-path causal intervention mechanism with source domain knowledge constraints. Inspired by the front-door adjustment concept, we design an Orthogonal Structural Decoupling (OSD) strategy to construct a pure structural mediator pathway and weaken style interference by learning mediating structural representations. Inspired by the back-door adjustment concept, we design an Anatomy-Guided Local Causal Intervention (AG-LCI) module to sever the spurious correlation paths established between structure and appearance by implicit anatomical attributes through mask-guided local style alteration. Meanwhile, we design a collaborative learning constraint mechanism containing local and global bias constraints; the local bias constrains the semantic consistency of different feature layers, while the global bias coordinates the optimization objectives between tasks, thereby maintaining the memory of source domain knowledge while generalizing. The main contributions of this paper are as follows:

- We investigate a challenging problem in the field of domain generalization: Catastrophic Forgetting in Single-Source Domain Generalization Segmentation (CF-SDG).
- We propose the Collaborative Causal Learning Network (CCL-Net). Through dual-path causal intervention, we eliminate spurious correlation paths caused by domain confounders and learn domain-invariant features from stable causal mechanisms. Simultaneously, a collaborative learning constraint mechanism is introduced to effectively maintain the model's memory of source domain knowledge while enhancing cross-domain generalization capability.
- Extensive experiments demonstrate that CCL-Net not only significantly improves the model's generalization performance on unseen domains but also effectively mitigates the catastrophic forgetting phenomenon, maintaining a lower forgetting rate on the source domain.

2 Related Work

2.1 Causal Domain Generalization in Medical Imaging

Traditional DG methods predominantly rely on statistical correlations for feature alignment or data augmentation [6, 23, 26]. However, statistical association is not equivalent to causation. Such assumptions render models prone to learning non-robust spurious correlations, leading to failure on unseen domains. Recently, Causal Inference has been introduced into medical image analysis, aiming to uncover the stable mechanisms underlying the data generation process via Structural Causal Models (SCMs) [19].

Existing work on medical causal generalization primarily focuses on eliminating acquisition-level confounders. For instance, some studies explicitly model the causal graph structure between domain variables and task features to identify and discard domain-specific or confounding features [14, 16]. Other approaches employ causal intervention based on contrastive learning or disentangled representation. These methods apply interventions in the latent space to compel the model to learn domain-independent causal-invariant representations [18,24]. Furthermore, guided by the principle of "intervention-augmentation equivariance," some studies construct virtual samples to simulate the impact of domain shifts on task outcomes, thereby enhancing generalization to unseen domains [4]. Unlike existing works, this paper proposes a dual-path causal intervention mechanism. It not only addresses front-door confounding arising from the imaging environment but also explicitly models and severs back-door spurious correlation paths induced by anatomical semantics.

2.2 Catastrophic Forgetting in Continual Learning

Catastrophic Forgetting (CF) is a classic challenge in the field of Continual Learning (CL), referring to the abrupt loss of knowledge from previous tasks due to drastic parameter updates when a neural network learns new tasks [25]. To address this issue, researchers have proposed regularization-based methods [5], replay-based methods [31], and dynamic architecture methods [12].

Although CF has been widely studied in continual learning, this problem has long been overlooked in Single-Source Domain Generalization. Traditional SDG research generally adopts style perturbation or feature transformation strategies [30,33]. We argue that such transformations in the feature space, pursued for the sake of generalization, intrinsically simulate the "task switching" process in continual learning but lead to Generalization-induced Catastrophic Forgetting. Specifically, this manifests as the destruction of originally clear and fine-grained discriminative features in the source domain data during the forced alignment with unknown domain distributions. In medical applications, source domain data is typically the only available high-quality annotated data (Ground Truth), and the degradation of source domain performance directly implies a loss of clinical reliability. Therefore, this paper introduces "anti-forgetting" into the SDG task, proposing a collaborative learning strategy that explicitly preserves source domain memory through local and global constraints.

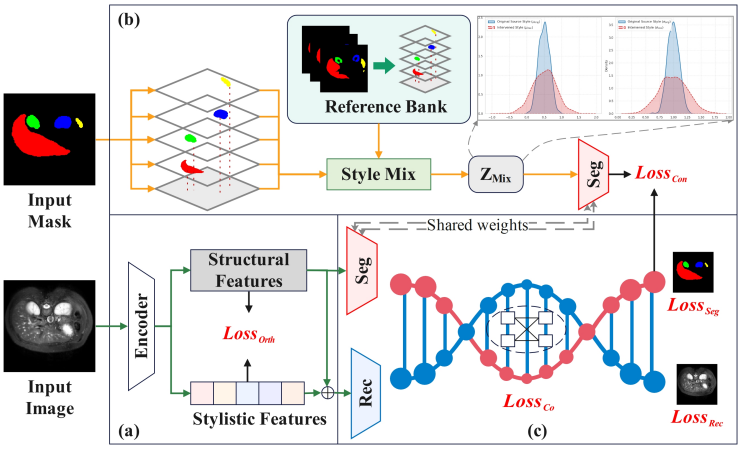

Fig. 2: Overall Overview of Our Method. (a) Orthogonal Structure Disentanglement. (b) Anatomy-Guided Local Causal Intervention. (c) Collaborative Learning Constraint Mechanism.

3 Method

3.1 Problem Definition

In the task of SDG, we are given a source domain dataset $D_S = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in X$ denotes a medical image and $y_i \in Y$ represents the corresponding pixel-level anatomical segmentation label (ground-truth). Our objective is to learn a causally stable segmentation model $f_\theta : X \to Y$ that is trained exclusively on D_S yet generalizes to unseen target domains D_T. We assume that the label Y is determined by the latent anatomical structure S, while the image X is jointly generated by the structure S, and domain-dependent appearance factors A, formulated as $Y \leftarrow S$ and $X \leftarrow g(S, A)$. The goal of this work is to learn a segmentation mechanism that remains invariant to interventions on A, ensuring that predictions are insensitive to appearance variations $do(A)$. Fig. 2 provides an overview of our proposed framework.

3.2 Dual-Path Causal Intervention Mechanism

To achieve generalized segmentation from a single source domain, we characterize the generation process of medical images using a Structural Causal Model (SCM) (see Fig. 3) and propose a dual-path intervention framework. Orthogonal Structure Disentanglement (OSD): Inspired by the concept of front-door adjustment, this module explicitly learns a mediating structural representation, aiming to extract a structural representation $s = \mathrm{Enc}_c(x)$ that is maximally domain-invariant. Anatomy-Guided Local Causal Intervention (AG-LCI): This module constructs counterfactual samples (or features) under appearance intervention to approximate the $do(A)$ operation. It enforces the segmentation head to rely

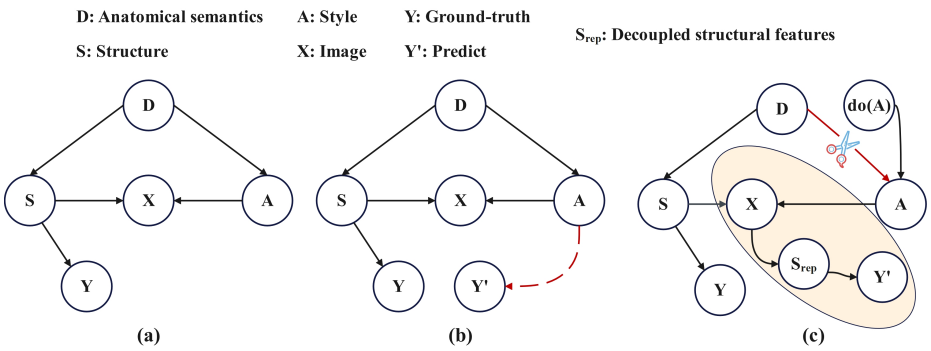

Fig. 3: Structural Causal Model (SCM) of our single-source domain generalization task. (a) The true causal generation process where domain confounder D dictates both structure S and appearance A. (b) In conventional ERM, the model tends to exploit the non-causal appearance shortcut ($A \to Y'$) due to the confounding backdoor path $S \leftarrow D \to A$. (c) Our CCL-Net utilizes dual-path intervention: AG-LCI performs $do(A)$ to physically sever the backdoor spurious correlation ($D \not\to A$), while OSD acts as a front-door adjustment to block A's intrusion by establishing a pure structural mediator pathway from X to Y'.

on S rather than A via consistency constraints. The final optimization objective prioritizes the supervised segmentation loss, combined with an intervention consistency regularization term to enforce invariance to $do(A)$:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}_S}\left[\mathcal{L}_{seg}\left(\mathcal{F}_{seg}\left(\mathrm{Enc}_c(x)\right), y\right)\right] + \lambda\, \mathbb{E}_{x\sim\mathcal{D}_S}\left[\mathcal{L}_{con}\left(\mathcal{F}_{seg}(F_i), \mathcal{F}_{seg}(Z_{mix})\right)\right],$$

where Z_{mix} denotes the counterfactual samples generated within the latent space.

Orthogonal Structure Disentanglement Inspired by the front-door adjustment principle, we explicitly learn a mediator-like structural representation. We employ a dual-branch encoder to explicitly disentangle structure from style. Specifically, the encoder comprises a shared shallow feature extractor E_{shared} and two parallel branches: a structure encoder E_s and a style encoder E_a. For an input image x:

$$s = E_s(E_{shared}(x)), \quad a = E_a(E_{shared}(x)) \tag{1}$$

Where $s \in \mathbb{R}^{C\times H\times W}$ is designed to encode spatial anatomical structures, while $a \in \mathbb{R}^L$ captures global styles. Following the classic paradigm of representation disentanglement and style-content decomposition [3,11], we encourage the decorrelation of structure and style representations in the projection space, serving as a tractable approximation for statistical independence. Specifically, we

first perform global pooling on the structural features and align their dimensions using projection heads:

$$\tilde{s} = p_s(\text{GAP}(s)) \in \mathbb{R}^d, \quad \tilde{a} = p_a(a) \in \mathbb{R}^d \tag{2}$$

Where GAP$(\cdot)$ denotes Global Average Pooling, and p_s, p_a are two-layer Multi-Layer Perceptron (MLP) projection heads. We define the soft orthogonality loss as:

$$\mathcal{L}_{orth} = \frac{1}{B} \sum_{i=1}^{B} \left(\frac{\tilde{s}_i^\top \tilde{a}_i}{\|\tilde{s}_i\|_2 \|\tilde{a}_i\|_2} \right)^2 \tag{3}$$

Minimizing $\mathcal{L}_{orth}$ promotes the decorrelation of structure and style representations in the projection space, thereby encouraging the structure branch to extract anatomical information that is more domain-invariant.

Anatomy-Guided Local Causal Intervention We weaken the shortcut connections caused by the correlation between structure and style through feature-level intervention on A and consistency learning. To sever this path and approximate the interventional distribution $P(Y|\text{do}(A))$, we leverage the batch-wise sample distribution to approximate the marginal distribution of style $P(A)$:

$$P(Y|\text{do}(A)) = \int P(Y|S, A = a)P(a)da \approx \mathbb{E}_{x_j \sim \mathcal{B}}[P(Y|S_i, A_j)] \tag{4}$$

Where x_j is a reference sample randomly sampled from the current batch $\mathcal{B}$. To address the heterogeneity of multi-organ appearances, we concretize the aforementioned intervention as anatomy-guided feature statistics transfer. For the source feature F_i and the reference feature F_j, we first extract local style statistics. To avoid spatial misalignment, we strictly utilize the mask M_j of the reference sample itself to compute the mean $\mu_k^{(j)}$ and standard deviation $\sigma_k^{(j)}$ for the k-th organ class. Based on this, we generate counterfactual samples Z_{mix} in the feature space:

$$Z_{mix} = \sum_{k=1}^{K} \mathbf{1}_{M_i=k} \odot \left(\sigma_k^{(j)} \frac{F_i - \mu_k^{(i)}}{\sigma_k^{(i)}} + \mu_k^{(j)} \right) + (1 - M_{fg}) \odot F_i \tag{5}$$

The above equation performs the intervention operation $\text{do}(A = A_j)$ via feature affine transformation, where $\mathbf{1}_{M_i=k}$ ensures the spatial accuracy of style injection. $M_{fg} = \sum_{k=1}^{K} M_k$ denotes the union of all organ masks (i.e., the foreground region). Finally, we introduce a consistency loss $\mathcal{L}_{con}$, enforcing the model to maintain consistent predictions for both the original feature F_i and the intervened feature Z_{mix}:

$$\mathcal{L}_{con} = \|\mathcal{F}_{seg}(F_i) - \mathcal{F}_{seg}(Z_{mix})\|_2^2 \tag{6}$$

By minimizing $\mathcal{L}_{con}$, we compel the model to output consistent predictions when the structure remains invariant while the style is intervened upon, thereby reducing dependency on non-causal stylistic factors.

3.3 Collaborative Learning Constraint Mechanism

To effectively mitigate CF-SDG, we introduce an auxiliary source domain reconstruction branch S_{rec} into CCL-Net to be collaboratively trained with the main segmentation branch S_{seg}.

Local Bias Constraint During feature encoding, network layers at different depths exhibit varying sensitivities to domain shifts and catastrophic forgetting. Shallow layers often carry low-level style information prone to disruption by stylization operations, while deep layers encode semantic information. To adaptively determine whether each layer necessitates "replaying" source domain memory from the reconstruction branch, we design a Local Bias Constraint Strategy. This strategy does not rely on static inter-layer connections; instead, it determines interaction behavior through the dynamic evolution of features between training steps t and $t+k$. Specifically, for the l-th layer, we compute the consistency between the segmentation features $F_{seg}^{(t,l)}$ and the reconstruction features $F_{rec}^{(t,l)}$. We utilize the Euclidean distance $d(\cdot,\cdot)$ to assess the optimization tendency of the current layer. Ultimately, based on the computed state, we generate a local interaction factor $f_{layer}^{(l)} \in \{0,1\}$, which dynamically enables or disables the feature interaction channel for that layer (see Algorithm 1 for the algorithmic procedure).

Algorithm 1: Local Bias Constraint Strategy.

Input: training step t; intermediate feature maps $F_{\text{seg}}^{(t,l)}$ and $F_{\text{rec}}^{(t,l)}$; feature maps after k update steps $F_{\text{seg}}^{(t+k,l)}$ and $F_{\text{rec}}^{(t+k,l)}$

$\text{Flag}_{\text{seg}} \leftarrow d\big(F_{\text{seg}}^{(t,l)}, F_{\text{seg}}^{(t+k,l)}\big) > d\big(F_{\text{seg}}^{(t,l)}, F_{\text{rec}}^{(t+k,l)}\big)$

$\text{Flag}_{\text{rec}} \leftarrow d\big(F_{\text{rec}}^{(t,l)}, F_{\text{rec}}^{(t+k,l)}\big) > d\big(F_{\text{rec}}^{(t,l)}, F_{\text{seg}}^{(t+k,l)}\big)$

▷ True → collaborate; False → independent

With probability δ: Flag $\leftarrow \neg$Flag

Compute state: State $\leftarrow \text{Flag}_{\text{seg}} \vee \text{Flag}_{\text{rec}}$

if *State* **then**

  $f_{\text{layer}}^{(l)} \leftarrow 1.0$

else

  $f_{\text{layer}}^{(l)} \leftarrow 0.0$

Return $f_{\textbf{layer}}^{(l)}$

To prevent the model from getting trapped in local optima or remaining in a single state for extended periods, we introduce a Stochastic Perturbation mechanism, which forcibly reverses the current state with a probability δ.

Global Bias Constraint If the auxiliary reconstruction task (memory) excessively dominates the direction of gradient updates, it may hinder the model's capacity to learn causal invariance (generalization). Consequently, it is necessary to dynamically balance the weights of both tasks at a global level. The Global Bias Constraint modulates task weights by analyzing the consistency of optimization directions between the two branches. We posit that effective optimization should drive the model to converge towards a more compact solution space. Accordingly, we define an "Effective Optimization Gain" based on distance reduction: When the update trend of the segmentation branch aligns with the objective of the reconstruction branch (i.e., favoring memory retention), we assign a positive gain to encourage the consolidation of source domain knowledge. Conversely, when the segmentation branch diverges from the reconstruction branch (i.e., focusing on generalization exploration), we assign a negative gain to prevent the memory task from interfering with the learning of causal features. Based on this mechanism, we derive the global interaction factor $f_{task}^{(l)}$ (see Algorithm 2 for the detailed calculation process).

Algorithm 2: Global Bias Constraint Strategy.

Input: training step t; feature maps at intermediate layer l, $F_{\text{seg}}^{(t,l)}$ and $F_{\text{rec}}^{(t,l)}$; feature maps after k update steps, $F_{\text{seg}}^{(t+k,l)}$ and $F_{\text{rec}}^{(t+k,l)}$

Compute dir_{seg}:

$$dir_{\text{seg}} \leftarrow \begin{cases} -d\big(F_{\text{seg}}^{(t,l)}, F_{\text{rec}}^{(t+k,l)}\big), & \text{if } Flag_{\text{seg}} \\ d\big(F_{\text{seg}}^{(t,l)}, F_{\text{seg}}^{(t+k,l)}\big), & \text{else} \end{cases}$$

Compute dir_{rec}:

$$dir_{\text{rec}} \leftarrow \begin{cases} d\big(F_{\text{rec}}^{(t,l)}, F_{\text{seg}}^{(t+k,l)}\big), & \text{if } Flag_{\text{rec}} \\ -d\big(F_{\text{rec}}^{(t,l)}, F_{\text{rec}}^{(t+k,l)}\big), & \text{else} \end{cases}$$

$f_{\text{task}}^{(l)} \leftarrow \text{Sigmoid}_{0-2}\big(dir_{\text{seg}} + dir_{\text{rec}}\big)$

Return $f_{\textbf{task}}^{(l)}$

3.4 Loss Function

To explicitly impose the bias-guided strategy, we propose a novel loss function, denoted as L_{Co}. This loss function employs the Mean Squared Error (MSE) to quantify the discrepancy between intermediate feature maps, modulated by

the layer-level guidance factor $f_{layer}^{(l)}$ and the task-level guidance factor $f_{task}^{(l)}$, formulated as follows:

$$\mathcal{L}_{Co} = \frac{1}{N}\sum_{l=1}^{N} f_{layer}^{(l)} \cdot f_{task}^{(l)} \cdot \frac{1}{M}\sum_{i=1}^{M}\left(F_{seg,i}^{(t,l)} - F_{rec,i}^{(t,l)}\right)^2 \tag{7}$$

Where N denotes the number of intermediate layers, and M represents the number of pixels in each feature map. $f_{layer}^{(l)}$ and $f_{task}^{(l)}$ serve as the weight adjustment factors for the l-th layer. $F_{seg,i}^{(t,l)}$ and $F_{rec,i}^{(t,l)}$ correspond to the i-th feature value of the l-th layer in S_{seg} and S_{rec} at stage t, respectively.

For the total segmentation loss, we combine the Dice Loss $\mathcal{L}_{\text{Dice}}$ with $\mathcal{L}_{\text{Co}}$ to enhance gradient stability and the model's pixel-level segmentation capability, formulated as follows:

$$\mathrm{L}_{\text{seg}} = \alpha \cdot \mathcal{L}_{\text{Dice}} + \beta \cdot \mathcal{L}_{\text{Co}} \tag{8}$$

Similarly, the total loss for $\mathcal{S}_{\text{rec}}$ integrates the Structural Similarity Loss $\mathcal{L}_{\text{SSIM}}$ with $\mathcal{L}_{\text{Co}}$:

$$\mathcal{L}_{\text{rec}} = \alpha' \cdot \mathcal{L}_{\text{SSIM}} + \beta' \cdot \mathcal{L}_{\text{Co}} \tag{9}$$

Combining the dual-path causal interventions, the final objective function is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{rec}} + \gamma\mathcal{L}_{\text{orth}} + \lambda\mathcal{L}_{\text{con}} \tag{10}$$

Where α, β, α', β', γ, and λ are the fusion weights for the losses.

4 Experiments

4.1 Experimental Setup

Datasets: We evaluate our method on two challenging datasets: the cross-modality abdominal dataset and the cross-sequence cardiac dataset. To ensure fair comparison, the detailed dataset splitting and preprocessing steps follow the settings established by Ouyang et al. [17]. The abdominal T2-SPIR MRI dataset contains 20 3D volumes, while the abdominal CT dataset consists of 30 3D volumes. The cardiac bSSFP and cardiac LGE datasets each comprise 45 3D volumes. For each experimental setting, the model is trained on a single source domain and tested on unseen target domains.

Implementation Details: We employ U-Net as the backbone network for our collaborative learning framework. All experiments are implemented using PyTorch 2.4 on four NVIDIA RTX 4090 GPUs. Regarding model training, we set the total number of epochs to 1000 and the batch size to 16, utilizing the Adam optimizer for optimization. The initial learning rate is set to 2×10^{-3} with a learning rate decay strategy. The stochastic perturbation probability δ

is set to 0.05. The hyperparameters $\alpha, \beta, \alpha', \beta', \gamma$, and λ in the loss function are determined via grid search to be 0.25, 0.25, 0.5, 0.5, 0.6, and 0.4, respectively. All reported results are the mean ± standard deviation obtained from five random initializations to ensure the stability of our method.

Evaluation Metrics: We primarily employ the Dice Similarity Coefficient (Dice) as the evaluation metric. To quantitatively assess the severity of catastrophic forgetting, we draw upon the concept of Forgetting Metric commonly used in Continual Learning (CL) [8]. Unlike standard CL settings where tasks arrive sequentially, we define the Source Forgetting Metric (SFM) as the performance gap on the source domain between the baseline model (Empirical Risk Minimization, ERM) trained solely on the source domain and our proposed generalized model. A lower SFM value indicates superior retention of source domain knowledge. An ideal generalized model should significantly improve performance on the target domain while maintaining an SFM close to zero (where a negative value implies positive transfer).

4.2 Comparison Experiment

We evaluate the performance of CCL-Net on two challenging medical image segmentation datasets and compare it with state-of-the-art Domain Generalization methods, including those based on Causal DG. To ensure fairness, the results for all comparison methods are reproduced using their officially released code. As shown in Table 1 and Table 2, our method consistently achieves optimal performance across all generalization scenarios.

Table 1: Quantitative comparison on the cross-modality abdominal dataset. (The best result in each column is **bolded**, and the second best is underlined.)

Method	Abdominal CT → Abdominal MRI					Abdominal MRI → Abdominal CT				
	Liver	R-Kidney	L-Kidney	Spleen	Average	Liver	R-Kidney	L-Kidney	Spleen	Average
Cutout [7]	79.32±2.87	81.21±2.41	83.03±3.12	75.78±3.95	79.84±3.01	85.89±2.64	64.21±4.32	72.46±3.58	55.98±4.71	69.64±3.82
MixStyle [32]	76.43±2.35	77.84±2.68	77.32±2.14	78.03±2.57	77.41±2.39	85.21±2.47	46.23±4.85	63.42±4.12	54.88±4.39	62.44±4.03
RandConv [27]	72.14±3.64	77.26±3.21	83.47±3.89	80.98±3.42	78.46±3.54	83.89±2.83	75.78±3.12	76.46±2.96	75.44±3.21	77.89±3.02
CSDG [18]	87.33±1.21	84.98±1.34	85.87±1.42	73.45±2.15	82.91±1.63	85.79±1.35	84.35±1.42	87.21±1.27	75.89±2.18	83.31±1.69
SLAug [23]	86.73±1.07	86.69±1.18	87.54±1.12	78.89±1.84	84.96±1.35	89.32±1.12	85.22±1.29	87.34±1.21	86.45±1.33	87.08±1.24
ICMSeg [4]	88.02±0.98	86.91±1.06	**87.34±1.01**	80.23±1.67	85.63±1.24	90.78±1.01	85.67±1.41	89.44±1.13	90.02±0.96	88.98±1.18
Ours	**89.67±0.51**	**87.32±0.56**	86.87±0.53	**81.04±0.69**	**86.23±0.59**	**90.89±0.48**	**86.97±0.63**	**91.03±0.45**	**91.32±0.42**	**90.05±0.54**

In the abdominal cross-modality dataset, the substantial domain shift leads to severe performance degradation in style-based methods (e.g., MixStyle drops to 62.44%), whereas CCL-Net maintains a robust Dice score of 90.05%. Similarly, in the cardiac segmentation task, CCL-Net demonstrates exceptional capability in preserving fine-grained structures. Specifically, for the myocardium (MYO)—a thin structure prone to segmentation fragmentation—our method achieves 81.32% (bSSFP → LGE), surpassing ICMSeg by 1.3% and MixStyle by

Table 2: Quantitative comparison on the cross-sequence cardiac dataset. (The best result in each column is **bolded**, and the second best is underlined.)

Method	Cardiac bSSFP → Cardiac LGE				Cardiac LGE → Cardiac bSSFP			
	LVC	MYO	RVC	Average	LVC	MYO	RVC	Average
Cutout [7]	87.89 ± 2.31	68.33 ± 2.87	78.32 ± 2.54	78.18 ± 2.63	89.88 ± 2.25	78.32 ± 2.51	87.73 ± 2.37	85.31 ± 2.41
MixStyle [32]	85.53 ± 2.12	63.75 ± 2.69	74.36 ± 2.37	74.55 ± 2.41	90.21 ± 2.08	78.45 ± 2.46	87.34 ± 2.28	85.33 ± 2.32
RandConv [27]	88.01 ± 2.48	72.97 ± 2.93	83.54 ± 2.61	81.51 ± 2.72	90.91 ± 2.37	80.78 ± 2.63	88.43 ± 2.44	86.71 ± 2.49
CSDG [18]	89.67 ± 1.24	76.51 ± 1.58	84.78 ± 1.37	83.65 ± 1.43	91.21 ± 1.29	80.01 ± 1.52	89.32 ± 1.34	86.85 ± 1.41
SLAug [23]	88.98 ± 1.11	78.33 ± 1.42	85.32 ± 1.29	84.21 ± 1.32	91.43 ± 1.17	80.59 ± 1.39	89.14 ± 1.26	87.05 ± 1.32
ICMSeg [4]	89.01 ± 1.03	80.02 ± 1.27	85.98 ± 1.21	85.00 ± 1.25	91.72 ± 1.06	80.93 ± 1.31	90.24 ± 1.19	87.63 ± 1.27
Ours	**90.87 ± 0.21**	**81.32 ± 0.24**	**87.03 ± 0.20**	**86.41 ± 0.23**	**92.03 ± 0.19**	**81.79 ± 0.23**	**91.45 ± 0.18**	**88.42 ± 0.22**

a significant margin of 17.5%. This indicates that OSD successfully disentangles shape information from complex modality noise. Crucially, CCL-Net exhibits the lowest standard deviation across all experiments, proving that our dual causal intervention strategy effectively eliminates spurious correlations. Visual comparisons of different methods are presented in Fig. 4.

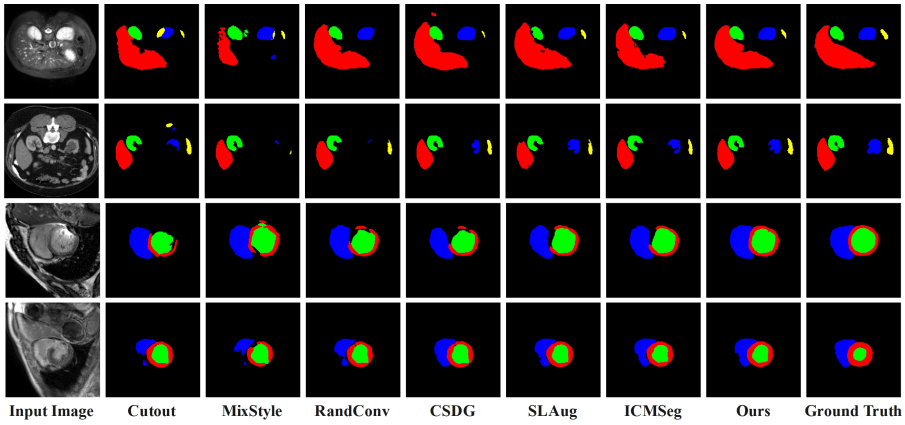

Fig. 4: Visual Comparison of Segmentation Results Using Different Methods

To further validate the capability of CCL-Net on the source domain, we tested the generalization results of comparison methods using source domain data. We simultaneously compared our method with a baseline model (Empirical Risk Minimization, ERM) trained solely on the source domain. As shown in Table 3, standard Domain Generalization methods face severe "Catastrophic Forgetting" issues. For instance, augmentation techniques such as Cutout and MixStyle exhibit source forgetting metrics (SFM) as high as 14.9% and 14.03% on the abdominal dataset. This confirms that aggressive augmentation strategies often destroy the discriminative features required by the source domain. In contrast, CCL-Net effectively mitigates this problem. Thanks to the constraints imposed on feature drift by our Collaborative Learning Constraint Mechanism,

our method maintains performance highly comparable to the ERM baseline. On the abdominal dataset, CCL-Net achieves an SFM as low as 1.07%, reducing the forgetting rate by over 13% compared to Cutout. Similarly, on the cardiac dataset, our lowest SFM is merely 1.26%, whereas the latest SOTA methods still exhibit forgetting rates of approximately 3-10%. These results indicate that CCL-Net does not forget source domain knowledge while extending to unseen domains.

Table 3: Quantitative comparison of Average Dice and Selection Frequency Metric (SFM) across different datasets.

Method	Cardiac bSSFP		Cardiac LGE		Abdominal CT		Abdominal MRI	
	Average	SFM	Average	SFM	Average	SFM	Average	SFM
ERM	92.31	–	89.55	–	92.54	–	93.24	–
Cutout [7]	80.32	11.99	83.82	5.73	82.78	9.76	78.34	14.90
MixStyle [32]	80.78	11.53	82.33	7.22	81.13	11.41	79.21	14.03
RandConv [27]	82.46	9.85	83.83	5.72	82.45	10.09	81.32	11.92
CSDG [18]	86.43	5.88	84.04	5.51	84.23	8.31	84.27	8.97
SLAug [23]	85.74	6.57	86.33	3.22	87.45	5.09	88.34	4.90
ICMSeg [4]	82.32	9.99	85.42	4.13	86.32	6.22	87.32	5.92
Ours	**89.02**	**3.29**	**88.29**	**1.26**	**90.32**	**2.22**	**92.17**	**1.07**

4.3 Ablation Experiment

To verify the effectiveness and complementarity of the core components in CCL-Net, we conducted an ablation study on two datasets. We progressively integrated Orthogonal Structure Disentanglement (OSD), Anatomy-Guided Local Causal Intervention (AG-LCI), and the Collaborative Learning Constraint (CL) into the baseline. Detailed quantitative results are presented in Table 4.

Compared to the baseline model (VA1), introducing either OSD or AG-LCI alone yields improvements in generalization performance, verifying the effectiveness of front-door disentanglement and back-door intervention in eliminating domain-specific noise. Notably, introducing CL alone leads to more significant performance gains. We attribute this to CL acting as a strong regularization term, thereby enhancing feature robustness. Combining different modules (VA5-VA7) results in further performance superposition, indicating complementary effects among components. We observe that the addition of the CL module consistently boosts the performance of different variants (VA4, VA6, VA7). This demonstrates that the collaborative learning constraint effectively promotes the learning of more robust domain-invariant representations while optimizing feature space consistency. Furthermore, combining OSD with CL yields a more pronounced performance leap compared to using OSD alone. We analyze that this is because OSD disentangles structural features from style noise, providing "cleaner" feature representations for the CL module. Ultimately, the integration of all modules (VA8) achieves optimal performance.

Table 4: Ablation study results of our method. The checkmark (✓) indicates the component is included, while the cross (×) indicates it is excluded.

Method	Baseline	OSD	AG-LCI	CL	CT→MRI (Avg. Dice)	MRI→CT (Avg. Dice)	bSSFP→LGE (Avg. Dice)	LGE→bSSFP (Avg. Dice)
VA1	✓	×	×	×	55.98	62.15	53.32	58.67
VA2	✓	✓	×	×	59.33	67.34	59.46	65.41
VA3	✓	×	✓	×	60.76	67.98	63.49	68.46
VA4	✓	×	×	✓	70.59	75.34	72.12	74.93
VA5	✓	✓	✓	×	76.45	78.30	76.45	77.83
VA6	✓	×	✓	✓	79.12	81.05	80.56	81.94
VA7	✓	✓	×	✓	85.34	86.88	85.43	87.35
VA8(Ours)	✓	✓	✓	✓	**86.23**	**90.05**	**86.41**	**88.42**

Table 5: Comparison of Source Forgetting Metric (SFM) with and without our CCL module.

Method	Cardiac bSSFP	Cardiac LGE	Abdominal CT	Abdominal MRI	Avg. SFM
w/o CL	10.45	6.82	8.54	9.68	8.87
Ours (CCL-Net)	**3.29**	**1.26**	**2.22**	**1.07**	**1.96**

The results presented in Table 5 reveal that relying solely on causal intervention leads to significant catastrophic forgetting. This confirms our hypothesis: destylization and counterfactual intervention, while "eliminating domain shifts," sacrifice the representational capacity for the original distribution. However, upon introducing the CL module, the SFM across all datasets is substantially suppressed. This result demonstrates that the CL module successfully anchors source domain knowledge within the feature space through explicit local-global consistency constraints.

5 Conclusion

To address the often-overlooked challenge of Catastrophic Forgetting in single-source domain generalization for multi-organ medical image segmentation, this paper proposes a novel Collaborative Causal Learning Network (CCL-Net). We designed a dual-path causal intervention mechanism with source domain knowledge constraints, which explicitly disentangles structure and style via the front-door path and severs spurious correlations between anatomical semantics and pseudo-styles via the back-door path, thereby extracting robust domain-invariant features. Simultaneously, we introduced a novel collaborative learning constraint mechanism that effectively resolves the CF-SDG problem by regularizing source domain memory through dynamic local-global consistency. Extensive experiments demonstrate that CCL-Net not only achieves SOTA segmentation performance on unseen target domains but also maintains a significantly lower forget-

ting rate on the source domain, providing a safe and robust solution for clinical medical image analysis.

Acknowledgements

This work was supported by the Central South University Research Programme of Advanced Interdisciplinary Studies (No. 2023QYJC041).

References

1. Arslan, M.F., Guo, W., Li, S.: Single-source domain generalization in deep learning segmentation via lipschitz regularization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 666–674. Springer (2024)
2. Atwany, M., Lashgari, M., Choudhury, R.P., Grau, V., Banerjee, A.: Single-source domain generalization for coronary vessels segmentation in x-ray angiography. In: International Workshop on Statistical Atlases and Computational Models of the Heart. pp. 1–11. Springer (2024)
3. Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., Erhan, D.: Domain separation networks. Advances in neural information processing systems **29** (2016)
4. Chen, B., Zhu, Y., Ao, Y., Caprara, S., Sutter, R., Rätsch, G., Konukoglu, E., Susmelj, A.: Generalizable single-source cross-modality medical image segmentation via invariant causal mechanisms. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3592–3602. IEEE (2025)
5. Chen, P.H., Wei, W., Hsieh, C.J., Dai, B.: Overcoming catastrophic forgetting by bayesian generative regularization. In: International conference on machine learning. pp. 1760–1770. PMLR (2021)
6. Cheng, S., Gokhale, T., Yang, Y.: Adversarial bayesian augmentation for single-source domain generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11400–11410 (2023)
7. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
8. Hayes, T.L., Kemker, R., Cahill, N.D., Kanan, C.: New metrics and experimental paradigms for continual learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 2031–2034 (2018)
9. Huang, R., Cai, H., Zhuo, W., Cai, S., Lin, H., Fan, W., Su, W.: Ada: An adaptive augmentation framework for single-source domain generalization in medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 45–54. Springer (2025)
10. Huang, Y., Han, L., Dou, H.: Generative feature style augmentation for domain generalization in medical image segmentation. Pattern Recognition **162**, 111416 (2025)
11. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: Proceedings of the European conference on computer vision (ECCV). pp. 35–51 (2018)
12. Li, X., Zhou, Y., Wu, T., Socher, R., Xiong, C.: Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In: International conference on machine learning. pp. 3925–3934. PMLR (2019)

13. Liu, C., Cao, Y., Zhang, Y., Su, X., Zhu, H.: Perturbating, tuning, and collaborating: Harnessing vision foundation models for single domain generalization on medical imaging. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 5370–5378 (2025)
14. Lv, F., Liang, J., Li, S., Zang, B., Liu, C.H., Wang, Z., Liu, D.: Causality inspired representation learning for domain generalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8046–8056 (2022)
15. Lv, X., Dong, X., Wang, L., Yang, J., Zhao, L., Pu, B., Jin, Z., Li, X.: Test-time domain generalization via universe learning: A multi-graph matching approach for medical image segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15621–15631 (2025)
16. Mahajan, D., Tople, S., Sharma, A.: Domain generalization using causal matching. In: International conference on machine learning. pp. 7313–7324. PMLR (2021)
17. Ouyang, C., Biffi, C., Chen, C., Kart, T., Qiu, H., Rueckert, D.: Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In: European conference on computer vision. pp. 762–780. Springer (2020)
18. Ouyang, C., Chen, C., Li, S., Li, Z., Qin, C., Bai, W., Rueckert, D.: Causality-inspired single-source domain generalization for medical image segmentation. IEEE Transactions on Medical Imaging **42**(4), 1095–1106 (2022)
19. Sheth, P., Moraffah, R., Candan, K.S., Raglin, A., Liu, H.: Domain generalization–a causal perspective. arXiv preprint arXiv:2209.15177 (2022)
20. Song, Y., Li, C., Ding, H., Liao, Z., Liao, Z.: Treefeddg: Alleviating global drift in federated domain generalization for medical image segmentation. arXiv preprint arXiv:2510.18268 (2025)
21. Song, Y., Ren, S., Lu, Y., Fu, X., Wong, K.K.: Deep learning-based automatic segmentation of images in cardiac radiography: a promising challenge. Computer Methods and Programs in Biomedicine **220**, 106821 (2022)
22. Spanos, N., Arsenos, A., Theofilou, P.A., Tzouveli, P., Voulodimos, A., Kollias, S.: Complex style image transformations for domain generalization in medical images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5036–5045 (2024)
23. Su, Z., Yao, K., Yang, X., Huang, K., Wang, Q., Sun, J.: Rethinking data augmentation for single-source domain generalization in medical image segmentation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 2366–2374 (2023)
24. Wang, Y., Liu, F., Chen, Z., Wu, Y.C., Hao, J., Chen, G., Heng, P.A.: Contrastive-ace: Domain generalization through alignment of causal mechanisms. IEEE Transactions on Image Processing **32**, 235–250 (2022)
25. Wang, Z., Yang, E., Shen, L., Huang, H.: A comprehensive survey of forgetting in deep learning beyond continual learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(3), 1464–1483 (2024)
26. Xu, Y., Xie, S., Reynolds, M., Ragoza, M., Gong, M., Batmanghelich, K.: Adversarial consistency for single domain generalization in medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 671–681. Springer (2022)
27. Xu, Z., Liu, D., Yang, J., Raffel, C., Niethammer, M.: Robust and generalizable visual representation learning via random convolutions. arXiv preprint arXiv:2007.13003 (2020)

28. Yi, J., Bi, Q., Zheng, H., Zhan, H., Ji, W., Huang, Y., Li, S., Li, Y., Zheng, Y., Huang, F.: Hallucinated style distillation for single domain generalization in medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 438–448. Springer (2024)
29. Yoon, J.S., Oh, K., Shin, Y., Mazurowski, M.A., Suk, H.I.: Domain generalization for medical image analysis: A review. Proceedings of the IEEE **112**(10), 1583–1609 (2024)
30. Zhang, C., Hu, Z., Dai, S., He, Q., Liu, D., Yan, K., Wang, P.: Boundary-aware contrastive learning for single-source domain generalization in medical image segmentation. In: 2024 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2024)
31. Zhou, F., Cao, C.: Overcoming catastrophic forgetting in graph neural networks with experience replay. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 4714–4722 (2021)
32. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Mixstyle neural networks for domain generalization and adaptation. International Journal of Computer Vision **132**(3), 822–836 (2024)
33. Zhou, Z., Qi, L., Yang, X., Ni, D., Shi, Y.: Generalizable cross-modality medical image segmentation via style augmentation and dual normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20856–20865 (2022)