COUNTERFACTUAL LOGIT DEBIASING FOR UNSUPERVISED CROSS-MODALITY SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised domain adaptation enables cardiac image segmentation across acquisition domains without target-domain annotations. Existing methods have improved cross-modality transfer through image translation, feature alignment, and prediction regularization, but acquisition-related context can still influence the final segmentation decision. We propose counterfactual logit debiasing (CLD), which compares predictions from observed spatial features and a structural reference under the same global domain context. Both predictions use a shared segmentation head, and their logit difference serves as the final segmentation output. CLD is trained with labeled source images and unlabeled target images. On bSSFP-to-LGE and CT-to-MR cardiac segmentation, it achieves mean Dice scores of 0.823 and 0.778, improving upon the strongest compared UDA baseline by 0.7 and 1.7 percentage points, respectively. Component studies further examine the contributions of the model design and prediction contrast.

1 INTRODUCTION

Predictors trained under one acquisition process often fail when that process changes. In image segmentation, appearance can shift substantially across modalities even when the semantic label space remains fixed. Cardiac computed tomography and magnetic resonance offer a concrete example: contrast and intensity differ markedly, but both require spatially coherent anatomical predictions. Encoder–decoder models are effective within a distribution (Ronneberger et al., 2015; Isensee et al., 2021); transferring them without target annotations remains difficult.

Unsupervised domain adaptation (UDA) uses labeled source examples and unlabeled target examples. Domain-adversarial learning aligns features (Ganin et al., 2016), structured-output methods align predictions (Tsai et al., 2018; Vu et al., 2019), and image translation addresses appearance mismatch (Zhu et al., 2017; Chen et al., 2020). Class-aware prototypes and contrastive objectives add semantic constraints (Liu et al., 2022; Lee et al., 2022; Gu et al., 2024). These methods improve cross-domain transfer, but spatial evidence for anatomy and acquisition-related context are learned from the same images and can both affect the final segmentation prediction. We focus on the prediction change associated with an image's spatial features when the global domain context is kept fixed. CLD uses separate branches to represent spatial structural features and global context. A shared segmentation head evaluates the observed structural features and a running-mean structural reference under the same context, and the difference between their logits is used for segmentation. During training, source annotations supervise both factual and difference-based predictions. Unlabeled target images provide entropy and consistency constraints, while domain-adversarial learning regularizes the structural representation.

Our work makes three contributions. First, we introduce a prediction-level comparison that uses a structural reference under matched global context and takes the factual–reference logit difference as the final segmentation output. Second, we develop a dual-branch architecture with a shared segmentation head and training objectives that combine labeled source and unlabeled target images. Third, we evaluate the method on bSSFP-to-LGE and CT-to-MR cardiac segmentation through quantitative comparisons, component and supervision ablations, sensitivity analysis, and qualitative results.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

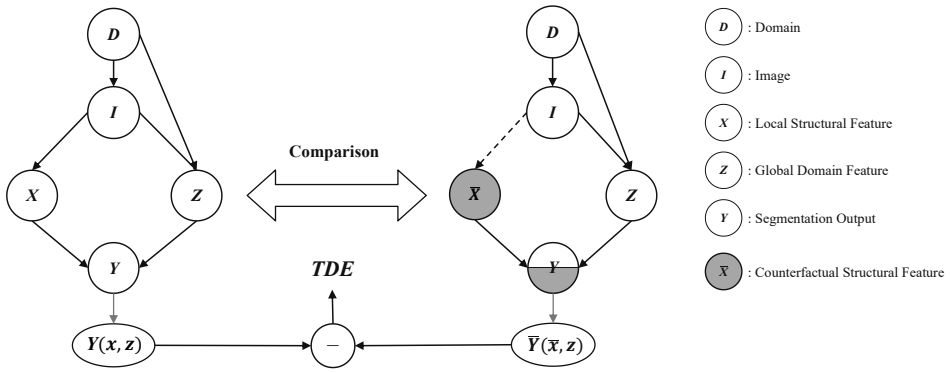


Figure 1: Causal formulation of the factual and counterfactual predictions. D, I, X, Z, and Y denote the domain, input image, local structural representation, global domain representation, and segmentation logits, respectively. The factual output $Y(x, z)$ is obtained from the observed structure and domain context. In the counterfactual view, X is replaced by the mean structural representation $\bar{X}$ while Z remains fixed, producing $\bar{Y}(\bar{x}, z)$. Their difference forms the TDE-inspired structural contribution. The dashed edge denotes the counterfactual replacement rather than an ordinary causal dependency.

2 Related Work

Unsupervised domain adaptation for dense prediction. Unsupervised domain adaptation (UDA) transfers knowledge from a labeled source domain to an unlabeled target domain. Domain-adversarial learning encourages the feature extractor to produce representations that cannot be distinguished by a domain classifier (Ganin et al., 2016). For semantic segmentation, AdaptSegNet performs adversarial alignment in the structured output space, while ADVENT aligns entropy-based prediction representations (Tsai et al., 2018; Vu et al., 2019). Maximum Classifier Discrepancy uses the disagreement between task classifiers to identify target samples outside the support of the source distribution (Saito et al., 2018). Other methods reduce appearance differences directly. CyCADA combines image-level translation with feature-level alignment under semantic consistency constraints (Hoffman et al., 2018), whereas Fourier Domain Adaptation transfers low-frequency spectral information between domains without an additional image-translation network (Yang & Soatto, 2020). These approaches demonstrate that domain discrepancy can be addressed at the image, feature, or output level, but they do not explicitly separate spatial anatomical evidence from global acquisition-related context within the segmentation prediction.

Cross-modality medical image segmentation. Domain shift is particularly pronounced in medical imaging because different scanners, acquisition protocols, and imaging modalities can produce substantially different appearances for the same anatomical structures. Early medical UDA methods used adversarial objectives to learn domain-invariant features for brain-lesion segmentation (Kamnitsas et al., 2017). For cross-modality segmentation, Dou et al. introduced adversarial alignment of convolutional representations (Dou et al., 2018), and PnP-AdaNet aligned multi-level features and segmentation outputs for unpaired CT–MR cardiac segmentation (Dou et al., 2019). These methods establish feature alignment as an effective strategy for transferring anatomical knowledge across medical domains.

A complementary direction translates images between modalities. SynSeg-Net jointly performs unpaired image synthesis and target-modality segmentation without target annotations (Huo et al., 2019), building on cycle-consistent image translation (Zhu et al., 2017). SIFA combines image-space and feature-space adaptation to preserve semantic information during cross-modality cardiac segmentation (Chen et al., 2020). Disentangled representation learning instead separates shared anatomical content from modality-specific appearance (Yang et al., 2019). AdaptEvery further adapts representations at multiple levels of the segmentation pipeline (Vesal et al., 2021). Related cross-modality learning methods also use mutual knowledge distillation to transfer anatomical information between modality-specific segmentors (Li et al., 2020). Although image translation and

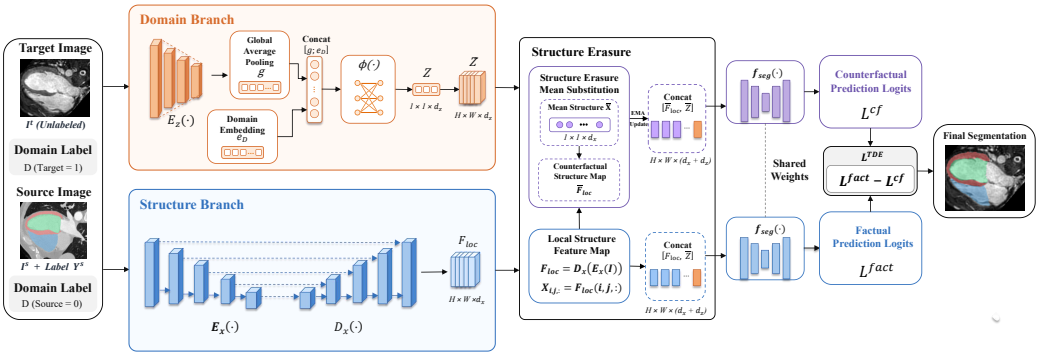


Figure 2: Overall architecture of the proposed method. The structure branch extracts the local feature map F_{loc}, and the domain branch combines globally pooled image features with a domain embedding to obtain the global representation Z. The factual prediction is computed from $[F_{\mathrm{loc}}; \widetilde{Z}]$. For the counterfactual prediction, the local feature map is replaced by $\bar{F}_{\mathrm{loc}}$, obtained by broadcasting the running mean structural vector $\bar{X}$, while the domain representation remains unchanged. The two predictions use the same segmentation head, and their logit difference forms $L^{\mathrm{TDE}} = L^{\mathrm{fact}} - L^{\mathrm{cf}}$.

representation disentanglement can reduce modality discrepancy, their effectiveness depends on preserving task-relevant anatomy while changing modality-dependent appearance.

Semantic and contrastive adaptation. Recent approaches introduce category-aware constraints to avoid aligning features from different anatomical classes. MPSCL combines margin-preserving contrastive learning with self-paced adaptation (Liu et al., 2022), while BCL constructs bidirectional contrastive objectives for domain-adaptive semantic segmentation (Lee et al., 2022). SLCL uses soft labels to provide more informative contrastive supervision for cross-modality cardiac segmentation (Gu et al., 2024). MAPSeg combines three-dimensional masked autoencoding with pseudo-labeling to exploit spatial context in heterogeneous medical images (Zhang et al., 2024). These methods strengthen semantic discrimination during adaptation, whereas our work focuses on separating the contributions of local anatomical structure and global domain context at the prediction level.

Counterfactual and causal representation learning. Structural causal models distinguish interventions from ordinary statistical conditioning and provide a framework for reasoning about counterfactual outcomes (Pearl, 2009). In visual recognition, counterfactual subtraction has been used to suppress prediction components associated with undesirable contextual bias (Tang et al., 2020). Causality-inspired medical segmentation methods similarly seek anatomical representations that remain stable across changes in acquisition conditions (Ouyang et al., 2023). We use a prediction-level comparison in which the observed and mean structural representations are evaluated through the same segmentation head while the domain representation is held fixed. Their logit difference is used directly as the final segmentation prediction.

3 Method

We present counterfactual logit debiasing (CLD), which separately models spatially resolved anatomical structure and global acquisition-dependent context, and constructs the final prediction by contrasting factual and counterfactual logits produced by a shared segmentation head. Figure 2 illustrates the overall architecture, while Figure 3 summarizes the training objectives used for labeled source and unlabeled target images.

3.1 Causal formulation and learning setting

Let $\mathcal{S} = \{(I_s^n, Y_s^n)\}_{n=1}^{N_s}$ denote the labeled source-domain dataset and $\mathcal{T} = \{I_t^m\}_{m=1}^{N_t}$ denote the unlabeled target-domain dataset. Source and target images share the same segmentation label space

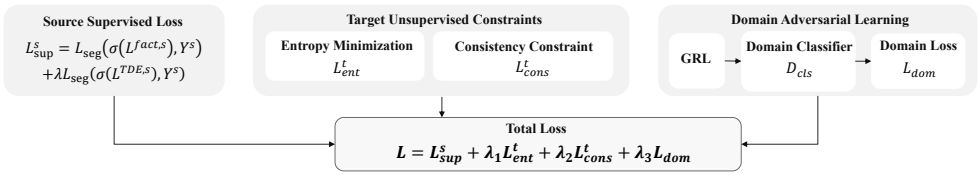


Figure 3: Overview of the training objective. Labeled source images supervise both the factual and TDE predictions. For unlabeled target images, entropy minimization and factual–TDE consistency regularize the segmentation predictions, while domain-adversarial learning aligns the structural representations through a gradient-reversal layer.

but differ in their acquisition-dependent appearance. During training, pixel-level annotations are available only for source-domain images.

We describe the prediction process using five variables. The domain variable D identifies the image-acquisition domain, and I denotes the observed medical image. The local structural representation X describes spatially resolved anatomical information, including organ shape, boundaries, and local texture. The global domain representation Z describes acquisition-dependent context, including intensity distribution, contrast, noise pattern, and imaging style. Finally, Y denotes the pixel-level segmentation logits.

The assumed dependencies are illustrated in Figure 1. The domain affects the observed image through the acquisition process, giving $D \to I$. The image determines the local structural and global domain representations, corresponding to $I \to X$ and $I \to Z$. The domain identity also provides explicit domain information to Z, producing $D \to Z$, while the segmentation logits are jointly determined by the structural and domain representations, giving $(X, Z) \to Y$.

The factual prediction uses the observed structural representation X together with the domain representation Z. To construct its counterfactual counterpart, we replace X with a mean structural representation $\bar{X}$ while holding Z fixed. Comparing the two predictions measures how the output changes when instance-specific structural evidence is replaced under the same domain context. This factual–counterfactual contrast provides the basis for the TDE-inspired prediction introduced below.

3.2 Overall architecture

The proposed architecture consists of a structure branch, a domain branch, and a shared segmentation head. Both labeled source images and unlabeled target images are processed by the same architecture. The structure branch preserves spatially resolved anatomical information, whereas the domain branch models global acquisition-dependent context.

Given an input image $I \in \mathbb{R}^{H\times W}$, the structure branch uses an encoder–decoder network to produce a local feature map F_{loc}. At each spatial position (i, j), the corresponding local structural vector is denoted by $X_{i,j}$. In parallel, the domain branch applies global average pooling to the feature map produced by E_z, concatenates the pooled representation with a learnable embedding e_D of the known domain identity, and maps the result to the global domain representation Z. The representations used by the segmentation head are defined jointly as

$$\begin{aligned} F_{\text{loc}} = D_x(E_x(I)) \in \mathbb{R}^{H\times W\times d_x}, \qquad X_{i,j} = F_{\text{loc}}(i,j,:) \in \mathbb{R}^{d_x}, \\ Z = \phi([\text{GAP}(E_z(I)); e_D]) \in \mathbb{R}^{d_z}, \\ \widetilde{Z}(i,j,:) = Z, \quad \forall (i,j), \qquad \widetilde{Z} \in \mathbb{R}^{H\times W\times d_z}. \end{aligned} \tag{1}$$

Here, E_x and D_x denote the structural encoder and decoder, respectively, E_z denotes the domain encoder, ϕ is a learnable mapping, and $[\cdot\,;\cdot]$ denotes concatenation. Broadcasting Z over the spatial dimensions provides every location with the same global domain context, while F_{loc} retains the spatial anatomical evidence required for pixel-level segmentation.

3.3 FACTUAL, COUNTERFACTUAL, AND TDE PREDICTIONS

The factual prediction retains both the observed local structural representation and the global domain context. To construct its counterfactual counterpart, we replace the observed local structural vectors with a position-independent mean representation. Let $\bar{X} \in \mathbb{R}^{d_x}$ denote the running structural mean. For each minibatch, the mean vector and its spatially broadcast feature map are obtained as

$$\begin{aligned} m = \frac{1}{N_{\mathrm{pix}}} \sum_{n=1}^{N_{\mathrm{pix}}} X_n, \qquad \bar{X} \leftarrow (1-\alpha)\bar{X} + \alpha m, \\ \bar{F}_{\mathrm{loc}}(i,j,:) = \bar{X}, \quad \forall (i,j), \qquad \bar{F}_{\mathrm{loc}} \in \mathbb{R}^{H \times W \times d_x}, \end{aligned} \tag{2}$$

where X_n indexes the local structural vectors participating in the minibatch statistic, N_{pix} is their total number, and α is the moving-average coefficient. Unlike F_{loc}, the resulting map contains the same structural vector at every spatial position and therefore suppresses the instance-specific spatial variation of the observed anatomy.

The factual and counterfactual inputs are processed by the same segmentation head f_{seg}. The domain representation $\widetilde{Z}$ is held fixed, while only the structural input is replaced. The factual, counterfactual, and TDE logits are therefore defined as

$$\begin{aligned} L^{\mathrm{fact}} &= f_{\mathrm{seg}}\left([F_{\mathrm{loc}}; \widetilde{Z}]\right), \\ L^{\mathrm{cf}} &= f_{\mathrm{seg}}\left([\bar{F}_{\mathrm{loc}}; \widetilde{Z}]\right), \\ L^{\mathrm{TDE}} &= L^{\mathrm{fact}} - L^{\mathrm{cf}}. \end{aligned} \tag{3}$$

The corresponding probability maps are $p^{\mathrm{fact}} = \mathrm{softmax}(L^{\mathrm{fact}})$ and $p^{\mathrm{TDE}} = \mathrm{softmax}(L^{\mathrm{TDE}})$. At inference, the final segmentation is obtained from $\widehat{Y}_{i,j} = \arg\max_c p^{\mathrm{TDE}}_{i,j}(c)$.

Because the factual and counterfactual predictions share the same domain representation and segmentation head, their difference is induced by replacing the mean structural reference with the observed structural representation. The logit contrast suppresses responses shared by the two evaluations and emphasizes the prediction change associated with the observed spatial structure. We use this contrast as the structure-oriented segmentation output.

3.4 TRAINING OBJECTIVE

The model is trained jointly using labeled source-domain images and unlabeled target-domain images. The complete objective contains source-domain supervision, target-domain entropy minimization, prediction consistency, and domain-adversarial learning.

For a labeled source image with annotation Y^s, we supervise both the factual prediction and the TDE prediction:

$$\mathcal{L}^s_{\mathrm{sup}} = \mathcal{L}_{\mathrm{seg}}\left(p^{\mathrm{fact},s}, Y^s\right) + \lambda\, \mathcal{L}_{\mathrm{seg}}\left(p^{\mathrm{TDE},s}, Y^s\right), \tag{4}$$

where $\mathcal{L}_{\mathrm{seg}} = \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{Dice}}$ combines the cross-entropy and Dice losses, and λ balances the two supervised predictions.

For an unlabeled target image, entropy minimization encourages confident TDE predictions, while the consistency term encourages the factual and TDE probability maps to remain close. In addition, a domain classifier $\mathcal{D}_{\mathrm{cls}}$ is connected to the structural representation Φ^{TDE} through a gradient-reversal layer. These three adaptation terms are

$$\begin{aligned} \mathcal{L}^t_{\mathrm{ent}} &= -\frac{1}{B_t HW} \sum_{b,i,j} \sum_{c=1}^{C} p^{\mathrm{TDE},t}_{b,i,j,c} \log p^{\mathrm{TDE},t}_{b,i,j,c}, \\ \mathcal{L}^t_{\mathrm{cons}} &= \frac{1}{B_t HW} \sum_{b,i,j} \left\| p^{\mathrm{fact},t}_{b,i,j} - p^{\mathrm{TDE},t}_{b,i,j} \right\|_1, \\ \mathcal{L}_{\mathrm{dom}} &= \mathcal{L}_{\mathrm{CE}}\left(\mathcal{D}_{\mathrm{cls}}\left(\mathrm{GRL}\left(\Phi^{\mathrm{TDE}}\right)\right), d\right). \end{aligned} \tag{5}$$

Here, B_t is the target-domain batch size, C is the number of segmentation classes, and d is the source or target domain label. The entropy and consistency terms regularize the target predictions

without target annotations. The domain classifier learns to distinguish the two domains, whereas gradient reversal encourages the structural representation to become domain invariant.

The complete training objective is

$$\mathcal{L} = \mathcal{L}^s_{\text{sup}} + \lambda_1 \mathcal{L}^t_{\text{ent}} + \lambda_2 \mathcal{L}^t_{\text{cons}} + \lambda_3 \mathcal{L}_{\text{dom}}, \tag{6}$$

where λ_1, λ_2, and λ_3 control the three adaptation terms.

4 Experiments

4.1 Datasets, metrics, and implementation details

We evaluate CLD on MS-CMRSeg and MMWHS, which represent cross-sequence and cross-modality cardiac segmentation, respectively. In both settings, source-domain annotations are available during training, whereas target-domain manual annotations are not used for optimization. The segmentation targets are the myocardium (MYO), left-ventricular cavity (LV), and right-ventricular cavity (RV).

MS-CMRSeg contains 45 paired multi-sequence cardiac MR cases (Zhuang et al., 2022), each including balanced steady-state free precession (bSSFP) and late gadolinium enhancement (LGE) images. We use bSSFP as the source domain and LGE as the target domain. Five cases are used for validation, and the remaining 40 cases are divided equally between training and testing. This task evaluates adaptation across MR sequences with substantial differences in intensity distribution and tissue contrast.

MMWHS contains CT and MR cardiac volumes (Zhuang et al., 2019), including 20 manually annotated CT volumes and 20 manually annotated MR volumes. An additional 32 CT and 25 MR volumes with automatically generated masks are used for training-data augmentation. Following the experimental protocol, CT is used as the source domain and MR as the target domain. Sixteen slices are uniformly sampled from each volume, and the manually annotated MR cases are evaluated using two-fold cross-validation. This setting provides a more pronounced cross-modality domain shift.

Evaluation metrics. We evaluate segmentation performance using the Dice similarity coefficient (DSC) (Dice, 1945; Crum et al., 2006) and the 95th-percentile Hausdorff distance (HD95) (Huttenlocher et al., 1993; Taha & Hanbury, 2015). DSC measures the overlap between the predicted and reference regions, whereas HD95 measures robust boundary error using the 95th percentile of the symmetric surface distance. Higher DSC and lower HD95 indicate better performance. We report results for MYO, LV, and RV together with their class average.

Implementation details. Our experiments were conducted using Python 3.9 and PyTorch 2.5 with CUDA 12.2 on an NVIDIA RTX 4090 GPU. The proposed model follows a 2D U-Net architecture

Table 1: Quantitative comparison on MS-CMRSeg and MMWHS. Higher DSC and lower HD95 indicate better performance. The best results among UDA methods are highlighted in bold. The target-supervised model uses target-domain annotations and is included as a reference.

	bSSFP → LGE		CT → MR	
Method	DSC ↑	HD95(mm) ↓	DSC ↑	HD95(mm) ↓
Source only	0.529	17.689	0.409	31.205
Target supervised	0.851	5.863	0.806	16.861
AdaptSegNet	0.724	9.945	0.657	22.717
ADVENT	0.744	9.581	0.740	14.166
AdaptEvery	0.775	9.002	0.759	13.003
MPSCL	0.761	9.104	0.736	14.307
BCL	0.798	8.226	0.683	25.074
SLCL	0.816	8.067	0.761	12.693
CLD (Ours)	**0.823**	**7.812**	**0.778**	**11.984**

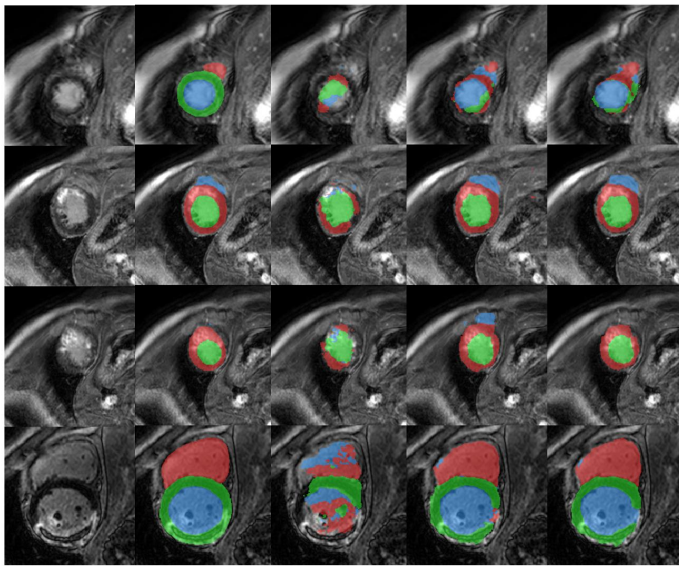


Figure 4: Qualitative segmentation results on target-domain LGE images from MS-CMRSeg (bSSFP-to-LGE). From left to right: input image, ground truth (GT), and predictions from Source Only, SLCL, and CLD (Ours). Green denotes MYO, blue LV, and red RV.

with a pretrained ResNet-50 encoder (He et al., 2016). The encoder–decoder produces spatial features, which are combined with the global domain representation by the shared segmentation head.

The model is optimized using stochastic gradient descent with a batch size of 32. Input preprocessing includes slice extraction, intensity normalization, and spatial resizing. Data augmentation consists of horizontal flipping, vertical flipping, and rotations by multiples of 90°. The loss weights are set to $\lambda_1 = 0.005$, $\lambda_2 = 0.30$, and $\lambda_3 = 0.005$, and the moving-average coefficient is set to $\alpha = 0.10$.

4.2 Comparison with existing methods

We compare CLD with the source-only baseline, the target-supervised reference, and representative UDA methods, including AdaptSegNet (Tsai et al., 2018), ADVENT (Vu et al., 2019), AdaptEvery (Vesal et al., 2021), MPSCL (Liu et al., 2022), BCL (Lee et al., 2022), and SLCL (Gu et al., 2024). The comparison covers both adaptation settings and reports mean DSC and HD95 averaged over MYO, LV, and RV. The target-supervised row uses target-domain annotations and is included only as a reference rather than as a UDA competitor.

On bSSFP-to-LGE adaptation, CLD achieves a mean DSC of 0.823 and an HD95 of 7.812 mm. Compared with SLCL, the strongest UDA baseline in this setting, CLD improves DSC by 0.7 percentage points and reduces HD95 by 0.255 mm. This result is consistent with the incremental benefit of the prediction-level contrast observed in the component ablation.

Figure 4 further shows that CLD produces more complete anatomical regions and more continuous boundaries than the source-only model and SLCL.

The CT-to-MR task presents a larger appearance discrepancy between the source and target domains. In this setting, CLD obtains a mean DSC of 0.778 and an HD95 of 11.984 mm, outperforming SLCL by 1.7 percentage points in DSC and 0.709 mm in HD95. The consistent improvements across the two adaptation settings demonstrate the effectiveness of CLD under both cross-sequence and cross-modality domain shifts.

Figure 5 shows a similar improvement on the larger CT-to-MR appearance shift, particularly in structural completeness and local boundary delineation.

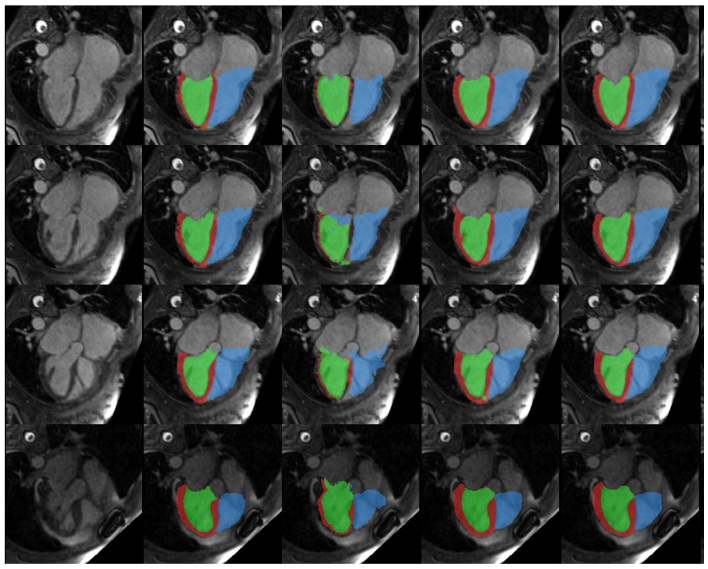


Figure 5: Qualitative segmentation results on target-domain MR images from MMWHS (CT-to-MR). From left to right: input image, ground truth (GT), and predictions from Source Only, SLCL, and CLD (Ours). Red denotes MYO, green LV, and blue RV.

4.3 Ablation studies

We conduct ablation studies on MS-CMRSeg using the same dataset partition and evaluation protocol. The baseline is a single-branch U-Net trained with source supervision. The two-branch variant uses the factual output for inference. The counterfactual variant additionally constructs the reference prediction but retains factual inference. The final variant uses the factual–counterfactual logit difference for inference. We also compare source-domain supervision strategies separately.

Table 2: Component ablation on MS-CMRSeg. Components are added progressively, and the best results are highlighted in bold.

Configuration	MYO	LV	RV	Mean DSC	HD95(mm)
Baseline	0.434	0.614	0.538	0.529	17.69
+ Two branches	0.742	0.879	0.768	0.796	8.28
+ Counterfactual branch	0.758	0.889	0.785	0.811	8.01
+ TDE prediction	**0.771**	**0.896**	**0.801**	**0.823**	**7.81**

Table 2 shows that the dual-branch architecture raises the mean DSC from 0.529 to 0.796 and reduces HD95 from 17.69 to 8.28 mm. With factual inference retained, the counterfactual branch raises mean DSC to 0.811; using the logit difference for TDE prediction yields the best mean DSC of 0.823 and HD95 of 7.81 mm. These improvements support the contributions of both branch separation and prediction-level logit contrast.

Table 3: Ablation of source-domain supervision strategies on MS-CMRSeg.

Supervision strategy	Mean DSC ↑	HD95(mm) ↓
Factual prediction only	0.813	8.28
TDE prediction only	0.817	8.08
Joint factual and TDE supervision	**0.823**	**7.81**

Table 3 further shows that supervising the TDE prediction outperforms factual-only supervision, while joint factual and TDE supervision achieves the best DSC and HD95. The two objectives therefore provide complementary optimization signals.

4.4 Sensitivity analysis

We conduct a one-factor sensitivity analysis on MS-CMRSeg by varying λ_1, λ_2, λ_3, and α individually while keeping the remaining training settings fixed.

Table 4: Hyperparameter sensitivity on MS-CMRSeg. Boldface marks the selected value and its corresponding DSC and HD95 in each parameter sweep.

Parameter	Value	DSC ↑	HD95(mm) ↓	Parameter	Value	DSC ↑	HD95(mm) ↓
λ_1	0	0.786	8.8	λ_2	0	0.798	8.3
	0.001	0.809	8.2		0.1	0.811	8.1
	0.005	**0.823**	**7.8**		**0.3**	**0.823**	**7.8**
	0.010	0.820	7.9		0.5	0.819	8.0
	0.050	0.804	8.2		1.0	0.807	8.2
λ_3	0	0.800	8.4	α	0.01	0.808	8.3
	0.001	0.812	8.2		0.05	0.816	8.1
	0.005	**0.823**	**7.8**		**0.10**	**0.823**	**7.8**
	0.010	0.818	8.1		0.20	0.819	8.1
	0.050	0.800	8.3		0.30	0.812	8.2

Performance peaks at $\lambda_1 = 0.005$, $\lambda_2 = 0.30$, $\lambda_3 = 0.005$, and $\alpha = 0.10$, yielding a mean DSC of 0.823 and an HD95 of 7.8 mm. The results show that moderate regularization and reference updating provide the most effective balance.

5 Discussion

The component ablation helps distinguish the effects of representation design and prediction contrast. Replacing the source-only baseline with two branches gives the largest gain in mean DSC, from 0.529 to 0.796. With factual prediction retained for inference, adding the counterfactual branch raises it to 0.811. Using the factual–counterfactual logit difference as the final prediction raises it further to 0.823. Together, these rows distinguish the gain from the two-branch representation from the additional gains associated with the counterfactual branch and the final logit contrast.

The shared segmentation head evaluates the observed and reference structural features under the same global context. This makes the contrast sensitive to the change in structural evidence within a given input, while the source and target objectives guide the representation during training. The running mean provides a consistent structural reference across images, and its update rate controls how quickly it follows the learned features. At inference, the learned mean is retained as the structural reference for each test image. In the sensitivity study, the best results occur at intermediate loss weights and structural-mean update rate; stronger constraints or faster updates do not improve the measured scores. The two benchmarks cover a shift between MR sequences and a shift between CT and MR. Appendix A provides the construction and training steps behind this comparison.

6 Conclusion

We presented counterfactual logit debiasing for unsupervised domain-adaptive cardiac image segmentation. By contrasting factual and counterfactual logits from a shared segmentation head, CLD emphasizes local structural evidence while accounting for domain-related context. It achieves mean DSC scores of 0.823 and 0.778 on bSSFP-to-LGE and CT-to-MR adaptation, outperforming the strongest compared UDA method by 0.7 and 1.7 percentage points, respectively. The quantitative, ablation, sensitivity, and qualitative results consistently support the proposed design.

AI USE STATEMENT

Generative AI tools, including OpenAI Codex, were used to assist with English-language editing and polishing, including improvements to grammar, clarity, manuscript organization, and LaTeX presentation. All scientific ideas, methodological decisions, experiments, analyses, and conclusions were developed by the authors. The authors reviewed and revised all AI-assisted text and take full responsibility for the final content of the manuscript.

ETHICS STATEMENT

This study uses publicly available cardiac imaging benchmarks and reports results on existing, de-identified data. No new participants were recruited and no clinical intervention was performed. The method and its evaluation are presented for research purposes, not for direct clinical decision-making.

REPRODUCIBILITY STATEMENT

The model architecture, counterfactual construction, and training objectives are described in Section 3. Section 4 reports the datasets, evaluation metrics, implementation settings, comparisons, and ablations. Appendix A gives the structural-reference construction and the source, target, and inference procedures in more detail.

REFERENCES

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. doi: 10.1109/TMI.2020.2972701.

William R. Crum, Oscar Camara, and Derek L. G. Hill. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Transactions on Medical Imaging*, 25(11):1451–1461, 2006. doi: 10.1109/TMI.2006.880587.

Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3): 297–302, 1945. doi: 10.2307/1932409.

Qi Dou, Cheng Ouyang, Cheng Chen, Hao Chen, and Pheng-Ann Heng. Unsupervised cross-modality domain adaptation of ConvNets for biomedical image segmentations with adversarial loss. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 691–697, 2018. doi: 10.24963/ijcai.2018/96.

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. doi: 10.1109/ACCESS.2019.2929258.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Mingxuan Gu, Mareike Thies, Siyuan Mei, Fabian Wagner, Mingcheng Fan, Yipeng Sun, Zhaoya Pan, Sulaiman Vesal, Ronak Kosti, Dennis Possart, Jonas Utz, and Andreas Maier. Unsupervised domain adaptation using soft-labeled contrastive learning with reversed Monte Carlo method for cardiac image segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2024*, volume 15009 of *Lecture Notes in Computer Science*, pp. 681–691, Cham, Switzerland, 2024. Springer Nature Switzerland. doi: 10.1007/978-3-031-72114-4_65.

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*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1989–1998. PMLR, 2018.

Yuankai Huo, Zhoubing Xu, Hyeonsoo Moon, Shunxing Bao, Albert Assad, Tamara K. Moyo, Michael R. Savona, Richard G. Abramson, and Bennett A. Landman. SynSeg-Net: Synthetic segmentation without target modality ground truth. *IEEE Transactions on Medical Imaging*, 38 (4):1016–1025, 2019. doi: 10.1109/TMI.2018.2876633.

Daniel P. Huttenlocher, Gregory A. Klanderman, and William J. Rucklidge. Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863, 1993. doi: 10.1109/34.232073.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. doi: 10.1038/s41592-020-01008-z.

Konstantinos Kamnitsas, Christian F. Baumgartner, Christian Ledig, Virginia F. J. Newcombe, Joanna P. Simpson, Andrew D. Kane, David K. Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, and Ben Glocker. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *Information Processing in Medical Imaging*, volume 10265 of *Lecture Notes in Computer Science*, pp. 597–609. Springer, 2017. doi: 10.1007/978-3-319-59050-9_47.

Geon Lee, Chanho Eom, Wonkyung Lee, Hyekang Park, and Bumsub Ham. Bi-directional contrastive learning for domain adaptive semantic segmentation. In *Computer Vision—ECCV 2022*, volume 13690 of *Lecture Notes in Computer Science*, pp. 38–55, Cham, Switzerland, 2022. Springer Nature Switzerland. doi: 10.1007/978-3-031-20056-4_3.

Kang Li, Lequan Yu, Shujun Wang, and Pheng-Ann Heng. Towards cross-modality medical image segmentation with online mutual knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 775–783, 2020. doi: 10.1609/aaai.v34i01.5421.

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. doi: 10.1109/JBHI.2022.3140853.

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*, 42(4):1095–1106, 2023. doi: 10.1109/TMI.2022.3224067.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, U.K., 2 edition, 2009. doi: 10.1017/CBO9780511803161.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241, Cham, Switzerland, 2015. Springer. doi: 10.1007/978-3-319-24574-4_28.

Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3723–3732, 2018. doi: 10.1109/CVPR.2018.00392.

Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Medical Imaging*, 15(1):29, 2015. doi: 10.1186/s12880-015-0068-x.

Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1513–1524, 2020.

Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7472–7481, 2018. doi: 10.1109/CVPR.2018.00780.

Sulaiman Vesal, Mingxuan Gu, Ronak Kosti, Andreas Maier, and Nishant Ravikumar. Adapt everywhere: Unsupervised adaptation of point-clouds and entropy minimization for multi-modal cardiac image segmentation. *IEEE Transactions on Medical Imaging*, 40(7):1838–1851, 2021. doi: 10.1109/TMI.2021.3066683.

Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2517–2526, 2019. doi: 10.1109/CVPR.2019.00262.

Junlin Yang, Nicha C. Dvornek, Fan Zhang, Julius Chapiro, MingDe Lin, and James S. Duncan. Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, volume 11765 of *Lecture Notes in Computer Science*, pp. 255–263. Springer, 2019. doi: 10.1007/978-3-030-32245-8_29.

Yanchao Yang and Stefano Soatto. FDA: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4085–4095, 2020. doi: 10.1109/CVPR42600.2020.00414.

Xuzhe Zhang, Yuhao Wu, Elsa Angelini, Ang Li, Jia Guo, Jerod M. Rasmussen, Thomas G. O'Connor, Pathik D. Wadhwa, Andrea Parolin Jackowski, Hai Li, Jonathan Posner, Andrew F. Laine, and Yun Wang. MAPSeg: Unified unsupervised domain adaptation for heterogeneous medical image segmentation based on 3D masked autoencoding and pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5851–5862, 2024. doi: 10.1109/CVPR52733.2024.00559.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223–2232, 2017. doi: 10.1109/ICCV.2017.244.

Xiahai Zhuang, Lei Li, Christian Payer, Darko Štern, Martin Urschler, Mattias P. Heinrich, Julien Oster, Chunliang Wang, Örjan Smedby, Cheng Bian, Xin Yang, Pheng-Ann Heng, Aliasghar Mortazi, Ulas Bagci, Guanyu Yang, Chenchen Sun, Gaetan Galisot, Jean-Yves Ramel, Thierry Brouard, Qianqian Tong, Weixin Si, Xiangyun Liao, Guodong Zeng, Zenglin Shi, Guoyan Zheng, Chengjia Wang, Tom MacGillivray, David Newby, Kawal Rhode, Sebastien Ourselin, Raad Mohiaddin, Jennifer Keegan, David Firmin, and Guang Yang. Evaluation of algorithms for multi-modality whole heart segmentation: An open-access grand challenge. *Medical Image Analysis*, 58:101537, 2019. doi: 10.1016/j.media.2019.101537.

Xiahai Zhuang, Jiahang Xu, Xinzhe Luo, Chen Chen, et al. Cardiac segmentation on late gadolinium enhancement MRI: A benchmark study from multi-sequence cardiac MR segmentation challenge. *Medical Image Analysis*, 81:102528, 2022. doi: 10.1016/j.media.2022.102528.

A Additional method details

A.1 Causal interpretation and structural reference

The causal formulation in Section 3.1 distinguishes local structural information X from global domain context Z. The former represents spatial evidence relevant to segmentation, such as anatomical shape and boundaries, while the latter represents acquisition-related properties such as intensity, contrast, and imaging style. Both can affect the segmentation logits. Because pixel annotations are available only in the source domain, the model may learn to associate source-domain appearance with the annotated structures. A prediction based on the observed image can therefore reflect both local anatomical evidence and a response to domain-related context.

The factual prediction evaluates the observed structural features together with the domain context. To examine the contribution of those features under the same context, we construct a second prediction in which the observed structural map is replaced while the domain branch output $\widetilde{Z}$ is kept fixed. Both predictions use the same segmentation head. The reference prediction thus captures the response produced with the retained context but without the observed image's position-specific structural features. Comparing it with the factual prediction highlights the change associated with supplying those features. This is the reasoning behind using their logit difference as the TDE-inspired output in Equation 3.

The replacement map is constructed from the running structural mean $\bar{X}$ in Equation 2. Local structural vectors from the current minibatch contribute to its update, and the resulting vector is broadcast over the spatial grid to form $\bar{F}_{\mathrm{loc}}$. This map has the same dimensions as the observed map F_{loc}, but every position receives the same reference vector. It therefore retains a representative value in the learned structural feature space without carrying the observed image's spatial variation in organ shape and boundaries.

The reference prediction is used as a baseline for the contrast, rather than as a standalone segmentation. At inference, the factual and reference logits are computed with the same domain context, and their difference produces the final class scores. The predicted mask is obtained by taking the pixelwise argmax of $\mathrm{softmax}(L^{\mathrm{fact}} - L^{\mathrm{cf}})$.

A.2 Training and inference

At each iteration, a labeled source minibatch and an unlabeled target minibatch are processed by the same structural and domain branches. The source annotations supervise the factual and TDE probability maps through the combined cross-entropy and Dice loss in Equation 4. Target predictions contribute to entropy minimization on the TDE map and an L_1 consistency term between factual and TDE maps. A domain classifier receives intermediate structure-oriented features from both domains through a gradient-reversal layer. Its classification loss supplies the cross-domain adversarial signal in Equation 5.

At inference, the input is passed through both feature branches and the shared segmentation head twice, using the observed and mean structural maps with the same domain context. The domain classifier and training losses are not used; prediction uses only the resulting TDE logits. The final label at each pixel is the highest-probability class among the background, myocardium, left-ventricular cavity, and right-ventricular cavity.

A.3 Evaluation protocol

The MS-CMRSeg evaluation uses bSSFP as source and LGE as target, with five validation cases and the remaining 40 cases divided equally between training and testing. For MMWHS, CT is the source and MR is the target; 16 slices are uniformly selected per volume, and the manually labeled MR cases are evaluated with two-fold cross-validation. Target manual labels do not contribute to the training losses.

For each foreground class, DSC measures the overlap between the predicted and reference regions. HD95 is the larger of the 95th percentiles of the directed surface distances from prediction to reference and from reference to prediction. The reported class average covers MYO, LV, and RV.