Article

Not peer-reviewed version

Domain Shift-Robust Tumor Segmentation Across Hospitals

Karl Andersson *

Posted Date: 17 March 2026

doi: 10.20944/preprints202603.1153.v1

Keywords: tumor segmentation; domain shift; medical image analysis; MRI; CT; multi-center validation; self-supervised learning; test-time adaptation; robustness

 doi:10.20944/preprints202603.1153.v1

Article

Domain Shift-Robust Tumor Segmentation Across Hospitals

A Preprint Manuscript on Reliable Multi-Center Medical Image Segmentation Under Cross-Site Distribution Shift

Karl Andersson

KTH Royal Institute of Technology, Sweden; karl.andersson.krit@gmail.com

Abstract

Deep learning-based tumor segmentation has achieved strong performance on benchmark datasets, yet models often degrade when deployed in new hospitals. This decline is largely driven by domain shift, including differences in scanners, acquisition protocols, reconstruction settings, patient populations, and annotation styles. In high-stakes clinical workflows, such instability limits real adoption because a model that performs well in one center may fail silently in another. This paper presents a preprint-ready methodological framework for domain shift-robust segmentation in multi-hospital MRI and CT tumor imaging. The proposed design combines four complementary ingredients: strong segmentation backbones from the U-Net family, domain-generalization through intensity, style, and frequency-based augmentation, self-supervised pretraining on unlabeled multi-site data, and optional label-free test-time adaptation for target hospitals. The manuscript emphasizes a deployment-oriented evaluation protocol that prioritizes worst-site reliability, boundary safety, calibration, and failure analysis rather than average Dice alone. We describe an experimental plan with leave-one-hospital-out validation, targeted ablations, uncertainty analysis, and stress tests under artifact corruption. The expected pattern is that self-supervised pretraining and frequency-aware augmentation reduce the gap between in-domain and out-of-domain performance, improve worst-site Dice, and lower extreme boundary errors measured by Hausdorff distance. The central argument is that robustness should be treated as a first-class objective in medical image segmentation and that multi-center validation, transparent reporting, and clinically meaningful error analysis are necessary before deployment.

Keywords: tumor segmentation; domain shift; medical image analysis; MRI; CT; multi-center validation; self-supervised learning; test-time adaptation; robustness

1. Introduction

Tumor segmentation is a foundational task in medical image analysis because it supports diagnosis, staging, treatment planning, longitudinal monitoring, and outcome assessment. In radiotherapy planning and image-guided surgery, clinicians frequently require accurate delineation of lesions and surrounding structures. Manual contouring, however, remains labor-intensive and is subject to substantial inter-observer and intra-observer variability. Deep learning has transformed this area by learning direct image-to-mask mappings, with U-Net and its variants emerging as dominant architectures in biomedical segmentation tasks (Ronneberger et al., 2015).

Despite major progress on public benchmarks, a critical translational gap remains: models trained in one institution often underperform when deployed in another. This phenomenon, commonly referred to as domain shift, is especially severe in medical imaging. Even when the underlying anatomy and pathology are similar, hospitals differ in scanner vendors, field strengths, acquisition parameters, reconstruction kernels, dose settings, slice thicknesses, use of contrast agents,

 doi:10.20944/preprints202603.1153.v1

preprocessing routines, and local clinical workflow conventions. These shifts alter image appearance in ways that can make learned representations brittle.

In clinical practice, the cost of brittle segmentation is not merely statistical. Boundary errors can propagate into downstream decisions such as tumor volume estimation, surgical margin design, radiotherapy targeting, treatment response tracking, and eligibility assessment for trials. A system that reports high average benchmark performance but fails unpredictably under cross-site variation is difficult to trust and difficult to integrate into real workflows.

The present paper addresses this gap by reframing segmentation performance around cross-hospital reliability. Rather than optimizing only mean Dice on a fixed dataset, we focus on training and evaluation strategies that improve worst-site behavior, reduce cross-site variance, and expose failure modes before deployment. The proposed direction is practical: it avoids assuming target-site labels, incorporates unlabeled multi-site data when available, and pairs algorithmic design with a more rigorous evaluation protocol.

2. Contributions of This Preprint

This manuscript makes four main contributions. First, it presents a unified research blueprint for robust tumor segmentation across hospitals, combining domain generalization, self-supervised pretraining, and optional label-free adaptation in a single experimental framework. Second, it shifts the evaluation focus from average performance to worst-site reliability, calibration, and clinically meaningful boundary risk. Third, it outlines structured ablations that isolate the contribution of each robustness component and help distinguish genuine robustness from incidental gains. Fourth, it highlights reporting practices that can improve reproducibility and translational relevance in future multi-center segmentation studies.

3. Literature Review

Early deep learning progress in medical imaging was driven by convolutional neural networks that learn hierarchical representations from data (LeCun et al., 2015). For segmentation, U-Net introduced an encoder-decoder structure with skip connections that preserve localization detail while enabling semantic abstraction, making it highly effective for biomedical images (Ronneberger et al., 2015). The extension to 3D U-Net enabled volumetric learning for CT and MRI and became especially important for lesion analysis in three-dimensional scans (Çiçek et al., 2016). Later work such as nnU-Net showed that much of segmentation success depends not only on architecture choice but also on careful self-configuration of preprocessing, patch sampling, augmentation, and inference heuristics (Isensee et al., 2021).

At the same time, surveys in medical image analysis have repeatedly noted that generalization remains one of the most important unresolved challenges in real-world use (Litjens et al., 2017; Shen et al., 2017). Many published models are evaluated under relatively controlled train-test settings, often within the same dataset or institutional source. As a result, reported scores may overstate readiness for deployment. The clinical reality is more heterogeneous: acquisition pipelines differ across institutions, and the statistical distribution seen during training rarely matches what is encountered after deployment.

Domain adaptation methods attempt to close this gap by aligning source and target feature distributions. Adversarial alignment, style transfer, feature normalization, and pseudo-labeling are common strategies. These methods can be effective but often require access to target-domain data during training or tuning, which is not always feasible. In clinical environments, privacy concerns, governance constraints, and workflow limitations often mean that a new hospital has unlabeled data at most, and sometimes only streaming test images at deployment time.

Domain generalization provides a complementary path by encouraging the model to learn hospital-invariant features from multiple source sites without depending on target labels. Practical tools include intensity perturbations, scanner-style randomization, histogram matching, noise and

 doi:10.20944/preprints202603.1153.v1

blur corruption, and frequency-domain manipulations that reduce reliance on institution-specific appearance shortcuts. Frequency-aware methods are particularly relevant because scanners and reconstruction routines influence spectral content and texture patterns that may otherwise become spurious signals for the model.

Self-supervised learning has also become increasingly important because medical datasets often contain much more unlabeled than labeled data. By pretraining an encoder with masked reconstruction or contrastive objectives, a model can learn richer anatomical and structural priors before segmentation fine-tuning. This may produce features that are less tied to a single hospital's appearance distribution and more robust to limited labeled data. In multi-site settings, self-supervised pretraining is attractive because it can leverage pooled or federated unlabeled scans even when segmentation masks remain scarce.

Recent work in neuroimaging further illustrates the difference between strong task performance and robust deployment. Kar et al. (2024) report excellent performance for intracranial hemorrhage detection using an attention-enhanced CNN on CT images, highlighting the value of clinically relevant localization. Although that study focuses on detection rather than tumor segmentation, it underscores a broader lesson: high performance in a controlled evaluation does not automatically guarantee reliable cross-site behavior. This motivates a shift toward robustness-centered segmentation research, especially in multi-hospital environments.

4. Problem Formulation

We assume K source hospitals with labeled segmentation data and one unseen target hospital used only for testing. Each hospital may differ in scanner vendor, imaging protocol, patient demographics, and annotation style. Let x denote an MRI or CT image volume and y the corresponding tumor mask. The goal is to learn a segmentation function $f\theta(x) \rightarrow \hat{y}$ that performs well on an unseen target hospital without requiring target labels.

The segmentation task may be binary, for example tumor versus background, or multi-class, such as enhancing tumor, necrotic core, and edema. Depending on computational resources and slice anisotropy, the framework may use either 2D slice-based learning or 3D patch-based learning. The core challenge is that empirical risk minimization on pooled source data may still encourage hospital-specific shortcuts rather than clinically meaningful features. Therefore, the training process must intentionally promote robustness.

5. Proposed Framework

The proposed framework has four layers: a strong segmentation backbone, domain-generalization augmentations, self-supervised pretraining, and optional test-time adaptation. The design is modular so that each component can be evaluated independently through ablation studies.

Backbone. A U-Net family model serves as the primary segmentation backbone. For computationally constrained settings or very large datasets, a 2D U-Net may be appropriate. Where volumetric context is critical and memory allows, a 3D U-Net can better capture inter-slice continuity and lesion morphology. In either case, nnU-Net-inspired preprocessing and training heuristics provide a strong and reproducible baseline.

Loss design. To balance region overlap and voxel-wise stability under class imbalance, training uses a composite loss consisting of Dice loss and cross-entropy loss. Dice loss is especially important for small tumor regions, while cross-entropy stabilizes optimization and improves class-wise discrimination. Additional boundary-sensitive losses may be considered when extreme contour errors are clinically unacceptable.

Domain-generalization. Robustness is encouraged using intensity augmentation, scanner-style perturbation, and frequency-domain mixing. Intensity transforms include gamma shifts, contrast changes, brightness variation, additive noise, blur, and bias-field simulation. Style augmentation includes histogram perturbation or appearance transfer inspired by adaptive instance normalization.

 doi:10.20944/preprints202603.1153.v1

Frequency augmentation randomizes or mixes spectral components across samples so that the network becomes less dependent on site-specific texture or reconstruction signatures.

Self-supervised pretraining. Before supervised fine-tuning, the encoder may be pretrained on unlabeled multi-site imaging data. Masked autoencoding teaches the network to reconstruct missing image regions, which encourages structural understanding. Contrastive objectives encourage invariance across augmented views of the same scan. Both strategies are useful when labeled segmentation masks are limited but unlabeled scans are abundant across hospitals.

Test-time adaptation. When target-hospital images arrive, safe label-free adaptation can update only a restricted subset of parameters. Examples include adapting batch normalization statistics or minimizing predictive entropy under strict safeguards. Because test-time adaptation can also harm performance if done aggressively, the framework includes rollback rules, confidence monitoring, and minimal-parameter updates to reduce the risk of negative adaptation.

6. Experimental Design

The experimental design is centered on deployment realism. Instead of random train-test splits across all data, we use leave-one-hospital-out validation. In each fold, the model is trained on K-1 hospitals and evaluated on the held-out site. Repeating this process across all sites yields a distribution of cross-hospital results rather than a single pooled number.

Preprocessing should be standardized but not overly aggressive. Resampling, intensity clipping, z-score normalization, and patch extraction should be applied consistently across sites while preserving realistic variability. Importantly, site identity should not leak through filename conventions, preprocessing differences, or fold construction.

Ablation studies isolate the effect of each component: baseline U-Net only, baseline plus conventional augmentation, baseline plus frequency/style augmentation, self-supervised pretraining plus fine-tuning, and the full framework with optional test-time adaptation. Additional stress tests should inject realistic corruptions such as motion blur, noise spikes, low-dose appearance, missing slices, or simulated bias fields. These perturbations help determine whether gains are robust or only apparent under clean benchmark conditions.

7. Evaluation Metrics and Reporting

Dice similarity coefficient remains the standard overlap metric, but it should not stand alone. In tumor segmentation, boundary-based metrics such as the 95th percentile Hausdorff distance capture extreme contour errors that can be clinically significant even when Dice appears acceptable. Sensitivity is important to quantify missed tumor regions, while precision and volumetric error can reveal over-segmentation tendencies.

Calibration and uncertainty should also be reported. Expected Calibration Error (ECE), reliability diagrams, confidence histograms, and uncertainty maps can indicate whether the model knows when it might be wrong. In deployment, this is crucial because uncertain cases can be routed for manual review or quality assurance rather than accepted silently.

Site-wise reporting is essential. Every experiment should present performance by hospital, together with mean, standard deviation, and worst-site values. A method that improves the mean but leaves one site catastrophically weak should not be considered robust. The proposed framework therefore treats worst-site Dice and worst-site Hausdorff distance as primary metrics, not optional supplements.

8. Expected Findings and Interpretation

As a research-framework manuscript, this paper presents the expected pattern of results rather than a completed benchmark table. The most likely trend is that a standard segmentation baseline performs well in-domain but loses accuracy and boundary stability in unseen hospitals. Frequency-

 doi:10.20944/preprints202603.1153.v1

aware and style-aware augmentations should reduce this gap by weakening the model's dependence on scanner-specific appearance cues.

Self-supervised pretraining is expected to improve worst-site generalization more consistently than simple augmentation alone, particularly when hospitals contribute heterogeneous unlabeled data. The likely mechanism is that the encoder learns broader anatomical features before becoming specialized for segmentation. In practice, this should narrow the in-domain versus out-of-domain gap and increase resilience to moderate artifact corruption.

Test-time adaptation may provide additional gains, but its contribution is likely to be conditional and more variable than that of augmentation or pretraining. Conservative approaches such as batch normalization updates may help alignment with limited risk, whereas more aggressive entropy-minimization strategies may require strong monitoring. Therefore, any claimed improvement from test-time adaptation should be accompanied by evidence that it does not introduce new failure modes.

An especially important expectation is that some of the clearest gains may appear not in average Dice but in boundary safety. A modest Dice increase paired with a meaningful reduction in Hausdorff distance can be more clinically valuable because it indicates fewer extreme contour failures. This shift in emphasis aligns the technical evaluation more closely with real clinical risk.

9. Failure Modes and Clinical Risk Analysis

Robustness claims are incomplete without explicit failure analysis. Likely residual error modes include very small tumors, poorly contrasted lesions, heavy motion artifacts, post-surgical anatomy, and rare subtypes that are weakly represented in the source hospitals. Another important challenge is annotation inconsistency: different institutions may contour boundaries according to slightly different conventions, making some apparent domain shift partly a labeling problem.

Failure analysis should therefore include qualitative review by case type, tumor size, modality, artifact burden, and hospital. Representative visual examples should compare the input image, reference mask, prediction, uncertainty map, and error overlay. Such visualization helps distinguish whether errors arise from under-segmentation, boundary leakage, false positive hotspots, or systematic bias toward certain anatomical locations.

From a clinical safety perspective, the most concerning scenario is silent failure: the model produces a confident but incorrect segmentation in a hospital whose image characteristics differ from training. This is why calibration, uncertainty, and external validation must accompany segmentation accuracy. A system intended for workflow support should include mechanisms to flag high-risk cases rather than forcing binary acceptance of every prediction.

10. Reproducibility and Deployment Considerations

To make cross-hospital robustness research useful to the community, reporting should include hospital split definitions, preprocessing rules, augmentation settings, pretraining configuration, model selection criteria, and failure-case summaries. Whenever privacy permits, code, trained weights, and synthetic or de-identified evaluation artifacts should be shared. Reproducibility is especially important because robustness claims can be sensitive to seemingly minor implementation decisions.

Deployment planning should also consider governance and systems issues. Multi-center collaboration requires data use agreements, privacy-preserving pipelines, and clear accountability for model updates. Integration into radiology or oncology workflows may require PACS compatibility, audit logging, latency constraints, human override paths, and post-deployment monitoring. A robust segmentation model is therefore not just a better neural network; it is part of a broader socio-technical system.

 doi:10.20944/preprints202603.1153.v1

11. Limitations and Future Research

This manuscript has several limitations. First, it presents a preprint-style methodological paper rather than a completed empirical benchmark. Second, access to truly diverse multi-hospital datasets can be difficult due to privacy, governance, and annotation cost. Third, annotation variability across hospitals may blur the line between domain shift and label inconsistency. Fourth, 3D training and self-supervised pretraining can be computationally expensive, which may limit reproducibility for smaller labs.

Future research should explore federated or split-learning strategies that enable robustness training without centralizing sensitive data. Uncertainty-guided quality assurance pipelines could route ambiguous cases to human review. Causal and shortcut analyses may help detect when the model relies on scanner signatures or acquisition artifacts instead of pathology. Finally, the field would benefit from stronger community standards that require worst-site metrics, calibration reporting, and external multi-center validation as default practice.

12. Conclusions

Deep learning has made tumor segmentation faster and more consistent, but real deployment across hospitals remains limited by domain shift. A clinically useful model must generalize across scanners, protocols, reconstruction settings, and patient populations rather than perform well only on one benchmark distribution. This paper presented a practical and preprint-ready research framework for domain shift-robust tumor segmentation that combines strong U-Net baselines, robustness-oriented augmentation, self-supervised multi-site pretraining, and cautious label-free adaptation.

The main message is that robustness should be treated as a first-class objective in medical image segmentation. Progress should be judged not only by higher average Dice but also by better worst-site performance, safer boundaries, improved calibration, and transparent failure analysis. With rigorous multi-center validation and safety-oriented design, robust segmentation methods can move closer to trustworthy clinical use.

Table 1. Robustness techniques and the cross-hospital variability they address.

Technique	Primary target	Expected robustness benefit
Intensity transforms	Scanner gain and contrast shift	Reduces sensitivity to brightness and intensity variation across hospitals
Noise and blur	Motion, low dose, reconstruction artifacts	Improves stability under degraded image quality
Style / histogram perturbation	Vendor and protocol appearance	Discourages shortcut learning from site-specific style cues
Frequency mixing	Reconstruction kernels and texture signatures	Promotes invariance to spectral differences linked to scanners
Self-supervised pretraining	Limited labels and site-specific features	Learns broader anatomical representations before segmentation
Test-time adaptation	Residual target mismatch	Allows cautious alignment to new hospitals without target labels

Table 2. Recommended evaluation metrics for robust multi-hospital tumor segmentation.

Metric	What it measures	Why it matters for deployment
Dice coefficient	Region overlap	Standard segmentation quality; easy to compare across studies
Hausdorff distance (95%)	Extreme boundary deviation	Captures clinically risky contour failures
Sensitivity / recall	Missed tumor burden	Important for under-segmentation risk

 doi:10.20944/preprints202603.1153.v1

Precision	False positive burden	Useful when over-segmentation affects treatment planning
Expected Calibration Error	Confidence reliability	Supports safe triage and human-in-the-loop review
Worst-site Dice	Minimum hospital performance	Prevents good averages from hiding one catastrophic site

Multi-hospital MRI/CT data (labeled + unlabeled)	**Self-supervised pretraining + robust augmentation**	**Segmentation fine-tuning (U-Net / 3D U-Net)**	**Site-wise evaluation, uncertainty, and TTA**

Practical takeaway. A robust medical segmentation model should be evaluated like a safety-critical system: by how consistently it behaves in the hardest hospitals, how large its worst boundary errors are, and whether it signals uncertainty when the input distribution changes.

Figure 1. Conceptual overview of the proposed robust segmentation pipeline.

References

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In Medical Image Computing and Computer-Assisted Intervention (MICCAI) (pp. 424–432).

Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods, 18(2), 203–211.

Kar, N. K., Jana, S., Rahman, A., Ashokrao, P. R., & Mangai, R. A. (2024). Automated intracranial hemorrhage detection using deep learning in medical image analysis. In 2024 International Conference on Data Science and Network Security (ICDSNS) (pp. 1–6). IEEE. https://doi.org/10.1109/ICDSNS62112.2024.10691276

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60–88.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention (MICCAI) (pp. 234–241).

Shen, D., Wu, G., & Suk, H.-I. (2017). Deep learning in medical image analysis. Annual Review of Biomedical Engineering, 19, 221–248.