Adversarial Perturbations Improve Generalization of Confidence Prediction in Medical Image Segmentation

Jonathan Lennartz
Thomas Schultz

University of Bonn, Lamarr Institute for Machine Learning and Artificial Intelligence

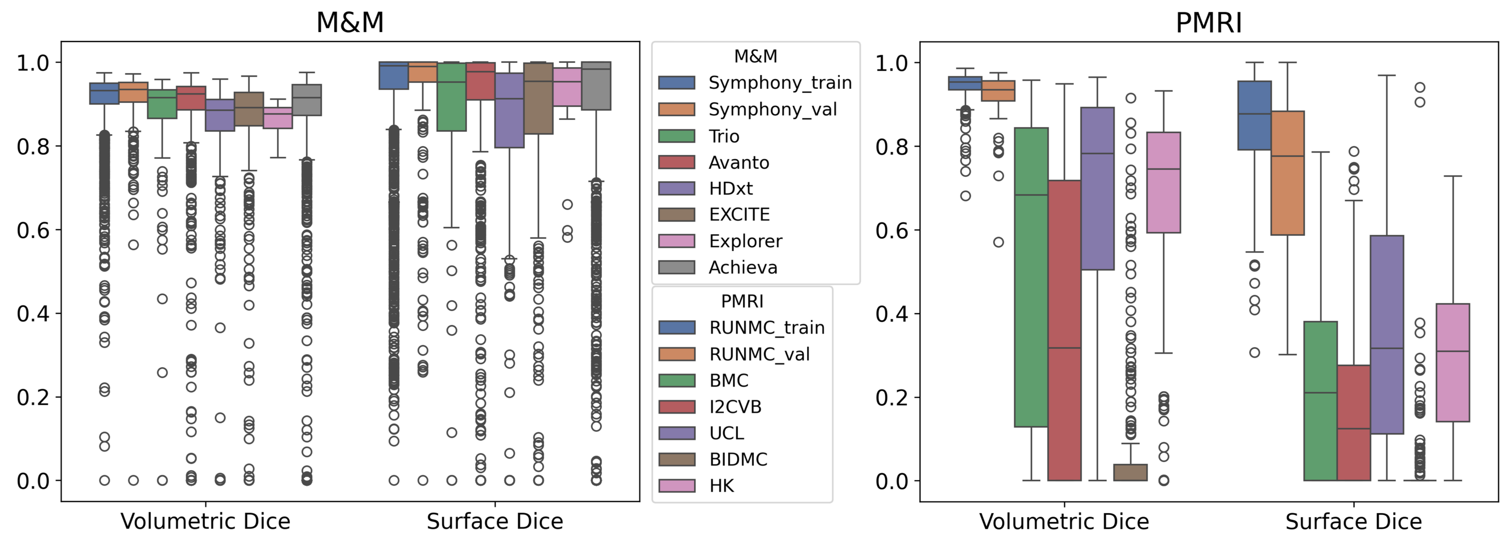
LAMARR
INSTITUTE FOR MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Abstract

We introduce a novel adversarial strategy for training confidence predictors for medical image segmentation [1]. By perturbing images to decrease the predicted accuracy, we generate hypotheses about out-of-distribution (OOD) image modifications that the predictor expects to degrade segmentation quality. Continuously including these adversarial perturbations in training improves generalization across scanners while adding negligible computational overhead.

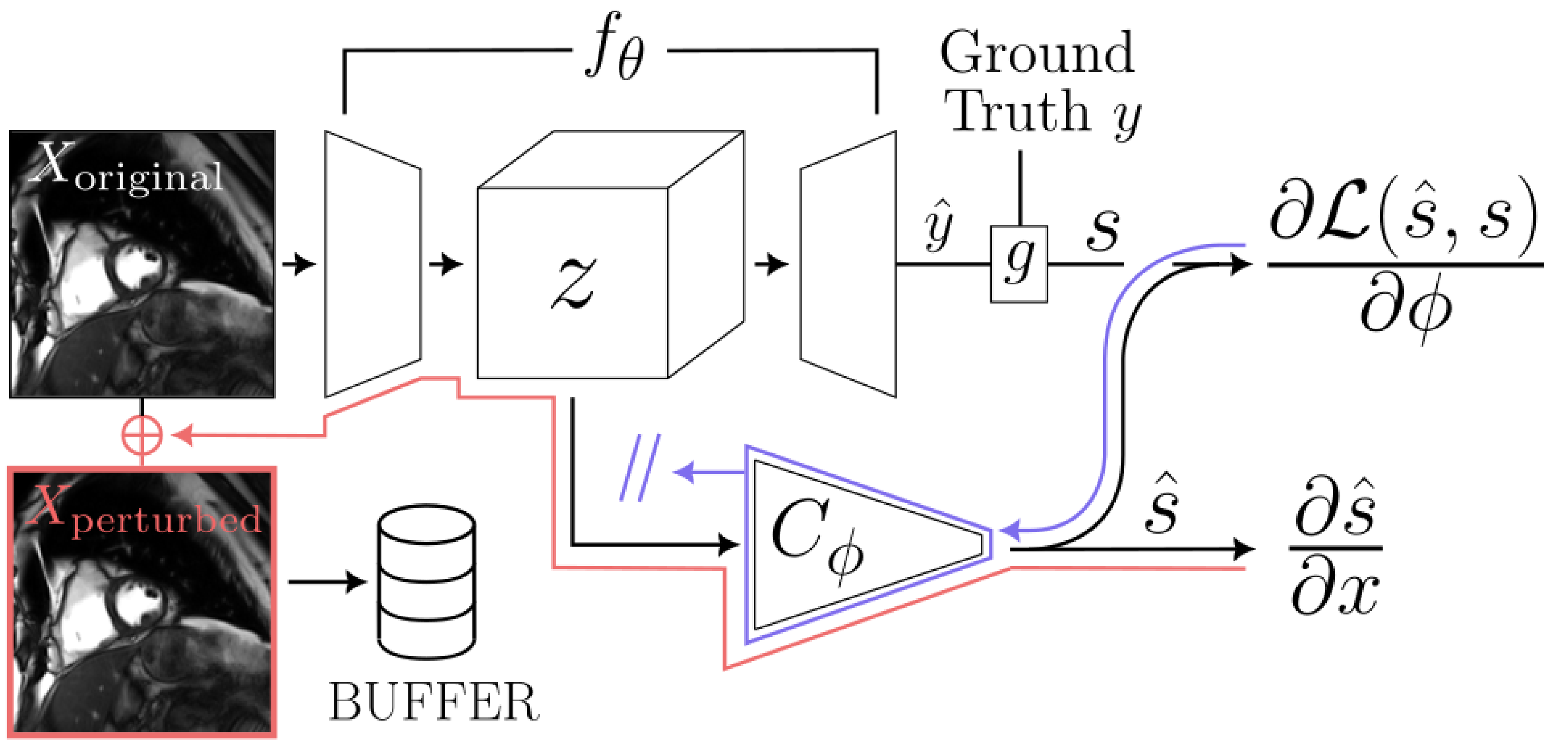
Domain Shift in Medical Imaging

We evaluate on cardiac MRI (M&Ms v2 [3]) and prostate MRI (PMRI [2]), each containing scans from diverse devices. Substantial performance drops occur on several target domains, motivating robust confidence estimation.

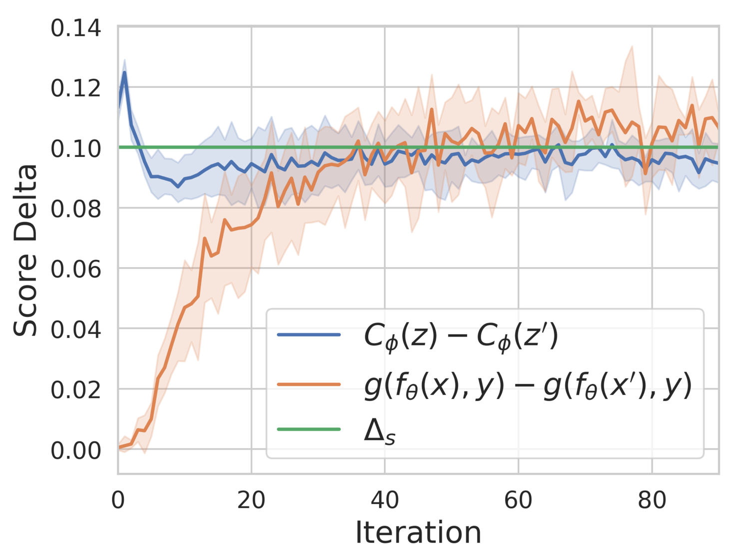


Methodology

From the activations of a frozen, fully trained U-Net f_θ, a confidence predictor C_ϕ learns to predict a confidence score s (volumetric or surface Dice).To align f_θ and C_ϕ outside the training distribution, we generate image perturbations via the gradient of the predicted score $\hat{s}$, scale them to achieve a desired average effect size Δ_s on C_ϕ, and include perturbed images in its training.



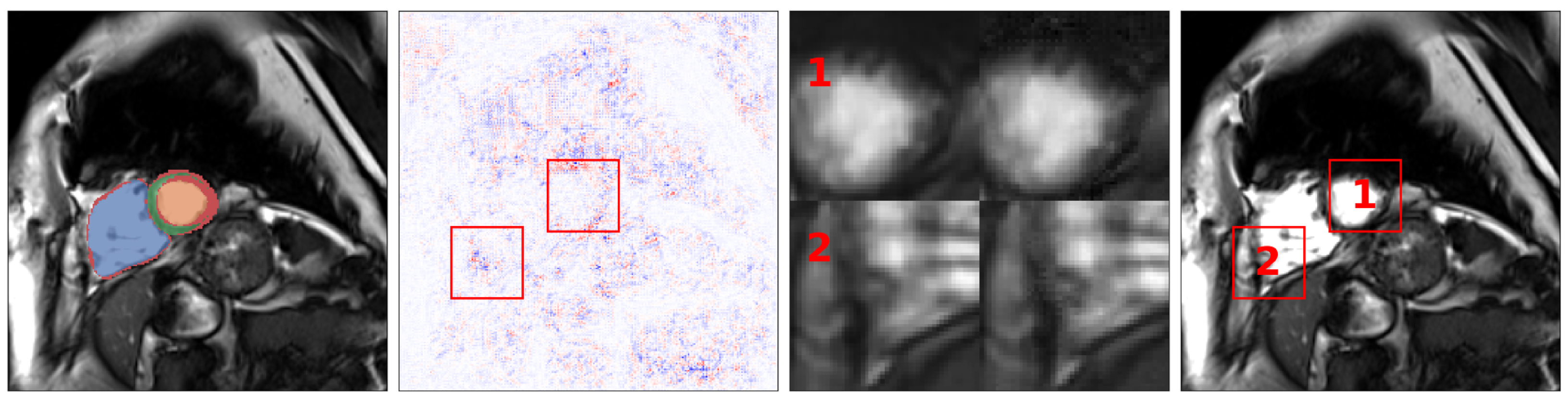
Early in training, perturbations hardly affect f_θ, revealing misalignment. Our iterative adversarial inclusion corrects this, yielding better OOD calibration.



x	Input image
f_θ	U-Net
z	Penultimate activations
C_ϕ	Confidence predictor
g	Confidence measure
Δ_s	Desired effect size
$\mathcal{L}$	Loss function

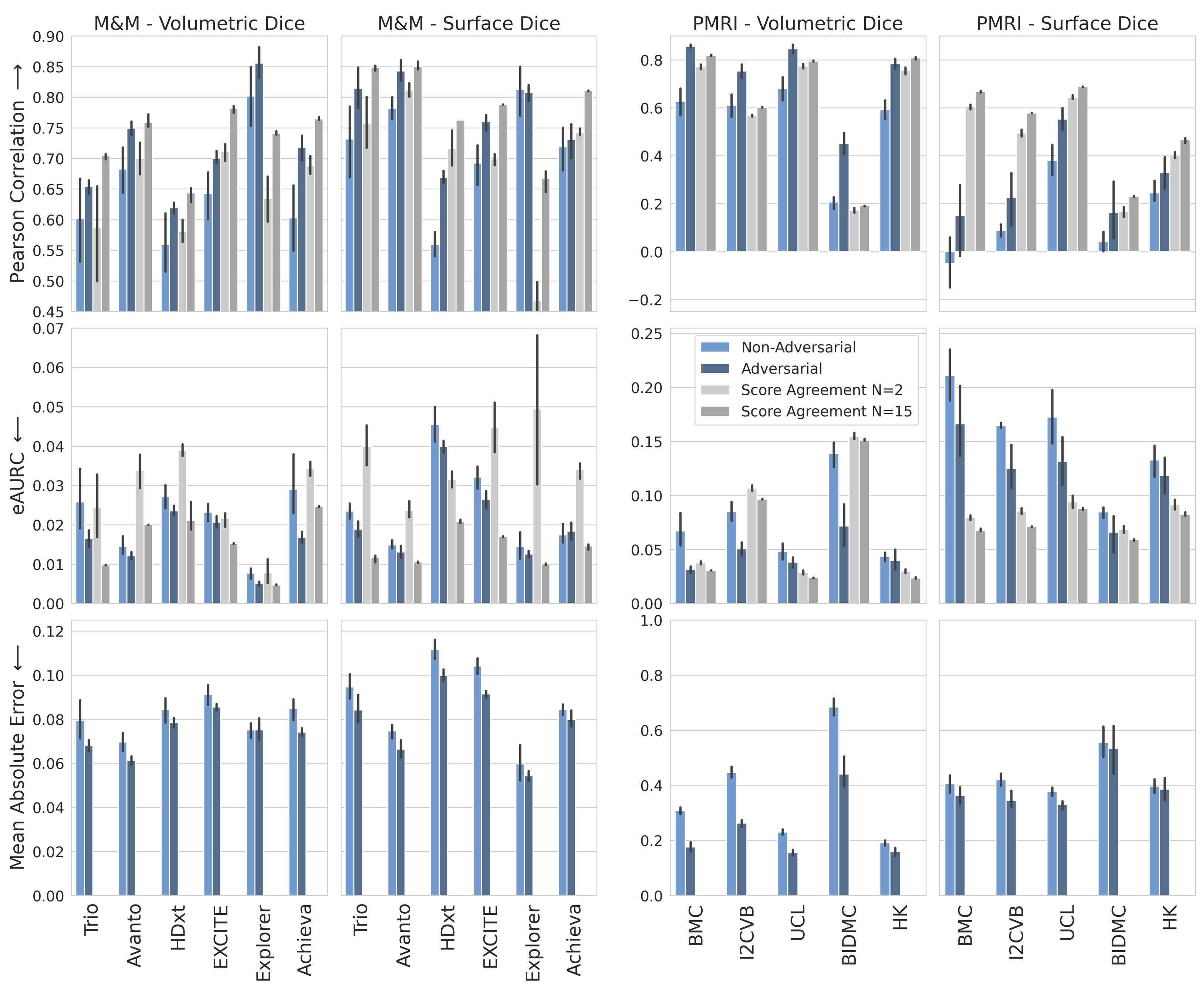
Results

Example (after alignment): adversarial perturbation reduces predicted confidence ($0.92 \rightarrow 0.70$) and real segmentation quality ($0.90 \rightarrow 0.81$). Errors appear in red.



Prediction after perturbation and error (red) | Applied perturbation | Original (left) and perturbed (right) crops | Image and crop indicator

Our adversarial training improves correlation with true OOD segmentation accuracy and reduces excess area under the risk-coverage curve (eAURC) and mean absolute error versus non-adversarial training. In several domains it rivals or surpasses score agreement [4], an ensemble-based baseline up to three orders of magnitude more expensive, while producing absolute confidence predictions.



Conclusion

Direct confidence prediction leaves the segmentation network unchanged, adds minimal overhead, and yields absolute quality estimates. Adversarial training aligns predictor and segmenter under domain shift, improving reliability. Future work: Ongoing work more clearly beats baseline with improved predictor architecture.

References and Code

[1] Lennartz, J. and Schultz, T.. Adversarial Perturbations Improve Generalization of Confidence Prediction in Medical Image Segmentation. *MIDL* (2025).

[2] Liu, Q. et al.. Shape-aware Meta-learning for Generalizing Prostate MRI Segmentation to Unseen Domains. *MICCAI* (2020).

[3] Martín-Isla, C. et al.. Deep Learning Segmentation of the Right Ventricle in Cardiac MRI: The M&Ms Challenge. *IEEE Journal of Biomedical and Health Informatics* (2023).

[4] Roy, A. G. et al.. Inherent Brain Segmentation Quality Control from Fully ConvNet Monte Carlo Sampling. *MICCAI* (2018).

Partner institutions:

technische universität dortmund
Fraunhofer IAIS
UNIVERSITÄT BONN
Fraunhofer IML

Institutionally funded by:

Bundesministerium für Forschung, Technologie und Raumfahrt
Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen