

Boundary-Selective Pseudo-Label Repair and Model Absorption for Semi-Supervised Medical Image Segmentation

Yuan Liu

College of Computer Science and Technology, Qingdao University, Qingdao 266071, China

Email: liuyuan84@qdu.edu.cn

How to cite this paper: Liu, Y. (2026). Boundary-Selective Pseudo-Label Repair and Model Absorption for Semi-Supervised Medical Image Segmentation. Academic Journal of Emerging Technologies, 3(3), 19–30. ISSN Print: 3104-4417, ISSN Online: 3104-4425.

<https://doi.org/10.63313/AJET.9070>

Published: 2026-09-26

Copyright © 2026 by author(s) and

Erytis Publishing Limited.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Abstract

Pseudo-labels can exploit unlabeled medical images, but high-overlap masks may still contain local contour errors and selected corrections may disappear during fine-tuning. We recast pseudo-label refinement as a repair-to-model problem and present Boundary-Selective Pseudo-Label Repair and Model Absorption (CPPA). A validation-selected frozen source supplies the no-edit reference. Wavelet-frequency evidence, spatial responses, and disagreement between source predictors constrain a morphological Snake to bounded add/remove candidates. A local policy scores each contour arc against the paired no-edit source and abstains when the estimated benefit is unsafe. The selected mask supervises an edit-aware absorption bank, after which a validation-only router promotes one checkpoint or retains the source. Across 27 internal settings spanning CVC-ClinicDB, BUSI-benign, and BUSI-malignant, three label ratios, and three seeds, the final single-model outputs improve the selected source by 0.0241 Dice, 0.0536 BF1, and 0.0402 Boundary IoU on average. Adapted checkpoints are promoted in 26 of 27 settings, and all nine dataset-ratio means improve Dice, BF1, and Boundary IoU. Frozen transfer to CVC-ColonDB improves Dice and BF1 at every label ratio. The evidence supports a bounded, abstaining repair policy whose local corrections can be absorbed into one deployable network.

Keywords

Semi-Supervised Segmentation; Pseudo-Label Repair; Boundary Quality; Active Contour; Model Absorption; Medical Image Analysis

1. Introduction

Medical image segmentation is often limited by annotation cost rather than by the availability of images. Semi-supervised methods address this imbalance with pseudo-labels, consistency regularization, and cross-model supervision, but they usually treat reliability as a property of a whole mask or a confidence threshold. This leaves a practical failure mode unresolved: a mask can have good overlap while

a small number of contour errors dominate boundary metrics and clinical shape interpretation.

The missing decision is local and comparative. For a particular boundary arc, should the model keep its current prediction, or apply a concrete add/remove intervention supported by the image and by model disagreement? A useful correction must also survive parameter transfer. Improving an intermediate pseudo-label is insufficient if ordinary fine-tuning overwrites stable regions or fails to store the correction in the final network.

We therefore formulate pseudo-label refinement as a staged repair-to-model problem. The frozen source is the no-edit action. A bounded candidate generator proposes only local contour interventions. A counterfactual arc policy estimates the operational benefit of each intervention relative to that source and can abstain. An edit-aware absorption objective then transfers accepted edits into a single model while distilling unchanged source behavior. Finally, a validation-only router provides a conservative promotion rule with the source checkpoint as a permanent fallback.

Our central claim is deliberately bounded: within the evaluated polyp and breast-lesion cohorts, local contour repair can be selected without labels on unlabeled images and can yield source-relative gains after absorption, especially for boundary metrics. We do not claim causal identification, universal state-of-the-art performance, or cross-modality generalization.

The contributions are fourfold. First, we separate candidate quality, edit selection, and parameter absorption, making each stage measurable. Second, we introduce a local-arc policy that compares an edited candidate with its paired no-edit source and abstains under unsafe evidence. Third, we use edit-weighted absorption and a validation-routed bank to convert sparse repairs into one deployable checkpoint. Fourth, we evaluate the full evidence chain over 27 frozen internal settings and an unchanged external CVC-ColonDB transfer, including failure cases and distance metrics.

2. Related Work

Consistency regularization and pseudo-labeling form the basis of modern semi-supervised segmentation. Mean Teacher uses an exponential-moving-average target, while MixMatch, FixMatch, and UniMatch combine confidence or weak-to-strong consistency with augmentation. Cross-Consistency Training, Cross Pseudo Supervision, and U2PL address decoder diversity, cross-model agreement, or unreliable regions. FlexMatch, FreeMatch, SoftMatch, and InstanT adapt confidence according to class, distribution, or instance context.

Medical segmentation variants add uncertainty estimates, dual-task consistency, shape constraints, mutual consistency, and copy-paste augmentation. These methods motivate risk-aware learning, but they generally do not score a concrete

local contour intervention against the corresponding no-edit outcome. CPPA makes this paired decision explicit and keeps the source as a safe action when evidence is insufficient.

Boundary-aware losses and metrics such as Boundary IoU, BF1, boundary loss, Hausdorff distance, and ASSD show why overlap alone is not enough. Our use of these measures is not an inference-time contour postprocessor. The candidate mask is an intermediate training target, and its value is tested by whether a single adapted network retains the correction.

U-Net remains a common backbone for lesion segmentation, while PraNet and the CVC datasets provide established polyp benchmarks. BUSI offers benign and malignant ultrasound cohorts with different boundary characteristics. We use these datasets to test a shared repair and absorption protocol rather than to claim an architecture-specific advance.

3. Method

CPPA has four stages: source-evidence localization, bounded candidate generation, counterfactual arc selection, and edit-aware student absorption. Inference uses only the final promoted network. Candidate generation and policy selection are training-time operations and never access test labels.

3.1. Frozen source and bounded candidates

Let x be an image, p_s the probability response of a frozen source, and $m_s = 1[p_s > t_s]$ its validation-calibrated mask. The source is selected from protocol-matched FixMatch/UniMatch-style and CPS-style models using validation data only. It is the no-edit action and remains available as a fallback.

We compute normalized wavelet-frequency evidence e_f , combined spatial-gradient/response evidence e_s , and a source-disagreement map d . The first two define an edge-stopping surface for a morphological geodesic active contour:

$$g(x) = \exp\left(-\beta (w_f e_f(x) + w_s e_s(x))^2\right). \quad (1)$$

Starting from m_s and conservative eroded or dilated variants, fixed Snake configurations evolve inside a narrow inner/outer envelope and disagreement-supported boundary bands. With $\Delta_k = m_k \text{ XOR } m_s$, the candidate set is

$$C(x) = \{m_s\} \cup \{m_k : m_k = A\phi_k(m_s; g), \text{supp}(\Delta_k) \subseteq \Omega_s, |\Delta_k| \leq \rho HW\}, \rho = 0.04. \quad (2)$$

The disagreement map restricts where edits may occur; it is not silently mixed into the edge-stopping image. This gives a measurable candidate budget and prevents a contour optimizer from rewriting stable interiors.

3.2. Counterfactual local-arc policy

For each candidate, Δ_k is split into connected, direction-consistent contour arcs. A

compact benefit network receives only image- and model-derived features: source confidence, candidate consensus, signed add/remove direction, frequency and spatial responses, disagreement, and local geometry. On labeled images, a changed pixel is beneficial when the candidate agrees with the ground-truth mask:

$$z_{k,u} = 1[m_k(u) = y(u)], u \in \Delta_k; L_{pol} = \frac{1}{|S|} \sum_{k,u \in S} (BCE(q_\psi(f_{k,u}), z_{k,u})). \quad (3)$$

The term counterfactual denotes this operational paired comparison, not a causal-identification claim. For arc a , the implemented score combines the mean and a calibrated lower-tail quantile:

$$s(a) = 0.5 \text{mean}(q_a) + 0.5 Q_\eta(q_a). \quad (4)$$

Accepted non-overlapping arcs overwrite the source only on their support; all other pixels remain m_s . The calibration configuration maximizes a validation utility with an edit-area penalty, minimum active rate, edit-precision and harmful-active constraints, and a Dice-drop tolerance:

$$U = \text{Dice} + 0.25 \text{BF1} + 0.40 \text{B} - \text{IoU} - 0.001 \text{HD95} - 0.002 \text{ASSD}. \quad (5)$$

If no feasible configuration improves the source utility, the accepted-arc set is empty and $m^* = m_s$. Thus policy application to unlabeled training images uses no ground truth and has an explicit abstention path.

3.3. Edit-aware absorption and guarded deployment

The selected mask m^* is intermediate supervision, not the final prediction. Each adapted network is initialized from the frozen source, and images containing accepted edits are oversampled. Stable pixels retain the source pseudo-probability; changed pixels receive smoothed selected targets and explicit add/remove weights.

With $E = m^* \text{ XOR } m_s$, the target is

$$t(u) = p_s(u), u \notin E; t(u) = (1 - \alpha)p_s(u) + \alpha m^*(u), u \in E. \quad (6)$$

The edit loss is weighted only on E and is combined with the segmentation, boundary, consistency, and optional source-distillation terms:

$$L_{abs} = L_{mask} + \lambda_b L_{boundary} + \lambda_e L_{edit} + \lambda_c L_{cons} + \lambda_d L_{src}. \quad (7)$$

We evaluate three fixed absorption routes in every setting: local-arc CPPA, directional add/remove, and candidate-vote consensus. A validation-only router compares each route with the frozen source and promotes one checkpoint only when its validation utility exceeds the source by $\delta = 0.001$ and every member of $\{\text{Dice}, \text{BF1}, \text{B-IoU}\}$ remains within $\epsilon = 0.002$ of the source. Otherwise, it deploys the source. Training may evaluate multiple routes, but inference always uses one network.

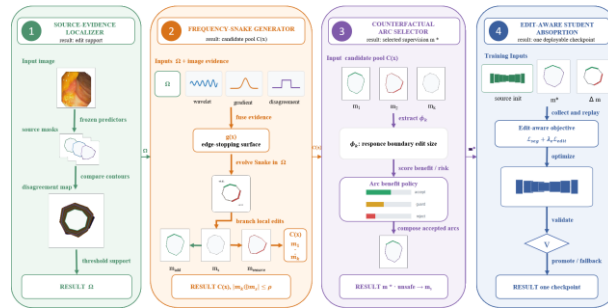


Figure 1. CPPA decomposes pseudo-label refinement into source-evidence localization, bounded candidate generation, counterfactual arc selection, and edit-aware absorption. The source remains the no-edit fallback, and the validation router deploys one checkpoint.

4. Experimental Setup

4.1. Datasets and protocol

We use three internal cohorts. Grouped CVC-ClinicDB contains 490 training and 122 test frames, with related frames kept in the same partition. BUSI-benign and BUSI-malignant are treated as separate cohorts with 349/88 and 168/42 train/test images, respectively. For each cohort, 15% of the training split is reserved for validation before sampling 5%, 10%, or 20% labeled subsets. Seeds 42-44 control labeled sampling and model training. Images are resized to 256 x 256.

For external validation, every CVC-trained checkpoint, threshold, post-processing choice, and promotion decision is frozen before evaluation on the 380-image CVC-ColonDB test set. Exact resized-image hashing found no overlap with the CVC-ClinicDB split. ColonDB labels are used only for metrics.

4.2. Baselines and implementation

We compare supervised U-Net, Mean Teacher, BCP-style boundary copy-paste, FixMatch/UniMatch-style confidence training, and CPS-style cross supervision. All common-protocol baselines use the same backbone, split manifest, eight-epoch budget, and validation operating-point search. We additionally port the official UniMatch recipe at repository commit 0324f3d to the common U-Net, split, and budget. The source for each CPPA setting is selected from the two strongest common-protocol sources, FixMatch/UniMatch-style and CPS-style.

Candidate generation uses a wavelet high-pass cutoff of 0.055, inner/outer boundary widths of 4/5 pixels, and a 4% maximum edit area. The benefit network is a two-layer, 96-wide MLP trained for 24 epochs. Contours are split into 16 angular bins before connected-arc scoring. The adapted U-Net is fine-tuned from the source for eight epochs with batch size 4 and learning rate 5×10^{-6} . Selector constraints, loss weights, and promotion margins are fixed before the three-cohort matrix.

4.3. Metrics and Reporting

We report Dice, IoU, BF1, Boundary IoU (B-IoU), HD95, and ASSD. Higher is better for Dice, IoU, BF1, and B-IoU; lower is better for HD95 and ASSD. Tables 1-3 use the complete internal matrix. Table 4 isolates the evidence chain on nine confirmation cells, and Table 5 reports the frozen external transfer. Qualitative examples are illustrative only; aggregate claims use the complete matrices.

5. Results

5.1. Repair Quality and Absorption Durability

Table 1 separates three questions that are often conflated: whether the candidate pool contains useful edits, whether the policy selects them, and whether a selected edit survives parameter transfer. Frequency-Snake repair has positive composite utility in 24 of 27 cells. CPPA selection raises the mean source gain to 0.0107 Dice, 0.0418 BF1, and 0.0290 B-IoU, with positive composite utility in 26 of 27 cells. The final routed single model reaches larger source-relative gains, indicating that edit-aware absorption can consolidate sparse corrections instead of merely reproducing an intermediate mask.

Table 1. Three-stage evidence chain over the same 27 dataset-ratio-seed cells. Positive distance changes indicate lower error.

Stage	Coverage	Overlap (Dice / IoU)	Boundary (BF1 / B-IoU)	Distance (HD95 / ASSD)
Frequency + Snake repair	24/27	+0.0091 / +0.0120	+0.0390 / +0.0272	+1.298 / +0.566
CPPA selected pseudo-label	26/27	+0.0107 / +0.0137	+0.0418 / +0.0290	+1.278 / +0.617
Final single model	26/27 deploy	+0.0241 / +0.0297	+0.0536 / +0.0402	+2.890 / +1.425

5.2. Internal comparison

Table 2 reports the final single checkpoint, not an oracle, selected pseudo-label, or ensemble. Adapted checkpoints are promoted in 26 of 27 settings. All nine dataset-ratio means improve Dice, BF1, and B-IoU relative to the selected source. BUSI-malignant retains low absolute boundary scores, so the positive deltas should be read as source-relative improvements rather than as a solution to the hardest ultrasound contours.

Table 2. Final single-model results on the three internal cohorts. Values are mean +/- standard deviation over seeds 42-44.

Dataset	Labels	Promote	Overlap (Dice / IoU)	Boundary (BF1 / B-IoU)	Distance (HD95 / ASSD)
CVC-ClinicDB	5%	3/3	0.7454+/-0.0226 / 0.6595+/-0.0203	0.5346+/-0.0127 / 0.3975+/-0.0103	39.30+/-6.84 / 13.43+/-2.57

CVC-ClinicDB	10%	3/3	0.8157+/-0.0094 / 0.7375+/-0.0082	0.6255+/-0.0129 / 0.4750+/-0.0106	28.89+/-1.03 / 8.76+/-0.59
CVC-ClinicDB	20%	3/3	0.8554+/-0.0104 / 0.7861+/-0.0080	0.6808+/-0.0113 / 0.5209+/-0.0056	22.37+/-0.93 / 6.61+/-0.65
BUSI-benign	5%	3/3	0.6789+/-0.0285 / 0.5921+/-0.0372	0.4333+/-0.0832 / 0.3537+/-0.0576	32.64+/-1.97 / 18.45+/-0.07
BUSI-benign	10%	3/3	0.7249+/-0.0295 / 0.6443+/-0.0333	0.5081+/-0.0451 / 0.4054+/-0.0327	29.49+/-4.03 / 15.79+/-2.94
BUSI-benign	20%	3/3	0.7472+/-0.0364 / 0.6697+/-0.0300	0.5726+/-0.0328 / 0.4519+/-0.0239	30.83+/-0.63 / 12.43+/-2.47
BUSI-malignant	5%	3/3	0.6558+/-0.0193 / 0.5246+/-0.0113	0.1728+/-0.0229 / 0.1324+/-0.0162	48.09+/-2.42 / 17.98+/-4.02
BUSI-malignant	10%	3/3	0.6614+/-0.0281 / 0.5323+/-0.0255	0.1917+/-0.0201 / 0.1461+/-0.0162	45.30+/-3.23 / 16.79+/-1.83
BUSI-malignant	20%	2/3	0.7085+/-0.0076 / 0.5829+/-0.0103	0.2177+/-0.0213 / 0.1674+/-0.0163	39.21+/-0.46 / 14.38+/-0.46

Against the metric-wise strongest non-CPPA baseline, CPPA forms an envelope over all four overlap and boundary metrics in 8 of 9 dataset-ratio groups (Table 3). The exception is BUSI-malignant at 10% labels, where CPS-style is ahead by 0.0047 Dice, 0.0013 IoU, 0.0050 BF1, and 0.0023 B-IoU. Distance metrics are less uniform: CPPA loses HD95 at CVC 5% and BUSI-benign 20%, and loses ASSD at CVC 5%, BUSI-benign 5%, and BUSI-malignant 10%.

Table 3. Difference from the metric-wise strongest non-CPPA baseline. Positive values favor CPPA; distance changes are baseline error minus CPPA error.

Dataset	Labels	Overlap (Dice / IoU)	Boundary (BF1 / B-IoU)	Distance (HD95 / ASSD)
CVC-ClinicDB	5%	+0.0086 / +0.0155	+0.0563 / +0.0336	-3.314 / -0.593
CVC-ClinicDB	10%	+0.0094 / +0.0117	+0.0229 / +0.0215	+1.010 / +0.316
CVC-ClinicDB	20%	+0.0061 / +0.0092	+0.0290 / +0.0196	+0.262 / +0.766
BUSI-benign	5%	+0.0510 / +0.0646	+0.1219 / +0.0988	+3.028 / +0.994
BUSI-benign	10%	+0.0421 / +0.0622	+0.1410 / +0.1006	+4.510 / +0.473
BUSI-benign	20%	+0.0084 / +0.0107	+0.0652 / +0.0393	-3.641 / +0.494
BUSI-malignant	5%	+0.0056 / +0.0058	+0.0095 / +0.0088	+1.359 / +1.130
BUSI-malignant	10%	-0.0047 / -0.0013	-0.0050 / -0.0023	+1.481 / -0.286
BUSI-malignant	20%	+0.0297 / +0.0349	+0.0332 / +0.0265	+3.313 / +1.080

5.3. Ablation of the Evidence Chain

Table 4 isolates the mechanism on three cohorts and three seeds. The candidate oracle shows substantial bounded-repair headroom. Selection recovers a smaller but positive fraction of that headroom, with stronger relative gains for BF1 and B-IoU than for Dice. Naive fine-tuning on the selected mask is harmful, whereas the fixed absorption route and the full validation router recover positive overlap and boundary gains. This supports the distinction between pseudo-label quality and parameter-transfer quality.

Table 4. Full-pipeline ablation on CVC-ClinicDB 20%, BUSI-benign 10%, and BUSI-malignant 10%. Values are mean changes from the validation-selected source.

Stage	Output	Overlap (Dice / IoU)	Boundary (BF1 / B-IoU)	Distance (HD95 / ASSD)
Frozen source	source mask	+0.0000 / +0.0000	+0.0000 / +0.0000	+0.000 / +0.000
Frequency-Snake candidates (oracle)	mask	+0.0276 / +0.0348	+0.0769 / +0.0545	+3.151 / +1.422
CPPA selection	mask	+0.0073 / +0.0102	+0.0274 / +0.0208	+0.517 / +0.294
Ordinary absorption	one network	-0.0138 / -0.0170	-0.0445 / -0.0325	-1.157 / -0.260
Absorption bank, fixed route	one network	+0.0160 / +0.0223	+0.0474 / +0.0345	+1.342 / +0.689
Full CPPA + router	one network	+0.0211 / +0.0298	+0.0600 / +0.0441	+2.500 / +0.443

5.4. Frozen External Validation

Table 5 evaluates the CVC-ClinicDB operating point unchanged on CVC-ColonDB. Deployable Dice improves by 0.0110, 0.0054, and 0.0066 at 5%, 10%, and 20% labels, while BF1 improves by 0.0090, 0.0292, and 0.0290. All seven internally promoted networks also improve external Dice, while the two rejected 5% networks retain their frozen source predictors. At 20% labels, five metrics improve but ASSD worsens by 1.722, showing that overlap and boundary gains can coexist with a few distant contour errors.

Table 5. Frozen external validation on 380 CVC-ColonDB images. All operating decisions are fixed on CVC-ClinicDB.

Train data	Labels	Output	Overlap (Dice / IoU)	Boundary (BF1 / B-IoU)	Distance (HD95 / ASSD)
CVC-ClinicDB	5%	Source	0.6425+/-0.0450 / 0.5481+/-0.0380	0.3804+/-0.0130 / 0.2964+/-0.0108	55.96+/-10.87 / 21.60+/-4.96

CVC-ClinicDB	5%	CPPA deployable	0.6536+/-0.0521 / 0.5610+/-0.0477	0.3894+/-0.0228 / 0.3055+/-0.0204	50.70+/-16.39 / 20.62+/-5.80
CVC-ClinicDB	10%	Source	0.7321+/-0.0109 / 0.6400+/-0.0112	0.4858+/-0.0063 / 0.3801+/-0.0026	37.81+/-2.38 / 14.24+/-0.82
CVC-ClinicDB	10%	CPPA deployable	0.7375+/-0.0097 / 0.6481+/-0.0114	0.5149+/-0.0142 / 0.3999+/-0.0118	36.83+/-1.65 / 13.97+/-0.93
CVC-ClinicDB	20%	Source	0.7356+/-0.0086 / 0.6511+/-0.0081	0.5174+/-0.0128 / 0.4052+/-0.0086	37.12+/-1.82 / 14.72+/-1.79
CVC-ClinicDB	20%	CPPA deployable	0.7421+/-0.0010 / 0.6604+/-0.0038	0.5464+/-0.0105 / 0.4261+/-0.0069	36.76+/-1.44 / 16.44+/-1.52

5.5. Qualitative Audit and Operating Cost

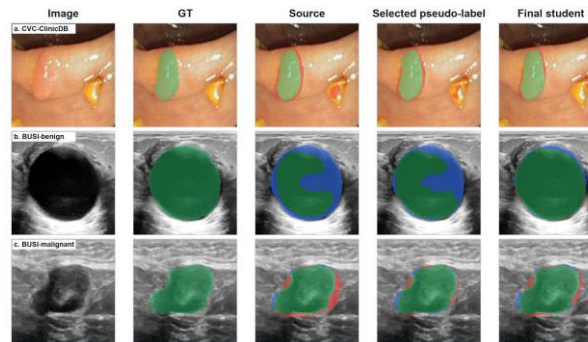


Figure 2. Representative repair-to-model examples from CVC-ClinicDB, BUSI-benign, and BUSI-malignant. The overlays show how selected contour edits are absorbed by the final student. Ground truth is used only for visualization.

The deployed network contains 34.07 million parameters and runs at 3.89 ms per 256 x 256 image on an RTX 4080. For CVC 20%, seed 42, candidate generation and policy selection require 270.0 s and 412.9 s offline. The bank contains 81 route outputs, with 80 retained checkpoints and an estimated 3.00 GPU hours. These costs are paid during training and not at deployment.

6. Discussion

The main result is a complete but staged repair-to-model pathway. Selected pseudo-labels improve source contours on average, and final single-model outputs amplify these gains after absorption. The nine-cell ablation shows why the final stage matters: ordinary fine-tuning degrades the source, whereas sparse edit-aware absorption and route diversity retain positive gains. The validation router is therefore part of the method, not an implementation detail.

The strongest evidence concerns boundary behavior. Selection captures 59.19% and

58.07% of the available BF1 and B-IoU oracle gains, compared with 41.35% of Dice oracle gain. This asymmetry is consistent with the method design: disagreement-supported, low-area arcs are suited to contour repair, but they cannot recover a large missed lesion. Improving Dice capture without sacrificing abstention is the clearest optimization target.

The comparison also argues for source-relative evaluation. UniMatch is stronger on CVC 20%, whereas CPPA is stronger on BUSI-benign 10%. We therefore interpret CPPA as a complementary repair protocol that can improve a strong validated source, not as a universal replacement or an unconditional state-of-the-art claim.

7. Limitations

The method remains source-dependent. A bounded Snake cannot recover a completely missed lesion outside the source envelope, and correlated source errors limit the candidate oracle. BUSI-malignant is small and retains low absolute boundary scores. The ColonDB test is a genuine frozen external evaluation, but it remains within colonoscopy and does not establish cross-modality generalization. External ASSD degradation at 20% labels further shows that a few distant contour errors can coexist with local overlap gains.

CPPA is also staged rather than jointly end-to-end. Source training, candidate generation, policy learning, and absorption are separate, and the absorption bank adds training cost even though deployment uses one network. More author-recipe baselines, label ratios, independent cohorts, and calibration studies are needed before making broader claims.

8. Conclusion

We presented CPPA, a boundary-selective pseudo-label repair and model-absorption protocol for semi-supervised medical image segmentation. Its innovation is the explicit separation of local edit selection from parameter transfer: a frozen source defines the no-edit action, bounded candidates expose measurable repair headroom, a counterfactual arc policy abstains when evidence is unsafe, and edit-aware absorption stores accepted corrections in one network. Across three internal cohorts and a frozen CVC-ColonDB transfer, the protocol improves source-relative overlap and boundary quality on average while retaining a conservative fallback. The remaining challenges are complete-miss recovery, higher Dice oracle capture, lower training cost, and broader external validation.

References

- [1] Al-Dhabyani, W., Gomaa, M., Khaled, H. and Fahmy, A. (2020) Dataset of Breast Ultrasound Images. Data in Brief, 28, 104863.
- [2] Bai, Y., Chen, D., Li, Q., Shen, W. and Wang, Y. (2023) Bidirectional Copy-Paste for Semi-Supervised Medical Image Segmentation. Proceedings of CVPR, 11514-11524.
- [3] Bernal, J., Sanchez, F. J., Fernandez-Esparrach, G., Gil, D., Rodriguez, C. and Vilarino, F.

- (2015) WM-DOVA Maps for Accurate Polyp Highlighting in Colonoscopy. *Computerized Medical Imaging and Graphics*, 43, 99-111.
- [4] Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A. and Raffel, C. (2019) MixMatch: A Holistic Approach to Semi-Supervised Learning. *NeurIPS*.
 - [5] Chen, H., Tao, R., Fan, Y., et al. (2023) SoftMatch: Addressing the Quantity-Quality Tradeoff in Semi-Supervised Learning. *ICLR*.
 - [6] Chen, X., Yuan, Y., Zeng, G. and Wang, J. (2021) Semi-Supervised Semantic Segmentation with Cross Pseudo Supervision. *CVPR*.
 - [7] Chen, Y., Yang, Z., Shen, C., et al. (2024) Evidence-Based Uncertainty-Aware Semi-Supervised Medical Image Segmentation. *Computers in Biology and Medicine*, 170, 108004.
 - [8] Cheng, B., Girshick, R., Dollar, P., Berg, A. C. and Kirillov, A. (2021) Boundary IoU: Improving Object-Centric Image Segmentation Evaluation. *CVPR*.
 - [9] Fan, D.-P., Ji, G.-P., Zhou, T., et al. (2020) PraNet: Parallel Reverse Attention Network for Polyp Segmentation. *MICCAI*.
 - [10] Karimi, D. and Salcudean, S. E. (2020) Reducing the Hausdorff Distance in Medical Image Segmentation with Convolutional Neural Networks. *IEEE TMI*, 39, 499-513.
 - [11] Kervadec, H., Bouchtiba, J., Desrosiers, C., et al. (2021) Boundary Loss for Highly Unbalanced Segmentation. *Medical Image Analysis*, 67, 101851.
 - [12] Laine, S. and Aila, T. (2017) Temporal Ensembling for Semi-Supervised Learning. *ICLR*.
 - [13] Li, M., Wu, R., Liu, H., et al. (2023) InstantT: Semi-Supervised Learning with Instance-Dependent Thresholds. *NeurIPS*.
 - [14] Li, S., Zhang, C. and He, X. (2020) Shape-Aware Semi-Supervised 3D Semantic Segmentation for Medical Images. *MICCAI*.
 - [15] Luo, X., Chen, J., Song, T. and Wang, G. (2021) Semi-Supervised Medical Image Segmentation through Dual-Task Consistency. *AAAI*, 8801-8809.
 - [16] Luo, X., Wang, G., Liao, W., et al. (2022) Semi-Supervised Medical Image Segmentation via Uncertainty Rectified Pyramid Consistency. *Medical Image Analysis*, 80, 102517.
 - [17] Olsson, V., Tranheden, W., Pinto, J. and Svensson, L. (2021) ClassMix: Segmentation-Based Data Augmentation for Semantic Segmentation. *WACV*, 1368-1377.
 - [18] Ouali, Y., Hudelot, C. and Tami, M. (2020) Semi-Supervised Semantic Segmentation with Cross-Consistency Training. *CVPR*.
 - [19] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. *MICCAI*, 234-241.
 - [20] Shi, Y., Zhang, J., Ling, T., et al. (2022) Inconsistency-Aware Uncertainty Estimation for Semi-Supervised Medical Image Segmentation. *IEEE TMI*, 41, 608-620.
 - [21] Sohn, K., Berthelot, D., Li, C.-L., et al. (2020) FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *NeurIPS*.
 - [22] Tarvainen, A. and Valpola, H. (2017) Mean Teachers are Better Role Models. *NeurIPS*.
 - [23] Vázquez, D., Bernal, J., Sanchez, F. J., et al. (2017) A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images. *Journal of Healthcare Engineering*, 2017, 4037190.
 - [24] Wang, Y., Chen, H., Heng, Q., et al. (2023) FreeMatch: Self-Adaptive Thresholding for Semi-Supervised Learning. *ICLR*.
 - [25] Wang, Y., Wang, H., Shen, Y., et al. (2022) Semi-Supervised Semantic Segmentation Using Unreliable Pseudo Labels. *CVPR*.
 - [26] Wu, Y., Ge, Z., Zhang, D., et al. (2022) Mutual Consistency Learning for Semi-Supervised Medical Image Segmentation. *Medical Image Analysis*, 81, 102530.
 - [27] Yang, L., Qi, L., Feng, L., Zhang, W. and Shi, Y. (2023) Revisiting Weak-to-Strong Consistency in Semi-Supervised Semantic Segmentation. *CVPR*.

- [28] Yang, L., Zhuo, W., Qi, L., Shi, Y. and Gao, Y. (2022) ST++: Make Self-Training Work Better for Semi-Supervised Semantic Segmentation. CVPR, 4268-4277.
- [29] Yu, L., Wang, S., Li, X., et al. (2019) Uncertainty-Aware Self-Ensembling Model for Semi-Supervised 3D Left Atrium Segmentation. MICCAI.
- [30] Zhang, B., Wang, Y., Hou, W., et al. (2021) FlexMatch: Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling. NeurIPS.
- [31] Zhao, X., Qi, Z., Wang, S., et al. (2024) RCPS: Rectified Contrastive Pseudo Supervision for Semi-Supervised Medical Image Segmentation. IEEE JBHI.
- [32] Zou, Y., Zhang, Z., Zhang, H., et al. (2021) PseudoSeg: Designing Pseudo Labels for Semantic Segmentation. ICLR.