Quality-Guided Semi-Supervised Learning for Medical Image Segmentation

Kumar Abhishek[0000−0002−7341−9617] and Ghassan Hamarneh[0000−0001−5040−7448]

School of Computing Science, Simon Fraser University, Canada
{kabhishe,hamarneh}@sfu.ca

Abstract. Training accurate medical image segmentation models requires large amounts of densely annotated data, which is costly and time-consuming to obtain. Semi-supervised learning (SSL) alleviates this by learning from both abundant unlabeled data and limited labeled data. However, most modern SSL methods rely on pseudolabels for unlabeled data, and typically assess their reliability through model confidence or uncertainty, measures that are self-referential and lack explicit grounding in segmentation quality. Instead, we propose a *quality*-guided SSL framework that trains a dedicated network to estimate segmentation quality from image-mask pairs. The predictor is trained on variable-quality masks generated through synthetic corruptions augmented with imperfect outputs from partially trained segmentation models, capturing realistic error patterns encountered during training. We integrate the quality predictor into SSL through two complementary mechanisms: a quality-aware regularization loss and a quality-based pseudolabel sample reweighting scheme. We show that our method serves as a drop-in enhancement to existing SSL frameworks. Extensive experiments across five datasets and multiple architectures demonstrate consistent improvements over competing SSL methods, advancing the state-of-the-art in semi-supervised medical image segmentation.

Keywords: Image segmentation · Segmentation quality · Semi-supervised learning.

1 Introduction

Accurate segmentation of medical images is fundamental to clinical workflows, yet dense pixelwise annotations, necessary for training deep learning-based segmentation models, remain costly and scarce [22,38,1]. Semi-supervised learning (SSL) addresses this annotation scarcity by leveraging abundant unlabeled data alongside limited labels, and has become a dominant paradigm for label-efficient medical image segmentation.

Most existing SSL approaches differ in how they leverage unlabeled data, falling into three broad categories: (i) consistency regularization, enforcing prediction invariance under perturbations, notably mean teacher (MT) [39], its

uncertainty-aware extension (UA-MT) [42], and interpolation consistency training (ICT) [41]; (ii) pseudolabel methods, using confident predictions as surrogate supervision [19,35], extended by cross-pseudo supervision (CPS) [9]; and (iii) contrastive learning, leveraging representation-level objectives on unlabeled data [7]. Across these methods, unlabeled samples are either treated uniformly regardless of prediction quality (MT, ICT), or filtered using model-derived confidence as a proxy for reliability (UA-MT, FixMatch [35]). Although calibration techniques can reduce overconfidence [15], medical segmentation networks remain poorly calibrated in practice [24]. More fundamentally, even perfectly calibrated confidence is self-referential, since it reflects the model's belief about its own prediction, and cannot catch systematic errors arising from the same representations that produced them. We argue that an independent assessment of segmentation quality is a better alternative to model certainty for guiding SSL training.

Predicting segmentation quality without ground truth has been studied for clinical quality control. Early work predicted Dice scores from hand-crafted features [18] and reverse classifiers to estimate quality [40]. More recent approaches directly regress quality metrics from image-segmentation pairs [30,10,28], and large-scale models now offer general-purpose quality prediction across diverse anatomies [33]. However, this entire line of work treats quality prediction as an end goal, filtering unreliable segmentations post-hoc. No prior work has leveraged learned quality prediction to guide semi-supervised training itself.

We bridge these two directions by training a quality predictor that estimates segmentation quality from image-mask pairs, then using it to provide learning signal for unlabeled data in SSL. Unlike confidence from a single forward pass, predicted quality provides a complementary signal by comparing mask structure against image evidence, independently of the segmentation network's own representations. While related to sample reweighting for noisy labels [29,34], our predictor directly estimates segmentation accuracy rather than inferring importance from per-sample loss values. However, a quality predictor trained on variable quality masks from labeled data must generalize to real network predictions of unlabeled data.

To provide a quality-based guidance for medical image segmentation in a semi-supervised setting, we contribute: (1) the first framework-agnostic method to leverage learned segmentation quality prediction for guiding SSL; (2) two complementary mechanisms for integrating quality predictions into SSL training: a differentiable quality regularizer and a pseudolabel reweighting scheme, applicable as drop-in enhancements to existing SSL methods; (3) a mask corruption strategy incorporating partially-trained model predictions that exhibit characteristic neural network errors [5,17] to bridge the distribution gap between synthetic and real network errors; and (4) we perform comprehensive experiments across five cross-dataset pairs spanning dermatology and colonoscopy, five SSL paradigms, and multiple model architectures, yielding consistent improvements over state-of-the-art baselines. Our code is publicly available at https://github.com/sfu-mial/QG-SSL.

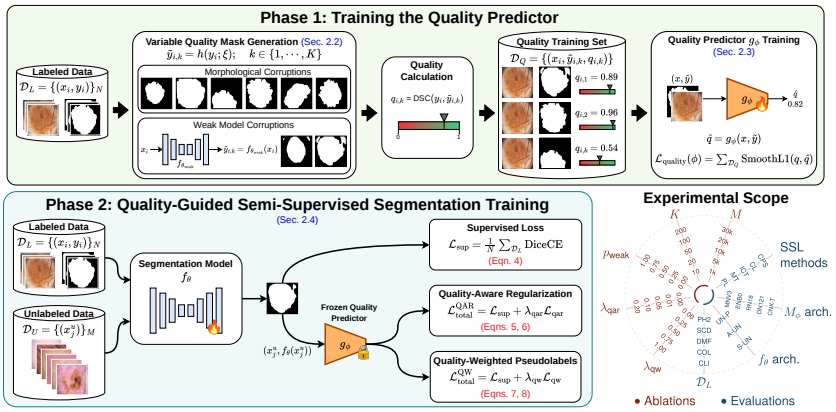

Fig. 1. An overview of the proposed quality-guided semi-supervised segmentation methods, along with the scope of experiments present in this paper.

2 Method

Our proposed approach has two phases: (Phase 1) training a quality predictor g_ϕ to estimate segmentation quality from image-mask pairs, and (Phase 2) using the frozen g_ϕ to guide semi-supervised segmentation training. Fig. 1 provides an overview and the scope of experiments presented in this paper.

2.1 Problem Definition

Let $\mathcal{D}_L = \{(x_i, y_i)\}_{i=1}^N$ denote a small labeled set, where $x_i \in \mathbb{R}^{H\times W\times C}$ is an image and $y_i \in \{0,1\}^{H\times W}$ is its ground truth binary segmentation mask, and let $\mathcal{D}_U = \{x_j^u\}_{j=1}^M$ ($M \gg N$) be a larger unlabeled set. Our goal is to train a segmentation network f_θ that leverages both $\mathcal{D}_L$ and $\mathcal{D}_U$. In our experiments, $\mathcal{D}_L$ and $\mathcal{D}_U$ come from related but distinct sources (e.g., PH2 [25] and ISIC2020 [32]), a challenging setting that better reflects clinical practice. We do not assume matched distributions between labeled and unlabeled data.

2.2 Variable Quality Mask Generation

To train g_ϕ, we require images paired with arbitrary masks, where each mask has a corresponding quality score. We construct a synthetic dataset $\mathcal{D}_Q$ from $\mathcal{D}_L$ by generating, for each $(x_i, y_i) \in \mathcal{D}_L$, a set of K degraded masks using a stochastic corruption function h:

$$\tilde{y}_{i,k} = h(y_i;\, \xi_k), \qquad q_{i,k} = \mathrm{DSC}(y_i,\, \tilde{y}_{i,k}), \qquad k \in \{1, \cdots K\}, \tag{1}$$

where ξ_k represents random perturbation parameters, and $q_{i,k} \in [0,1]$ is the Dice score (DSC) of each corrupted mask. We sample two types of degradations randomly. First (Type 1), we use random morphological operations (erosion/dilation with different kernel sizes), translations, elastic deformations, additive noise, and boundary perturbations. However, morphological corruptions

alone may not capture error patterns produced by real neural networks during semi-supervised training. To bridge this distribution gap, (Type 2) we augment our corruption strategy with predictions from partially trained (weak) segmentation models $f_{\theta_{\text{weak}}}$. We train a U-Net [31] on $\mathcal{D}_L$ from random initialization and collect checkpoints at early epochs (epochs 1, 3, 5, 10, 15, 20). These weak models produce segmentation predictions exhibiting characteristic early training failure patterns [17,5]. Previous work on learned quality prediction has also noted that limited corruption diversity restricts the predictor's sensitivity to fine-grained quality differences [30], further motivating the inclusion of real network outputs of imperfect segmentation models. Therefore, when training g_ϕ, we sample with probability p_{weak} from these weak-model predictions (Type 2) and $1 - p_{\text{weak}}$ from Type 1 corruptions. The resulting dataset $\mathcal{D}_Q = \{(x_i, \tilde{y}_{i,k}, q_{i,k})\}$ contains image-mask-quality triplets spanning the full range of Dice scores. We explore various settings of p_{weak} and K.

2.3 Segmentation Quality Predictor

The quality predictor g_ϕ takes an image-mask pair, outputs a scalar quality estimate (Eqn. 2) and is trained to minimize a regression loss ℓ_{reg} (Eqn. 3):

$$\hat{q} = g_\phi(x, \tilde{y}), \tag{2}$$

$$\mathcal{L}_{\text{quality}}(\phi) = \sum_{(x,\, \tilde{y},\, q)\, \in\, \mathcal{D}_Q} \ell_{\text{reg}}(g_\phi(x, \tilde{y}),\ q)\,. \tag{3}$$

A key property distinguishing quality prediction from proxy signals such as model confidence is that it is designed to be contextually grounded: to assess whether $\tilde{y}$ is a good segmentation of x, g_ϕ must compare mask structure against visual evidence in the image, rather than relying on the mask or the model's internal state alone. Once trained, g_ϕ is frozen and acts as a differentiable quality assessment function for any image-mask pair without requiring ground truth.

2.4 Quality-Guided Semi-Supervised Training

We now train f_θ using both $\mathcal{D}_L$ and $\mathcal{D}_U$. For all labeled samples, we minimize:

$$\mathcal{L}_{\text{sup}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell_{\text{seg}}(f_\theta(x_i),\ y_i)\,. \tag{4}$$

For unlabeled data, we propose two alternative mechanisms for leveraging the frozen g_ϕ, differing in whether the segmentation loss ℓ_{seg} gradients propagate through g_ϕ (QAR) or g_ϕ serves only to compute per-sample weights (PL-QW).

A: Quality-Aware Regularization (QAR): For each unlabeled sample x_j^u, the soft prediction $f_\theta(x_j^u)$ is passed into g_ϕ. No explicit pseudolabels are generated; instead, gradients of the loss $\mathcal{L}_{\text{qar}}$ propagate from the scalar quality output

back through the predicted mask into segmentation model parameters θ, encouraging f_θ to produce segmentations that g_ϕ judges as high quality:

$$\mathcal{L}_{\text{qar}}(\theta) = \frac{1}{M}\sum_{j=1}^{M}\left(1 - g_\phi\big(x_j^u, f_\theta(x_j^u)\big)\right). \tag{5}$$

The complete objective is a weighted sum of the two losses:

$$\mathcal{L}_{\text{total}}^{\text{QAR}} = \mathcal{L}_{\text{sup}} + \lambda_{\text{qar}}\,\mathcal{L}_{\text{qar}}. \tag{6}$$

B: Quality-Weighted Pseudolabels (PL-QW): Given pseudolabels $\hat{y}_j^u$ for unlabeled samples $x_j^u \in \mathcal{D}_U$, we weight the per-sample loss by predicted pseudolabel quality:

$$\mathcal{L}_{\text{qw}}(\theta) = \frac{1}{M}\sum_{j=1}^{M} w_j \cdot \ell_{\text{seg}}\big(f_\theta(x_j^u),\ \hat{y}_j^u\big), \tag{7}$$

where $w_j = g_\phi(x_j^u, \hat{y}_j^u)$ is computed with g_ϕ frozen and detached from the computational graph. Unlike QAR, no gradients flow through g_ϕ; it serves purely as a sample weighting function that upweights high-quality pseudolabels and downweights unreliable ones. The complete objective is:

$$\mathcal{L}_{\text{total}}^{\text{QW}} = \mathcal{L}_{\text{sup}} + \lambda_{\text{qw}}\,\mathcal{L}_{\text{qw}}. \tag{8}$$

A key property of this formulation is its orthogonality to the choice of semi-supervised method: any approach generating pseudolabels $\hat{y}_j^u$ can be augmented by weighting per-sample losses with w_j, without requiring architectural changes.

3 Results and Discussion

Datasets: We study 5 medical image segmentation datasets from two modalities as labeled data $\mathcal{D}_L$: PH2 (N=200) [25], Skin Cancer Detection (SCD; N=206) [14], and DermoFit (DMF; N=1,300) [2] for skin lesion segmentation, and CVC-ColonDB (COL; N=380) [3] and CVC-ClinicDB (CLI; N=612) [4] for polyp segmentation in colonoscopy. All datasets are split 70:10:20 for train, validation, and test. As unlabeled data $\mathcal{D}_U$, we use ISIC2020-Train (M=33,126) [32] and Polyp-Box-Seg (M=4,070) [8] for dermatology and colonoscopy respectively: both drawn from different sources than $\mathcal{D}_L$. We use 5,000 ISIC2020 images for main experiments (Table 1) and the remainder for hyperparameter sensitivity analyses (Table 2). All models were trained on Ubuntu 22.04 with Intel Core i9-14900K, 64GB RAM, NVIDIA RTX4090, Python 3.10.19, and PyTorch 2.9.0.
Quality predictor training and evaluation: We implement g_ϕ as a ResNet-18 encoder, that takes the channel-wise concatenation of image-mask pairs as input, with a regression head (dropout of 0.15), trained for 150 epochs with AdamW [23] (learning rate=3e-4, weight decay=5e-4, batch size 32) with cosine

Table 1. Quantitative results (DSC and IoU; mean$_{std.err.}$) for all SSL baselines, their quality-weighted versions, and QAR. Using a quality predictor consistently improves segmentation performance across 5 datasets and 3 segmentation model architectures.

$\mathcal{D}_L$	f_θ Arch.		*Pseudolabels*				*Student-Teacher Models*			*Interp. Consistency*		*Contrastive*		*Cross Pseudo Sup.*		QAR
			SUP	PL-T [19]	PL-C [35]	PL-QW (Ours)	MT [39]	UA-MT [42]	MT-QW (Ours)	ICT [41]	ICT-QW (Ours)	CL [7]	CL-QW (Ours)	CPS [9]	CPS-QW (Ours)	(Ours)
PH2	UN-P	DSC	92.93$_{0.52}$	93.21$_{0.88}$	93.92$_{1.10}$	95.46$_{0.45}$	93.53$_{1.62}$	94.33$_{0.62}$	95.32$_{0.43}$	93.85$_{0.82}$	95.18$_{0.46}$	92.37$_{1.82}$	94.48$_{0.56}$	94.20$_{0.71}$	95.19$_{0.52}$	95.48$_{0.48}$
		IoU	86.96$_{0.89}$	87.75$_{1.41}$	89.19$_{1.58}$	91.45$_{0.79}$	89.04$_{2.00}$	89.51$_{1.04}$	91.18$_{0.76}$	88.81$_{1.30}$	90.95$_{0.81}$	87.23$_{2.12}$	89.75$_{0.97}$	89.46$_{1.16}$	91.01$_{0.91}$	91.50$_{0.84}$
	A-UN	DSC	91.73$_{0.69}$	92.90$_{0.67}$	93.34$_{0.81}$	95.40$_{0.45}$	93.18$_{1.60}$	93.70$_{0.64}$	95.38$_{0.44}$	94.16$_{0.58}$	95.17$_{0.49}$	92.91$_{1.88}$	95.09$_{0.48}$	93.91$_{0.70}$	94.99$_{0.52}$	95.29$_{0.47}$
		IoU	85.01$_{1.12}$	87.17$_{1.12}$	87.92$_{1.32}$	91.33$_{0.80}$	88.40$_{1.98}$	88.41$_{1.08}$	91.31$_{0.78}$	89.18$_{0.99}$	90.95$_{0.86}$	88.28$_{2.20}$	90.80$_{0.84}$	88.83$_{1.16}$	90.64$_{0.89}$	91.15$_{0.83}$
	S-UN	DSC	93.02$_{0.55}$	94.50$_{0.75}$	94.86$_{0.50}$	95.57$_{0.41}$	94.90$_{0.47}$	94.87$_{0.46}$	95.36$_{0.44}$	94.69$_{0.50}$	95.26$_{0.40}$	93.63$_{1.65}$	95.70$_{0.39}$	94.32$_{0.49}$	95.45$_{0.44}$	95.58$_{0.42}$
		IoU	87.14$_{0.94}$	90.05$_{1.22}$	90.39$_{0.87}$	91.63$_{0.74}$	90.63$_{0.84}$	90.30$_{0.81}$	91.27$_{0.78}$	90.08$_{0.89}$	91.05$_{0.72}$	89.21$_{1.95}$	91.86$_{0.70}$	89.42$_{0.85}$	91.42$_{0.78}$	91.66$_{0.75}$
SCD	UN-P	DSC	89.88$_{0.95}$	91.39$_{1.84}$	91.87$_{1.16}$	93.18$_{0.73}$	91.46$_{1.19}$	92.20$_{0.93}$	92.91$_{0.88}$	91.86$_{1.31}$	93.61$_{0.50}$	92.37$_{1.13}$	92.67$_{1.02}$	90.06$_{0.90}$	93.26$_{0.75}$	93.24$_{0.69}$
		IoU	82.15$_{1.46}$	85.70$_{2.21}$	85.73$_{1.71}$	87.57$_{1.18}$	85.08$_{1.76}$	86.05$_{1.45}$	87.22$_{1.35}$	85.87$_{1.84}$	88.16$_{0.87}$	86.52$_{1.59}$	86.96$_{1.54}$	82.39$_{1.40}$	87.72$_{1.21}$	87.63$_{1.11}$
	A-UN	DSC	90.89$_{0.89}$	91.71$_{1.68}$	91.83$_{1.02}$	93.15$_{0.76}$	92.23$_{1.32}$	92.40$_{0.90}$	93.06$_{0.81}$	92.34$_{0.98}$	93.04$_{0.73}$	92.42$_{0.82}$	92.64$_{0.88}$	92.64$_{0.72}$	93.62$_{0.65}$	93.11$_{0.72}$
		IoU	83.78$_{1.39}$	85.99$_{2.02}$	85.50$_{1.54}$	87.54$_{1.21}$	86.41$_{1.73}$	86.35$_{1.39}$	87.42$_{1.27}$	86.34$_{1.51}$	87.31$_{1.17}$	86.33$_{1.50}$	86.76$_{1.38}$	86.61$_{1.16}$	88.26$_{1.04}$	87.44$_{1.17}$
	S-UN	DSC	91.54$_{0.58}$	92.60$_{0.87}$	92.69$_{0.70}$	93.84$_{0.47}$	92.71$_{0.53}$	93.30$_{0.55}$	93.69$_{0.46}$	92.94$_{0.45}$	93.64$_{0.52}$	93.07$_{0.65}$	93.71$_{0.47}$	93.30$_{0.54}$	93.72$_{0.49}$	93.77$_{0.46}$
		IoU	84.62$_{0.96}$	86.70$_{1.36}$	86.60$_{1.15}$	88.55$_{0.81}$	86.59$_{0.89}$	87.80$_{0.93}$	88.27$_{0.80}$	86.90$_{0.78}$	88.22$_{0.89}$	87.31$_{1.07}$	88.31$_{0.81}$	87.70$_{0.91}$	88.35$_{0.84}$	88.41$_{0.79}$
DMF	UN-P	DSC	90.00$_{0.52}$	90.10$_{0.51}$	90.30$_{0.50}$	91.34$_{0.44}$	90.53$_{0.50}$	90.61$_{0.47}$	91.24$_{0.44}$	90.91$_{0.47}$	91.35$_{0.43}$	91.04$_{0.46}$	91.01$_{0.45}$	90.44$_{0.49}$	91.22$_{0.43}$	91.34$_{0.43}$
		IoU	82.87$_{0.73}$	82.89$_{0.75}$	83.31$_{0.74}$	84.76$_{0.67}$	83.56$_{0.74}$	83.62$_{0.71}$	84.58$_{0.67}$	84.10$_{0.69}$	84.75$_{0.65}$	84.29$_{0.68}$	84.24$_{0.68}$	83.36$_{0.71}$	84.54$_{0.66}$	84.74$_{0.66}$
	A-UN	DSC	90.06$_{0.47}$	90.35$_{0.46}$	90.54$_{0.45}$	91.31$_{0.42}$	90.60$_{0.44}$	90.67$_{0.46}$	91.27$_{0.42}$	90.53$_{0.50}$	91.21$_{0.43}$	90.38$_{0.48}$	91.46$_{0.42}$	90.56$_{0.46}$	91.25$_{0.43}$	91.37$_{0.44}$
		IoU	82.68$_{0.69}$	83.15$_{0.69}$	83.44$_{0.68}$	84.65$_{0.65}$	83.68$_{0.68}$	83.69$_{0.70}$	84.59$_{0.64}$	83.57$_{0.74}$	84.50$_{0.66}$	83.26$_{0.71}$	84.92$_{0.65}$	83.49$_{0.69}$	84.58$_{0.66}$	84.81$_{0.67}$
	S-UN	DSC	90.47$_{0.45}$	91.17$_{0.45}$	91.21$_{0.42}$	91.94$_{0.41}$	91.24$_{0.43}$	91.17$_{0.43}$	91.67$_{0.42}$	91.21$_{0.45}$	91.84$_{0.41}$	91.22$_{0.45}$	91.67$_{0.42}$	91.60$_{0.40}$	92.08$_{0.40}$	91.89$_{0.40}$
		IoU	83.30$_{0.67}$	84.50$_{0.68}$	84.49$_{0.65}$	85.72$_{0.64}$	84.57$_{0.65}$	84.43$_{0.65}$	85.28$_{0.65}$	84.56$_{0.68}$	85.54$_{0.64}$	84.59$_{0.68}$	85.27$_{0.65}$	85.23$_{0.62}$	85.92$_{0.62}$	85.59$_{0.62}$
COL	UN-P	DSC	89.17$_{1.82}$	89.59$_{1.81}$	88.89$_{1.49}$	91.73$_{1.19}$	90.56$_{1.86}$	90.10$_{1.85}$	91.64$_{1.21}$	90.72$_{1.45}$	91.03$_{1.59}$	89.57$_{1.83}$	90.73$_{1.41}$	89.60$_{1.71}$	90.62$_{1.69}$	92.01$_{1.02}$
		IoU	82.93$_{2.05}$	83.53$_{1.95}$	81.99$_{1.96}$	86.01$_{1.56}$	85.23$_{1.97}$	84.40$_{1.94}$	85.93$_{1.60}$	84.81$_{1.81}$	85.46$_{1.78}$	83.69$_{2.08}$	84.79$_{1.82}$	83.45$_{1.99}$	85.04$_{1.90}$	86.24$_{1.45}$
	A-UN	DSC	89.25$_{1.94}$	88.38$_{1.91}$	89.19$_{1.88}$	91.37$_{1.47}$	91.40$_{1.01}$	89.76$_{1.25}$	91.42$_{1.60}$	90.10$_{1.90}$	90.96$_{1.73}$	89.00$_{2.28}$	90.57$_{1.74}$	89.84$_{1.71}$	90.29$_{1.68}$	91.20$_{1.50}$
		IoU	83.28$_{2.06}$	81.91$_{2.15}$	83.03$_{2.02}$	85.80$_{1.70}$	85.15$_{1.43}$	82.90$_{1.71}$	86.19$_{1.82}$	84.79$_{2.08}$	85.64$_{1.88}$	83.74$_{2.32}$	85.06$_{1.94}$	83.84$_{1.97}$	84.51$_{1.93}$	85.57$_{1.72}$
	S-UN	DSC	90.44$_{1.07}$	91.02$_{1.46}$	91.44$_{1.37}$	92.63$_{0.83}$	91.90$_{1.13}$	91.81$_{1.15}$	92.70$_{0.87}$	91.77$_{0.91}$	91.72$_{1.10}$	90.63$_{1.33}$	92.10$_{1.37}$	91.52$_{1.44}$	92.55$_{1.06}$	92.71$_{0.74}$
		IoU	83.64$_{1.49}$	85.17$_{1.68}$	85.68$_{1.56}$	86.96$_{1.22}$	86.14$_{1.45}$	86.01$_{1.45}$	87.16$_{1.25}$	86.54$_{1.34}$	85.81$_{1.46}$	84.35$_{1.63}$	86.76$_{1.32}$	86.00$_{1.67}$	87.18$_{1.42}$	87.01$_{1.15}$
CLI	UN-P	DSC	92.01$_{1.12}$	92.62$_{1.03}$	92.67$_{1.00}$	93.80$_{0.63}$	93.34$_{0.91}$	92.98$_{0.71}$	93.81$_{0.69}$	92.53$_{0.96}$	93.48$_{0.76}$	92.27$_{1.05}$	93.33$_{0.66}$	92.55$_{0.81}$	93.72$_{0.75}$	93.64$_{0.71}$
		IoU	86.80$_{1.28}$	87.70$_{1.25}$	87.78$_{1.27}$	88.97$_{0.89}$	88.61$_{1.09}$	87.69$_{1.00}$	89.09$_{0.95}$	87.30$_{1.13}$	88.64$_{1.01}$	87.23$_{1.32}$	88.21$_{0.94}$	87.15$_{1.11}$	89.00$_{0.98}$	88.85$_{0.98}$
	A-UN	DSC	92.42$_{1.08}$	92.56$_{1.08}$	92.74$_{0.73}$	93.61$_{0.70}$	93.04$_{1.01}$	93.12$_{0.82}$	93.98$_{0.68}$	93.00$_{1.01}$	93.67$_{0.73}$	92.42$_{1.02}$	93.57$_{0.71}$	92.56$_{0.88}$	93.36$_{0.75}$	94.01$_{0.57}$
		IoU	87.38$_{1.21}$	87.70$_{1.28}$	88.15$_{0.99}$	88.76$_{0.96}$	88.34$_{1.19}$	88.16$_{1.08}$	89.39$_{0.95}$	88.26$_{1.19}$	88.87$_{0.94}$	87.31$_{1.22}$	88.70$_{0.96}$	87.29$_{1.15}$	88.41$_{1.01}$	89.23$_{0.82}$
	S-UN	DSC	91.08$_{1.08}$	92.09$_{0.94}$	92.34$_{0.95}$	93.38$_{0.80}$	92.59$_{1.00}$	92.84$_{0.94}$	93.49$_{0.85}$	93.00$_{0.95}$	93.47$_{0.83}$	92.70$_{0.93}$	93.03$_{0.92}$	91.51$_{1.07}$	93.20$_{0.88}$	93.36$_{0.83}$
		IoU	85.17$_{1.29}$	86.60$_{1.19}$	87.09$_{1.22}$	88.53$_{1.03}$	87.61$_{1.24}$	87.90$_{1.19}$	88.83$_{1.09}$	88.20$_{1.19}$	88.75$_{1.07}$	87.70$_{1.13}$	88.17$_{1.15}$	85.90$_{1.50}$	88.41$_{1.14}$	88.58$_{1.09}$

annealing with warm restarts (initial period 10 epochs, doubles after each restart) and early stopping (validation loss; patience 25 epochs) to minimize SmoothL1 loss (Eqn. 3) [13]. We set $p_{\text{weak}} = 0.05$ (Sec. 2.2) and generate $K = 50$ degraded masks per sample (Eqn. 1). On test sets, g_ϕ achieves MAE in [0.043, 0.088] and Pearson's correlation coefficient $\rho > 0.92$ across all 5 datasets (Table 2 $\mathcal{A}$). Zeroing the image input in Eqn. 2, i.e., $g_\phi(x = \mathbf{0}, \tilde{y})$, increases MAE by an average of 0.399 ± 0.137 across all 5 datasets (MAE $\in [0, 1]$), strongly confirming that g_ϕ leverages image content (contextual grounding), rather than relying on the mask alone.

SSL segmentation training and evaluation: Next, we leverage our trained quality predictor to improve segmentation using semi-supervised learning (SSL). We evaluate 3 architectures for f_θ: a purely convolutional model, U-Net++ [44] (UN-P; 26.08M parameters, 14.14 GFLOPs), a convolutional model with attention gates, Attention U-Net [26] (A-UN; 24.71M params, 6.16 GFLOPs), and a pure Transformer-based architecture, Swin-U-net [6] (S-UN; 34.27M params, 7.55 GFLOPs), optimizing Dice + cross-entropy loss (Eqn. 4) [16,37] for 200 epochs with AdamW (lr and weight decay set to 1e-4), a cosine annealing scheduler, and early stopping (validation DSC; patience of 30 epochs). We set $\lambda_{\text{qar}} = 0.01$ and $\lambda_{\text{qw}} = 0.25$ (Eqn. 6,8) and follow a ramp-up schedule during initial training epochs [39], and report Dice (DSC) and Jaccard (IoU) averaged

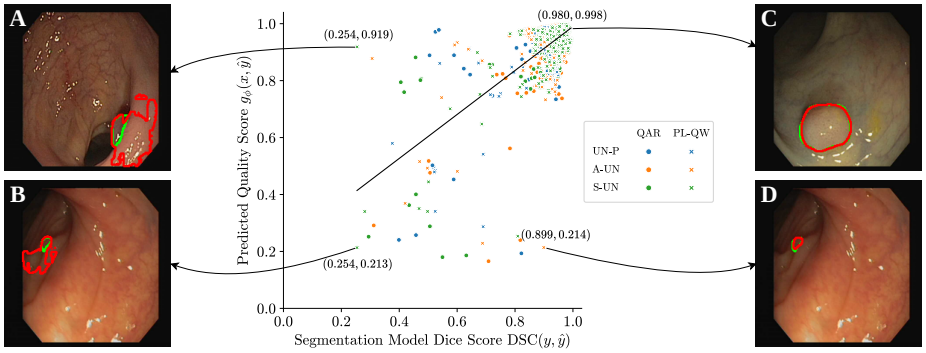

Fig. 2. Scatter plot of the segmentation predictions' Dice DSC$(y, \hat{y})$ and the corresponding predicted quality estimates $g_\phi(x, \hat{y})$ on the test set of CLI dataset, and four representative images (A-D) with the ground truth (green) and predicted (red) segmentations. We observe a strong, stat. sig., positive linear correlation (ρ=0.69; p=1e-314).

over 3 runs with different seeds. We compare QAR (Sec. 2.4 A) against existing popular and widely used SSL paradigms: pseudolabels [19] with pixel-level confidence thresholded at 0.9 (PL-T), sample-level confidence-weighted pseudolabels [35] (PL-C), mean teacher [39] (MT) and its uncertainty-aware extension [42] (UA-MT), interpolation consistency training [41] (ICT), pseudolabel-based contrastive learning [7] (CL), and cross-pseudo supervision [9] (CPS). We do not compare against Zheng et al. [43], despite them incorporating quality estimation into SSL, because their code is not available. Nevertheless, we note that their estimator is trained end-to-end with the segmentation model and coupled to a specific self-training pipeline, whereas ours is the first framework-agnostic approach: an independently trained quality predictor that enhances any pseudolabel-generating method without architectural changes or retraining. Table 1 shows that QAR outperforms all competing SSL paradigms across all datasets and models, indicating that segmentation quality, even when predicted, may still be a better training signal than model confidence or uncertainty. It is important to note that the last two competing methods, CL and especially CPS (as it trains two segmentation models in tandem), are computationally more demanding. Despite no architectural modifications, QAR matches or surpasses state-of-the-art results on these datasets [11,36,12,21,27,20], many of which employ architectural innovations orthogonal to ours, suggesting further gains from combination.

Quality weighting as a drop-in module: Next, we show how our quality-weighted pseudolabels can be integrated with any approach that generates pseudolabels $\hat{y}_j^u$ by weighting per-sample losses with w_j (Sec. 2.4 B). Concretely, for MT [39]: $\mathcal{L}_{\text{MT-QW}} = \frac{1}{M}\sum_j w_j \|f_\theta(x_j^u) - \hat{y}_j^u\|^2$; for CPS [9]: $\mathcal{L}_{\text{CPS-QW}} = \frac{1}{M}\sum_j [w_{j,2}\,\ell_{\text{seg}}(f_{\theta_1}(x_j^u), \hat{y}_{j,2}^u) + w_{j,1}\,\ell_{\text{seg}}(f_{\theta_2}(x_j^u), \hat{y}_{j,1}^u)]$; for ICT [41]: $\mathcal{L}_{\text{ICT-QW}} = \frac{1}{M}\sum_j w_j\,\ell_{\text{seg}}(f_\theta(\tilde{x}_j^u), \tilde{y}_j^u)$; and for CL [7]: $\mathcal{L}_{\text{CPL-QW}} = \frac{1}{M}\sum_j w_j \big[\ell_{\text{seg}}(f_\theta(x_j^u), \hat{y}_j^u) +$

Table 2. Ablation and hyperparameter sensitivity experiments. The values used for the main experiments table (Table 1) are highlighted. Results reported as mean$_{std.err.}$. Unless specified otherwise, $\mathcal{D}_L$ is PH2, $\mathcal{D}_U$ is ISIC2020-Train, g_ϕ is ResNet-18, f_θ is Swin-Unet, $p_{\text{weak}} = 0.05$, $K = 50$, $\lambda_{\text{qar}} = 0.01$, $\lambda_{\text{qw}} = 0.25$, and $M = 5,000$. e_{val96} denotes the number of training epochs for the validation DSC to reach 96%.

$\mathcal{A}$. g_ϕ's metrics for all datasets					
Dataset	PH2	SCD	DMF	COL	CLI
MAE / ρ	$0.043_{0.007}$ / 0.972	$0.052_{0.007}$ / 0.973	$0.060_{0.004}$ / 0.926	$0.074_{0.007}$ / 0.955	$0.088_{0.013}$ / 0.927
$\mathcal{B}$. Varying g_ϕ's backbones (Sec. 2.3)					
Backbone	MobileNetV3L	EfficientNet-B0	ResNet-18	DenseNet-121	ConvNeXt-T
Params / FLOPs	4.53M / 0.24G	4.34M / 0.42G	11.31M / 1.87G	7.22M / 2.95G	28.02M / 4.47G
MAE / ρ	$0.088_{0.012}$ / 0.849	$0.054_{0.009}$ / 0.961	$0.043_{0.007}$ / 0.972	$0.089_{0.011}$ / 0.837	$0.055_{0.008}$ / 0.955
$\mathcal{C}$. Varying p_{weak} (i.e., probability of sampling from weak models' predictions; Sec. 2.2)					
p_{weak}	0.00	0.05	0.10	0.15	0.20
MAE / ρ	$0.053_{0.008}$ / 0.960	$0.043_{0.007}$ / 0.972	$0.048_{0.007}$ / 0.965	$0.049_{0.007}$ / 0.953	$0.051_{0.009}$ / 0.948
$\mathcal{D}$. Varying K (i.e., num. augmented training samples; Eqn. 1)					
K	10	20	50	100	200
MAE / ρ	$0.061_{0.009}$ / 0.937	$0.048_{0.006}$ / 0.956	$0.043_{0.007}$ / 0.972	$0.047_{0.006}$ / 0.969	$0.050_{0.005}$ / 0.977
$\mathcal{E}$. Impact of $p_{\text{weak}} > 0$ on segmentation performance					
p_{weak}	0.00	0.05	p_{weak}	0.00	0.05
QAR: DSC	$94.91_{0.49}$	$95.58_{0.42}$	PL-QW: DSC	$95.10_{0.46}$	$95.57_{0.41}$
$\mathcal{F}$. Varying λ_{qar}, i.e., relative weight of $\mathcal{L}_{\text{qar}}$ (Eqn. 6)					
λ_{qar}	0.00	0.01	0.05	0.10	0.20
DSC	$93.02_{0.55}$	$95.58_{0.42}$	$95.52_{0.42}$	$95.14_{0.48}$	$95.03_{0.49}$
$\mathcal{G}$. Varying λ_{qw}, i.e., relative weight of $\mathcal{L}_{\text{qw}}$ (Eqn. 8)					
λ_{qw}	0.00	0.25	0.50	0.75	1.00
DSC	$93.02_{0.55}$	$95.57_{0.41}$	$95.25_{0.44}$	$95.31_{0.43}$	$95.14_{0.44}$
$\mathcal{H}$. Varying M, i.e., number of unlabeled samples (Eqns. 5–8)					
M	1,000	5,000	10,000	20,000	30,000
QAR: DSC / e_{val96}	$95.48_{0.43}$ / 13	$95.58_{0.42}$ / 2	$95.36_{0.43}$ / 1	$95.50_{0.43}$ / 1	$95.36_{0.45}$ / 1
PL-QW: DSC / e_{val96}	$95.48_{0.43}$ / 12	$95.57_{0.41}$ / 2	$95.35_{0.43}$ / 2	$95.48_{0.43}$ / 1	$95.35_{0.44}$ / 1

$\lambda_c \mathcal{L}_{\text{contrast}}(x_j^u, \hat{y}_j^u)\big]$, where quality weights both the pseudolabel loss and the contrastive loss. In all cases, $w_j = g_\phi(x_j^u, \hat{y}_j^u)$ (with $w_{j,k}$ evaluating the pseudolabel from network k in CPS). Table 1 shows that quality-weighted variants ($*$-QW) consistently outperform original versions across all but one setting (DMF + UN-P), confirming the general applicability of our quality-weighting. A scatter plot of the calculated DSC and the predicted quality (Fig. 2), on CLI's test set across all 3 runs of both QAR and PL-QW, shows that the two are strongly correlated. Example images (A-D) show two successful g_ϕ predictions for near-perfect (B) and poor (C) segmentations, and two failure cases for g_ϕ (A, D).

Ablation studies (Table 2)**:** Among 5 backbones from different architecture families, ResNet-18 performs the best for g_ϕ ($\mathcal{B}$). Incorporating weak-model corruptions (i.e., $p_{\text{weak}} > 0$) improves g_ϕ's performance ($\mathcal{C}$), and increasing corrupted masks per sample (i.e., K) helps up to a saturation point ($\mathcal{D}$). g_ϕ trained with weak-model corruptions helps improve segmentation model performance for both QAR and PL-QW ($\mathcal{E}$). Varying λ_{qar} and λ_{qw} ($\mathcal{F}, \mathcal{G}$) shows that even subop-

timal weights outperform the zero-weight baseline. Finally, increasing unlabeled data ($\mathcal{H}$) has minimal impact on final DSC, but substantially accelerates convergence: e_{val96} (i.e., the number of epochs to reach 96% validation DSC) decreases considerably with larger M.

4 Conclusion

We presented a contextually-grounded deep learning-based approach to estimating the quality of medical image segmentations. Our quality predictor is trained on corrupted masks generated using synthetic degradations and weak segmentation models' predictions. We then integrated our quality predictor into existing semi-supervised learning (SSL)-based segmentation frameworks through two complementary mechanisms: either as a regularization loss or as a sample reweighting mechanism, without any architectural modifications to the segmentation network. Extensive experiments across multiple datasets and model architectures demonstrated consistent improvements over existing SSL paradigms, confirming that learned quality prediction provides an effective training signal for leveraging unlabeled data. Future work could explore extending quality-guided SSL to multi-class segmentation, and leveraging quality predictions for active learning to identify unlabeled samples to be prioritized for expert annotation.

Acknowledgments. The authors thank Darren Sutton for initial discussions and acknowledge computational support from NVIDIA Corporation and the Digital Research Alliance of Canada. Partial funding for this project was provided by the Natural Sciences and Engineering Research Council of Canada (NSERC RGPIN-2020-06752).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Asgari Taghanaki, S., et al.: Deep semantic segmentation of natural and medical images: a review. Artif Intell Rev (2021)
2. Ballerini, L., et al.: A Color and Texture Based Hierarchical K-NN Approach to the Classification of Non-melanoma Skin Lesions. In: Color Medical Image Analysis (2013)
3. Bernal, J., et al.: Towards automatic polyp detection with a polyp appearance model. Pattern Recognit (2012)
4. Bernal, J., et al.: WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Comput Med Imaging Graph (2015)
5. Byrne, N., et al.: A persistent homology-based topological loss function for multi-class CNN segmentation of cardiac MRI. In: MICCAI STACOM (2021)
6. Cao, H., et al.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: ECCV (2022)
7. Chaitanya, K., et al.: Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. Med Image Anal (2023)
8. Chen, S., et al.: Weakly supervised polyp segmentation in colonoscopy images using deep neural networks. J Imaging (2022)

9. Chen, X., et al.: Semi-Supervised Semantic Segmentation with Cross Pseudo Supervision. In: CVPR (2021)
10. DeVries, T., et al.: Leveraging uncertainty estimates for predicting segmentation quality. arXiv preprint arXiv:1807.00502 (2018)
11. Du, S., et al.: MDViT: Multi-domain vision transformer for small medical image segmentation datasets. In: MICCAI (2023)
12. Dzieniszewska, A., Garbat, P., Piramidowicz, R.: Improving skin lesion segmentation with self-training. Cancers (2024)
13. Girshick, R.: Fast R-CNN. In: ICCV (2015)
14. Glaister, J., et al.: MSIM: Multistage illumination modeling of dermatological photographs for illumination-corrected skin lesion analysis. IEEE Trans Biomed Eng (2013)
15. Guo, C., et al.: On calibration of modern neural networks. In: ICML (2017)
16. Isensee, F., et al.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods (2021)
17. Kirillov, A., et al.: PointRend: Image segmentation as rendering. In: CVPR (2020)
18. Kohlberger, T., et al.: Evaluating Segmentation Error without Ground Truth. In: MICCAI (2012)
19. Lee, D.H.: Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. In: ICML CRLW (2013)
20. Li, Y., et al.: SUTrans-NET: a hybrid transformer approach to skin lesion segmentation. PeerJ Comput Sci (2024)
21. Liang, P., et al.: A multilevel alignment and cross-fusion knowledge distillation framework for vision transformer-based medical image segmentation. Future Gener Comput Syst (2026)
22. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med Image Anal (2017)
23. Loshchilov, I., et al.: Decoupled weight decay regularization. In: ICLR (2019)
24. Mehrtash, A., et al.: Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE Trans Med Imaging (2020)
25. Mendonça, T., et al.: PH^2 - A dermoscopic image database for research and benchmarking. In: EMBC (2013)
26. Oktay, O., et al.: Attention u-net: Learning where to look for the pancreas. In: MIDL (2018)
27. Perera, S., et al.: MobileUNETR: A lightweight end-to-end hybrid vision transformer for efficient medical image segmentation. In: ECCVW (2025)
28. Qiu, P., et al.: QCResUNet: Joint Subject-Level and Voxel-Level Prediction of Segmentation Quality. In: MICCAI (2023)
29. Ren, M., , et al.: Learning to reweight examples for robust deep learning. In: ICML (2018)
30. Robinson, R., et al.: Real-time prediction of segmentation quality. In: MICCAI (2018)
31. Ronneberger, O., et al.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
32. Rotemberg, V., et al.: A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Sci Data (2021)
33. Senbi, A., et al.: Towards ground-truth-free evaluation of any segmentation in medical images. arXiv preprint arXiv:2409.14874 (2024)
34. Shu, J., et al.: Meta-weight-net: Learning an explicit mapping for sample weighting. NeurIPS (2019)

35. Sohn, K., et al.: FixMatch: Simplifying semi-supervised learning with consistency and confidence. NeurIPS (2020)
36. Song, P., et al.: SU-RMT: Toward bridging semantic representation and structural detail modeling for medical image segmentation. Information Fusion (2026)
37. Taghanaki, S.A., et al.: Combo loss: Handling input and output imbalance in multi-organ segmentation. Comput Med Imaging Graph (2019)
38. Tajbakhsh, N., et al.: Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. Med Image Anal (2020)
39. Tarvainen, A., et al.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: NeurIPS (2017)
40. Valindria, V.V., et al.: Reverse Classification Accuracy: Predicting Segmentation Performance in the Absence of Ground Truth. IEEE Trans Med Imaging (2017)
41. Verma, V., et al.: Interpolation consistency training for semi-supervised learning. Neural Netw (2022)
42. Yu, L., et al.: Uncertainty-Aware Self-ensembling Model for Semi-supervised 3D Left Atrium Segmentation. In: MICCAI (2019)
43. Zheng, Z., et al.: Semi-supervised segmentation with self-training based on quality estimation and refinement. In: MICCAI MLMI (2020)
44. Zhou, Z., et al.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans Med Imaging (2019)