DOI: 10.65322/j.srd.06202026.54

RELIABILITY-AWARE POLYP SEGMENTATION WITH THRESHOLD OPTIMIZATION AND TEMPERATURE SCALING

Arabboyev M.M., *Associate Professor,*
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi

Abstract. *This study evaluated mask overlap and pixel probability quality in endoscopic polyp segmentation. Validation-based threshold optimization and temperature scaling were examined on Kvasir-SEG. Three configurations were compared using an 80:10:10 split and random seeds 42, 7, and 2024. The UNet352 configuration with both procedures achieved mean Dice of 0.7900 and IoU of 0.6534. The fixed-threshold UNet256 baseline achieved 0.7405 and 0.5880, respectively. The Brier score decreased from 0.0591 to 0.0455 and negative log-likelihood from 0.2235 to 0.1719. Expected calibration error changed from 0.0224 to 0.0215.*

Annotatsiya. *Polip segmentatsiyasida niqobning etalon bilan ustma-ust tushishi va piksel ehtimollarining kalibrlanishi baholandi. Kvasir-SEG to‘plamida validatsiya asosida qaror chegarasini optimallashtirish va harorat bo‘yicha masshtablash tekshirildi. Uchta konfiguratsiya 80:10:10 nisbatda bo‘lingan ma’lumotlarda 42, 7 va 2024 tasodifiy boshlang‘ich qiymatlari bilan baholandi. Ikkala amal qo‘llangan UNet352 da o‘rtacha Dice 0.7900 va IoU 0.6534 bo‘ldi. Qaror chegarasi 0.5 bo‘lgan UNet256 da bu qiymatlar mos ravishda 0.7405 va 0.5880 ni tashkil etdi. Brier ko‘rsatkichi 0.0591 dan 0.0455 gacha, manfiy log-ehtimollik 0.2235 dan 0.1719 gacha kamaydi. Kutilayotgan kalibrlash xatosi 0.0224 dan 0.0215 gacha o‘zgardi.*

Аннотация. *В работе оценивались совпадение предсказанной маски полипа с эталонной и калибровка попиксельных вероятностей. На наборе Kvasir-SEG исследованы оптимизация порога и температурное масштабирование по валидационной выборке. Три конфигурации оценивались при разбиении 80:10:10 и трёх начальных значениях генератора случайных чисел (42, 7 и 2024). Для UNet352 с обеими процедурами средние значения Dice и IoU составили 0.7900 и 0.6534. Для UNet256 с фиксированным порогом 0.5 получены 0.7405 и 0.5880 соответственно. Показатель Брайера снизился с 0.0591 до 0.0455, а отрицательное логарифмическое правдоподобие — с 0.2235 до 0.1719. Ожидаемая ошибка калибровки изменилась с 0.0224 до 0.0215.*

Keywords: *polyp segmentation, U-Net, threshold optimization, temperature scaling, calibration, Kvasir-SEG.*

Kalit so‘zlar: *polip segmentatsiyasi, U-Net, qaror chegarasi, harorat bo‘yicha masshtablash, kalibrlash, Kvasir-SEG.*

Ключевые слова: *сегментация полипов, U-Net, оптимизация порога, температурное масштабирование, калибровка, Kvasir-SEG.*

Introduction. Endoscopic polyp segmentation assigns a foreground or background label to each pixel of a colonoscopy image. Kvasir-SEG provides paired images and reference masks for evaluating this task [1]. The U-Net encoder–decoder with skip connections provides a baseline for mapping image features to pixel scores [2]. The binary mask used for evaluation is obtained by applying a decision threshold to a map of foreground probabilities.

Dice and intersection over union (IoU) quantify the agreement between binary masks [3]. The probabilities behind those masks are a second evaluation target. Two pixels can receive the same binary label at a given threshold and carry very different probabilities of belonging to the polyp. A check of calibration compares model confidence with empirical correctness. This distinction matters when a segmentation system displays the confidence attached to a region [4].

DOI: 10.65322/j.srd.06202026.54

This study compares mask overlap and probability quality across three complete Kvasir-SEG configurations. Results are summarized across three seeds; the configurations differ in several components, so their comparison does not isolate the effect of any one component.

Literature review. Polyp segmentation research has developed several alternatives within the encoder–decoder family. ResUNet++ adds residual, attention and multiscale components to a U-Net-like design [5]. PraNet uses parallel aggregation and reverse attention to refine regions and boundaries [6]. Evaluating these types of changes requires a metric suited to the target of the evaluation. Metrics Reloaded describes how properties of the segmentation task and data affect the choice and interpretation of validation measures [7].

Threshold selection is a decision problem. For binary predictions, the F1 measure has the same count-based form as Dice, and its maximizing cutoff depends on the scores and the evaluation population [8]. Temperature scaling addresses the probability estimates through a single fitted scalar on the logits [9]. The two operations have different validation targets. Maximizing Dice selects a binary mask operating point. Minimizing negative log-likelihood (NLL) fits probabilities against the reference labels.

Mehrtash et al. showed that medical segmentation networks with strong overlap can produce overconfident predictions [4]. Yeung et al. studied the connection between Dice-based training and overconfidence and evaluated a modified loss for biomedical segmentation [10]. The present experiment studies calibration after training by adjusting logits at inference. Both an error based on confidence bins and scores that use the full probabilities are reported. Estimation of a binned calibration error is sensitive to the binning procedure [11]. Dataset shift can also change calibration, which motivates identifying the domain in which the test scores were obtained [12].

Materials and methods. ***Dataset and segmentation models.*** Kvasir-SEG contains 1,000 endoscopic polyp images and corresponding segmentation masks [1]. The dataset was partitioned into training, validation and test subsets in the ratio 80:10:10 (800, 100 and 100 image–mask pairs). Training data were used to fit the segmentation models, the validation subset supplied the threshold and temperature criteria, and the test subset supplied the reported outcomes.

The experimental models were UNet256, UNet352 and EfficientNetB0-UNet256. The numbers 256 and 352 denote the square input dimensions. The comparison involving EfficientNetB0 changes the encoder within the 256 × 256 setting [13]. The evaluated output of each model is a pixel-level logit map for the polyp class. An annotated pixel is foreground when its reference label is one and is background when its reference label is zero.

Temperature scaling and threshold optimization. Let z_i denote the logit for pixel i and q_i the corresponding foreground probability. For temperature T, the sigmoid mapping is given by Eq. (1). The unscaled probability is obtained at T = 1.

$$q_i = \sigma\left(\frac{z_i}{T}\right),\ T > 0,\ \sigma(s) = \frac{1}{1+\exp(-s)} \quad (1)$$

Temperature scaling fits T by minimizing NLL on the validation predictions, as in Eq. (2) [9]. A positive T preserves the ordering of the logits. Values above one move the sigmoid probabilities closer to 0.5. The fitted scalar is then used for the test probabilities.

$$T^* = \arg\min_{T>0} \mathrm{NLL}_{\mathrm{val}}(T) \quad (2)$$

At a decision threshold τ, the model assigns a pixel to the mask when its probability reaches τ. Validation optimization selects the cutoff producing the largest Dice score over the candidate thresholds τ. The fixed-cutoff configuration uses τ = 0.5. Figure 1 summarizes how validation-selected parameters are applied to produce the test segmentation mask.

$$\hat{y}_i(\tau) = 1[q_i \geq \tau],\quad \tau^* = \arg\max_{\tau \in \mathcal{T}} D_{\mathrm{val}}(\tau) \quad (3)$$

DOI: 10.65322/j.srd.06202026.54

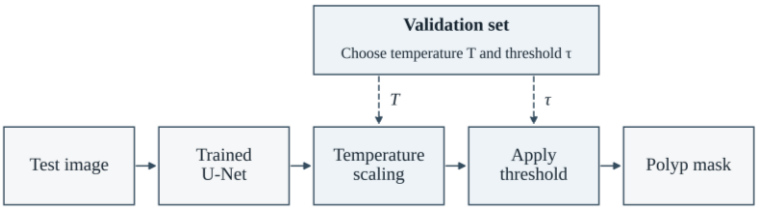

Figure 1. Segmentation workflow with temperature and threshold selected on the validation set.

Evaluation measures. The predicted and reference masks were compared using Dice (D) and IoU (I). In Eq. (4), TP, FP and FN denote the numbers of true-positive, false-positive and false-negative pixels, respectively [3]. Both measures increase as the overlap of the predicted and reference foreground regions improves.

$$D = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \quad I = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{4}$$

Brier score and NLL were evaluated using the foreground probabilities and reference labels. For N evaluated pixels and labels $y_i \in \{0, 1\}$, the Brier score is the average squared probability error [14]. NLL is the average negative logarithm of the probability assigned to the observed label. These scores use the probabilities before conversion to a binary mask.

$$\mathrm{Brier} = \frac{1}{N}\sum_{i=1}^{N}(q_i - y_i)^2 \tag{5}$$

$$\mathrm{NLL} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(q_i) + (1 - y_i)\log(1 - q_i)] \tag{6}$$

Expected calibration error (ECE) is the weighted mean gap between empirical accuracy and mean confidence across the confidence bins [9], [11]. B_m denotes a bin and M is the number of bins. Lower ECE, Brier score and NLL indicate better performance under the respective measure.

$$\mathrm{ECE} = \sum_{m=1}^{M}\frac{|B_m|}{N}\,|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)| \tag{7}$$

Experimental setup.

Table 1 lists the three configurations retained for comparison. A0 uses UNet256 with a fixed threshold of 0.5 and unscaled probabilities. A3 uses UNet352 with a validation-selected threshold and temperature scaling. A4 uses an EfficientNetB0 encoder in a 256 × 256 U-Net with the same types of validation post-processing. The A0–A3 comparison changes input size and post-processing together.

Table 1

Experimental configurations on Kvasir-SEG.

Setting	Segmentation model	Decision threshold	Temperature scaling
A0	UNet256	Fixed at 0.5	No
A3	UNet352	Validation-tuned	Yes
A4	EfficientNetB0-UNet256	Validation-tuned	Yes

The configurations were evaluated with seeds 42, 7 and 2024. Test measurements are reported as mean and standard deviation over the three runs. The A3–A0 comparison uses paired t-tests by seed and reports two-sided p-values and 95% confidence intervals for the overlap and probability measures. These comparisons are exploratory because each estimate uses three pairs and the intervals are reported per measure.

Results and discussion. ***Configuration comparison.***

Table 2 gives the test results. Among the three configurations, A3 gave the highest mean Dice and IoU and the lowest mean ECE, Brier score and NLL. Relative to A0, mean Dice and IoU increased by 0.0495 and 0.0654, respectively. The Brier score decreased by 0.0136 and NLL by 0.0516. The ECE difference was 0.0009.

DOI: 10.65322/j.srd.06202026.54

Table 2

Test results across three seeds on Kvasir-SEG.

Setting	Dice ↑	IoU ↑	ECE ↓	Brier ↓	NLL ↓
A0	0.7405 ± 0.0121	0.5880 ± 0.0152	0.0224 ± 0.0021	0.0591 ± 0.0088	0.2235 ± 0.0373
A3	0.7900 ± 0.0273	0.6534 ± 0.0377	0.0215 ± 0.0034	0.0455 ± 0.0070	0.1719 ± 0.0228
A4	0.5346 ± 0.0117	0.3649 ± 0.0108	0.0428 ± 0.0222	0.1288 ± 0.0261	0.4012 ± 0.0750

Values are mean ± standard deviation. Arrows indicate the direction of better performance.

The UNet352 configuration A3 had higher mean overlap and lower Brier score and NLL than the fixed-threshold UNet256 baseline A0. A4 had the lowest mean Dice and IoU and the highest Brier score and NLL. These comparisons describe the complete configurations on the stated test split; they do not isolate the contribution of input size, threshold selection or temperature scaling.

Paired comparison with the baseline.

Table 3 presents the reported 95% confidence intervals and p-values for A3 minus A0. The confidence interval for the Dice difference lies above zero and that for the Brier score lies below zero. The intervals for IoU, ECE and NLL include zero. In particular, the ECE data provide little evidence of a systematic change between A0 and A3.

Table 3

Paired differences between A3 and A0 across the three seeds.

Measure	Difference	95% confidence interval	p-value
Dice	+0.0495	[+0.0018, +0.0971]	0.0467
IoU	+0.0654	[−0.0021, +0.1328]	0.0530
ECE	−0.0009	[−0.0143, +0.0125]	0.7990
Brier score	−0.0136	[−0.0244, −0.0026]	0.0334
NLL	−0.0516	[−0.1059, +0.0027]	0.0549

The test scores come from a held-out partition of one dataset. Under a shift in image source, acquisition conditions or case mix, calibration can change [12]. The inference here is limited to Kvasir-SEG and the three reported runs. The paired contrasts establish the direction and uncertainty of the observed configuration-level differences. Reproducing individual runs would require the fitted temperatures, selected cutoffs and binning details in addition to the aggregate scores.

Conclusion. On Kvasir-SEG, the UNet352 configuration with validation-based threshold optimization and temperature scaling achieved mean Dice of 0.7900 and IoU of 0.6534. Compared with the fixed-threshold UNet256 baseline, this complete configuration had higher mask overlap and lower Brier score and NLL. Its ECE differed little. Three-seed paired comparisons showed the clearest differences for Dice and Brier score ($p = 0.0467$ and $p = 0.0334$); after Holm correction for the five measures, none of the differences remains significant at the 0.05 level, so these results should be interpreted with caution.

References

1. Jha, D., Smedsrud, P. H., Riegler, M. A., Halvorsen, P., de Lange, T., Johansen, D., & Johansen, H. D. (2020). Kvasir-SEG: A segmented polyp dataset. In *MultiMedia Modeling* (Lecture Notes in Computer Science, Vol. 11962, pp. 451–462). Springer.

DOI: 10.65322/j.srd.06202026.54

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

3. Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Medical Imaging, 15*, Article 29.

4. Mehrtash, A., Wells, W. M., Tempany, C. M., Abolmaesumi, P., & Kapur, T. (2020). Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Transactions on Medical Imaging, 39*(12), 3868–3878.

5. Jha, D., Smedsrud, P. H., Riegler, M. A., Johansen, D., de Lange, T., Halvorsen, P., & Johansen, H. D. (2019). ResUNet++: An advanced architecture for medical image segmentation. In *Proceedings of the IEEE International Symposium on Multimedia (ISM)* (pp. 225–230). IEEE.

6. Fan, D.-P., Ji, G.-P., Zhou, T., Chen, G., Fu, H., Shen, J., & Shao, L. (2020). PraNet: Parallel reverse attention network for polyp segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2020* (pp. 263–273). Springer.

7. Maier-Hein, L., Reinke, A., Godau, P., et al. (2024). Metrics reloaded: Recommendations for image analysis validation. *Nature Methods, 21*, 195–212.

8. Lipton, Z. C., Elkan, C., & Narayanaswamy, B. (2014). Optimal thresholding of classifiers to maximize F1 measure. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2014)* (pp. 225–239). Springer.

9. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (PMLR, Vol. 70, pp. 1321–1330).

10. Yeung, M., Rundo, L., Nan, Y., Sala, E., Schönlieb, C.-B., & Yang, G. (2023). Calibrating the Dice loss to handle neural network overconfidence for biomedical image segmentation. *Journal of Digital Imaging, 36*, 739–752.

11. Roelofs, R., Cain, N., Shlens, J., & Mozer, M. C. (2022). Mitigating bias in calibration error estimation. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics* (PMLR, Vol. 151, pp. 4036–4054).

12. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems* (Vol. 32).

13. Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning* (PMLR, Vol. 97, pp. 6105–6114).

14. Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1–3.