BALANCING THE FALSE POSITIVE-NEGATIVE TRADE-OFF TO ENHANCE IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Image segmentation is a fundamental task in computer vision with applications spanning diverse domains, particularly in medical imaging. However, the effectiveness of segmentation techniques often varies across datasets and tasks. For instance, methods like SRL and cLDice focus on segmenting thin tubular structures, while models such as IC-Net are tailored for brain tumor segmentation in MRI scans. Despite the availability of such specialized approaches, there remains a need for a unified framework that can generalize well across different segmentation challenges. In this study, we work on the observation that most of the strategies disproportionately emphasize reducing either False Negatives (FN) or False Positives (FP) and fail to achieve an optimal balance between the two. Building on this observation, we propose a novel method, Supervised Mask Modulation (SMM), that enhances segmentation performance by conditioning the ground truth masks during training to keep a balance between both highly important metrics. Our approach is architecture-agnostic and has been validated on a range of benchmark datasets, consistently outperforming state-of-the-art methods, often achieving significantly better results than the baseline.

1 INTRODUCTION

In the ever-evolving field of computer vision, image segmentation remains a critical task with wide-ranging applications in domains such as autonomous driving, industrial inspection, and medical imaging. To tackle the diverse challenges posed by different segmentation problems, a multitude of architectural innovations have been proposed over the years. Classical designs like U-Net (Azad et al., 2024) laid the groundwork for encoder-decoder based segmentation, and have since been extended in numerous directions. Several methods have been developed to address domain-specific challenges, particularly in medical image segmentation. Innovations in loss function such as centerlineDice (clDice) (Shit et al., 2021) and Skeleton Recall Loss (SRL) (Kirchhoff et al., 2024), specifically target segmentation of thin curvilinear structures. On similar lines, Kervadec (Kervadec et al., 2019) utilizes a novel loss function focusing on enhanced boundary predictions. Apart from novel loss functions, architectural novelties have also been introduced, catering to specific segmentation tasks. ICNet (Li et al., 2020) integrated a convolutional network with multiple resolution inputs, designed for accurate tumor core and enhancing region segmentation in brain tumor segmentation tasks. DeepMedic (Kamnitsas et al., 2016) introduced a dual-pathway 3D CNN with dense inference for effective segmentation of small lesions in brain MRI. DUNet (Sheng et al., 2024) introduced constant resolution U-blocks and dense feature connections to effectively detect fine-grained cracks with high accuracy and generalization, even in cluttered scenes.

Amidst the wide array of available segmentation methodologies, selecting an approach well-suited to a specific task can be non-trivial. To address this challenge, we propose **Supervised Mask Modulation (SMM)**, a unified architecture-agnostic strategy designed to enhance segmentation performance across diverse tasks, including those involving complex and irregular structures. In medical image segmentation, a false negative corresponds to a missed abnormality, such as a tumor or lesion not being detected, whereas a false positive refers to incorrectly identifying a healthy region as diseased. Minimizing false negatives is especially critical in clinical settings, as they may lead to missed diagnoses or delayed treatments. Our framework is based on the observation that in medical image segmentation tasks, the False Negative Rate (FNR) is often significantly higher than the False Positive Rate (FPR). This observation is supported by both prior experimental data (see Section 3)

and empirical evidence from our current results (see Section 6). But working merely to reduce the FNR can take be equally damaging in terms of the toll bore by other generic metrics(Dice Similarity Coefficient (DSC), Jaccard Similarity Index (JSI), clDice) caused by the imbalance of the FP.

The proposed methodology builds upon the integration of two well-established paradigms: controlling FN-FP balance and mask modulation. While each of these strategies has independently demonstrated effectiveness in improving segmentation performance, their isolated application often limits generalizability across diverse tasks (see Section 2). To overcome these limitations, we introduce a unified framework that synergistically combines false negative suppression with mask modulation, thereby enabling a more versatile and robust solution adaptable to a broader spectrum of segmentation challenges.

The proposed framework has been rigorously evaluated on a diverse set of publicly available datasets, each selected to represent distinct and challenging segmentation scenarios. These datasets span a broad spectrum of applications, ranging from irregular and complex structures in histopathology images to object delineation in real-world aerial imagery. This diversity underscores the generalizability and robustness of our approach across varied domains and segmentation tasks. With the proposed SMM framework we make the following contributions:

- **Exploitation of the FP**: Our framework exploits the hypothesis that the number of FN is significantly higher than the number of FP. We attempt to improve the model's performance by introducing intended FP, conditioned by model performance, into the ground truth masks for enhanced training, thereby penalizing the model for missing out class pixels in smaller regions or some structures entirely. The results validate that this strategy tends to bring an overall improvement in the model performance.
- **Multi-Class Compatibility**: The framework has proved effective across diverse datasets, demonstrating improved performance on both binary and multi-class segmentation tasks. Its versatility makes it applicable to a wide range of imaging scenarios.
- **Architecture Agnostic**: The framework does not include making alterations to any particular architecture. rather, it proposes in a change in the training paradigm of models and can therefore integrate seamlessly with a wide range of pipelines across different model architectures.

2 Related Work

Enhancing segmentation performance by mitigating FN, often reflected as improved recall, has been a central focus of recent studies. A common approach involves modifying loss functions to increase model sensitivity, particularly in imbalanced or complex medical datasets. For instance, Xiang et al. (2019a;b) designed loss functions to enhance reliability, while Chan et al. (2020) employed maximum likelihood estimation with Bayesian decision theory to better handle underrepresented classes. Other methods, including Zhong et al. (2021) and Kervadec et al. (2019), introduce pixel-wise or contour-based losses specifically aimed at reducing FN in fine structures. Beyond loss design, architectural innovations such as PatchRefineNet (Nagendra & Kifer, 2024) refine outputs by correcting spatial biases in logits, and depth-based strategies (Maag, 2021; Maag & Rottmann, 2022) further improve recall. However, these architectural solutions often incur additional computational complexity, rendering loss-based approaches a more lightweight and widely adopted alternative.

Mask transformations have also been explored to enhance segmentation. Skeletonization transforms have been utilized for the detection of fine tubular structures, with a focus on preserving topological integrity (Kirchhoff et al., 2024; Shit et al., 2021). Similarly, Kats et al. (2019) introduced a soft-labeled mask combined with soft dice loss for lesion segmentation, while Vasudeva et al. (2024) employed geodesic distance transforms to assign soft labels near boundaries.

Building on these efforts, we propose a novel mask transformation strategy guided by model-predicted false negatives, complemented by tailored training mechanisms. This approach broadens the applicability of our methodology while directly addressing the FNR-FPR trade-off in segmentation tasks.

3 THE EXAGGERATED FALSE NEGATIVES

While numerous studies have proposed segmentation techniques tailored to specific tasks, several underlying principles emerge that are broadly applicable across diverse segmentation problems:

1. De Rosa et al. (2024) observed that although their U-Net-based ensemble achieved high precision, it exhibited low recall, indicating that while false positives (FP) were reduced, false negatives (FN) remained predominant.
2. In a teacher-student weakly supervised setup for colon polyp segmentation, Jia et al. (2024) reported that the segmentation outputs "present a quite high FNR inside the polyp area."
3. By modulating the Tversky-loss parameter β, Do et al. (2020) highlighted the FNR-FPR trade-off, noting that "as β increased, the false-positive rate systematically decreased while the false-negative rate systematically increased."

These observations, corroborated by additional studies (Delgado et al., 2024; Luo et al., 2023), demonstrate that FNR often substantially exceeds FPR in many segmentation tasks. Our own evaluation of U-Net models, reported in Table 2, further confirms this trend.

This phenomenon can be intuitively explained in medical imaging tasks such as brain tumor or lesion segmentation, where the foreground region typically occupies only a small fraction of the image relative to the background. Such an imbalance hampers the model's ability to comprehensively capture the foreground. From a theoretical standpoint, standard objectives such as cross-entropy minimize the expected misclassification rate under maximum likelihood estimation, implicitly assuming equal costs for false positives and false negatives. Consequently, they fail to emphasize recall in settings where false negatives are more critical. Evaluation metrics in many domains—including medical imaging—are based on precision and recall rather than accuracy, and these metrics are not aligned with the likelihood training criterion (Goodfellow et al., 2016, p. 265). This misalignment often leads to models that prioritize precision over recall, exacerbating the FNR. Our method leverages this insight by guiding the model to predict additional positives, thereby reducing FNR while minimally affecting FPR, ultimately achieving an optimal trade-off between these inversely related metrics.

4 METHODOLOGY

Given that FPR values are consistently lower than FNR in medical imaging tasks, we leverage this asymmetry to guide our approach. Specifically, the framework introduces controlled FP regions in the ground truth masks to encourage the model to predict positives in previously missed areas. Implementation details are discussed in subsequent sections.

4.1 MISS-AWARE MASK MODULATION (MAMM)

Algorithm 1 Miss-Aware Mask Modulation (MAMM)

Require: Prediction $\hat{\mathbf{Y}}$, Ground truth $\mathbf{Y}$

1: $\mathbf{FN} \leftarrow (\mathbf{Y} - \hat{\mathbf{Y}}) > 0$ ▷ Extract false negatives
2: $\mathbf{U} \leftarrow \text{Dilate}(\mathbf{FN})$ ▷ Expand missed regions
3: $\mathbf{Y^M} \leftarrow \mathbf{U} \cup \mathbf{Y}$ ▷ Generate modulated mask
4: **return** $\mathbf{Y^M}$

FN are defined as regions in the ground truth mask that belong to the foreground but were incorrectly predicted as background by the segmentation model. To extract these missed regions, we compute the difference between the ground truth mask, $\mathbf{Y}$, and the predicted mask, $\hat{\mathbf{Y}}$:

$$\mathbf{FN} = (\mathbf{Y} - \hat{\mathbf{Y}}) > 0,$$

where positive values correspond to false negatives, negative values correspond to false positives, and correctly classified pixels (true positives and true negatives) are reduced to zero. We then retain only the positive entries corresponding to the FN.

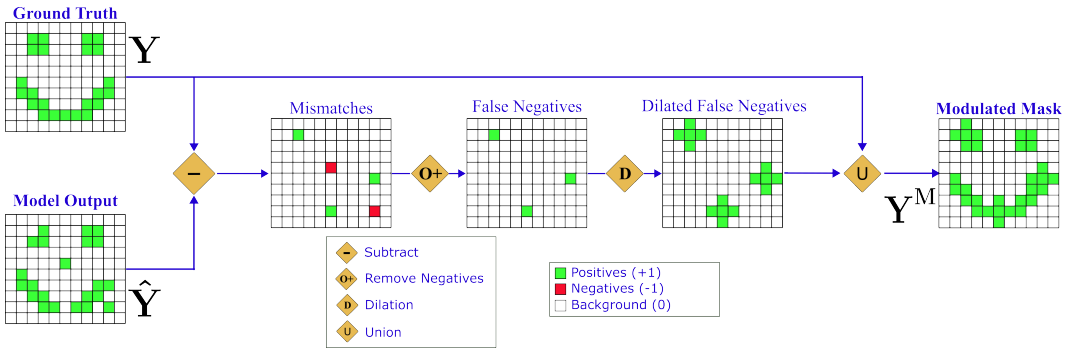

Figure 1: **Miss-Aware Mask Modulation** The mask modulation process begins by subtracting the predicted mask from the ground truth to identify misclassified pixels. Only the false negatives are retained, dilated, and combined with the original ground truth to generate the updated modulated mask.

Dilation: To emphasize regions overlooked by the model, the FN mask is dilated with a diamond-shaped kernel of radius 2, and its union with the ground truth yields the updated modulated mask. This operation is performed independently for each class to construct the final mask.

This modulation strategy, termed **Miss-Aware Mask Modulation (MAMM)**, adaptively updates the mask to reflect the model's current errors while remaining anchored to the original ground truth. By refreshing the modulated masks at each epoch, the procedure ensures that training consistently targets the most recent false negatives (Figure 1).

4.2 Training Algorithm

Algorithm 1 provides a mechanism to enhance focus on regions prone to being missed. We propose two strategies to leverage this transformation during training, differing in the degree of penalization applied to the model. These can be interpreted as 'hard' and 'soft' approaches to mask modulation, which are described in detail in the subsequent subsections.

Algorithm 2 Supervised Mask Modulation *v1*

Require: Input $\mathbf{X}$, Ground Truth $\mathbf{Y}$, and Modulated Mask $\mathbf{Y_0^M} = \mathbf{Y}$

1: **for** each epoch (t) **do**
2: $\quad \hat{\mathbf{Y}}_\mathbf{t} \leftarrow \text{Model}(\mathbf{X})$
3: $\quad \mathscr{L} \leftarrow \mathscr{L}_{vanilla}(\hat{\mathbf{Y}}_\mathbf{t}, \mathbf{Y}) + \mathscr{L}_{ESL}(\hat{\mathbf{Y}}_\mathbf{t}, \mathbf{Y_t^M})$
4: $\quad$ Backpropagate loss $\mathscr{L}$
5: $\quad \tilde{\mathbf{Y}}_{\mathbf{t+1}} \leftarrow \text{MAMM}(\hat{\mathbf{Y}}_\mathbf{t}, \mathbf{Y})$
6: **end for**

4.2.1 Supervised Mask Modulation *v1*

In this variant, we propose a specialized loss function, Elevated Senstivity Loss (ESL), explicitly designed to impose a strong penalty on FN in the model's predictions. Its primary objective is to ensure that small or subtle regions are accurately detected and not overlooked. This is achieved by incorporating the count of FN directly into the denominator of the loss formulation. To maintain scale invariance and normalize the contribution of each pixel, the total number of pixels, being a constant, is also included in the denominator. The explicit focus on penalizing missed detections characterizes this approach as a hard penalization strategy, motivating its designation as the hard training algorithm for SMM.

Let Y denote the ground truth mask and $\hat{Y}$ denote the predicted mask, each consisting of N pixels indexed by $i \in \Omega$, where Ω is the set of all pixel locations. Let $y_i \in \{0, 1\}$ and $\hat{y}_i \in [0, 1]$ denote the values of the i^{th} pixel in Y and $\hat{Y}$, respectively. We define the **Elevated Sensitivity Loss (ESL)** as:

$$\mathscr{L}_{\text{ESL}} = -\frac{\sum_{i\in\Omega} y_i\,\hat{y}_i}{N + \sum_{i\in\Omega} y_i(1-\hat{y}_i)} \tag{1}$$

where:

- $N = |\Omega|$ is the total number of pixels in the mask.
- $y_i \in \{0, 1\}$ is the ground truth value of the i^{th} pixel.
- $\hat{y}_i \in [0, 1]$ is the predicted value of the i^{th} pixel.

The multiplicative factor of N acts as a normalization term for the loss, ensuring scale consistency. Its significance is discussed in further detail in Appendix B.

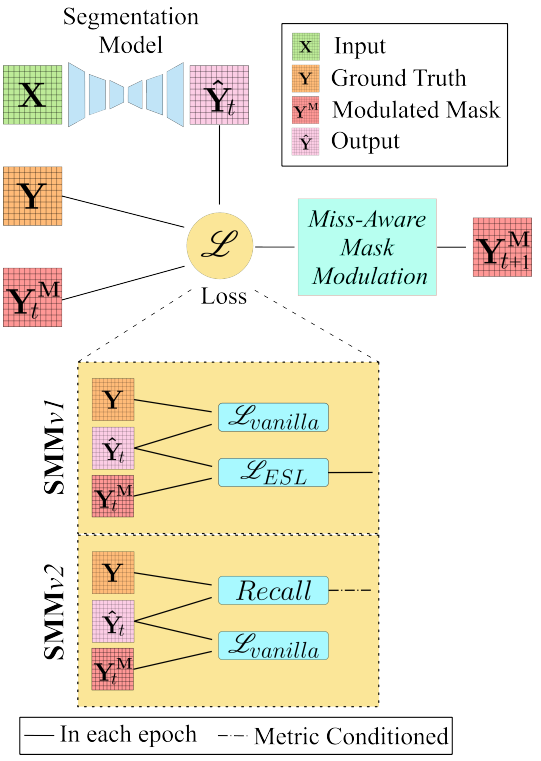

Figure 2: **Supervised Mask Modulation.** Given an input $\mathbf{X}$, the segmentation model predicts $\hat{\mathbf{Y}}$, which is compared with the ground truth $\mathbf{Y}$ and modulated mask $\mathbf{Y_t^M}$ to compute the loss. This loss updates the modulated mask for subsequent epochs, yielding $\mathbf{Y_{t+1}^M}$. Both SMM variants employ MAMM to generate these masks.

The ESL loss is applied in conjunction with modulated masks generated from Algorithm 1. Training begins with a warm-up phase, defaulted to 20% of total epochs, during which neither mask updates nor ESL computation occurs, allowing the model to learn global structures. Post pretraining, masks are updated each epoch, and ESL is computed using the modulated masks, while standard loss functions operate on the original ground truth. The final loss is the sum of both, ensuring the model follows the general learning trajectory while explicitly penalizing FN.

4.2.2 Supervised Mask Modulation *v2*

We propose an adaptive training strategy that modulates the mask based on the model's performance, measured via the *recall metric* after each epoch. Let the model first undergo pretraining for 20% of the total epochs. We consistently store the recall values in a fixed-length queue of size L.

To assess the trend of model performance, we compute the gradient of the best-fit line for recall over epochs, referred to as β:

$$\beta = \frac{\text{Cov}(\mathbf{x}, \mathbf{y})}{\text{Var}(\mathbf{y})}, \tag{2}$$

with

$$\mathbf{x} = [x_1, x_2, \ldots, x_L]^\top, \quad x_i: \text{ recall at epoch } i,$$
$$\mathbf{y} = [1, 2, \ldots, L]^\top, \quad y_i: \text{ epoch index } i,$$

where $\text{Var}(\mathbf{y})$ is the variance of $\mathbf{y}$ and $\text{Cov}(\mathbf{x}, \mathbf{y})$ the covariance between $\mathbf{x}$ and $\mathbf{y}$.

A small or negative β indicates stagnation or decline in performance, prompting updates to the modulated mask. When β falls below a threshold γ, controlled FP are introduced near the true boundary via dilation (Algorithm 3), while FN are consistently computed against the original ground truth. To prevent excessive mask expansion, prior modulations are cleared before new ones are applied. This thresholding mechanism balances exploration and consolidation by adapting to gradient magnitude, drawing inspiration from adaptive thresholding in semi-supervised learning (Xu et al., 2021): gradients above γ suppress modulation, reinforcing confident regions, whereas those below γ trigger updates to redirect learning toward uncertain or overlooked areas. Recall is employed solely for

evaluation and excluded from optimization, and γ is linearly decayed during training to reflect the model's evolving identification of novel positives. Unlike earlier variants, this approach dispenses with explicit penalization, motivating its designation as the soft training algorithm for SMM (Algorithm 4).

Algorithm 3 UpdateMask

Require: Queue Q, Prediction $\hat{\mathbf{Y}}_\mathbf{t}$, Ground truth $\mathbf{Y}$, and Modulated Mask $\mathbf{Y}_\mathbf{t}^\mathbf{M}$

Retrieve threshold parameter γ
Compute recall r between $\mathbf{Y}$ and $\hat{\mathbf{Y}}_\mathbf{t}$:

$$r = \frac{\sum(\mathbf{Y} \wedge \hat{\mathbf{Y}}_\mathbf{t})}{\sum \mathbf{Y}}$$

Append r to queue Q
Compute gradient β from queue Q using Eq. 2
if $\beta < \gamma$ **then**
    $\mathbf{Y}_{\mathbf{t+1}}^\mathbf{M} \leftarrow \text{MAMM}(\hat{\mathbf{Y}}_\mathbf{t}, \mathbf{Y})$
else
    $\mathbf{Y}_{\mathbf{t+1}}^\mathbf{M} \leftarrow \mathbf{Y}_\mathbf{t}^\mathbf{M}$
end if
return $\mathbf{Y}_{\mathbf{t+1}}^\mathbf{M}$

Algorithm 4 Supervised Mask Modulation *v2*

Require: Input $\mathbf{X}$, Ground Truth $\mathbf{Y}$, and Modulated Mask $\mathbf{Y}_\mathbf{0}^\mathbf{M} = \mathbf{Y}$

Initialize: Empty queue Q with fixed length L
for each epoch t **do**
    $\hat{\mathbf{Y}}_\mathbf{t} \leftarrow \text{Model}(\mathbf{X})$
    $\mathscr{L} \leftarrow \text{Loss}(\hat{\mathbf{Y}}_\mathbf{t}, \mathbf{Y}_\mathbf{t}^\mathbf{M})$
    Backpropagate using $\mathscr{L}$
    $\mathbf{Y}_{\mathbf{t+1}}^\mathbf{M} \leftarrow \textsc{UpdateMask}(Q, \hat{\mathbf{Y}}_\mathbf{t}, \mathbf{Y}, \mathbf{Y}_\mathbf{t}^\mathbf{M})$
end for

5 Experimental Setup

Table 1: **Dataset Summary** Characteristics of the datasets used for training and evaluation, covering multiple 2D segmentation tasks ranging from binary to multi-class segmentation. Tr and Ts are abbreviations for Train and Test, respectively. # denotes "Number of".

Dataset	Image Dims	# Classes	# Images (Tr + Ts)
BoMBR (Raina et al., 2024)	512×512	4	201 + 50
DRIVE (Hassan et al., 2015)	512×512	2	80 + 20
Cracks (Tomaszkiewicz & Owerko, 2023)	224×224	2	572 + 143
Drone[0]	512×512	5	320 + 80

5.1 Dataset Description

We validated the proposed framework on four publicly available datasets encompassing diverse image domains and segmentation challenges, including both binary and multi-class tasks. The **BoMBR** dataset (Raina et al., 2024) involves segmentation of fat globules, reticulin fibers, and bone marrow from biopsy images for reticulin quantification. To assess performance on tubular structures, we used the **DRIVE** dataset (Hassan et al., 2015) for retinal vessel segmentation. Beyond medical imaging, we evaluated on two real-world datasets: fine crack segmentation in concrete surfaces

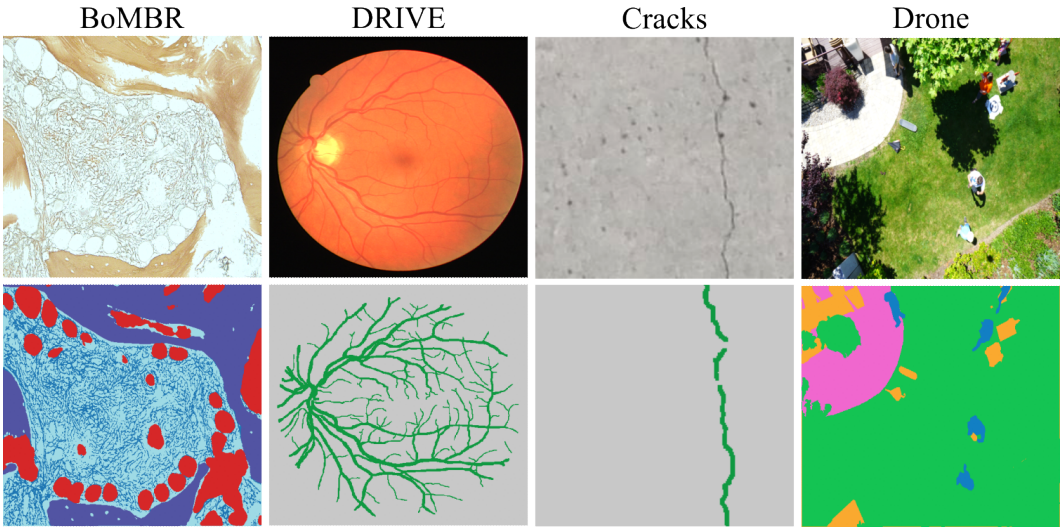

Figure 3: **Dataset Samples** Representative samples from the four datasets used to validate our framework, covering diverse 2D segmentation tasks across medical, industrial, and natural image domains, and including both binary and multi-class settings.

(Tomaszkiewicz & Owerko, 2023) and object segmentation in aerial drone imagery[1]. Representative samples are shown in Figure 3.

5.2 Model Configurations

We evaluated model performance using a standardized U-Net (Ronneberger et al., 2015) pipeline with four encoder-decoder stages. Each stage consists of two convolutional layers with batch normalization and ReLU activation. Decoder features are upsampled via transposed convolutions and fused with encoder outputs through skip connections, preserving spatial detail.

Models were trained with an initial learning rate of 0.1 and linear decay, using a batch size of 4. Epochs were dataset-specific to ensure convergence. All experiments were run on NVIDIA Tesla T4 and GeForce GTX 1080 Ti GPUs.

5.3 Training Setup

We evaluate our models using U-Net as the base architecture, comparing against strong, architecture-agnostic baselines. All experiments were repeated over five random seeds for significance analysis. The baselines are: **Vanilla U-Net** (trained with Dice loss (Dice)+Categorical Cross Entropy loss (CCE)), **U-Net+SRL** (Vanilla U-Net with SRL), and **U-Net+BL** (Vanilla U-Net with Boundary Loss (BL)).

Since CCE fails under overlapping class regions induced by MAMM, SMM*v2* replaces it with class-wise Binary Cross-Entropy to support multi-label pixels. We set queue length $L = 15$ and γ to the mean of β values from pretraining epochs.

Evaluation employed complementary metrics: (i) **Overlap** (DSC, JSI), (ii) **Topology** (clDice), and (iii) **Error** (FNR, FPR), covering accuracy, structural preservation, and under/over-segmentation tendencies.

[1] `http://dronedataset.icg.tugraz.at/`

Table 2: **Test set metrics** of U-Net models trained using different strategies. SMM*v1* and SMM*v2* present results for both versions of our proposed framework. An asterisk ($*$) indicates statistical significance at $p < 0.05$, based on t-tests comparing our best-performing method with the strongest baseline.

Method	DSC ↑	clDice ↑	JSI ↑	FNR ↓	FPR ↓
BoMBR (Raina et al., 2024)					
Vanilla U-Net	66.02 ± 2.11	63.57 ± 1.76	56.42 ± 2.44	26.04 ± 1.82	7.58 ± 0.81
U-Net + SRL	66.84 ± 2.21	63.54 ± 1.36	57.24 ± 2.41	25.78 ± 2.07	7.29 ± 0.95
U-Net + BL	67.09 ± 1.06	64.15 ± 1.29	57.80 ± 1.04	26.11 ± 1.15	7.23 ± 0.38
SMM*v1*	66.82 ± 1.16	64.12 ± 0.95	57.37 ± 1.38	25.92 ± 0.77	7.27 ± 0.39
SMM*v2*	$\mathbf{67.46 \pm 1.24}$	$\mathbf{64.42 \pm 0.64}$	$\mathbf{57.96 \pm 1.13}$	$\mathbf{24.73 \pm 1.14}$	$\mathbf{7.09 \pm 0.39}$
DRIVE (Hassan et al., 2015)					
Vanilla U-Net	79.63 ± 1.45	83.48 ± 1.84	66.21 ± 1.98	21.51 ± 2.00	2.55 ± 0.10
U-Net + SRL	80.01 ± 0.47	84.27 ± 0.81	66.72 ± 0.65	$\mathbf{18.85 \pm 0.97}$	2.97 ± 0.08
U-Net + BL	79.72 ± 1.74	83.36 ± 1.48	66.35 ± 2.39	23.40 ± 1.13	$\mathbf{2.12 \pm 0.39}$
SMM*v1*	$\mathbf{80.64 \pm 1.30}$	$\mathbf{84.42 \pm 1.51}$	$\mathbf{67.62 \pm 1.83}$	20.98 ± 2.08	2.31 ± 0.26
SMM*v2*	78.93 ± 0.68	82.71 ± 0.94	65.24 ± 0.92	21.53 ± 1.08	2.79 ± 0.09
Cracks (Tomaszkiewicz & Owerko, 2023)					
Vanilla U-Net	64.57 ± 0.87	74.92 ± 1.15	51.20 ± 0.80	31.39 ± 1.22	0.33 ± 0.01
U-Net + SRL	62.51 ± 3.31	71.93 ± 5.15	49.12 ± 3.36	$\mathbf{29.69 \pm 0.80}$	0.44 ± 0.13
U-Net + BL	64.05 ± 0.93	74.73 ± 0.82	50.82 ± 0.93	33.15 ± 0.66	$\mathbf{0.31 \pm 0.01}$
SMM*v1*	$\mathbf{64.74 \pm 0.20}$	$\mathbf{75.35 \pm 0.34}^*$	$\mathbf{51.44 \pm 0.21}^*$	31.16 ± 0.58	0.33 ± 0.01
SMM*v2*	62.93 ± 2.73	72.64 ± 4.50	49.56 ± 2.80	33.08 ± 2.82	0.33 ± 0.02
Drone[1]					
Vanilla U-Net	49.58 ± 2.54	44.44 ± 2.34	39.95 ± 2.04	29.47 ± 2.76	6.32 ± 0.25
U-Net + SRL	48.92 ± 1.45	43.92 ± 1.38	39.16 ± 1.02	29.69 ± 1.23	6.40 ± 0.24
U-Net + BL	45.37 ± 8.06	40.13 ± 7.81	37.76 ± 6.60	36.23 ± 7.17	6.73 ± 1.28
SMM*v1*	50.49 ± 1.72	45.45 ± 1.91	40.89 ± 1.46	29.20 ± 1.79	6.07 ± 0.28
SMM*v2*	$\mathbf{51.34 \pm 2.39}$	$\mathbf{46.21 \pm 2.09}$	$\mathbf{41.61 \pm 2.31}$	$\mathbf{27.70 \pm 2.20}$	$\mathbf{5.93 \pm 0.33}^*$

6 Results and Discussion

6.1 Evaluation Protocol

To assess the robustness and generalizability of our approach, we validated the method across a diverse set of benchmark datasets. Following the recommendations of *The Machine Learning Reproducibility Checklist* (Pineau et al., 2021), we attempted to mitigate stochastic effects in training and ensure reproducibility by repeating each experimental configuration using five fixed random seeds. Reported results are expressed as mean $\pm$ standard deviation across these runs, providing a reliable estimate of model performance while attributing observed differences to methodological improvements rather than random variations in initialization or data shuffling.

6.2 Metric-Level Insights

Table 2 shows that SRL consistently reduces FNR, often with a moderate increase in FPR relative to Vanilla U-Net, while BL shows the opposite trend. As discussed in Section 3, striking a balance between FNR and FPR is critical—our model achieves this, yielding simultaneous reductions in both rates and improved complementary metrics. Consequently, our method surpasses all baselines in Dice, clDice, and JSI, evidencing superior overlap and structural segmentation across datasets.

Statistical significance was evaluated as described in Appendix A, using one-sided t-tests ($p < 0.05$) for pairwise comparisons with the strongest baselines. Distinct variant-specific trends across dataset categories are further analyzed in Appendix D.

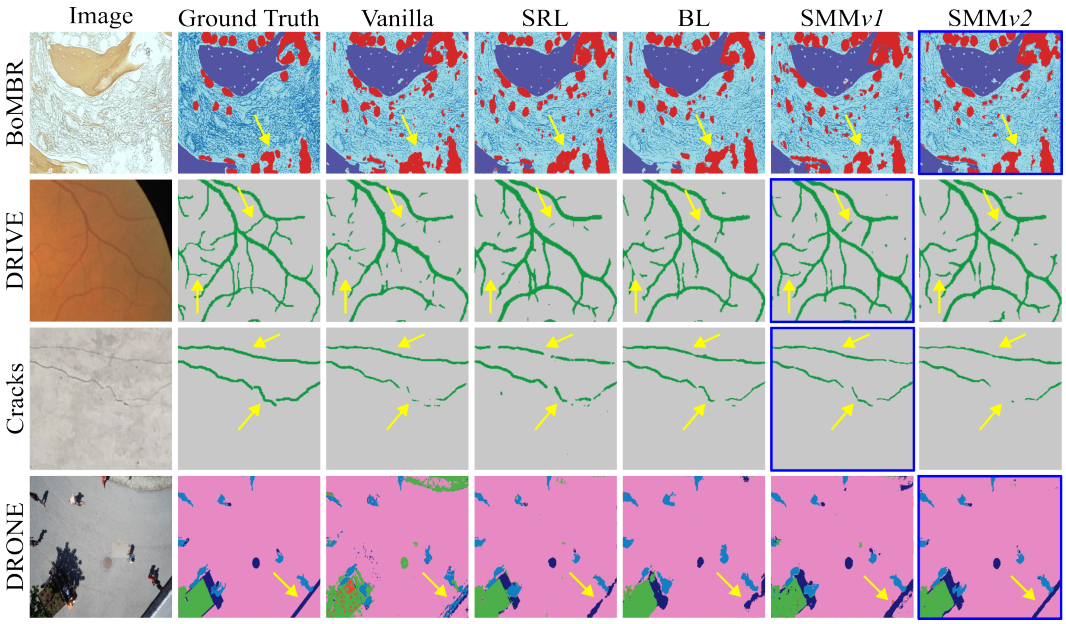

Figure 4: **Visual Results** The figure presents a sample test output of each model for all the utilized datasets. The better-performing version of SMM is marked by a blue bounding box. It may be noticed that in cases of Category 1 datasets, SMM*v1* is able to segregate the regions missed by the baselines. SMM*v2*, on the other hand, efficiently balances under-prediction and over-prediction in the case of Category 2 datasets, thus ensuring accurate semantic segregation of separate classes.

6.3 Architecture-Agnostic Deployment

Our training framework is designed to be independent of architecture-specific features, enabling robust deployment across diverse segmentation networks. Both variants of SMM employ a unified mask modulation strategy with a generalizable training procedure that can be applied to any segmentation architecture. The effectiveness of this approach is further demonstrated in Appendix C, where we report results on SegNet (Badrinarayanan et al., 2017), showing that SMM maintains strong performance even on architectures not seen during primary experiments.

7 Conclusion

Despite the proliferation of task-specific segmentation techniques, there remains a pressing need for a unified training paradigm capable of delivering consistent performance across heterogeneous tasks and application domains. In this work, we propose a segmentation model training strategy, denoted Supervised Mask Modulation (SMM), which is architecture-agnostic and demonstrates efficacy across a broad spectrum of segmentation challenges. The principal motivation of SMM is to optimize the balance between FN and FP, thereby enhancing segmentation fidelity.

Central to our framework is a novel mask transformation, Miss-Aware Mask Modulation (MAMM), derived from model-predicted FN regions, which is leveraged alongside two complementary training strategies to reinforce model learning. This consolidation reduces the otherwise fragmented landscape of segmentation methodologies into two generalizable strategies. We extensively validate our approach on publicly available datasets, benchmarking against state-of-the-art baselines, where our strategies consistently yield superior performance across diverse segmentation scenarios. Evaluation using multiple network architectures further highlights the generality and robustness of the proposed methodology, underscoring its practical utility for real-world segmentation pipelines.

REFERENCES

Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

Robin Chan, Matthias Rottmann, Fabian Hüger, Peter Schlicht, and Hanno Gottschalk. Controlled false negative reduction of minority classes in semantic segmentation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2020.

Alessandro Pasquale De Rosa, Marco Benedetto, Stefano Tagliaferri, Francesco Bardozzo, Alessandro D'Ambrosio, Alvino Bisecco, Antonio Gallo, Mario Cirillo, Roberto Tagliaferri, and Fabrizio Esposito. Consensus of algorithms for lesion segmentation in brain mri studies of multiple sclerosis. *Scientific Reports*, 14(1):21348, 2024.

Emilio Delgado, Roberto Rodriguez-Echeverria, Antonio Jesús Fernández-García, Juan D Gutiérrez, and Miguel Ángel Suero-Rodrigo. Advancing precision in medical image segmentation: A performance analysis of loss functions for covid-19 lung infection segmentation in computed tomography images. *IET Image Processing*, 18(13):4047–4065, 2024.

Hung P Do, Yi Guo, Andrew J Yoon, and Krishna S Nayak. Accuracy, uncertainty, and adaptability of automatic myocardial asl segmentation using deep cnn. *Magnetic resonance in medicine*, 83 (5):1863–1874, 2020.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Gehad Hassan, Aboul Ella Hassanien, Nashwa El-Bendary, and Ali Fahmy. Blood vessel segmentation approach for extracting the vasculature on retinal fundus images using particle swarm optimization. In *2015 11th international computer engineering conference (ICENCO)*, pp. 290–296. IEEE, 2015.

Yiwen Jia, Guangming Feng, Tang Yang, Siyuan Chen, and Fu Dai. Self-adaptive teacher-student framework for colon polyp segmentation from unannotated private data with public annotated datasets. *Plos one*, 19(8):e0307777, 2024.

Konstantinos Kamnitsas, Enzo Ferrante, Sarah Parisot, Christian Ledig, Aditya V Nori, Antonio Criminisi, Daniel Rueckert, and Ben Glocker. Deepmedic for brain tumor segmentation. In *International workshop on Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries*, pp. 138–149. Springer, 2016.

Eytan Kats, Jacob Goldberger, and Hayit Greenspan. Soft labeling by distilling anatomical knowledge for improved ms lesion segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 1563–1566. IEEE, 2019.

Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. Boundary loss for highly unbalanced segmentation. In *International conference on medical imaging with deep learning*, pp. 285–296. PMLR, 2019.

Yannick Kirchhoff, Maximilian R Rokuss, Saikat Roy, Balint Kovacs, Constantin Ulrich, Tassilo Wald, Maximilian Zenk, Philipp Vollmuth, Jens Kleesiek, Fabian Isensee, et al. Skeleton recall loss for connectivity conserving and resource efficient segmentation of thin tubular structures. In *European Conference on Computer Vision*, pp. 218–234. Springer, 2024.

Gongyang Li, Zhi Liu, and Haibin Ling. Icnet: Information conversion network for rgb-d based salient object detection. *IEEE Transactions on Image Processing*, 29:4873–4884, 2020.

Xiao Luo, Yadi Yang, Shaohan Yin, Hui Li, Weijing Zhang, Guixiao Xu, Weixiong Fan, Dechun Zheng, Jianpeng Li, Dinggang Shen, et al. False-negative and false-positive outcomes of computer-aided detection on brain metastasis: Secondary analysis of a multicenter, multireader study. *Neuro-oncology*, 25(3):544–556, 2023.

Kira Maag. False negative reduction in video instance segmentation using uncertainty estimates. In *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 1279–1286. IEEE, 2021.

Kira Maag and Matthias Rottmann. False negative reduction in semantic segmentation under domain shift using depth estimation. *arXiv preprint arXiv:2207.03513*, 2022.

Savinay Nagendra and Daniel Kifer. Patchrefinenet: Improving binary segmentation by incorporating signals from optimal patch-wise binarization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1361–1372, 2024.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Journal of machine learning research*, 22(164):1–20, 2021.

Panav Raina, Satyender Dharamdasani, Dheeraj Chinnam, Praveen Sharma, and Sukrit Gupta. Bombr: an annotated bone marrow biopsy dataset for segmentation of reticulin fibers. *bioRxiv*, pp. 2024–10, 2024.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Shibo Sheng, Hui Yin, Ying Yang, Aixin Chong, and Hua Huang. Dunet: Dense u-blocks network for fine-grained crack detection. *Signal, Image and Video Processing*, 18(2):1929–1938, 2024.

Suprosanna Shit, Johannes C Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien PW Pluim, Ulrich Bauer, and Bjoern H Menze. cldice-a novel topology-preserving loss function for tubular structure segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16560–16569, 2021.

Karolina Tomaszkiewicz and Tomasz Owerko. A pre-failure narrow concrete cracks dataset for engineering structures damage classification and segmentation. *Scientific Data*, 10(1):925, 2023.

Sukesh Adiga Vasudeva, Jose Dolz, and Herve Lombaert. Geols: geodesic label smoothing for image segmentation. In *Medical Imaging with Deep Learning*, pp. 468–478. PMLR, 2024.

Kaite Xiang, Kaiwei Wang, and Kailun Yang. A comparative study of high-recall real-time semantic segmentation based on swift factorized network. In *Artificial Intelligence and Machine Learning in Defense Applications*, volume 11169, pp. 105–118. SPIE, 2019a.

Kaite Xiang, Kaiwei Wang, and Kailun Yang. Importance-aware semantic segmentation with efficient pyramidal context network for navigational assistant systems. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pp. 3412–3418. IEEE, 2019b.

Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. Dash: Semi-supervised learning with dynamic thresholding. In *International conference on machine learning*, pp. 11525–11536. PMLR, 2021.

Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7273–7282, 2021.