RankSEG-RMA: An Efficient Segmentation Algorithm via Reciprocal Moment Approximation

Zixun Wang
Department of Statistics and Data Science
The Chinese University of Hong Kong
1155225012@link.cuhk.edu.hk

Ben Dai
Department of Statistics and Data Science
The Chinese University of Hong Kong
bendai@cuhk.edu.hk

Abstract

Semantic segmentation labels each pixel in an image with its corresponding class, and is typically evaluated using the Intersection over Union (IoU) and Dice metrics to quantify the overlap between predicted and ground-truth segmentation masks. In the literature, most existing methods estimate pixel-wise class probabilities, then apply argmax or thresholding to obtain the final prediction. These methods have been shown to generally lead to inconsistent or suboptimal results, as they do not directly maximize segmentation metrics. To address this issue, a novel consistent segmentation framework, RankSEG, has been proposed, which includes RankDice and RankIoU specifically designed to optimize the Dice and IoU metrics, respectively. Although RankSEG almost guarantees improved performance, it suffers from two major drawbacks. First, it is its computational expense—RankDice has a complexity of $\mathcal{O}(d \log d)$ with a substantial constant factor (where d represents the number of pixels), while RankIoU exhibits even higher complexity $\mathcal{O}(d^2)$, thus limiting its practical application. For instance, in LiTS, prediction with RankSEG takes 16.33 seconds compared to just 0.01 seconds with the argmax rule. Second, RankSEG is only applicable to overlapping segmentation settings, where multiple classes can occupy the same pixel, which contrasts with standard benchmarks that typically assume non-overlapping segmentation. In this paper, we overcome these two drawbacks via a *reciprocal moment approximation* (RMA) of RankSEG with the following contributions: (i) we improve RankSEG using RMA, namely RankSEG-RMA, reduces the complexity of both algorithms to $\mathcal{O}(d)$ while maintaining comparable performance; (ii) inspired by RMA, we develop a pixel-wise score function that allows efficient implementation for non-overlapping segmentation settings. We illustrate the effectiveness of our method across various datasets and state-of-the-art models. The code of our method is available in: https://github.com/ZixunWang/RankSEG-RMA.

1 Introduction

Semantic segmentation is a fundamental task in computer vision that assigns each pixel in an image to a specific class, serving as a cornerstone for applications such as autonomous driving [Cordts et al., 2016, Feng et al., 2020], medical image analysis [Heller et al., 2019, Bilic et al., 2023], and augmented reality [Ko and Lee, 2020].

Evaluating the performance of segmentation models naturally requires appropriate metrics that accurately reflect segmentation quality. Specifically, pixel-wise accuracy (Acc) is often biased toward classes that occupy large image regions and fails to account for false positives [Everingham et al., 2010, Wang et al., 2023a]. Consequently, the Intersection over Union (IoU) and Dice metrics have emerged as the standard evaluation measures for semantic segmentation [Cordts et al., 2016,

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

 https://doi.org/10.52202/085713-5452

Zhou et al., 2017]. However, regardless of the metrics employed, most existing works adhere to a classification-based segmentation procedure: (i) training-step: training a model to estimate pixel-wise class probabilities using a strictly proper loss [Gneiting and Raftery, 2007] (e.g., cross-entropy loss [Mao et al., 2023]); (ii) prediction-step: followed by applying argmax or thresholding to these probabilities for the final prediction [Chen et al., 2017, Zhao et al., 2017, Xie et al., 2021]. Yet, as demonstrated by Dai and Li [2023], the prediction-step by argmax and thresholding are inconsistent, meaning that even with an infinite number of data and perfect probability estimation, those approaches still cannot achieve optimal performance in terms of IoU and Dice metrics. Therefore, these methods are typically suboptimal in practical applications.

An alternative direction is designing surrogate loss functions which attempt to optimize IoU or Dice directly, with the most popular approaches being soft-IoU/Dice loss [Rahman and Wang, 2016, Sudre et al., 2017, Eelbode et al., 2020] and Lovász extension loss [Yu and Blaschko, 2018, Berman et al., 2018]. However, Lovász hinge loss has been shown to be inconsistent by Finocchiaro et al. [2022], and consequently, its empirical performance improvements remain controversial [Ma et al., 2021, Dai and Li, 2023]. For soft-IoU/Dice loss, the consistency remains unclear. Nevertheless, soft IoU/Dice loss functions are non-convex, making optimization challenging and unstable in practice. Perhaps for this reason, soft-IoU/Dice loss is typically used in combination with cross-entropy through ad hoc training strategies, with final segmentation predictions made using argmax and thresholding operations. These approaches generally require tuning an additional hyperparameter—the weight between cross-entropy and soft-IoU/Dice loss, resulting in high computational costs and inconvenience in practice.

To this end, a ranking-based consistent segmentation rule (RankSEG; Dai and Li [2023]) is specifically developed to directly optimize IoU and Dice metrics. Unlike the argmax rule and surrogate loss functions, RankSEG offers provable consistency and practical performance improvement. Furthermore, compared to surrogate loss functions, RankSEG only modifies the prediction-step and can serve as a *plug-and-play* module by directly utilizing a model trained with cross-entropy loss, simply replacing the argmax operation in prediction-step.

While theoretically sound, their approach exhibits notable limitations: (1) the algorithms are computationally intensive for high dimensional data—with RankDice, the less demanding of the two, having a time complexity of $\mathcal{O}(d \log d)$ with a large constant factor, where d is the number of pixels. For example, it requires 16.33 seconds on the LiTS [Bilic et al., 2023] dataset, compared to only 0.01 seconds by the argmax rule. (2) In multiclass segmentation, the algorithms are only applicable in overlapping settings where multiple classes can occupy the same pixel, which deviates from standard benchmarks [Everingham et al., 2010, Cordts et al., 2016], and also restricts the application of RankSEG in certain scenarios, such as panoptic segmentation [Kirillov et al., 2019].

Contribution. In this paper, we leverage **reciprocal moment approximation (RMA)** in segmentation to address the aforementioned disadvantages with the following contributions:

- We propose RankSEG-RMA, which reduces the computational complexity of RankSEG (both IoU and Dice) to $\mathcal{O}(d)$ while preserving comparable performance.
- We develop a pixel-wise score function based on RMA, enabling efficient adaptation to non-overlapping segmentation settings, in line with standard benchmarks.
- We have theoretically established the quality of the proposed RMA (Theorem 2), and empirical evidence demonstrate that our method not only outperforms the conventional argmax rule but also significantly reduces computational costs compared to existing RankSEG algorithms.

2 Background

In this section, we begin by distinguishing between two different definitions of the IoU and Dice metrics: IoU^{D}/Dice^{D} and IoU^{I}/Dice^{I} [Wang et al., 2023a], advocating for the latter in practical applications. Building upon IoU^{I}/Dice^{I}, we review RankSEG and its approximated algorithms.

For clarity, our discussion starts with binary segmentation, with extensions to multiclass segmentation presented in Section 3.2. Let $\mathbf{X} \in \mathbb{R}^d, \mathbf{Y} \in \{0,1\}^d$ represent the random variables for an image and its corresponding segmentation mask, respectively. Consider a dataset $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ consisting of n realizations. The segmentation function $\boldsymbol{\delta} : \mathbb{R}^d \to \{0,1\}^d$ produces a predicted mask $\boldsymbol{\delta}(\mathbf{x}) \in \{0,1\}^d$

for a test image $\mathbf{x} \in \mathbb{R}^d$. We denote $p_j(\mathbf{x}) = \mathbb{P}(Y_j = 1|\mathbf{x})$ as the conditional probability of pixel j being a foreground pixel given the image $\mathbf{x}$. Index set $\{1, \cdots, d\}$ is denoted as $[d]$.

2.1 Dice/IoU metrics and its variations in implementation

Dice/IoU metrics are defined based on true positives (TP), false positives (FP), and false negatives (FN). However, in practical implementations, the calculation of these components (TP, FP, FN) can be specifically defined at either the dataset or image level, yielding two different metric implementations. For example, the dataset-level and image-level TP are computed as follows:

$$\mathrm{TP}^{\mathrm{D}}(\boldsymbol{\delta}) = (\mathbf{y}_1^\intercal, \cdots, \mathbf{y}_n^\intercal)^\intercal \Big(\boldsymbol{\delta}^\intercal(\mathbf{x}_1), \cdots, \boldsymbol{\delta}^\intercal(\mathbf{x}_n)\Big) = \sum_{i=1}^n \mathbf{y}_i^\intercal \boldsymbol{\delta}(\mathbf{x}_i), \quad \mathrm{TP}_i^{\mathrm{I}} = \mathbf{y}_i^\intercal \boldsymbol{\delta}(\mathbf{x}_i).$$

Specifically, dataset-level TP aggregates values across the entire dataset, while image-level TP is computed separately for each image. This distinction leads to different averaging strategies when calculating Dice and IoU metrics. Furthermore, dataset-level and image-level Dice are defined as:

$$\mathrm{Dice}^{\mathrm{D}}(\boldsymbol{\delta}) = \frac{2\mathrm{TP}^{\mathrm{D}}(\boldsymbol{\delta})}{\mathrm{TP}^{\mathrm{D}}(\boldsymbol{\delta}) + \mathrm{FP}^{\mathrm{D}}(\boldsymbol{\delta}) + \mathrm{FN}^{\mathrm{D}}(\boldsymbol{\delta})}, \quad \mathrm{Dice}^{\mathrm{I}}(\boldsymbol{\delta}) = \frac{1}{n}\sum_{i=1}^n \frac{2\mathrm{TP}_i^{\mathrm{I}}(\boldsymbol{\delta})}{\mathrm{TP}_i^{\mathrm{I}}(\boldsymbol{\delta}) + \mathrm{FP}_i^{\mathrm{I}}(\boldsymbol{\delta}) + \mathrm{FN}_i^{\mathrm{I}}(\boldsymbol{\delta})},$$

where $\mathrm{FP}^{\mathrm{D}}, \mathrm{FP}^{\mathrm{I}}$ and $\mathrm{FN}^{\mathrm{D}}, \mathrm{FN}^{\mathrm{I}}$ are defined analogously at the dataset-level or image-level.

Although $\mathrm{IoU}^{\mathrm{D}}$/$\mathrm{Dice}^{\mathrm{D}}$ are more prevalent in the literature [Everingham et al., 2010, Cordts et al., 2016], a growing trend [Liu et al., 2023, Kirillov et al., 2023, Wang et al., 2023a] recognizes $\mathrm{IoU}^{\mathrm{I}}$/$\mathrm{Dice}^{\mathrm{I}}$ as more favorable for two key reasons. Firstly, $\mathrm{IoU}^{\mathrm{D}}$/$\mathrm{Dice}^{\mathrm{D}}$ exhibit a bias toward large objects [Yang et al., 2022], which dominate the confusion matrix. This is particularly concerning given the size imbalance in existing datasets [Wang et al., 2023a]. In safe-critical applications, such as medical image analysis or autonomous driving, failing to detect small but critical objects can be catastrophic. Secondly, $\mathrm{IoU}^{\mathrm{I}}$/$\mathrm{Dice}^{\mathrm{I}}$ offers statistical information at the image-level. For instance, the variance of $\mathrm{IoU}^{\mathrm{I}}$/$\mathrm{Dice}^{\mathrm{I}}$ quantifies robustness, and the lower quantile measures worst-case performance [Wang et al., 2023a]. Consequently, we adopt $\mathrm{IoU}^{\mathrm{I}}$/$\mathrm{Dice}^{\mathrm{I}}$ as the focus in this paper. Notably, these two types of metrics differ significantly at the population level. RankSEG-based methods are designed to optimize image-level metrics, which in turn may consequently result in decreased performance on dataset-level metrics.

2.2 RankSEG and its blind approximation

For simplicity, we will omit the dependence on $\mathbf{x}$ hereafter, but it is important to note that all following notations are conditional on $\mathbf{X} = \mathbf{x}$. RankSEG [Dai and Li, 2023] establishes a novel segmentation framework that directly (or consistently) maximizes Dice/IoU metrics. Specifically, it first ranks the pixel-wise class probabilities and then selects the top τ^* pixels as segmented pixels, where τ^* is so-called the optimal volume. This framework is primarily motivated by the optimal rule outlined in the following theorem; a similar result for $\mathrm{IoU}^{\mathrm{I}}$ is omitted for brevity.

Theorem 1 (The Bayes rule for $\mathrm{Dice}^{\mathrm{I}}$-segmentation [Dai and Li, 2023])**.** *Assume that $Y_i \perp Y_j|\mathbf{X}$. A segmentation rule $\boldsymbol{\delta}^*$ is a global maximizer of $\mathbb{E}(\mathrm{Dice}^I(\boldsymbol{\delta}))$ if and only if $\delta_j^* = \mathbb{1}(p_j \geq p_{j_{\tau^*}})$, where j_τ is the index with the τ-th largest probability. The optimal volume τ^* is given by:*

$$\tau^* = \operatorname*{argmax}_{\tau \in \{0,1,\cdots,d\}} \pi(\mathcal{J}_\tau) \quad \textit{with} \quad \pi(\mathcal{J}_\tau) = \sum_{j \in \mathcal{J}_\tau} \mathbb{E}\left(\frac{2p_j}{\tau + \Gamma_{-j} + 1}\right), \tag{1}$$

where $\mathcal{J}_\tau = \{j' : p_{j'} \geq p_{j_\tau}\}$ is the index set of the top τ conditional probabilities with $\mathcal{J}_0 = \emptyset$, and $\Gamma_{-j} = \sum_{j' \neq j} B_{j'}$ is a Poisson-binomial random variable with $B_{j'}$ being a Bernoulli random variable with success probability $p_{j'}$.

An intuitive interpretation of Theorem 1 is that $p_{j_{\tau^*}}$ serves as an adaptive threshold that varies across different input images, in contrast to the fixed threshold (0.5) commonly used in binary segmentation framework. This adaptation, in return, indicates that a fixed threshold framework leads to suboptimal performance in terms of $\mathrm{Dice}^{\mathrm{I}}$. This is illustrated by the following example.

Example. *Consider $d = 2$ with $p_1 = 0.7, p_2 = 0.4$. The Bayes rule produces $\boldsymbol{\delta}^* = (1,1)^\intercal$, whereas the conventional thresholding or argmax rule yields $\widetilde{\boldsymbol{\delta}} = (1,0)^\intercal$. Since $\mathrm{Dice}^I((1,1)^\intercal) \approx 0.827 > 0.607 \approx \mathrm{Dice}^I((1,0)^\intercal)$, the thresholding or argmax rule is suboptimal.*

https://doi.org/10.52202/085713-5452

Blind approximation (BA). The primary computational bottleneck in RankDice is the optimization of the optimal volume. Specifically, computing $\pi(\mathcal{J}_\tau)$ for all $\tau \in \{0, 1, \cdots, d\}$ in (1) has a complexity of $\mathcal{O}(d^2)$. To mitigate this, Dai and Li [2023] proposed *RankDice-BA*, which replaces Γ_{-j} with Γ to make the expectation independent with index j, yielding an approximation for $\pi(\mathcal{J}_\tau)$:

$$\pi_{\text{BA}}(\mathcal{J}_\tau) = \mathbb{E}\left(\frac{2}{\tau + \Gamma + 1}\right)\Big(\sum_{j \in \mathcal{J}_\tau} p_j\Big) = \left(\sum_{l=0}^{d} \frac{2\mathbb{P}(\Gamma = l)}{\tau + l + 1}\right)\Big(\sum_{j \in \mathcal{J}_\tau} p_j\Big). \tag{2}$$

Fast Fourier transform (FFT) is then used to reduce the overall complexity in evaluating $\pi_{\text{BA}}(\mathcal{J}_\tau)$ for all $\tau \in \{0, 1, \cdots, d\}$ to $\mathcal{O}(d \log d)$. While this achieves a significant improvement, BA method still exhibits the following limitations: (1) the constant factor associated with FFT is generally non-negligible in practice; (2) it is challenging to apply in non-overlapping settings, as shown in Dai and Li [2023, Lemma 7]; and (3) BA is not readily applicable to RankIoU due to the large approximation errors, which therefore remains $\mathcal{O}(d^2)$ time complexity. To address these limitations, we propose a reciprocal moment approximation that further reduces the complexity of both RankDice and RankIoU to $\mathcal{O}(d)$ and enables efficient solution for non-overlapping segmentation.

3 RankSEG-RMA

3.1 Reciprocal moment approximation

We begin by introducing the reciprocal moment approximation, which is a technique for approximating the reciprocal moment (or negative first moment) of a Poisson-binomial random variable.

Theorem 2 (Reciprocal moment approximation to RankSEG)**.** *Let Γ be a Poisson-binomial random variable, then for any $\tau \geq 1$, we have*

$$(\mathbb{E}\Gamma + \tau)^{-1} \leq \mathbb{E}(\Gamma + \tau)^{-1} \leq (\frac{d+1}{d}\mathbb{E}\Gamma + \tau - 1)^{-1}. \tag{3}$$

Therefore, we propose the following $\pi_{RMA}(\mathcal{J}_\tau)$ to approximate $\pi(\mathcal{J}_\tau)$ in (1):

$$\pi_{RMA}(\mathcal{J}_\tau) = \frac{2}{\tau + \mathbb{E}\Gamma + 1}\Big(\sum_{j \in \mathcal{J}_\tau} p_j\Big), \tag{4}$$

and its approximation error for any set $\mathcal{I} \subseteq [d]$ and $\tau = |\mathcal{I}|$ is bounded by:

$$|\pi_{RMA}(\mathcal{I}) - \pi(\mathcal{I})| \leq 2(\mathbb{E}\Gamma + \tau)^{-1}. \tag{5}$$

Theorem 2 provides two main results: (i) the RMA approximation form (4) for approximating $\pi(\mathcal{J}_\tau)$, inspired by the exchange of expectation and reciprocal in (3); and (ii) a provable error bound that characterize the quality of the RMA approximation. The primary advantage of using RMA is that it avoids expanding the reciprocal moment (RM) into a sum of d terms, which is computationally expensive. Specifically, $\pi_{\text{RMA}}(\mathcal{J}_\tau)$ converts such a nonlinear expectation into a linear one, allowing the evaluation of $\pi_{\text{RMA}}(\mathcal{J}_\tau)$ for all $\tau \in [d]$ to be performed in $\mathcal{O}(1)$ time, once $\mathbb{E}\Gamma$ and $\sum_{j \in \mathcal{J}_\tau} p_j$ are precomputed. In a sharp contrast, evaluating $\pi(\mathcal{J}_\tau)$ for any $\tau \in [d]$ requires $\mathcal{O}(d)$ operations each time. Notably, the first result, (3) in Theorem 2, credited to Dai and Li [2023] and built upon more fundamental results of reciprocal moments Chao and Strawderman [1972], Wooff [1985], is quite general and may be of independent interest for other applications.

The approximation error bound (5), particularly when $\mathcal{I} = \mathcal{J}_\tau$, decreases as the expected volume of predicted mask increases, which typically occurs when d is large. Even for small objects occupying only a 30×30 region in a 256×256 image, with an expected volume $\mathbb{E}(\Gamma) = \tau = 1000$, the approximation error remains below 0.1%, which is generally acceptable in practice.

We now summarize RankDice-RMA for binary segmentation in Algorithm 1. RankIoU-RMA is developed analogously in Section B, with the same approximation. The first two steps prepare and store intermediate values for evaluating $\widehat{\pi}_{\text{RMA}}(\widehat{\mathcal{J}}_\tau)$ based on an estimated probabilities $\widehat{\mathbf{p}}$. After that, we identify the optimal volume $\widehat{\tau}^*$ and make prediction by selecting the top $\widehat{\tau}^*$ pixels. Neglecting the sorting operation, the time complexity of the RankDice-RMA and RankIoU-RMA is reduced to $\mathcal{O}(d)$, compared to $\mathcal{O}(d \log d)$ for RankDice-BA and $\mathcal{O}(d^2)$ for RankIoU. For example, RankDice-RMA achieves 48x speedup for RankDice-BA in LiTS dataset [Bilic et al., 2023].

Algorithm 1 RankDice-RMA-Binary

Input: Estimated probability map $\widehat{\mathbf{p}} \in [0,1]^d$ for a given input image.

Output: The predicted segmentation mask $\widehat{\boldsymbol{\delta}} \in \{0,1\}^d$.

1: Rank probabilities $\widehat{\mathbf{p}}$ in descending order, yielding $\widehat{p}_{j_1} \geq \cdots \geq \widehat{p}_{j_d}$.

2: Prepare cumulative sum of top probabilities and mean of Poisson-binomial

$$\widehat{q}_\tau = \sum_{k=1}^{\tau} \widehat{p}_{j_k} \quad \text{for } \tau \in [d], \quad \widehat{\mu} = \sum_{j=1}^{d} \widehat{p}_j.$$

3: Compute $\widehat{\pi}_{\text{RMA}}(\widehat{\mathcal{J}}_\tau) = \frac{2\widehat{q}_\tau}{\tau + \widehat{\mu} + 1}$ for $\tau \in [d]$, according to (4).

4: Determine optimal volume $\widehat{\tau}^* = \operatorname{argmax}_{\tau \in [d]} \widehat{\pi}_{\text{RMA}}(\widehat{\mathcal{J}}_\tau)$.

5: Make prediction by $\widehat{\delta}_j = \mathbb{1}(p_j \geq \widehat{p}_{j_{\widehat{\tau}^*}})$ for $j \in [d]$.

3.2 RMA-score for non-overlapping multiclass segmentation

To extend RankSEG to non-overlapping multiclass segmentation, a natural approach is first applying binary RankSEG to each class independently, and then address any overlaps. As discussed in the introduction, perfectly addressing overlaps is currently beyond the capabilities of RankSEG, as the non-overlapping constraint leads to a nonlinear assignment problem [Kuhn, 1955], which is generally computationally intractable. Therefore, the focus of this section is on utilizing RMA to solve overlapping pixels provided by RankSEG, ultimately producing non-overlapping segmentation.

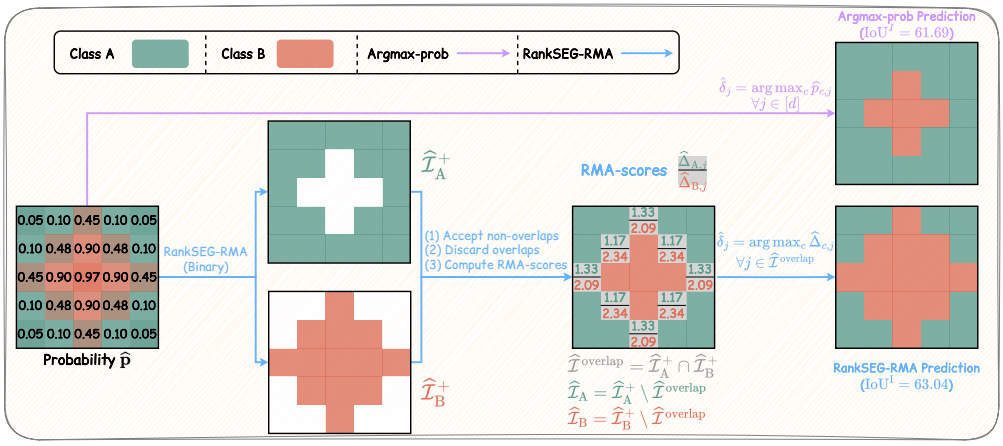

Figure 1: Comparison of Argmax-prob and RankSEG-RMA. (a) Argmax-prob: each pixel is predicted to the class with the highest probability. (b) RankSEG-RMA: segmentation masks $\widehat{\mathcal{I}}_c^+$ for each class are obtained independently; non-overlapping parts $\widehat{\mathcal{I}}_c$ are accepted, while overlaps $\widehat{\mathcal{I}}^{\text{overlap}}$ are resolved by applying argmax over RMA-scores.

Our solution draws inspiration from *Argmax-prob* method, which efficiently resolves non-overlapping constraint by assigning pixels to their highest-probability classes, that is,

$$\widehat{\delta}_j = \operatorname*{argmax}_{c \in [C]} \widehat{p}_{c,j} \quad \forall j \in [d],$$

where $\widehat{p}_{c,j}$ is the estimated probability of pixel j belonging to class c. This approach is computationally efficient but does not guarantee optimal performance. One reason is that merely examining probability values does not accurately reflect how individual pixel assignments contribute to segmentation metrics. To address this issue, we extend the probability in the argmax framework to a Dice/IoU-related score. Unlike the probability score, our proposed score function is grounded on the Bayes rule in Theorem 1 and can be efficiently computed by RMA. We refer to the score function as *RMA-score*.

https://doi.org/10.52202/085713-5452

Algorithm 2 RankDice-RMA-Multiclass

Input: Estimated probability map $\widehat{\mathbf{p}} \in [0,1]^{C \times d}$.

Output: The predicted segmentation mask $\widehat{\boldsymbol{\delta}} \in [C]^d$.

```
/* Obtain overlapping segmentation mask */
for c = 1 to C do
    ψ̂_c = RankDice-RMA-Binary(p̂_c), Î_c^+ = {j : ψ̂_{c,j} = 1}.
end for
/* Resolve overlapping by argmax over RMA-scores */
Identify overlapping indices, Î^overlap = ⋃_{c≠c'} (Î_c^+ ∩ Î_{c'}^+).
for c = 1 to C do
    Discard assignments for overlapping pixels, Î_c = Î_c^+ \ Î^overlap.
    Accept prediction for not overlapping pixels, δ̂_j = c for j ∈ Î_c.
end for
Compute RMA-scores, Δ̂_{c,j} via (7) for j ∈ Î^overlap and c ∈ [C].
Resolve overlapping by argmax, δ̂_j = argmax_{c∈[C]} Δ̂_{c,j} for j ∈ Î^overlap.
Return δ̂
```

To proceed, let $\widehat{\mathcal{I}}_c^+$ denote the index set of pixels assigned to class c by RankSEG, $\widehat{\mathcal{I}}^{\text{overlap}} = \bigcup_{c \neq c'} (\widehat{\mathcal{I}}_c^+ \cap \widehat{\mathcal{I}}_{c'}^+)$ the index set of overlapping pixels, and $\widehat{\mathcal{I}}_c = \widehat{\mathcal{I}}_c^+ \setminus \widehat{\mathcal{I}}^{\text{overlap}}$ the non-overlapping part, as illustrated in Figure 1. We resolve $\widehat{\mathcal{I}}^{\text{overlap}}$ to ensure segmentation masks are non-overlapping:

$$\widehat{\delta}_j = \operatorname*{argmax}_{c \in [C]} \widehat{\Delta}_{c,j}, \quad \forall j \in \widehat{\mathcal{I}}^{\text{overlap}}, \tag{6}$$

where $\widehat{\Delta}_{c,j}$ is increment of Dice-RMA by adding pixel j for class c, which is defined as:

$$\widehat{\Delta}_{c,j} = \widehat{\pi}_{\text{RMA}}(\widehat{\mathcal{I}}_c \cup \{j\}) - \widehat{\pi}_{\text{RMA}}(\widehat{\mathcal{I}}_c) = \frac{2\left(\widehat{p}_{c,j} + \sum_{i \in \widehat{\mathcal{I}}_c} \widehat{p}_{c,i}\right)}{|\widehat{\mathcal{I}}_c| + \widehat{\mu}_c + 2} - \frac{2\sum_{i \in \widehat{\mathcal{I}}_c} \widehat{p}_{c,i}}{|\widehat{\mathcal{I}}_c| + \widehat{\mu}_c + 1}, \tag{7}$$

where $\widehat{\mu}_c = \sum_{j=1}^d \widehat{p}_{c,j}$ represents the estimated mean volume of class c. Intuitively, (6) maximizes the immediate improvement by choosing the class that yields the highest marginal gain in the Dice-RMA objective. While this greedy solution does not guarantee a globally optimal assignment over all overlapping pixels simultaneously, it is computationally efficient and empirically effective for reducing overlap and improving final segmentation performance.

To summarize, the procedure of RankDice-RMA for multiclass segmentation is presented in Algorithm 2. After applying binary RankDice-RMA, the predicted set $\widehat{\mathcal{I}}_c^+$ of each class c is obtained. The overlapping pixels $\widehat{\mathcal{I}}^{\text{overlap}}$ are then identified, and the assignments for these pixels are discarded. The non-overlapping pixels $\widehat{\mathcal{I}}_c$ are assigned to their respective classes. Finally, we compute the RMA-scores for the overlapping pixels and resolve overlaps by selecting the class with the highest score. The complexity for addressing overlapping is $\mathcal{O}(Cd)$, which is no worse than Argmax-prob.

4 Experiments

4.1 Setup

Datasets. We conduct experiments on five datasets: (1) PASCAL VOC [Everingham et al., 2010], (2) Cityscapes [Cordts et al., 2016], (3) ADE20K [Zhou et al., 2017], (4) LiTS [Bilic et al., 2023], and (5) KiTS [Heller et al., 2021]. These datasets cover a diverse range of scenarios, including urban scenes (Cityscapes), "thing" and "stuff" (PASCAL VOC and ADE20K), as well as medical images (LiTS and KiTS). The datasets contain between 200 images (LiTS) and 20,000 images (ADE20K), and the number of classes varies from binary segmentation (LiTS) to over a hundred (ADE20K). We only segment tumors in LiTS and KiTS, treating them as binary segmentation tasks.

Models. We employ following six segmentation models: (1) UNet [Ronneberger et al., 2015], (2) DeepLabV3+ [Chen et al., 2018], (3) PSPNet [Zhao et al., 2017], (4) UPerNet [Xiao et al.,

2018], (5) SegFormer [Xie et al., 2021], and (6) CPT [Tang et al., 2025]. The first four models are CNN-based and utilize backbones such as ResNet [He et al., 2016] or ConvNeXt [Liu et al., 2022], whereas SegFormer and CPT are transformer-based models. The models are trained using the cross-entropy loss, and we compare the proposed RankSEG-RMA with the conventional argmax or thresholding rule for multiclass or binary segmentation, respectively. The training details can be found in Section D.

Evaluation. As discussed in Section 2.1, we evaluate the segmentation models using both mIoU^{I}/mDice^{I} and mIoU^{C}/mDice^{C}, which are straightforward extensions of binary metric IoU^{I}/Dice^{I} to multiclass segmentation. The metrics with superscripts ^{I} and ^{C} differ when not all classes are present in every image (see Wang et al. [2023a] for details).

4.2 Overall performance

Table 1: Performance for different prediction methods with various models on PASCAL VOC, Cityscapes, and ADE20K.

Model	Prediction	PASCAL VOC				Cityscapes				ADE20K			
		mIoU^{I}	mIoU^{C}	mDice^{I}	mDice^{C}	mIoU^{I}	mIoU^{C}	mDice^{I}	mDice^{C}	mIoU^{I}	mIoU^{C}	mDice^{I}	mDice^{C}
PSPNet	Argmax-prob	83.59	72.59	87.69	78.22	71.33	63.38	78.96	71.34	49.78	33.83	56.89	40.36
(ResNet50)	RankDice-RMA	**84.21**	**73.91**	**88.42**	**79.75**	**72.00**	**64.20**	**79.68**	**72.28**	**50.70**	**36.30**	**58.52**	**43.67**
PSPNet	Argmax-prob	85.48	75.57	89.18	80.78	73.07	65.89	80.45	73.55	51.32	37.42	58.66	44.44
(ResNet101)	RankDice-RMA	**85.98**	**76.64**	**89.74**	**81.94**	**73.72**	**66.53**	**81.14**	**74.28**	**51.57**	**38.09**	**59.17**	**45.29**
DeepLabV3+	Argmax-prob	84.19	73.96	88.11	79.31	73.55	65.98	80.80	73.63	49.78	33.83	56.89	40.36
(ResNet50)	RankDice-RMA	**84.79**	**75.26**	**88.84**	**80.88**	**74.05**	**66.68**	**81.38**	**74.49**	**49.82**	**34.28**	**57.19**	**40.92**
DeepLabV3+	Argmax-prob	86.40	77.25	89.83	82.08	73.37	66.17	80.59	73.71	52.53	37.52	59.57	44.13
(ResNet101)	RankDice-RMA	**86.80**	**78.14**	**90.32**	**83.14**	**73.92**	**66.68**	**81.24**	**74.33**	**52.64**	**38.14**	**59.95**	**44.85**
SegFormer	Argmax-prob	85.40	75.85	89.21	81.13	73.24	65.57	80.49	73.16	53.03	38.19	60.06	44.83
(MiTB2)	RankDice-RMA	**85.85**	**76.01**	**89.44**	**81.04**	**73.81**	**66.41**	**81.14**	**74.13**	**53.67**	**39.09**	**61.09**	**46.11**
SegFormer	Argmax-prob	86.86	77.57	90.11	82.15	73.32	66.13	80.53	73.65	54.09	40.00	61.03	46.50
(MiTB4)	RankDice-RMA	**87.28**	**78.59**	**90.56**	**83.22**	**74.10**	**67.14**	**81.38**	**74.74**	**54.72**	**40.82**	**61.92**	**47.57**
UPerNet	Argmax-prob	87.82	79.52	91.03	84.11	75.66	68.83	82.61	76.08	56.94	42.86	63.98	49.61
(ConvNeXt)	RankDice-RMA	**88.25**	**80.31**	**91.48**	**84.98**	**76.17**	**69.57**	**83.21**	**76.97**	**57.67**	**43.84**	**64.93**	**50.85**
CPT	Argmax-prob	88.56	80.74	91.62	85.18	75.33	68.39	82.25	75.74	57.85	44.59	64.75	51.27
(Swin-Large)	RankDice-RMA	**88.89**	**81.53**	**92.01**	**86.08**	**75.86**	**69.29**	**82.85**	**76.76**	**58.63**	**45.56**	**65.83**	**52.58**

Table 2: Performance for different prediction methods with various models on LiTS and KiTS.

Prediction	Model	LiTS		KiTS		Model	LiTS		KiTS	
		IoU^{I}	Dice^{I}	IoU^{I}	Dice^{I}		IoU^{I}	Dice^{I}	IoU^{I}	Dice^{I}
Argmax-prob	DeepLabV3+ (ResNet18)	34.31	42.81	54.61	47.20	UNet (ResNet18)	36.40	45.18	56.03	49.28
RankDice-BA		**36.11**	**45.04**	**58.00**	**50.57**		**38.34**	**47.54**	**59.10**	**52.08**
RankDice-RMA		**36.12**	**45.04**	**58.00**	**50.57**		**38.34**	**47.54**	**59.10**	**52.08**
Argmax-prob	DeepLabV3+ (ResNet50)	38.45	47.58	61.16	54.19	UNet (ResNet50)	38.45	47.58	57.36	51.00
RankDice-BA		**40.09**	**49.50**	**63.56**	**56.22**		**40.71**	**50.08**	**60.07**	**50.34**
RankDice-RMA		**40.09**	**49.50**	**63.56**	**56.22**		**40.70**	**50.07**	**60.07**	**53.54**

Table 3: Time consumption (in seconds) of model forward and different prediction rules with single A100 GPU. DeepLabV3+ (ResNet50) is used for the medical datasets, while UPerNet (ConvNeXt) is used for the others. The mean and standard deviation over 10 runs are reported. ✗ indicates that the method is not applicable due to non-overlapping benchmark setups.

	Pascal VOC	Cityscapes	ADE20K	LiTS	KiTS
Argmax-prob	0.05 (±0.01)	0.22 (±0.01)	0.43 (±0.08)	0.01 (±0.00)	0.01 (±0.00)
RankDice-RMA	6.93 (±1.14)	10.15 (±1.77)	58.00 (±3.44)	0.34 (±1.19)	0.26 (±0.15)
RankDice-BA	✗	✗	✗	16.33 (±1.19)	9.99 (±0.15)
Model forward	40.77 (±5.15)	175.81 (±3.02)	324.59 (±13.42)	14.15 (±0.64)	11.86 (±0.20)

Results for PASCAL VOC, Cityscapes, and ADE20k are presented in Table 1, while those for LiTS and KiTS are shown in Table 2. The best performance within each model is highlighted in **bold**, while the best across all models is highlighted in pink. If two performances are very close and both are the best, we highlight both. Three observations can be drawn from these results.

https://doi.org/10.52202/085713-5452

- **Our proposed method significantly outperforms the conventional Argmax-prob across all datasets and models**, irrespective of light or heavy backbones, demonstrating its effectiveness and robustness. For instance, on Cityscapes with SegFormer (MiTB4), RankDice-RMA improves mDice^{I} and mDice^{C} by 0.85% and 1.09%, respectively. In addition, on LiTS with UNet (ResNet50), RankDice-RMA outperforms Argmax-prob by 2.49% in Dice^{I}.
- As shown in Tables 2 and 3, RankDice-RMA achieves significant time efficiency improvements over RankDice-BA while maintaining similar performance on the LiTS (48x speedup) and KiTS datasets (38x speedup). Hence, we conclude that **RankDice-RMA is a strict improvement over RankDice-BA**. Although RankDice-RMA is slower than Argmax-prob, the absolute time consumption is negligible compared to the model forward time. In contrast, such argument can not be applied to RankDice-BA, whose time consumption is comparable to the model forward.
- **RankDice-RMA simultaneously boost IoU performance**, even though it is originally motivated by the Bayes rule for Dice. Furthermore, RankDice-RMA and RankIoU-RMA achieve nearly identical performance across all experiments (results for RankIoU-RMA are omitted for simplicity), suggesting that the two metrics are closely related and that either RankDice-RMA or RankIoU-RMA can serve as a unified prediction method for both metrics.

4.3 Class-wise performance

To further evaluate the performance of RankDice-RMA, we report class-wise results on PASCAL VOC in Table 4. Improvements over Argmax-prob are highlighted in **green** for positive changes and in **red** for negative ones. The results indicate that RankDice-RMA consistently enhances performance across most classes. Two key observations can be made:

- The performance gains are more pronounced for classes with lower baseline performance, suggesting that **RankDice-RMA is particularly effective for difficult classes**. For instance, the `Chair` class, which exhibits a low IoU of 48.05% under Argmax-prob, is boosted to 50.52%, an enhancement of 2.47%; whereas the `Areoplane` class, with a high initial IoU of 90.39%, only sees a marginal improvement of 0.45%. This trend may result in negative changes for classes like `Bird` and `Sheep`, where Argmax-prob already performs well, leaving limited room for improvement with the Bayes rule.
- Although the error bound in Theorem 2 implies a larger approximation error for classes with smaller volume, the results show that **RankDice-RMA still achieves substantial performance gains for these small objects**. For example, our analysis indicates that `Bottle` and `Chair` are among the smallest objects in the dataset. Nonetheless, these classes exhibit significant improvements, possibly because the benefits of the Bayes rule outweigh the approximation error.

Table 4: Class-wise IoU on PASCAL VOC with UPerNet (ConvNeXt).

Prediction	Aeroplane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow
Argmax-prob	90.39	50.33	91.18	81.19	69.21	89.55	78.78	92.24	48.05	92.47
RankDice-RMA	90.84	51.76	90.86	81.70	71.77	90.31	80.46	92.43	50.52	92.69
(Improvement)	**+0.45**	**+1.43**	**-0.32**	**+0.51**	**+2.56**	**+0.76**	**+1.68**	**+0.19**	**+2.47**	**+0.22**
Prediction	**Dining Table**	**Dog**	**Horse**	**Motorbike**	**Person**	**Potted Plant**	**Sheep**	**Sofa**	**Train**	**TV Monitor**
Argmax-prob	54.12	93.55	91.21	89.51	82.26	57.11	92.59	62.09	91.97	77.25
RankDice-RMA	55.28	93.56	91.34	89.80	83.55	59.20	92.54	62.88	92.17	78.05
(Improvement)	**+1.16**	**+0.01**	**+0.13**	**+0.29**	**+1.29**	**+2.09**	**-0.05**	**+0.79**	**+0.20**	**+0.80**

4.4 Worst-case analysis

For safe-critical applications, it is crucial to evaluate the worst-case performance of segmentation models. In this context, image-level metrics provide more detailed insights than dataset-level metrics for assessing worst-case scenarios [Wang et al., 2023a]. Without loss of generality, consider $\text{mIoU}_1^I \leq \text{mIoU}_2^{\text{I}} \leq \cdots \leq \text{mIoU}_n^{\text{I}}$ denote the sorted image-level mIoU values for n images in a test set. We define the average mIoU over those below the lowest q-th quantile as:

$$\text{mIoU}^{\text{I}_q} = \frac{1}{\lfloor nq \rfloor} \sum_{i=1}^{\lfloor nq \rfloor} \text{mIoU}_i^{\text{I}}.$$

By definition, this metric quantifies performance of the worst q-th quantile images. Table 5 presents the mIoU^{I_5} and $\text{mIoU}^{\text{I}_{10}}$ results, where UPerNet(ConvNeXt) is used for PASCAL VOC, Cityscapes, and ADE20K, while DeepLabV3+(ResNet50) is used for LiTS and KiTS. The results demonstrate that **our method also improves the worst-case performance across all datasets**.

Table 5: mIoU^{I_5} and $\text{mIoU}^{\text{I}_{10}}$ on PASCAL VOC, Cityscapes, ADE20K, LiTS, and KiTS.

Prediction	mIoU^{I_5}					$\text{mIoU}^{\text{I}_{10}}$				
	VOC	Cityscapes	ADE20K	LiTS	KiTS	VOC	Cityscapes	ADE20K	LiTS	KiTS
Argmax-prob	44.80	59.02	24.99	2.13	6.72	52.10	61.72	29.54	4.05	8.39
RankDice-RMA	**46.21**	**59.96**	**25.56**	**2.70**	**8.49**	**53.17**	**62.51**	**30.35**	**4.81**	**10.38**
(Improvement)	**+1.41**	**+0.94**	**+0.57**	**+0.57**	**+1.77**	**+1.07**	**+0.79**	**+0.81**	**+0.76**	**+1.89**

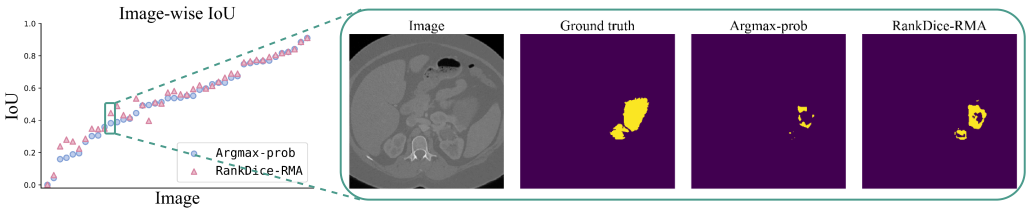

Figure 2: Image-wise performance and an example of a worst-case segmentation on KiTS. The left plot presents the IoU for each image, with indices sorted in ascending order according to IoU under Argmax-prob. The right plot displays the segmentation results for a slice of a worst-case image.

Figure 2 shows image-level IoU for a KiTS validation fold and a worst-case segmentation example. The left plot shows IoU values for each image. It is evident that **our method outperforms Argmax-prob across most images, especially on difficult cases**. The right plot displays segmentation results for one worst-case image. The tumor consists of two adjacent segments, but Argmax-prob captures only a small portion of the larger segment and almost misses the smaller one. In contrast, **our method not only produces a more complete segmentation but also successfully identifies the smaller segment**. This example highlights the potential of our method for challenging clinical scenarios.

4.5 Ablation studies

Effect of the RMA-score. We have introduced the RMA-score to address $\widehat{\mathcal{I}}^{\text{overlap}}$ that occurs when applying RankDice for each class independently. We now demonstrate that the scores are indeed crucial for improving performance by comparing them with two ad-hoc alternatives in (6):

- **Prob-scores.** The predicted probability $\widehat{p}_{c,j}$ is directly used as score of pixel j for class c.
- **WProb-scores.** As inspired by the RMA-scores or intuitive reasoning, classes with more already predicted pixels should have lower preference when resolving overlaps. Hence, a weighted version of the predicted probabilities is considered, i.e., $\widehat{s}_{c,j} = \widehat{p}_{c,j}/|\widehat{\mathcal{I}}_c|$.

As shown in Table 6, WProb-scores outperform Prob-scores on Pascal VOC and Cityscapes, supporting the intuition to account for predicted volume. However, WProb-scores underperform on ADE20K, indicating that simple weighting fails when many classes are present. In contrast, RMA-scores consistently perform best, particularly on ADE20K, where overlapping phenomena are more complex due to the large number of classes. This superiority is due to that RMA-scores are derived from the Bayes rule, making them more principled than heuristic methods. These results support our claim that RMA-scores are essential for improved performance.

Table 6: mIoU^{I} of using different scores.

	Pascal VOC	Cityscapes	ADE20K
Prob-scores	87.83	75.75	<u>56.96</u>
WProb-scores	<u>88.17</u>	<u>75.89</u>	56.75
RMA-scores	**88.25**	**76.17**	**57.67**

Effect of different bounds in RMA. Recall that Theorem 2 provides both lower and upper bounds under RMA, with the lower bound being preferred for its simplicity. As a complement, we further find that using the upper bound as an alternative approximation yields same performance. This suggests that **the choice of different bounds does not bother**. More importantly, this confirms that the bounds are tight, aligning with our theoretical analysis in Theorem 2.

https://doi.org/10.52202/085713-5452

5 Conclusion

In this paper, we propose RankSEG-RMA, a novel segmentation algorithm that grounds on the Bayes rule, and enjoys computational efficiency by using reciprocal moment approximation (RMA). Extensive experiments across various datasets and models demonstrate that RankSEG-RMA outperforms the conventional Argmax-prob and significantly reduces computational cost compared to the existing RankSEG-BA. Nevertheless, two limitations are noteworthy for future improvement. First, the proposed overlap resolution method predicts each pixel independently, which may not be optimal; future work could explore more global approaches while maintaining computational efficiency. Second, our work builds upon the assumption of conditional independence in Bayes rule, which could be relaxed in subsequent research.

Acknowledgments

We thank the anonymous Area Chair and reviewers for their valuable feedback, suggestions and support. This work was supported by the Hong Kong RGC-ECS Grant 24302422 and Hong Kong RGC Grant 14304823.

References

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4413–4421, 2018.

Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023.

Min-Te Chao and WE Strawderman. Negative moments of positive random variables. *Journal of the American Statistical Association*, 67(338):429–431, 1972.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

Ben Dai and Chunlin Li. Rankseg: a consistent ranking-based framework for segmentation. *Journal of Machine Learning Research*, 24(224):1–50, 2023.

Krzysztof Dembczynski, Arkadiusz Jachnik, Wojciech Kotlowski, Willem Waegeman, and Eyke Hüllermeier. Optimizing the f-measure in multi-label classification: Plug-in rule approach versus structured loss minimization. In *International Conference on Machine Learning*, pages 1130–1138. PMLR, 2013.

Tom Eelbode, Jeroen Bertels, Maxim Berman, Dirk Vandermeulen, Frederik Maes, Raf Bisschops, and Matthew B Blaschko. Optimization for medical image segmentation: theory and practice when evaluating with dice score or jaccard index. *IEEE Transactions on Medical Imaging*, 39(11): 3679–3690, 2020.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88: 303–338, 2010.

Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360, 2020.

Jessica J Finocchiaro, Rafael Frongillo, and Enrique B Nueve. The structured abstain problem and the lovász hinge. In *Conference on Learning Theory*, pages 3718–3740. PMLR, 2022.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*, 2019.

Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical Image Analysis*, 67:101821, 2021.

Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

Tae-young Ko and Seung-ho Lee. Novel method of semantic segmentation applicable to augmented reality. *Sensors*, 20(6):1737, 2020.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

Yi Lin. A note on margin-based loss functions in classification. *Statistics & Probability Letters*, 68 (1):73–82, 2004.

Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23592–23601, 2023.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

Jun Ma, Jianan Chen, Matthew Ng, Rui Huang, Yu Li, Chen Li, Xiaoping Yang, and Anne L Martel. Loss odyssey in medical image segmentation. *Medical Image Analysis*, 71:102035, 2021.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *International Conference on Machine learning*, pages 23803–23828. PMLR, 2023.

Sebastian Nowozin. Optimal decisions from probabilistic models: the intersection-over-union case. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 548–555, 2014.

 https://doi.org/10.52202/085713-5452

Dian Qin, Jia-Jun Bu, Zhe Liu, Xin Shen, Sheng Zhou, Jing-Jun Gu, Zhi-Hua Wang, Lei Wu, and Hui-Fen Dai. Efficient medical image segmentation based on knowledge distillation. *IEEE Transactions on Medical Imaging*, 40(12):3820–3831, 2021.

Md Atiqur Rahman and Yang Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International Symposium on Visual Computing*, pages 234–244. Springer, 2016.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017.

Quan Tang, Chuanjian Liu, Fagui Liu, Jun Jiang, Bowen Zhang, CL Philip Chen, Kai Han, and Yunhe Wang. Rethinking feature reconstruction via category prototype in semantic segmentation. *IEEE Transactions on Image Processing*, 2025.

Ambuj Tewari and Peter L Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8(5), 2007.

Zifu Wang, Maxim Berman, Amal Rannen-Triki, Philip Torr, Devis Tuia, Tinne Tuytelaars, Luc V Gool, Jiaqian Yu, and Matthew Blaschko. Revisiting evaluation metrics for semantic segmentation: Optimization and evaluation of fine-grained intersection over union. *Advances in Neural Information Processing Systems*, 36:60144–60225, 2023a.

Zifu Wang, Teodora Popordanoska, Jeroen Bertels, Robin Lemmens, and Matthew B Blaschko. Dice semimetric losses: Optimizing the dice score with soft labels. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 475–485. Springer, 2023b.

David A Wooff. Bounds on reciprocal moments with applications and developments in stein estimation and post-stratification. *Journal of the Royal Statistical Society: Series B (Methodological)*, 47 (2):362–371, 1985.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022.

Jiaqian Yu and Matthew B Blaschko. The lovász hinge: A novel convex surrogate for submodular losses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(3):735–748, 2018.

Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5(Oct):1225–1251, 2004.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2881–2890, 2017.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 633–641, 2017.

Table of Contents

A Segmentation calibration and RankSEG Framework **13**

B RankIoU-RMA **14**

C Proof of Theorem 2 **15**

D Training Details **16**

E Additional Results **16**

E.1 Statistical Significance Test 16

E.2 More Qualitative Visualizations 17

A Segmentation calibration and RankSEG Framework

Given a training dataset $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ where $\mathbf{x}_i \in \mathcal{X},\ \mathbf{y}_i \in \mathcal{Y}$, a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, and a hypothesis class $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}$, the empirical risk and population risk are defined as:

$$\widehat{\mathcal{R}}_\ell(h) = \frac{1}{n}\sum_{i=1}^n \ell(h(\mathbf{x}_i), \mathbf{y}_i) \quad \text{and} \quad \mathcal{R}_\ell(h) = \mathbb{E}_{\mathbf{X},\mathbf{Y}}[\ell(h(\mathbf{X}), \mathbf{Y})].$$

The empirical risk minimizer $\widehat{h}_n = \operatorname{argmin}_{h\in\mathcal{H}} \widehat{\mathcal{R}}_\ell(h)$ is used for making predictions. However, it is often the case that the target loss function is neither differentiable and nor convex, such as the zero-one loss in classification or the negative of IoU^{I}/Dice^{I} in our case, making direct optimization infeasible. Therefore, a surrogate loss function $\phi : \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}$, combined with a surrogate hypothesis class $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Z}\}$ and a decoding function (also known as link function) $d : \mathcal{Z} \to \mathcal{Y}$, is typically employed:

$$\widehat{f}_n = \operatorname*{argmin}_{f\in\mathcal{F}} \widehat{\mathcal{R}}_\phi \quad \text{and} \quad \bar{h}_n = d \circ \widehat{f}_n.$$

Note that the surrogate loss is designed to be easier to optimize than the original loss, the output space of the surrogate hypothesis may not align with the label space, and the decoding function maps the surrogate prediction back to the original label space. The desired property of the surrogate loss is calibrated, as specified in Theorem A.1.

Definition A.1 (Calibration)**.** A surrogate loss ϕ, associated with a decoding function d, is calibrated with respect to a target loss ℓ if, for any distribution over $\mathcal{X} \times \mathcal{Y}$ and any sequences $\{f_n\}_{n\in\mathbb{N}} \subset \mathcal{F}$, the following holds:

$$\left(\mathcal{R}_\phi(\widehat{f}_n) \to \inf_{f\in\mathcal{F}} \mathcal{R}_\phi(f)\right) \quad \Longrightarrow \quad \left(\mathcal{R}_\ell(d \circ \widehat{f}_n) \to \inf_{h\in\mathcal{H}} \mathcal{R}_\ell(h)\right) \quad \text{as} \quad n \to \infty.$$

For example, hinge loss with sign as decoding function and cross entropy loss with argmax as decoding function are calibrated with respect to zero-one loss in binary classification and multiclass classification, respectively [Lin, 2004, Zhang, 2004, Bartlett et al., 2006, Tewari and Bartlett, 2007, Mao et al., 2023].

In general, there are two principled approaches to achieving calibration or consistency: (1) designing a consistent surrogate loss function and making predictions via a suitable decoding function [Bartlett et al., 2006, Tewari and Bartlett, 2007], and (2) directly deriving the Bayes rule for target metrics and plugging in the estimated probabilities for prediction [Nowozin, 2014, Dembczynski et al., 2013, Dai and Li, 2023].

RankSEG [Dai and Li, 2023] belongs to the latter category. It does not require a carefully designed surrogate loss function and can be directly applied to models trained with cross-entropy loss; however, the decoding step is more involved. Nevertheless, Dai and Li [2023] demonstrate a ranking property of the Bayes rule, stating that the optimal prediction is to select the top τ^* pixels with the highest conditional probabilities, which significantly simplifies the decoding step.

 https://doi.org/10.52202/085713-5452

B RankIoU-RMA

Theorem B.1 (The Bayes rule for IoU^{I}-segmentation [Dai and Li, 2023]). *Assume that $Y_i \perp Y_j | \mathbf{X}$. A segmentation rule $\boldsymbol{\delta}^*$ is a global maximizer of $\mathbb{E}(IoU^I(\boldsymbol{\delta}))$ if and only if $\delta_j^* = \mathbb{1}(p_j \geq p_{j_{\tau^*}})$, and j_τ is the index with τ-th largest probability. The optimal volume τ^* is given by:*

$$\tau^* = \operatorname*{argmax}_{\tau \in \{0,1,\cdots,d\}} \nu(\mathcal{J}_\tau) \quad \textit{with} \quad \nu(\mathcal{J}_\tau) = \left(\sum_{j \in \mathcal{J}_\tau} p_j \right) \mathbb{E}\left(\frac{1}{\tau + \Gamma_{-\mathcal{J}_\tau}} \right) \tag{8}$$

where $\mathcal{J}_\tau = \{j' : p_{j'} \geq p_{j_\tau}\}$ is the index set of the top τ conditional probabilities with $\mathcal{J}_0 = \emptyset$, and $\Gamma_{-\mathcal{J}_\tau} = \sum_{j' \notin \mathcal{J}_\tau} B_{j'}$ is Poisson-binomial random variable with $B_{j'}$ being a Bernoulli random variable with success probability $p_{j'}$.

According to Theorem B.1, the Bayes rule for IoU^{I} shares substantial similarity with that of Dice^{I}, both of which consist of two parts: (1) ranking the conditional probabilities and (2) selecting the top τ^* pixels as positives. The primary difference lies in the computation of score functions when determining of the optimal volume τ^*, which is tailored to the respective metric. The consistency of RankSEG [Dai and Li, 2023, Lemma 10] is established by plugging in the estimated probabilities $\widehat{p}_j(\mathbf{x}; \theta)$, where θ is the model parameter trained by minimizing a strictly proper loss [Gneiting and Raftery, 2007].

Note that replacing $\Gamma_{-\mathcal{J}_\tau}$ with Γ in Theorem B.1 leads to large approximation error, especially when τ is large. Therefore, Blind approximation is no longer applicable in this case. However, RMA technique can still be employed to approximate $\nu(\mathcal{J}_\tau)$:

$$\nu_{\mathrm{RMA}}(\mathcal{J}_\tau) = \left(\sum_{j \in \mathcal{J}_\tau} p_j \right) \frac{1}{\tau + \mathbb{E}(\Gamma_{-\mathcal{J}_\tau})} \tag{9}$$

Based on this, we develop RankIoU-RMA for binary segmentation, as described in Algorithm 3. This algorithm is highly similar to RankDice-RMA, with the only difference being the use of the target function $\widehat{\nu}(\widehat{\mathcal{J}}_\tau)$.

Algorithm 3 RankIoU-RMA-Binary

Input: Estimated probability map $\widehat{\mathbf{p}} \in [0,1]^d$.
Output: The predicted segmentation mask $\widehat{\boldsymbol{\delta}} \in \{0,1\}^d$.

1: Rank probabilities $\widehat{\mathbf{p}}$ in descending order, yielding $\widehat{p}_{j_1} \geq \cdots \geq \widehat{p}_{j_d}$.
2: Prepare cumulative sum of top probabilities and mean of Poisson-binomial

$$\widehat{q}_\tau = \sum_{k=1}^{\tau} \widehat{p}_{j_k} \quad \text{for } \tau \in [d], \quad \widehat{\mu} = \sum_{j=1}^{d} \widehat{p}_j.$$

3: Compute $\widehat{\nu}_{\mathrm{RMA}}(\widehat{\mathcal{J}}_\tau) = \frac{\widehat{q}_\tau}{\tau + (\widehat{\mu} - \widehat{q}_\tau)}$ for $\tau \in [d]$, according to (9).
4: Determine optimal volume $\widehat{\tau}^* = \operatorname{argmax}_{\tau \in [d]} \widehat{\nu}_{\mathrm{RMA}}(\widehat{\mathcal{J}}_\tau)$.
5: Make prediction by $\widehat{\delta}_j = \mathbb{1}(p_j \geq \widehat{p}_{j_{\widehat{\tau}^*}})$ for $j \in [d]$.

In order to extend RankIoU-RMA to non-overlapping multiclass segmentation, it suffices to use the following RMA-scores for IoU, followed by an argmax to resolve overlaps:

$$\widehat{\Omega}_{c,j} = \widehat{\nu}(\widehat{\mathcal{I}}_c \cup \{j\}) - \widehat{\nu}(\widehat{\mathcal{I}}_c) = \frac{\widehat{p}_{c,j} + \sum_{k \in \widehat{\mathcal{I}}_c} \widehat{p}_{c,k}}{|\widehat{\mathcal{I}}_c| + (\widehat{\mu}_c - \widehat{p}_{c,j} - \sum_{k \in \widehat{\mathcal{I}}_c} \widehat{p}_{c,k})} - \frac{\sum_{k \in \widehat{\mathcal{I}}_c} \widehat{p}_{c,k}}{|\widehat{\mathcal{I}}_c| + (\widehat{\mu}_c - \sum_{k \in \widehat{\mathcal{I}}_c} \widehat{p}_{c,k})}, \tag{10}$$

where $\widehat{\mathcal{I}}_c$ is the index set of pixels assigned to class c and $\widehat{\mu}_c = \sum_{j=1}^{d} \widehat{p}_{c,j}$. The second term in (10) approximates the IoU when predicting mask by $\widehat{\mathcal{I}}_c$, while the first term approximates the IoU when pixel j is further included. Similarly, Algorithm 4 can be obtained by simply replacing the RMA-scores used in RankDice-RMA-Multiclass.

Algorithm 4 RankIoU-RMA-Multiclass

Input: Estimated probability map $\widehat{\mathbf{p}} \in [0,1]^{C\times d}$.

Output: The predicted segmentation mask $\widehat{\boldsymbol{\delta}} \in [C]^d$.

```
/* Obtain overlapping segmentation mask */
```

for $c = 1$ to C **do**

  $\widehat{\boldsymbol{\psi}}_c = \texttt{RankIoU-RMA-Binary}(\widehat{\mathbf{p}}_c)$, $\widehat{\mathcal{I}}_c^+ = \{j : \widehat{\psi}_{c,j} = 1\}$.

end for

```
/* Resolve overlapping by argmax over RMA-scores */
```

Identify overlapping indices, $\widehat{\mathcal{I}}^{\text{overlap}} = \bigcup_{c\neq c'} (\widehat{\mathcal{I}}_c^+ \cap \widehat{\mathcal{I}}_{c'}^+)$.

for $c = 1$ to C **do**

  Discard assignments for overlapping pixels, $\widehat{\mathcal{I}}_c = \widehat{\mathcal{I}}_c^+ \setminus \widehat{\mathcal{I}}^{\text{overlap}}$.

  Accept prediction for not overlapping pixels, $\widehat{\delta}_j = c$ for $j \in \widehat{\mathcal{I}}_c$.

end for

Compute RMA-scores, $\widehat{\Omega}_{c,j}$ via (10) for $j \in \widehat{\mathcal{I}}^{\text{overlap}}$ and $c \in [C]$.

Resolve overlapping by argmax, $\widehat{\delta}_j = \text{argmax}_{c\in[C]} \widehat{\Omega}_{c,j}$ for $j \in \widehat{\mathcal{I}}^{\text{overlap}}$.

Return $\widehat{\boldsymbol{\delta}}$

C Proof of Theorem 2

The following two lemmas are used in the proof of (3) in Theorem 2.

Lemma C.1 (Chao and Strawderman [1972]). *Let $a \in \mathbb{R}$ and X be a random variable such that $X + a > 0$ a.s. Define the probability generating function of X as $G_X(t) = \mathbb{E}(t^X)$ for $0 \le t \le 1$. Then,*

$$\mathbb{E}\left(\frac{1}{X+a}\right) = \int_0^1 G_X(u) t^{a-1} dt.$$

Lemma C.2 (Wooff [1985]). *Let $\Lambda \sim \text{Bin}(d,p)$ be a binomial random variable. Then, for any $a > 0$, the following inequalities hold:*

$$\mathbb{E}\left(\frac{1}{\Lambda+a}\right) \le \frac{1}{(d+1)p + a - 1}.$$

Note that binomial random variable $\Lambda \sim \text{Bin}(d,p)$ and Poisson-binomial random variable $\Gamma \sim \text{PB}(p_1, p_2, \cdots, p_d)$ have probability generating functions:

$$G_\Lambda(t) = (1-p+pt)^d \quad \text{and} \quad G_\Gamma(t) = \prod_{j=1}^d (1 - p_j + p_j t).$$

Now we are ready to prove Theorem 2.

Proof. We first prove (3). The lower bound that $(\mathbb{E}\Gamma + \tau)^{-1} \le \mathbb{E}(\Gamma+\tau)^{-1}$ follows from the Jensen's inequality. Let $\Lambda \sim \text{Bin}(d, \bar{p})$, where $\bar{p} = d^{-1}\sum_{j=1}^d p_j$. To prove the upper bound, we have:

$$\begin{aligned}
\mathbb{E}(\frac{1}{\Gamma+\tau}) &= \int_0^1 t^{\tau-1} G_\Gamma(t) dt = \int_0^1 t^{\tau-1} \left(\prod_{j=1}^d (1-p_j+p_j t)\right) dt \\
&\le \int_0^1 t^{\tau-1}(1-\bar{p}+\bar{p}t)^d dt = \int_0^1 t^{\tau-1} G_\Lambda(t) dt = \mathbb{E}(\frac{1}{\Lambda+\tau}) \\
&\le \left(\frac{1}{(d+1)\bar{p}+\tau-1}\right).
\end{aligned}$$

The first and last equalities follow from Theorem C.1. The first inequality is due to the arithmetic and geometric means inequality, and the last inequality follows from Theorem C.2.

https://doi.org/10.52202/085713-5452

To proceed with (5), we first establish an error bound for RMA. Let Γ be a Poisson-binomial random variable and let $\gamma \geq 1$. Then, we have:

$$\begin{aligned}\mathbb{E}(\Gamma+\tau)^{-1} - (\mathbb{E}\Gamma+\tau)^{-1} &\leq (\frac{d+1}{d}\mathbb{E}\Gamma+\tau-1)^{-1} - (\mathbb{E}\Gamma+\tau)^{-1} \\ &\leq (\mathbb{E}\Gamma+\tau-1)^{-1} - (\mathbb{E}\Gamma+\tau)^{-1} = \frac{1}{(\mathbb{E}\Gamma+\tau-1)(\mathbb{E}\Gamma+\tau)} \leq (\mathbb{E}\Gamma+\tau)^{-2}. \quad (11)\end{aligned}$$

For any $\mathcal{I} \subseteq [d]$, the error bound of RankDice-RMA is then given by:

$$\begin{aligned}|\pi_{\text{RMA}}(\mathcal{I}) - \pi(\mathcal{I})| &\leq \sum_{j\in\mathcal{I}} 2p_j \left|\mathbb{E}(\frac{1}{\tau+\Gamma_{-j}+1}) - \frac{1}{\tau+\mathbb{E}\Gamma+1}\right| \\ &= \sum_{j\in\mathcal{I}} 2p_j \left|\mathbb{E}(\frac{1}{\tau+\Gamma_{-j}+1}) - \frac{1}{\tau+\mathbb{E}\Gamma_{-j}+1} + \frac{1}{\tau+\mathbb{E}\Gamma_{-j}+1} - \frac{1}{\tau+\mathbb{E}\Gamma+1}\right| \\ &\leq \sum_{j\in\mathcal{I}} 2p_j \left|\mathbb{E}(\frac{1}{\tau+\Gamma_{-j}+1}) - \frac{1}{\tau+\mathbb{E}\Gamma_{-j}+1}\right| + \left|\frac{1}{\tau+\mathbb{E}\Gamma_{-j}+1} - \frac{1}{\tau+\mathbb{E}\Gamma+1}\right| \\ &\leq \sum_{j\in\mathcal{I}} 2p_j \left(\frac{1}{(\tau+\mathbb{E}\Gamma_{-j}+1)^2} + \frac{p_j}{(\tau+\mathbb{E}\Gamma_{-j}+1)(\tau+\mathbb{E}\Gamma+1)}\right) \\ &\leq \sum_{j\in\mathcal{I}} 2p_j \left(\frac{1}{(\tau+\mathbb{E}\Gamma)^2} + \frac{p_j}{(\tau+\mathbb{E}\Gamma)^2}\right) = \frac{\sum_{j\in\mathcal{I}} 2p_j(1+p_j)}{(\tau+\mathbb{E}\Gamma)^2} \leq \frac{2}{\tau+\mathbb{E}\Gamma}.\end{aligned}$$

Here, the third inequality follows from (11). The last inequality is because $\sum_{j\in\mathcal{I}} p_j \leq |\mathcal{I}| = \tau$ and $\sum_{j\in\mathcal{I}} p_j^2 \leq \sum_{j\in\mathcal{I}} p_j = \mathbb{E}\Gamma$. □

D Training Details

The training settings mainly follow Wang et al. [2023a,b]. For Pascal VOC, Cityscapes and ADE20K, AdamW optimizer with a weight decay of 0.01 is used. The learning rate starts from $1e-6$ and linearly warms up during the first 1% iterations to the initial learning rate $6e-5$. The learning rate is then decayed in a "poly" policy with an exponent of 1. The number of warm-up iterations is 400 for Pascal VOC and Cityscapes, and 800 for ADE20K. The total number of training iterations is 40,000 for Pascal VOC and Cityscapes, and 80,000 for ADE20K. Data augmentations including (i) random scaling in the range of $[0.5, 2.0]$, and (ii) random horizontal flipping with a probability of 0.5.

For LiTS and KiTS, we train the models using SGD with an initial learning rate of 0.01, momentum of 0.9, and weight decay of 0.0005. The learning rate is decayed in a "poly" policy with an exponent of 0.9. The batch size is 8 and the number of epochs is 60. These two datasets are originally multi-class segmentation tasks, but we convert them into binary segmentation by only treating the tumor as foreground. This is because we want to compare our method with RankDice-BA, which is only applicable to binary segmentation. Furthermore, since LiTS and KiTS do not include designated test sets, we employ 5-fold cross-validation to evaluate performance, following existing literature [Qin et al., 2021].

E Additional Results

E.1 Statistical Significance Test

Table 7: Mean performance in mIoU^{I} and p-values from t-tests between RankDice-RMA and Argmax-prob.

	Pascal VOC	Cityscapes	ADE20K
Argmax-prob	87.80 ± 0.12	75.63 ± 0.04	56.88 ± 0.09
RankDice-RMA	88.16 ± 0.11	76.11 ± 0.07	57.67 ± 0.11
p-value	1.12e-6	2.30e-13	5.96e-13

To validate the statistical significance of the performance improvement achieved by RankDice-RMA over Argmax-prob, we conduct 10 independent runs with different random seeds using UPerNet on VOC, Cityscapes, and ADE20K datasets.

We report mean and standard deviation of mIoU^{I}, along with the p-values from t-tests, as shown in Table 7. The results indicate that the improvements are statistically significant, with p-values far below 0.01 across all datasets.

More importantly, our method not only achieves a substantial improvement in the sense of mean performance, but also consistently outperforms Argmax-prob in every single run. For example, the results of the ten runs on ADE20K are presented in Table 8. This consistency arises because our method is deterministic and introduces no inherent randomness. It is applied to trained models by simply replacing Argmax-prob in the prediction step. Consequently, the comparison is highly stable as they share the same model.

Table 8: mIoU^{I} of 10 independent runs on ADE20K.

Run	1	2	3	4	5	6	7	8	9	10
Argmax-prob	56.84	56.86	56.93	57.01	56.77	56.94	57.01	56.80	56.84	56.77
RankDice-RMA	57.56	57.66	57.71	57.86	57.59	57.73	57.84	57.59	57.68	57.52

E.2 More Qualitative Visualizations

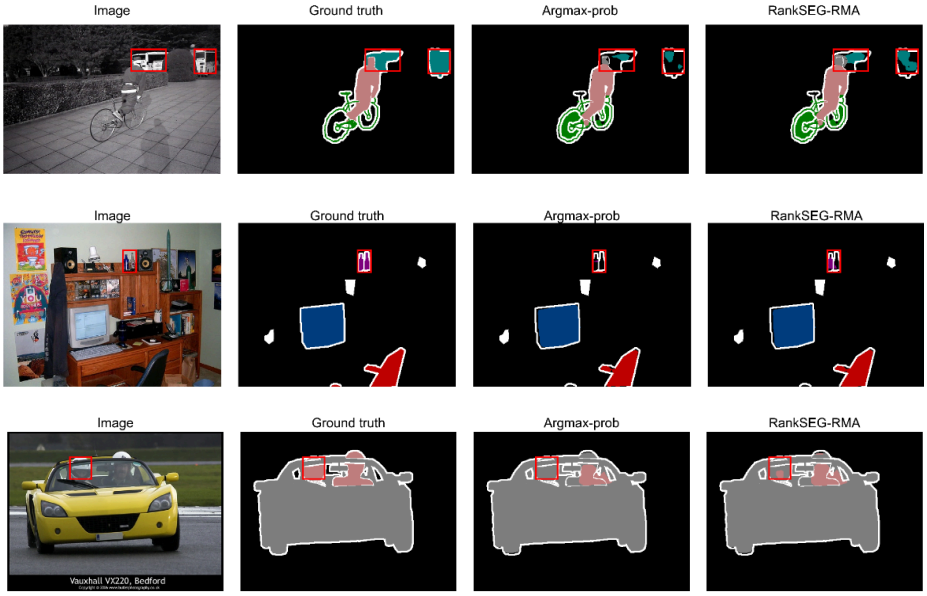

Figure 3: Qualitative visualizations on Pascal VOC. From left to right: input image, ground truth, prediction by Argmax-prob, and prediction by RankSEG-RMA. The key differences are highlighted in red boxes.

We provide additional qualitative visualizations in Figure 3 to compare the proposed RankSEG-RMA with the conventional Argmax-prob method, offering further insights into how our approach enhances segmentation quality. As highlighted by the red boxes in the figure, RankSEG-RMA outperforms Argmax-prob primarily in capturing complete regions of challenging objects and in detecting small objects.

For instance, in the first row, where the buses are partially occluded, Argmax-prob only sparsely identifies a small portion of the buses, whereas RankSEG-RMA achieves more complete segmentation. In the second example, Argmax-prob fails to detect the small bottles on the table, but RankSEG-RMA successfully identifies them. Similarly, in the third example, Argmax-prob completely misses one human face, while RankSEG-RMA detects it. These examples, together with the discussions in Sections 4.3 and 4.4, demonstrate that RankSEG-RMA is particularly effective for segmenting small and challenging objects.

 https://doi.org/10.52202/085713-5452

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in our abstract and introduction that proposed RankSEG-RMA significantly reduce computation cost and enable non-overlapping segmentation accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations have been discussed in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions are clearly stated in the corresponding theorems. The proofs are provided in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The implementation details are provided in Section 4 and Section D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

 https://doi.org/10.52202/085713-5452

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The datasets we used are open-source datasets. The code is available in https://anonymous.4open.science/r/RankSEG-RMA-4C14.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The implementation details are provided in Section 4 and Section D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Table 3 provides the standard deviation of 10 runs. Other experiments do not suffer from randomness.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The time consumption and used resources are provided in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have ensured that our research conforms to the code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

https://doi.org/10.52202/085713-5452

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper does not pose such a risk.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all external sources of assets and their licenses permit our use case.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

https://doi.org/10.52202/085713-5452

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs are not used for method development. They are only used for writing polishing.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.