BackSplit: The Importance of Sub-dividing the Background in Biomedical Lesion Segmentation

Rachit Saluja[1,2,3]*, Asli Cihangir[1], Ruining Deng[2,3],
Johannes C. Paetzold[2,3], Fengbei Liu[2,3], Mert R. Sabuncu[1,2,3]

[1]Cornell University [2]Cornell Tech [3]Weill Cornell Medicine

Abstract

Segmenting small lesions in medical images remains notoriously difficult. Most prior work tackles this challenge by either designing better architectures, loss functions, or data augmentation schemes; and collecting more labeled data. We take a different view, arguing that part of the problem lies in how the background is modeled. Common lesion segmentation collapses all non-lesion pixels into a single "background" class, ignoring the rich anatomical context in which lesions appear. In reality, the background is highly heterogeneous—composed of tissues, organs, and other structures that can now be labeled manually or inferred automatically using existing segmentation models.

*In this paper, we argue that training with fine-grained labels that sub-divide the background class, which we call **BackSplit**, is a simple yet powerful paradigm that can offer a significant performance boost without increasing inference costs. From an information theoretic standpoint, we prove that BackSplit increases the expected Fisher Information relative to conventional binary training, leading to tighter asymptotic bounds and more stable optimization. With extensive experiments across multiple datasets and architectures, we empirically show that BackSplit consistently boosts small-lesion segmentation performance, even when auxiliary labels are generated automatically using pretrained segmentation models. Additionally, we demonstrate that auxiliary labels derived from interactive segmentation frameworks exhibit the same beneficial effect, demonstrating its robustness, simplicity, and broad applicability.*

1. Introduction

Quantitative assessment of tumors and lesions represents one of the fundamental steps in clinical diagnosis and treatment planning. Despite substantial progress in medical image segmentation, particularly for anatomical structures such as organs — lesion segmentation remains a persistent challenge. Lesions are typically small, spatially sporadic, and underrepresented in dataset distributions, making them difficult to model reliably. These factors often lead to false positives and unstable predictions, ultimately limiting clinical deployment.

Prior research has primarily addressed this challenge through the development of improved network architectures, often tailored to specific tasks [6, 7, 9, 22, 23, 28, 45, 50, 54], or by designing specialized loss functions optimized for particular scenarios [1, 32, 38, 44, 51, 64], and task-specific data augmentation techniques [18, 61]. This focus is understandable, as medical image segmentation tasks vary considerably across imaging modalities, naturally motivating researchers to pursue modality- and task-specific architectural and optimization strategies.

In this work, we approach the problem from a different perspective by addressing a fundamental limitation shared across most existing architectures: the tendency to produce false positives and unstable lesion predictions due to their small size, irregular boundaries, and heterogeneous appearance. In conventional setups, all non-lesion pixels are collapsed into a single "background" class, disregarding the rich anatomical context in which lesions occur. Yet, this background is far from homogeneous—it consists of diverse tissues, organs, and structures that can now be labeled manually or inferred automatically using modern segmentation models.

To address this limitation, we propose **BackSplit** (as illustrated in Fig. 1), a simple and intuitive training paradigm that incorporates structured background supervision. BackSplit decomposes the background into multiple auxiliary classes and jointly optimizes them alongside the lesion target, thereby improving the supervision signal for the primary segmentation objective. This approach markedly reduces false positive detections and yields more accurate lesion boundaries. During inference, the model predicts the target class while implicitly leveraging the contextual knowledge learned from the auxiliary background classes, even though they are not explicitly used.

In summary, BackSplit redefines lesion segmentation by

*Corresponding Author, rs2492@cornell.edu

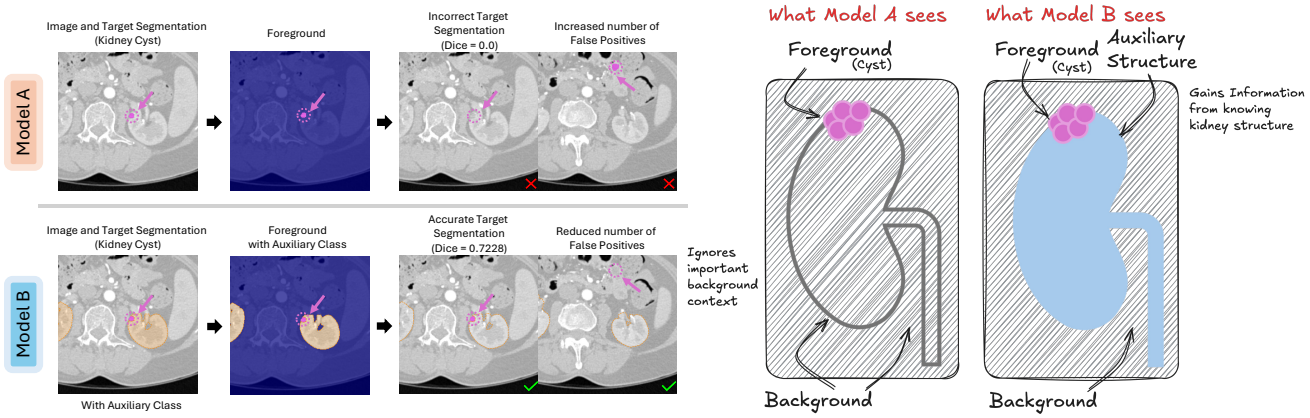

Figure 1. Comparison between conventional binary segmentation (Model A) and our BackSplit paradigm (Model B). Conventional lesion segmentation collapses all non-lesion regions into a single background, discarding the anatomical context and often producing false positives (yielding a Dice overlap score of 0.0). In contrast, BackSplit refines the background into semantically meaningful auxiliary structures (e.g., organ parenchyma) that are learned jointly with the lesion target. This structured background supervision enriches contextual understanding, yielding sharper lesion boundaries, fewer false detections (e.g., Dice = 0.72), and theoretically more stable predictions—consistent with higher expected Fisher Information and reduced estimator variance compared to conventional binary training.

introducing structure into what is conventionally treated as a uniform background, improving both precision and generalization.

Our contributions are as follows:

1. We show that training with BackSplit yields higher expected Fisher information compared to binary training (Theorem 1), leading to tighter asymptotic convergence bounds (Corollary 1).
2. We validate the paradigm across five diverse datasets spanning multiple imaging modalities and anatomical regions, consistently improving performance metrics and reducing false positives across all architectures (Tab. 1 and Tab. 2).
3. We show that BackSplit remains effective even when trained with automatically generated or noisy auxiliary segmentations, making it more practical and easier to implement in real-world settings (Tab. 3 and Tab. 4).

2. Related Work

Binary vs Multi-Class Segmentation in Medical Imaging. Medical image segmentation methods are commonly categorized into *binary* and *multi-class* paradigms, depending on how anatomical and pathological regions are delineated. In the binary setting, the task is formulated as distinguishing lesions from the background. This configuration remains prevalent in clinical pipelines, particularly for small or localized targets such as tumors [21], nodules [26], or infarcts [34]. Such datasets typically provide lesion annotations without explicit organ or tissue masks, enforcing a binary formulation that limits contextual awareness of surrounding anatomy [12, 15, 16, 62].

Multi-head architectures have been proposed to approximate multi-class segmentation within a binary framework, where each head predicts an independent class mask [8, 41]. Although this setup allows training with incomplete labels, it neglects inter-class dependencies and often yields inconsistent anatomical boundaries.

In contrast, multi-class segmentation explicitly models multiple anatomical entities (e.g., organ parenchyma, lesions, and residual background) within a unified softmax formulation. By incorporating auxiliary contextual labels, the model learns to delineate lesion boundaries relative to their anatomical environment. Prior studies have shown that such contextual supervision enhances feature discriminability and improves generalization [27, 33, 60].

Our work advances this direction by introducing a theoretical justification and comprehensive empirical analysis of structured background supervision in lesion segmentation. For the first time, we provide a rigorous information-theoretic justification showing that multi-class contextual supervision yields higher expected Fisher Information, and therefore more stable estimators than conventional binary training. Complementing this theory, we perform extensive experiments across multiple architectures and datasets, including those with automatically and interactively generated auxiliary segmentations.

Context-aware and Auxiliary Supervision. Contextual information plays a crucial role in distinguishing lesions from their surrounding anatomy, as many pathologies manifest through subtle appearance changes relative to neighboring tissues. Early studies leveraged organ masks or anatomical priors to stabilize lesion detection and constrain predictions within plausible regions [42, 55, 63]. More recent

approaches explicitly model cross-class context through multi-scale or anatomy-guided reasoning, demonstrating that learning relationships among organs, tissues, and lesions substantially improves discriminability and generalization [13, 14, 31, 53, 58]. These context-aware frameworks employ strategies such as hierarchical decomposition, anatomy-prompted supervision, and collaborative feature sharing across semantic levels to capture global anatomical coherence in addition to local lesion appearance. Multi-task or shape-aware regularization [24, 36, 39, 40, 52] further enhance boundary precision by enforcing consistency between related anatomical structures. A similar intuition has been explored in [37], where an auxiliary network is introduced to predict background classes that enhance contextual understanding for target segmentation. In contrast, our work provides the theoretical foundation explaining why such an approach is effective.

While these context-aware methods achieve impressive gains, they rely on additional architectural branches, explicit multi-task losses, or handcrafted priors. In contrast, our BackSplit paradigm is simply based on multi-class supervision: by decomposing the background into semantically meaningful auxiliary support structures, the model learns lesion–context relationships intrinsically through the label space rather than through network design. This supervision-level contextualization remains architecture-agnostic, scales naturally to different backbones, and consistently enhances precision, boundary stability.

Label Coarsening. To develop our theoretical framework, we draw inspiration from prior work analyzing the effects of label coarsening across different supervision regimes. Existing studies primarily investigate methods that allow models trained on coarse labels to achieve performance comparable to those trained on fine-grained annotations, provided the available data is sufficiently informative [17]. Conversely, other approaches treat each example as its own class to capture fine-grained patterns and increase inter-class separability when the label set lacks sufficient granularity [59]. Notably, [10] demonstrates that selecting an appropriate level of label granularity during training can yield greater performance gains than choosing a more sophisticated model architecture.

These prior works largely lack a theoretical foundation to understand the role of label coarsening and how it affects statistical efficiency. In this paper, we present a theoretically grounded analysis that connects label granularity to the expected Fisher Information of the learning problem. Specifically, we show that sub-dividing the background into semantically meaningful auxiliary structures enriches the contextual representation and provably increases the expected Fisher Information, thereby improving the stability and accuracy of lesion segmentation.

3. Methods

Let us formally demonstrate that multi-class training yields higher expected Fisher information than coarsened binary training, therefore providing more statistically efficient predictions for the target class.

3.1. Setup and Notation

Let (X, Y) be a pair of random data points where $X \in \mathcal{X}$ is the input, which in our case is an image (patch or pixel), and $Y \in \{1, \dots, K\}$ is the class label at some pixel. We assume a joint distribution $p_\theta(X, Y) = p_\theta(Y \mid X)\, p(X)$ parameterized by θ, which denotes the model parameters.

Coarsened Label Definition. We focus on a particular foreground (target) class $c \in \{1, \dots, K\}$, (e.g. a lesion of interest) and define the binary coarsened label indicating class c as:

$$Z = \mathbb{1}\{Y = c\} \in \{0, 1\}$$

where $\mathbb{1}$ is the indicator function, so that $Z = 1$ if Y is the target class and $Z = 0$ otherwise.

In simple words: we are focusing on a single class c and comparing two training scenarios: (i) the full *multiclass* case with K classes, vs. (ii) a *collapsed binary* case distinguishing c vs. not-c.

Posterior Class Probabilities. Let, $\eta_k(x, \theta) = \mathbb{P}_\theta(Y = k \mid X = x)$ and $\eta_c(x, \theta) = \mathbb{P}_\theta(Z = 1 \mid X = x)$ denote the per-class and target-class posterior probabilities, respectively.

Likelihoods and Score Functions. For a dataset $\{(X_i, Y_i)\}_{i=i}^n$, the two corresponding log-likelihoods are

$$\text{(multiclass)} \quad \ell_Y(\theta) = \sum_{i=1}^{n} \log p_\theta(Y_i \mid X_i)$$

$$\text{(collapsed binary)} \quad \ell_Z(\theta) = \sum_{i=1}^{n} \log p_\theta(Z_i \mid X_i)$$

where $p_\theta(Z = 1 \mid x) = \eta_c(x; \theta)$ and $p_\theta(Z = 0 \mid x) = 1 - \eta_c(x; \theta) = \sum_{k \neq c} \eta_k(x; \theta)$. We define the per-sample score (gradient of the log-likelihood) and the observed information (negative Hessian) as

$$s_Y(\theta) = \nabla_\theta \log p_\theta(Y \mid X)$$
$$s_Z(\theta) = \nabla_\theta \log p_\theta(Z \mid X)$$
$$I_Y(\theta) = -\nabla^2_\theta \log p_\theta(Y \mid X)$$
$$I_Z(\theta) = -\nabla^2_\theta \log p_\theta(Z \mid X)$$

Expected Fisher Information. Under standard regularity assumptions, the expected Fisher information matrices—equivalently the expectations of the observed information, are given by

$$\mathcal{I}_Y(\theta) = \mathbb{E}_\theta[s_Y(\theta)s_Y(\theta)^T] = \mathbb{E}_\theta[I_Y(\theta)]$$
$$\mathcal{I}_Z(\theta) = \mathbb{E}_\theta[s_Z(\theta)s_Z(\theta)^T] = \mathbb{E}_\theta[I_Z(\theta)]$$

Conditioning on a fixed input $X = x$ gives:

$$\mathcal{I}_Y(\theta \mid X = x) = \mathbb{E}_\theta[s_Y(\theta)s_Y(\theta)^T \mid X = x]$$

where the expectation is taken over the posterior $p_\theta(Y \mid X)$. The unconditional Fisher Information integrates this quantity over $p(X)$.

3.2. Expected Fisher information under label coarsening

Having defined the score functions and expected Fisher Information for both the multiclass and coarsened (binary) formulations, we now formalize their relationship. The following lemma shows that the coarsened score is the conditional expectation (or orthogonal projection) of the full multiclass score given the observed variables.

Lemma 1 (Score Projection [43, 47])**.** *Let $Z = g(Y)$ be a deterministic coarsening of the label Y. Then, under regularity conditions ensuring that differentiation and summation interchange,*

$$\mathbb{E}_\theta[s_Y(\theta) \mid Z, X] = s_Z(\theta).$$

[Proof in Supplementary Sec. A.2]. Intuitively, Lemma 1 (derived from the *Missing Information Principle* [43, 47]) implies that $s_Z(\theta)$ is the best L^2 approximation of $s_Y(\theta)$ measurable with respect to (Z, X). The coarsened score thus preserves only the gradient information recoverable from the coarsened labels while discarding variation across the collapsed classes.

Building on Lemma 1, we can now express the Fisher information of the complete labels Y as the sum of the information contained in the coarsened labels $Z = g(Y)$ and a *non-negative missing-information* term. This decomposition formalizes how label coarsening (e.g., collapsing multiclass labels into a binary mask) can only reduce the available expected Fisher information.

Theorem 1 (Label coarsening reduces expected Fisher information.)**.** *For every $\theta \in \Theta$,*

$$\mathcal{I}_Y(\theta) = \mathcal{I}_Z(\theta) + \mathbb{E}_\theta\big[\operatorname{Var}(s_Y(\theta) \mid Z, X)\big] \succeq \mathcal{I}_Z(\theta)$$

and therefore $\mathcal{I}_Z(\theta) \preceq \mathcal{I}_Y(\theta)$ in the Loewner (positive–semidefinite) order. Equality holds iff $s_Y(\theta)$ is completely determined by (Z, X), i.e. no variation in the complete–data score remains once Z is known.

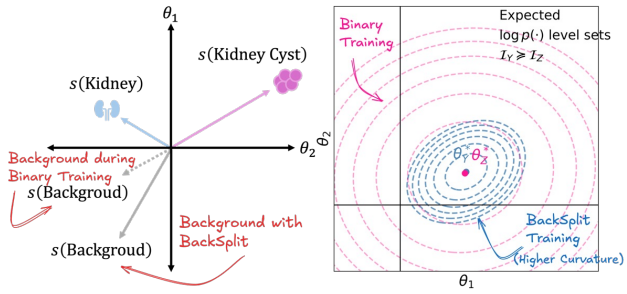

Figure 2. (Left) In binary training, the background gradient $s(\text{Background})$ conflates signals from multiple anatomical structures. Decomposing it into semantically distinct support structures (e.g., kidney) yields more disentangled and informative gradients. (Right) The corresponding log-likelihood level sets show that the decomposed formulation ($\mathcal{I}_Y$, blue) has sharper curvature and higher Fisher Information than the collapsed binary case ($\mathcal{I}_Z$, pink), leading to tighter and more stable parameter estimates.

[Proof in Supplementary Sec. A.3] The coarsened Fisher information $\mathcal{I}_Z$ captures the portion of curvature in the log-likelihood that is recoverable from the observed variables (Z, X), while the conditional variance term quantifies the information lost by collapsing fine-grained labels.

The equality condition in Theorem 1 is particularly instructive. It states that if the complete score $s_Y(\theta)$ is fully determined by the coarsened variables (Z, X) , then no Fisher information is lost. Intuitively, one can view $s_Y(\theta)$ as the gradient update that a sample (X, Y) contributes during training. Suppose instead that only the coarsened pair (X, Z) is observed. If every non-target label produces the same gradient direction given X , then the gradient obtained from the coarsened label is identical to that from the complete label—no information gap arises, and $\mathcal{I}_Y(\theta) = \mathcal{I}_Z(\theta)$.

In contrast, when different non-target classes induce different gradient directions (for instance, "organ A" and "organ B" yield distinct updates even though both are "not lesion"), collapsing these labels to a single "background" class effectively averages those gradients. This averaging removes curvature directions in parameter space, reducing the total expected Fisher information (as shown in Fig. 2). Consequently, for any input X with non-zero target probability $p_\theta(Z = 1 \mid X)$, if the non-target components of the multiclass gradient vary across classes, the collapsed binary model necessarily loses information. In other words, whenever non-target labels carry discriminative structure relevant to the target, the inequality in Theorem 1 becomes strict.

3.3. Implication for Max Likelihood Estimator (MLE)

We now analyze the implications of Theorem 1 for the maximum-likelihood estimators (MLEs) trained with multiclass and coarsened (binary) labels.

Corollary 1 (Asymptotic Efficiency of the Multiclass MLE.)**.** *Under standard regularity and correct-specification assumptions at the true parameter θ^*, the multiclass MLE is (weakly) more statistically efficient than the collapsed-binary MLE for any smooth functional of θ. Equivalently, for any differentiable quantity $g(\theta)$ of interest, the asymptotic estimation variance of $g(\hat{\theta}_Y)$ is no greater than $g(\hat{\theta}_Z)$.*

[Proof in Supplementary Sec. A.4] At the optimum θ^*, both estimators satisfy the classical MLE central-limit theorem:

$$\sqrt{n}(\hat{\theta_Y} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_Y(\theta^*)^{-1})$$
$$\sqrt{n}(\hat{\theta_Z} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_Z(\theta^*)^{-1})$$

Since Theorem 1, establishes the expected Fisher information ordering, $\mathcal{I}_Y(\theta) \succeq \mathcal{I}_Z(\theta)$, matrix inversion reverses this order for positive-definite matrices, and we obtain $\mathcal{I}_Y(\theta)^{-1} \preceq \mathcal{I}_Z(\theta)^{-1}$. Thus, the multiclass estimator has (weakly) smaller asymptotic covariance in parameter space.

The results thus far pertain to the quality of the parameter (weight) estimates. To relate this to the quality of model predictions, we employ the Delta method, which enables us to approximate the variance of prediction functions derived from these parameter estimates. Consider any smooth function $g(\theta)$ - for instance, the target-class posterior $\tau(\theta) = \eta_c(x_0; \theta)$, where x_0 denotes an arbitrary fixed input. By the delta method,

$$\sqrt{n}(g(\hat{\theta}) - g(\theta_0)) \xrightarrow{d} \mathcal{N}(0, G\Sigma G^T)$$

where $G = \nabla_\theta g(\theta_0)$, θ_0 is a parameter with non-singular Fisher information, and Σ is the asymptotic covariance of the corresponding MLE. As $\Sigma_Y = \mathcal{I}_Y(\theta)^{-1} \preceq \Sigma_Z = \mathcal{I}_Z(\theta)^{-1}$, the variance of any differentiable prediction such as $p_\theta(Y = c \mid x_0)$ is strictly smaller (or equal) under multiclass training.

$$\mathrm{Var}\Big(\sqrt{n}\,(\tau(\hat{\theta}_Y) - \tau(\theta^*))\Big) \leq \mathrm{Var}\Big(\sqrt{n}\,(\tau(\hat{\theta}_Z) - \tau(\theta^*))\Big)$$

This formalizes the notion that multiclass training makes more efficient use of data. Both estimators are consistent and improve as $n \to \infty$ but the multiclass model converges faster and its confidence regions and prediction variances shrink more rapidly. Only when the equality condition of Theorem 1 holds (the coarsened score fully determined by (Z, X) do the two MLEs achieve identical efficiency. In all other cases, where non-target classes induce distinct gradient directions, the collapsed-binary estimator discards curvature information and remains asymptotically less precise.

3.4. Special case: Softmax

We now instantiate Theorem 1 for the softmax layer which is commonly used for multi-class segmentation. This yields a closed-form expression for the information matrices in both the multiclass and the coarsened binary formulations.

Proposition 1 (Softmax Expected Fisher Information Decomposition.)**.** *Consider a softmax model with logits $f(x;\theta) \in \mathbb{R}^K$ and class probabilities $\eta = softmax(f)$. Let $J_f(x;\theta) = \partial f(x;\theta)/\partial\theta \in \mathcal{R}^{K\times p}$ denote the Jacobian of the logits with respect to the model parameters and let $e_c \in \mathbb{R}^K$ denote the c-th standard basis (one-hot) vector. Then for any fixed input $X = x$:*

$$\mathcal{I}_Y(\theta \mid X = x) = J_f^T\,(Diag(\eta) - \eta\eta^T)J_f$$

is the conditional expected Fisher information for the multiclass likelihood, and

$$\mathcal{I}_Z(\theta \mid X = x) = J_f^T \left(\frac{\eta_c}{(1-\eta_c)}(e_c - \eta)(e_c - \eta)^T\right) J_f$$

is the corresponding quantity for the collapsed-binary label. Hence,

$$\mathcal{I}_Y(\theta \mid X = x) = \mathcal{I}_Z(\theta \mid X = x) + (1-\eta_c)J_f^T\,(Diag(\pi) - \pi\pi^T)J_f \succeq \mathcal{I}_Z(\theta \mid X = x)$$

where $\pi_k = \eta_k/(1-\eta_c)$ for $k \neq c$ and $\pi_c = 0$.

[Proof in Supplementary Sec. A.5] Proposition 1 makes the information gap explicit for the softmax parameterization. The first term captures the curvature along the target-vs-rest direction $(e_c - \eta)$, identical to the expected Fisher information of the corresponding Bernoulli model. While the second term accounts for curvature within the non-target subspace of the probability simplex. Whenever at least two non-target classes have non-zero probability and the model parameters influence their relative proportions, this additional term is non-zero, yielding a strict inequality $\mathcal{I}_Y(\theta) \succ I_Z(\theta)$. Equality occurs only in degenerate cases (e.g., $K = 2$ or a deterministic non-target distribution).

Geometrically, $\mathrm{Diag}(\eta) - \eta\eta^T$ forms an ellipsoid of curvatures spanning the $K-1$ tangent directions of the simplex, while the collapsed term retains only the single direction $e_c - \eta$. Thus, label coarsening projects out all curvature within the non-target face, explaining the empirical loss of statistical efficiency observed in Sec. 3.3.

The softmax parameterization is particularly relevant to medical image segmentation, where most architectures employ a softmax output layer. Under this model, the Fisher decomposition exposes the information gap geometrically in logit space and directly relates it to practical scenarios such as lesion–background training. Collapsing non-target tissues into a single background class eliminates curvature directions that differentiate organs, thereby reducing Fisher information and impeding convergence.

Task	U-Net [28, 50] #Params	Dice ↑	HD-95 ↓	NSD ↑	ResEncU-Net [29] #Params	Dice ↑	HD-95 ↓	NSD ↑	SegResNet [46] #Params	Dice ↑	HD-95 ↓	NSD ↑
(a) Dataset: KiTS23 with Target Class = {Cyst}. Support Structures = {Kidney, Tumors} Total Sample Size (N = 489) — Imaging Modality = {CT}												
Cyst	31.19M	0.1787	428.4097	0.1695	102.35M	0.2532	412.4152	0.2483	18.79M	0.2914	374.4864	0.3554
Cyst (+BackSplit)	31.19M	**0.4573**	**267.2722**	**0.6004**	102.35M	**0.4614**	**259.6994**	**0.6100**	18.79M	**0.4290**	**287.6765**	**0.5580**
(b) Dataset: PANTHER-MR with Target Class = {Tumor}. Support Structures = {Pancreas} Total Sample Size (N = 92) — Imaging Modality = {MRI}												
Tumor	30.70M	0.4784	84.3689	0.2751	101.85M	0.516	66.2637	0.2987	18.79M	0.4744	**59.0499**	0.2623
Tumor (+BackSplit)	30.70M	**0.5251**	**54.2576**	**0.304**	101.85M	**0.5165**	**70.5463**	**0.303**	18.79M	**0.4995**	73.1859	**0.2778**
(c) Dataset: NSCLC-Radiomics with Target Class = {GTV}. Support Structures = {Multiple Organs} Total Sample Size (N = 415) — Imaging Modality = {CT}												
GTV	30.70M	0.4969	156.4725	0.3803	101.86M	0.502	145.0108	0.3879	18.79M	0.5061	142.8023	**0.3921**
GTV (+BackSplit)	30.70M	**0.5256**	**142.9498**	**0.4136**	101.86M	**0.5386**	**125.9383**	**0.4268**	18.79M	**0.5131**	**141.5307**	0.3842

Table 1. Quantitative comparison of segmentation performance across three architectures (U-Net [28, 50], ResEncU-Net [29], and SegResNet [46]) under standard and BackSplit training paradigms. Results are reported using 5-fold cross-validation on (a) KiTS (Cyst) [25], (b) PANTHER-MR (Tumor) [3], and (c) NSCLC-Radiomics (GTV) [2] datasets, with respective support structures indicated. BackSplit consistently improves Dice and NSD while reducing HD-95 across models and imaging modalities, without increasing model parameters (reported in millions).

4. Experiments

We evaluate BackSplit on five diverse and challenging datasets spanning CT, MRI, and PET modalities, and covering abdominal, thoracic, and whole-body regions. These datasets encompass a wide range of lesion segmentation tasks, architectures, and auxiliary structure labels, enabling a comprehensive assessment of robustness and generalizability even under label noise.

4.1. Data

The datasets collections comprise of:

1. **KiTS23** [25]: This dataset is designed to facilitate the development of automated systems for semantic segmentation of kidneys and associated tumors and cysts. We specifically focus on kidney cysts as the target class. The dataset comprises 489 CT scans.
2. **PANTHER** [3]: It comprises 92 annotated T1-weighted contrast-enhanced arterial phase MRI scans, each including manual annotations for both the pancreas and associated tumors.
3. **NSCLC-Radiomics** [2]: Dataset includes 415 CT scans from patients with non–small cell lung cancer (NSCLC), each containing manual segmentations of both thoracic organs and tumors. In our study, the gross tumor volume (GTV) serves as the target label.
4. **AutoPET 3** [19]: This dataset consists of 1,611 paired CT and PET scans with corresponding lesion segmentations, encompassing a variety of cancer types. Our target labels are the tumors.
5. **MSWAL** [57]: It comprises 484 abdominal CT scans with multi-lesion annotations spanning seven classes, including stones, tumors, and cysts. All lesion categories are treated as target segmentations in our analysis.

4.2. BackSplit is Agnostic to Network Architectures

In our experiments, we trained two models for each setting: (i) a binary baseline trained only on the target class as foreground, and (ii) a BackSplit-based model trained jointly on the target and its auxiliary support structures.

To assess architectural generality, we implemented BackSplit across three widely adopted segmentation backbones that have achieved competitive performance in benchmark challenges: (1) U-Net [50], (2) ResEnc U-Net [29] implemented via the nnU-Net framework [28], and (3) SegResNet [46].

Since nnU-Net heuristically configures both architecture and hyperparameters, we ensured identical configurations between baseline and BackSplit models for a fair comparison. All training followed the standard nnU-Net protocol, with detailed hyperparameters provided in the Supplementary Material.

Results. Our initial set of evaluations on the KiTS23, PANTHER, and NSCLC-Radiomics datasets were conducted with available ground-truth labels for the auxiliary classes, ensuring a controlled comparison. Each dataset was trained using five-fold cross-validation with all three architectures. Quantitative results are summarized in Tab. 1. Across all datasets and architectures, BackSplit consistently improves performance (measured with Dice, normalized surface distance, and the 95th-percentile Hausdorff distance). The auxiliary classes typically correspond to organs or tissues adjacent to the target structure (e.g., the pancreas for pancreatic tumors), demonstrating that structured background supervision improves lesion delineation without modifying the network architecture.

Task	U-Net [28, 50] #Params	Dice ↑	HD-95 ↓	NSD ↑
(a) Dataset: AutoPET with Target Class = {Tumor} Support Structures = {Multiple Organs} Total Sample Size (N = 1611) — Imaging Modality = {PET, CT}				
Tumor	30.79M	0.3881	637.6150	0.3478
Tumor (+BackSplit)	30.79M	**0.4435**	**594.8048**	**0.4128**
(b) Dataset: MSWAL with Target Class = {Multiple Lesions} Support Structures = {Multiple Organs} Total Sample Size (N = 484) — Imaging Modality = {CT}				
Gallstone	30.79M	0.2665	418.338	0.6664
Gallstone (+BackSplit)	30.79M	**0.3497**	**337.2602**	**0.7531**
Kidney Stone	30.79M	**0.195**	**471.4653**	**0.6247**
Kidney Stone (+BackSplit)	30.79M	0.1639	484.3217	0.5401
Liver Tumor	30.79M	0.2666	405.4476	0.3219
Liver Tumor (+BackSplit)	30.79M	**0.3312**	**338.2031**	**0.4683**
Kidney Tumor	30.79M	0.1855	436.8838	0.6233
Kidney Tumor (+BackSplit)	30.79M	**0.1979**	**407.2282**	**0.6394**
Pancreatic Cancer	30.79M	0.1836	448.5034	0.6018
Pancreatic Cancer (+BackSplit)	30.79M	**0.3228**	**308.2553**	**0.7746**
Liver Cyst	30.79M	0.317	384.3436	0.4586
Liver Cyst (+BackSplit)	30.79M	**0.3571**	**341.4273**	**0.546**
Kidney Cyst	30.79M	0.4931	234.848	0.6072
Kidney Cyst (+BackSplit)	30.79M	**0.5104**	**213.0923**	**0.6311**
All lesions mean	30.79M	0.2724	399.9756	0.5577
All lesions mean (+BackSplit)	30.79M	**0.3190**	**347.1125**	**0.6218**

Table 2. Quantitative results on (a) AutoPET and (b) MSWAL datasets using U-Net under standard and BackSplit training paradigms. For both datasets, support structures are automatically derived from a pre-trained organ segmentation model (U-Net trained on the AbdomenAtlas1.0Mini dataset). BackSplit yields consistent improvements in Dice and NSD, along with reductions in HD-95, across diverse lesion types and imaging modalities.

4.3. BackSplit works with model-derived auxiliary segmentations

A key requirement for BackSplit is the availability of auxiliary segmentations that can be used alongside the target class to enrich supervision. However, obtaining such annotations is often impractical as manual delineation of organ structures can require several hours per case. To overcome this limitation, we leverage existing organ segmentation models trained on large-scale datasets to automatically generate auxiliary labels through inference.

Results. We implemented this strategy with AutoPET [19] and MSWAL [57], which provide only lesion annotations. Following the same setup as the previous experiment, we trained a U-Net on AbdomenAtlas1.0Mini [49] to obtain a pretrained organ segmentation model, then used it to automatically generate auxiliary masks for the CT scans of AutoPET and MSWAL. These model-derived organ labels were spatially consistent with the lesion regions and jointly used with the original lesion masks during training.

Despite the absence of manual auxiliary annotations, BackSplit consistently improved Dice and surface-distance metrics, demonstrating robustness to automatically generated segmentations. As summarized in Tab. 2, performance gains were observed across all metrics. Notably, BackSplit achieved higher accuracy on the multi-modal AutoPET dataset (CT + PET) and on MSWAL, where multiple lesion types were trained jointly with multiple organ structures. All experiments used a U-Net implemented via nnU-Net and were evaluated with five-fold cross-validation.

Method	(a) AutoPET Dice ↑	HD-95 ↓	NSD ↑	(b) MSWAL Dice ↑	HD-95 ↓	NSD ↑
Regular Training	0.3921	677.2053	0.3415	0.2518	415.4836	0.5335
BackSplit	0.4537	618.5841	0.4199	0.2978	362.9077	0.6138
BackSplit w/ TS [56]	0.4456	642.5579	0.4036	0.2843	374.2040	0.5932
BackSplit w/ VS [20]	0.4314	682.6587	0.3761	0.2646	391.8890	0.5638

Table 3. Evaluation of BackSplit on (a) AutoPET (tumors) and (b) MSWAL (all lesions) for a single fold using a U-Net backbone. Organ masks from TotalSegmentator (TS) [56] and VIBE-Segmentator (VS) [20] were used as automatically extracted auxiliary structures. BackSplit consistently improves Dice, HD-95, and NSD scores, demonstrating that even automatically inferred support segmentations yield measurable gains.

4.4. BackSplit using pre-trained large models

BackSplit can also leverage organ segmentations extracted from large pretrained models such as TotalSegmentator [11, 56] and VIBE-Segmentator [20], which support both CT and MR modalities. Using the same experimental setup as in the previous subsection, we generated auxiliary organ masks with these models and incorporated them during training to evaluate whether performance gains persist when auxiliary labels are obtained entirely through automated inference.

Results. As summarized in Tab. 3, BackSplit continues to improve performance across Dice, HD-95, and NSD metrics for both AutoPET and MSWAL. This demonstrates that even simple, pretrained model inferences performed prior to training can yield a measurable and consistent performance gain.

4.5. BackSplit using Interactive Segmentation

With recent advances in interactive segmentation foundation models, we further evaluate BackSplit under realistic conditions where auxiliary segmentations are noisy or partially accurate. To simulate this scenario, we employ nnInteractive [30], a 3D interactive segmentation model. For each auxiliary structure, seven and ten random positive clicks are sampled from the ground-truth masks and used as seed inputs to nnInteractive. This procedure is applied to the KiTS23, NSCLC-Radiomics, and PANTHER datasets. For AutoPET and MSWAL, we follow the same protocol using the automatically generated auxiliary segmentations from

Method	(a) KiTS			(b) PANTHER-MR			(c) NSCLC-Rad			(d) AutoPET			(e) MSWAL		
	Dice ↑	HD-95 ↓	NSD ↑	Dice ↑	HD-95 ↓	NSD ↑	Dice ↑	HD-95 ↓	NSD ↑	Dice ↑	HD-95 ↓	NSD ↑	Dice ↑	HD-95 ↓	NSD ↑
Regular Training	0.2033	425.3273	0.1906	0.3828	157.3470	0.2320	0.5279	140.2698	0.4049	0.3921	677.2053	0.3415	0.2518	415.4836	0.5335
BackSplit	0.5297	249.5371	0.6703	0.4906	54.2010	0.2796	0.5862	124.2557	0.4533	0.4537	618.5841	0.4199	0.2978	362.9077	0.6138
BackSplit with ⊕7	0.4919	267.1777	0.6303	0.5079	74.6472	0.2995	0.5869	112.1321	0.4559	0.4562	620.5639	0.4217	0.2714	385.2560	0.5748
BackSplit with ⊕10	0.4921	276.7345	0.6195	0.4866	74.6971	0.2799	0.5746	113.6748	0.4470	0.4649	602.8102	0.4422	0.2805	378.3928	0.5860

Table 4. Single-fold evaluation of BackSplit using nnInteractive-derived support structures. Results are reported for (a) KiTS23, (b) PANTHER-MR, (c) NSCLC-Rad, (d) AutoPET, and (e) MSWAL. Even when auxiliary structures are generated from noisy nnInteractive segmentations with 7- and 10-click simulated user inputs (see Supplementary), BackSplit consistently outperforms standard training across Dice, HD-95, and NSD metrics.

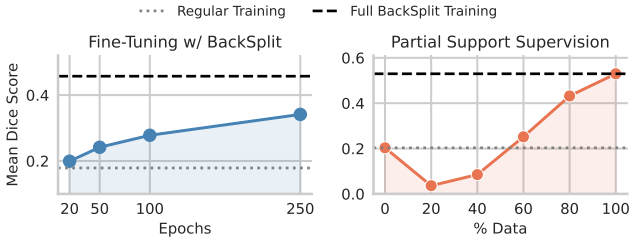

Figure 3. **(Left)** Fine-tuning a pretrained binary model with auxiliary structures steadily improves performance across epochs. **(Right)** Under partial support supervision, limited auxiliary data initially reduce performance but later yield consistent gains as more auxiliary structures are added, approaching full BackSplit performance. Evaluated on KiTS23 (Target is Cyst).

our pretrained model as input to nnInteractive. The resulting pseudo-annotations are spatially aligned with the original lesion masks and used for joint training following the same setup as the previous experiments. This experiment evaluates the robustness of BackSplit when trained with imperfect, interactively generated auxiliary supervision.

Results. As shown in the Supplementary Material, the auxiliary segmentations generated using nnInteractive exhibit modest accuracy. Nevertheless, when incorporated into U-Net training, we observe consistent performance improvements across all five datasets for both the 7-click and 10-click configurations (Tab. 4). These findings demonstrate that even highly noisy, interactively generated auxiliary cues can enhance target segmentation performance under the BackSplit paradigm, highlighting its robustness to imperfect supervision.

4.6. Behavior of BackSplit under Fine-Tuning and Partial Supervision

To explore how researchers can readily adopt the BackSplit paradigm, we conduct two complementary experiments using the KiTS23 dataset with cyst segmentation as the target task. First, we fine-tune a pretrained U-Net by incorporating auxiliary support structures. Second, we analyze the effect of training from scratch when only a subset of samples includes auxiliary annotations. Together, these experiments illustrate how existing models and datasets can be easily adapted to achieve improved performance with minimal modification.

Results. As shown in Fig. 3 (Left), fine-tuning a pretrained binary model with auxiliary support structures yields immediate performance improvements in Dice score, even after just 50 epochs of training. The performance continues to improve with additional training, nearly doubling by 250 epochs compared to the original binary model. In Fig. 3 (Right), we observe an interesting trend: when only a small fraction of auxiliary labels is included, models trained from scratch initially underperform, likely due to confusion between target and auxiliary structures. However, as the proportion of auxiliary labels increases linearly, the model gradually recovers and begins to exhibit the expected performance improvements characteristic of the BackSplit paradigm.

5. Conclusion

We introduced BackSplit, a simple and theoretically grounded training paradigm that improves lesion segmentation by decomposing the background into semantically meaningful auxiliary structures. From an information-theoretic standpoint, we showed that structured supervision increases the expected Fisher Information over conventional binary training, leading to more stable estimators. Experiments across five datasets, spanning multiple imaging modalities and model architectures, confirm BackSplit's effectiveness and generality.

We further validate BackSplit under realistic conditions using automatically derived and noisy auxiliary segmentations from large pretrained models and interactive frameworks. Across all settings, BackSplit maintains consistent performance gains. Although our theoretical analysis assumes a large-sample regime—where finer label granularity increases Fisher curvature—we acknowledge that, in small-data scenarios, this could amplify sampling noise and risk overfitting. Empirically, however, we did not observe such effects within typical medical segmentation dataset sizes.

In future work, researchers may explore which specific auxiliary structures contribute most to performance improvements. Another promising direction is the use of language-driven or proxy representations of anatomical context (e.g., textual cues such as "liver"

or "kidney") to provide structural information without requiring explicit segmentations. Overall, BackSplit offers a strong and extensible paradigm that enhances lesion segmentation and supports clinical decision-making.

References

[1] Nabila Abraham and Naimul Mefraz Khan. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, pages 683–687. IEEE, 2019. 1

[2] Hugo JWL Aerts, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, et al. Data from nsclc-radiomics-genomics. *(No Title)*, 2015. 6, 17, 18

[3] Amparo Soeli Betancourt Tarifa, Faisal Mahmood, Uffe Bernchou, and Peter Jan Koopmans. Panther challenge: Public training dataset, 2025. 6, 17, 18, 20

[4] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022. 16

[5] Binghui Chen, Weihong Deng, and Haifeng Shen. Virtual class enhanced discriminative embedding learning. *Advances in Neural Information Processing Systems*, 31, 2018. 19

[6] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 1

[7] 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. 1

[8] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019. 2

[9] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016. 1

[10] Elijah Cole, Kimberly Wilber, Grant Van Horn, Xuan Yang, Marco Fornoni, Pietro Perona, Serge Belongie, Andrew Howard, and Oisin Mac Aodha. On label granularity and object localization. In *European Conference on Computer Vision*, pages 604–620. Springer, 2022. 3

[11] Tugba Akinci D'Antonoli, Lucas K Berger, Ashraya K Indrakanti, Nathan Vishwanathan, Jakob Weiß, Matthias Jung, Zeynep Berkarda, Alexander Rau, Marco Reisert, Thomas Küstner, et al. Totalsegmentator mri: Robust sequence-independent segmentation of multiple anatomic structures in mri. *arXiv preprint arXiv:2405.19492*, 2024. 7, 17

[12] Ruining Deng, Quan Liu, Can Cui, Tianyuan Yao, Jun Long, Zuhayr Asad, R Michael Womick, Zheyu Zhu, Agnes B Fogo, Shilin Zhao, et al. Omni-seg: A scale-aware dynamic network for renal pathological image segmentation. *IEEE Transactions on Biomedical Engineering*, 70(9):2636–2644, 2023. 2

[13] Ruining Deng, Quan Liu, Can Cui, Tianyuan Yao, Juming Xiong, Shunxing Bao, Hao Li, Mengmeng Yin, Yu Wang, Shilin Zhao, et al. Hats: Hierarchical adaptive taxonomy segmentation for panoramic pathology image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 155–166. Springer, 2024. 3

[14] Ruining Deng, Quan Liu, Can Cui, Tianyuan Yao, Jialin Yue, Juming Xiong, Lining Yu, Yifei Wu, Mengmeng Yin, Yu Wang, et al. Prpseg: Universal proposition learning for panoramic renal pathology segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11736–11746, 2024. 3

[15] Konstantin Dmitriev and Arie E Kaufman. Learning multi-class segmentations from single-class datasets. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9501–9511, 2019. 2

[16] Xi Fang and Pingkun Yan. Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. *IEEE Transactions on Medical Imaging*, 39(11):3619–3629, 2020. 2

[17] Dimitris Fotakis, Alkis Kalavasis, Vasilis Kontonis, and Christos Tzamos. Efficient algorithms for learning from coarse labels. In *Conference on Learning Theory*, pages 2060–2079. PMLR, 2021. 3

[18] Fabio Garcea, Alessio Serra, Fabrizio Lamberti, and Lia Morra. Data augmentation for medical imaging: A systematic literature review. *Computers in biology and medicine*, 152:106391, 2023. 1

[19] Sergios Gatidis, Tobias Hepp, Marcel Früh, Christian La Fougère, Konstantin Nikolaou, Christina Pfannenberg, Bernhard Schölkopf, Thomas Küstner, Clemens Cyran, and Daniel Rubin. A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. *Scientific Data*, 9(1):601, 2022. 6, 7, 19, 20

[20] Robert Graf, Paul Platzek, Evamaria Olga Riedel, Constanze Ramschütz, Sophie Starck, Hendrik K Möller, Matan Atad, Henry Völzke, Robin Bülow, Carsten Oliver Schmidt, et al. Vibesegmentator: full body mri segmentation for the nako and uk biobank. *European Radiology*, pages 1–15, 2025. 7, 17

[21] Charley Gros, Andreanne Lemay, and Julien Cohen-Adad. Softseg: Advantages of soft versus binary training for image segmentation. *Medical image analysis*, 71:102038, 2021. 2

[22] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pages 272–284. Springer, 2021. 1, 16, 19

[23] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022. 1

[24] Tao He, Junjie Hu, Ying Song, Jixiang Guo, and Zhang Yi. Multi-task learning for the segmentation of organs at risk with label dependence. *Medical Image Analysis*, 61:101666, 2020. 3

[25] Nicholas Heller, Fabian Isensee, Dasha Trofimova, Resha Tejpaul, Zhongchen Zhao, Huai Chen, Lisheng Wang, Alex Golts, Daniel Khapun, Daniel Shats, Yoel Shoshan, Flora Gilboa-Solomon, Yasmeen George, Xi Yang, Jianpeng Zhang, Jing Zhang, Yong Xia, Mengran Wu, Zhiyang Liu, Ed Walczak, Sean McSweeney, Ranveer Vasdev, Chris Hornung, Rafat Solaiman, Jamee Schoephoerster, Bailey Abernathy, David Wu, Safa Abdulkadir, Ben Byun, Justice Spriggs, Griffin Struyk, Alexandra Austin, Ben Simpson, Michael Hagstrom, Sierra Virnig, John French, Nitin Venkatesh, Sarah Chan, Keenan Moore, Anna Jacobsen, Susan Austin, Mark Austin, Subodh Regmi, Nikolaos Papanikolopoulos, and Christopher Weight. The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct, 2023. 6, 17, 18, 20

[26] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul J Kennedy. Atrous convolution for binary semantic segmentation of lung nodule. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1015–1019. IEEE, 2019. 2

[27] He Huang, Guang Yang, Wenbo Zhang, Xiaomei Xu, Weiji Yang, Weiwei Jiang, and Xiaobo Lai. A deep multi-task learning framework for brain tumor segmentation. *Frontiers in oncology*, 11:690244, 2021. 2

[28] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 1, 6, 7, 16, 17, 18, 19, 20, 21

[29] Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F Jaeger. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 488–498. Springer, 2024. 6, 16, 19

[30] Fabian Isensee, Maximilian Rokuss, Lars Krämer, Stefan Dinkelacker, Ashis Ravindran, Florian Stritzke, Benjamin Hamm, Tassilo Wald, Moritz Langenberg, Constantin Ulrich, et al. nninteractive: Redefining 3d promptable segmentation. *arXiv preprint arXiv:2503.08373*, 2025. 7, 17

[31] Alexander Jaus, Constantin Seibold, Simon Reiß, Lukas Heine, Anton Schily, Moon Kim, Fin Hendrik Bahnsen, Ken Herrmann, Rainer Stiefelhagen, and Jens Kleesiek. Anatomy-guided pathology segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–13. Springer, 2024. 3

[32] 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*, pages 285–296. PMLR, 2019. 1

[33] Zhuo Kuang, Zengqiang Yan, and Li Yu. Weakly supervised learning for multi-class medical image segmentation via feature decomposition. *Computers in Biology and Medicine*, 171:108228, 2024. 2

[34] Erwan Lecesne, Antoine Simon, Mireille Garreau, Gilles Barone-Rochette, and Céline Fouard. Segmentation of cardiac infarction in delayed-enhancement mri using probability map and transformers-based neural networks. *Computer Methods and Programs in Biomedicine*, 242:107841, 2023. 2

[35] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pages 562–570. Pmlr, 2015. 18

[36] Bicao Li, Panpan Li, Zhoufeng Liu, Bei Wang, Chunlei Li, Xuwei Guo, Jing Wang, and Wei Li. A tri-path boundary preserving network via context-aware aggregator for multi-organ medical image segmentation. *Biomedical Signal Processing and Control*, 104:107593, 2025. 3

[37] Zeju Li, Konstantinos Kamnitsas, Cheng Ouyang, Chen Chen, and Ben Glocker. Context label learning: Improving background class representations in semantic segmentation. *IEEE Transactions on Medical Imaging*, 42(6):1885–1896, 2023. 3

[38] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1

[39] Jinhua Liu, Christian Desrosiers, Dexin Yu, and Yuanfeng Zhou. Semi-supervised medical image segmentation using cross-style consistency with shape-aware and local context constraints. *IEEE Transactions on Medical Imaging*, 43(4): 1449–1461, 2023. 3

[40] Shasha Liu, Yan Li, Xiaohu Li, and Guitao Cao. Shape-aware multi-task learning for semi-supervised 3d medical image segmentation. In *2021 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 1418–1423. IEEE, 2021. 3

[41] Shangke Liu, Mohamed Amgad, Deeptej More, Muhammad A Rathore, Roberto Salgado, and Lee AD Cooper. A panoptic segmentation dataset and deep-learning approach for explainable scoring of tumor-infiltrating lymphocytes. *NPJ Breast Cancer*, 10(1):52, 2024. 2

[42] Zhihua Liu, Lei Tong, Long Chen, Feixiang Zhou, Zheheng Jiang, Qianni Zhang, Yinhai Wang, Caifeng Shan, Ling Li, and Huiyu Zhou. Canet: Context aware network for brain glioma segmentation. *IEEE Transactions on Medical Imaging*, 40(7):1763–1777, 2021. 2

[43] Thomas A Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 44(2):226–233, 1982. 4, 12

[44] Laurin Lux, Alexander H. Berger, Alexander Weers, Nico Stucki, Daniel Rueckert, Ulrich Bauer, and Johannes C. Paetzold. Topograph: An efficient graph-based framework for

strictly topology preserving image segmentation. In *ICLR*, 2025. 1

[45] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 1

[46] Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *International MICCAI brainlesion workshop*, pages 311–320. Springer, 2018. 6, 16

[47] David Oakes. Direct calculation of the information matrix via the em. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):479–482, 1999. 4, 12

[48] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 16

[49] Chongyu Qu, Tiezheng Zhang, Hualin Qiao, Yucheng Tang, Alan L Yuille, and Zongwei Zhou. Abdomenatlas-8k: Annotating 8,000 ct volumes for multi-organ segmentation in three weeks. *Advances in Neural Information Processing Systems*, 36, 2023. 7, 17, 19

[50] 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*, pages 234–241. Springer, 2015. 1, 6, 7, 16, 17, 18, 19, 20, 21

[51] 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*, pages 16560–16569, 2021. 1

[52] Nico Stucki, Johannes C Paetzold, Suprosanna Shit, Bjoern Menze, and Ulrich Bauer. Topologically faithful image segmentation via induced matching of persistence barcodes. In *International Conference on Machine Learning*, pages 32698–32727. PMLR, 2023. 3

[53] Dingjie Su, Yihao Liu, Lianrui Zuo, and Benoit Dawant. Mesh-prompted anatomy segmentation. In *Medical Imaging with Deep Learning*, 2025. 3

[54] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 1

[55] Yan Wang, Yuyin Zhou, Wei Shen, Seyoun Park, Elliot K Fishman, and Alan L Yuille. Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. *Medical image analysis*, 55:88–102, 2019. 2

[56] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5): e230024, 2023. 7

[57] Zhaodong Wu, Qiaochu Zhao, Ming Hu, Yulong Li, Haochen Xue, Zhengyong Jiang, Angelos Stefanidis, Qiufeng Wang, Imran Razzak, Zongyuan Ge, et al. Mswal: 3d multi-class segmentation of whole abdominal lesions dataset. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 378–388. Springer, 2025. 6, 7

[58] Qing Xu, Wenting Duan, and Zhen Chen. Co-seg: Mutual prompt-guided collaborative learning for tissue and nuclei segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 130–140. Springer, 2025. 3

[59] Yuanhong Xu, Qi Qian, Hao Li, Rong Jin, and Juhua Hu. Weakly supervised representation learning with coarse labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10593–10601, 2021. 3

[60] Fatima Yousaf, Sajid Iqbal, Nosheen Fatima, Tanzeela Kousar, and Mohd Shafry Mohd Rahim. Multi-class disease detection using deep learning and human brain medical imaging. *Biomedical Signal Processing and Control*, 85: 104875, 2023. 2

[61] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 1

[62] Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1195–1204, 2021. 2

[63] Yongtao Zhang, Haimei Li, Jie Du, Jing Qin, Tianfu Wang, Yue Chen, Bing Liu, Wenwen Gao, Guolin Ma, and Baiying Lei. 3d multi-attention guided multi-task learning network for automatic gastric tumor segmentation and lymph node classification. *IEEE transactions on medical imaging*, 40(6): 1618–1631, 2021. 2

[64] Wentao Zhu, Yufang Huang, Liang Zeng, Xuming Chen, Yong Liu, Zhen Qian, Nan Du, Wei Fan, and Xiaohui Xie. Anatomynet: deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Medical physics*, 46(2):576–589, 2019. 1

BackSplit: The Importance of Sub-dividing the Background in Biomedical Lesion Segmentation

Supplementary Material

Contents

A. Full Proofs and Theoretical Insights

Our goal is to prove that using a full K-class training regime we yield a greater Fisher information than training on a coarsened/collapsed binary label. This in turn leads to more statistically efficient predictions for the target class.

A.1. Setup and Notation

The notations used to prove the following sections are used from Sec. 3.1.

Assumption 1 (Regularity and Optimization / Compute)**.** *We work with a parametric family $\{p_\theta(Y \mid X) : \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^p$ is an open set containing the true parameter $\theta^\star$. We assume:*

1. ***Model regularity.*** *For each fixed input x the conditional density $p_\theta(Y \mid X = x)$ is correctly specified at $\theta^\star$ and is twice continuously differentiable in θ in a neighbourhood of $\theta^\star$. Differentiation and summation over Y can be interchanged, and the Fisher information matrices*

$$I_Y(\theta) = \mathbb{E}_\theta\big[s_Y(\theta)s_Y(\theta)^\top\big], \qquad I_Z(\theta) = \mathbb{E}_\theta\big[s_Z(\theta)s_Z(\theta)^\top\big]$$

   exist and are finite. At $\theta^\star$, $I_Y(\theta^\star)$ and $I_Z(\theta^\star)$ are non–singular.
2. ***Common architecture.*** *The multiclass and collapsed–binary models share the same network architecture and parameterization up to the final classification layer, so that both likelihoods are defined on the same parameter space Θ.*
3. ***Optimization / compute parity.*** *In practice, both models are trained with the same optimization algorithm and comparable compute budgets (e.g., same number of epochs, batch sizes, and learning–rate schedules), and we treat the resulting estimators as approximate maximum–likelihood estimators. In particular, we assume that training converges to stationary points that lie in a neighbourhood of the (local) maximizers of the corresponding log–likelihoods, so that the usual MLE asymptotics apply.*

Under Assumption 1, the standard MLE central limit theorem and delta method results used in Sec. 3 and Sec. A hold for both the multiclass and collapsed–binary estimators.

A.2. Proof of Lemma 1

Lemma 1 (Score Projection [43, 47])**.** *Let $Z = g(Y)$ be a deterministic coarsening of the label Y. Then, under regularity conditions ensuring that differentiation and summation interchange,*

$$\mathbb{E}_\theta[s_Y(\theta) \mid Z, X] = s_Z(\theta).$$

Proof: Since $Z = g(Y)$, we can write:

$$\begin{aligned}\mathbb{E}[s_Y(\theta) \mid Z, X] &= \sum_{y:g(y)=Z} \nabla_\theta \log p_\theta(y \mid X) \frac{p_\theta(y \mid X)}{p_\theta(Z \mid X)} \\ &= \frac{\sum_{y:g(y)=Z} \nabla_\theta p_\theta(y \mid X)}{p_\theta(Z \mid X)} \\ &= \nabla_\theta \log p_\theta(Z \mid X) \\ &= s_Z(\theta)\end{aligned}$$

□

A.3. Proof of Theorem 1

Theorem 1 (Label coarsening reduces expected Fisher information.)**.** *For every $\theta \in \Theta$,*

$$\mathcal{I}_Y(\theta) = \mathcal{I}_Z(\theta) + \mathbb{E}_\theta\big[\operatorname{Var}(s_Y(\theta) \mid Z, X)\big] \succeq \mathcal{I}_Z(\theta)$$

and therefore $\mathcal{I}_Z(\theta) \preceq \mathcal{I}_Y(\theta)$ in the Loewner (positive–semidefinite) order. Equality holds iff $s_Y(\theta)$ is completely determined by (Z, X), i.e. no variation in the complete–data score remains once Z is known.

Proof: From the law of total variance, for any random vector U and any conditional variable V,

$$\mathbb{E}[UU^T] = \mathbb{E}\Big[\mathbb{E}[U \mid V]\,\mathbb{E}[U \mid V]^T\Big] + \mathbb{E}[\operatorname{Var}(U \mid V)]$$

substituting $U = s_Y(\theta)$ and $V = (Z, X)$ we get.

$$\begin{aligned}\mathbb{E}_\theta[s_Y(\theta) s_Y(\theta)^T] = \\ \mathbb{E}_\theta\Big[\mathbb{E}_\theta[s_Y(\theta) \mid (Z,X)]\,\mathbb{E}_\theta[s_Y(\theta) \mid (Z,X)]^T\Big] \\ + \mathbb{E}_\theta[\operatorname{Var}(s_Y(\theta) \mid (Z,X))]\end{aligned}$$

From Lemma 1, we get that $\mathbb{E}[s_Y(\theta) \mid Z, X] = s_Z(\theta)$.

$$\begin{aligned}\mathbb{E}_\theta[s_Y(\theta) s_Y(\theta)^T] = \mathbb{E}_\theta[s_Z(\theta) s_Z(\theta)^T] \\ + \mathbb{E}_\theta[\operatorname{Var}(s_Y(\theta) \mid (Z,X))]\end{aligned}$$

From the definitions of expect Fisher Information (from Sec. 3.1), we get:

$$\mathcal{I}_Y(\theta) = \mathcal{I}_Z(\theta) + \mathbb{E}_\theta[\operatorname{Var}(s_Y(\theta) \mid (Z,X))]$$

Additionally, we can define the missing information as

$$\mathcal{I}_{\text{missing}} = \mathbb{E}_\theta[\operatorname{Var}(s_Y(\theta) \mid (Z,X))]$$

□

A.4. Proof of Corollary 1

Corollary 1 (Asymptotic Efficiency of the Multiclass MLE.)**.** *Under standard regularity and correct-specification assumptions at the true parameter θ^*, the multiclass MLE is (weakly) more statistically efficient than the collapsed-binary MLE for any smooth functional of θ. Equivalently, for any differentiable quantity $g(\theta)$ of interest, the asymptotic estimation variance of $g(\hat{\theta}_Y)$ is no greater than $g(\hat{\theta}_Z)$.*

Proof: Under correct specification and standard regularity at θ^*, the MLE is asymptotically normal with covariance equal to the Fisher Information as MLE Central Limit Theorem converges in distribution:

$$\begin{aligned}\sqrt{n}(\hat{\theta}_Y - \theta^*) &\xrightarrow{d} \mathcal{N}(0, \mathcal{I}_Y(\theta^*)^{-1}) \\ \sqrt{n}(\hat{\theta}_Z - \theta^*) &\xrightarrow{d} \mathcal{N}(0, \mathcal{I}_Z(\theta^*)^{-1})\end{aligned}$$

From Theorem 1, we have $\mathcal{I}_Y(\theta) \succeq \mathcal{I}_Z(\theta)$, if $A \succeq B \succ 0$ then $A^{-1} \preceq B^{-1}$ Loewner order reverses under inversion. So,

$$\mathcal{I}_Y(\theta^*)^{-1} \preceq \mathcal{I}_Z(\theta^*)^{-1}$$

The result so far is about quality of the parameter (weight) estimates. We are interested in the quality of our predictions. To make that final connection, we use the delta method. Consider any smooth function $g(\theta)$ - for instance, the target-class posterior $\tau(\theta) = \eta_c(x_0; \theta)$, where x_0 denotes an arbitrary fixed input. By the delta method,

$$\sqrt{n}(g(\hat{\theta}) - g(\theta_0)) \xrightarrow{d} \mathcal{N}(0, G\Sigma G^T)$$

where $G = \nabla_\theta g(\theta_0)$, θ_0 is a parameter with non-singular Fisher information, and Σ is the asymptotic covariance of the corresponding MLE. Therefore, each of the models can be written as:

$$\begin{aligned}\sqrt{n}(\tau(\hat{\theta}_Y) - \tau(\theta^*)) &\xrightarrow{d} \mathcal{N}(0, \nabla_\tau^T \mathcal{I}_Y(\theta^*)^{-1} \nabla_\tau) \\ \sqrt{n}(\tau(\hat{\theta}_Z) - \tau(\theta^*)) &\xrightarrow{d} \mathcal{N}(0, \nabla_\tau^T \mathcal{I}_Z(\theta^*)^{-1} \nabla_\tau)\end{aligned}$$

From Theorem 1, we can show that,

$$\nabla_\tau^T \mathcal{I}_Y(\theta^*)^{-1} \nabla_\tau \leq \nabla_\tau^T \mathcal{I}_Z(\theta^*)^{-1} \nabla_\tau$$

So,

$$\begin{aligned}\operatorname{Var}\Big(\sqrt{n}\,(\tau(\hat{\theta}_Y) - \tau(\theta^*))\Big) \\ \leq \operatorname{Var}\Big(\sqrt{n}\,(\tau(\hat{\theta}_Z) - \tau(\theta^*))\Big)\end{aligned}$$

Hence the estimates for $p_\theta(Y = c \mid x_0)$ are asymptotically tighter under multiclass training.

□

A.5. Proof of Proposition 1

Proposition 1 (Softmax Expected Fisher Information Decomposition.)**.** *Consider a softmax model with logits $f(x;\theta) \in \mathbb{R}^K$ and class probabilities $\eta = softmax(f)$. Let $J_f(x;\theta) = \partial f(x;\theta)/\partial\theta \in \mathcal{R}^{K\times p}$ denote the Jacobian of the logits with respect to the model parameters and let $e_c \in \mathbb{R}^K$ denote the c-th standard basis (one-hot) vector. Then for any fixed input $X = x$:*

$$\mathcal{I}_Y(\theta \mid X = x) = J_f^T(Diag(\eta) - \eta\eta^T)J_f$$

is the conditional expected Fisher information for the multiclass likelihood, and

$$\mathcal{I}_Z(\theta \mid X = x) = J_f^T\left(\frac{\eta_c}{(1-\eta_c)}(e_c - \eta)(e_c - \eta)^T\right)J_f$$

is the corresponding quantity for the collapsed-binary label. Hence,

$$\mathcal{I}_Y(\theta \mid X = x) = \mathcal{I}_Z(\theta \mid X = x) + (1-\eta_c)J_f^T(Diag(\pi) - \pi\pi^T)J_f \succeq \mathcal{I}_Z(\theta \mid X = x)$$

where $\pi_k = \eta_k/(1-\eta_c)$ for $k \neq c$ and $\pi_c = 0$.

Proof: Let $f(x;\theta) \in \mathbb{R}^K$ be the logits and $\eta = \text{softmax(f)}$. For one observation $(X = x, Y = y)$ the log-likelihood for the datapoint is

$$\ell(\theta; y; x) = \log \eta_y = f_y - \log\sum_{j=1}^K e^{f_j}$$

To get the conditional expected fisher information let $J_f(x;\theta) = \nabla_\theta f(x;\theta)$ be the Jacobian of the logits with respect to θ. Then by the chain rule,

$$s_Y(\theta; y; x) = \nabla_\theta \ell = \left(\frac{\partial f}{\partial \theta}\right)^T \nabla_f \ell = J_f^T \nabla_f \log \eta_y$$

The softmax derivatives are,

$$\frac{\partial \eta_y}{\partial f_j} = \eta_y(\delta_{yj} - \eta_j)$$
$$\implies \nabla_f \log \eta_y = \frac{1}{\eta_y}\nabla_f \eta_y = \delta_y - \eta = e_y - \eta$$

δ_{yj} is the Kronecker delta and e_y is the y-th standard basis vector. So the score function can be written as

$$s_Y(\theta; y; x) = J_f^T(e_y - \eta)$$

So the conditional expectation at a point $(X = x, Y = y)$ is

$$\mathcal{I}_Y(\theta \mid x) = \sum_{y=1}^k \eta_y[s_Y(\theta; y; x)s_Y(\theta; y; x)^T]$$
$$\mathcal{I}_Y(\theta \mid x) = \sum_{y=1}^k \eta_y J_f^T(e_y - \eta)(e_y - \eta)^T J_f$$

The inner sum is the covariance of the one-hot vector e_Y with mean η

$$\sum_{y=1}^k \eta_y e_y e_y^T - \eta\eta^T = \text{Diag}(\eta) - \eta\eta^T$$

Therefore,

$$\mathcal{I}_Y(\theta \mid x) = J_f^T(\text{Diag}(\eta) - \eta\eta^T)J_f$$

For the conditional expected fisher information for the collapsed Bernoulli case we start with the log-likelihood.

$$\ell_Z(\theta; z; x) = z\log q + (1-z)\log(1-q)$$

Where $q = q(x;\theta)$. Differentiating w.r.t. θ we get

$$\nabla_\theta \ell_Z = \left(\frac{z}{q} - \frac{1-z}{1-q}\right)\nabla_\theta q = \frac{z-q}{q(1-q)}\nabla_\theta q$$

Therefore the score function for the collapsed Bernoulli is

$$s_Z(\theta; z; x) = \frac{z-q}{q(1-q)}\nabla_\theta q$$

Therefore the score function for the collapsed Bernoulli is

$$s_Z(\theta; z; x) = \frac{z-q}{q(1-q)}\nabla_\theta q$$

So the expected Fisher Information at x is

$$\mathcal{I}_Z(\theta \mid X = x) = \mathbb{E}_\theta[s_Z(\theta)s_Z(\theta)^T \mid X = x]$$

$$\mathcal{I}_Z(\theta \mid X = x) = \frac{\mathbb{E}_\theta[(Z-q)^2 \mid x]}{q^2(1-q)^2}(\nabla_\theta q)(\nabla_\theta q)^T$$

The numerator term is $\mathrm{Var}(Z \mid X = x) = \mathbb{E}[Z \mid X] - (\mathbb{E}[Z \mid X])^2 = q - q^2 = q(1-q)$

$$\mathcal{I}_Z(\theta \mid X = x) = \frac{1}{q(1-q)} (\nabla_\theta q)(\nabla_\theta q)^T$$

To get $\nabla_\theta q$, we take $q = \eta_c(f)$

$$\frac{\partial \eta_c}{\partial f_j} = \eta_c(\delta_{cj} - \eta_j) \implies \nabla_f \eta_c = \eta_c(e_c - \eta)$$

Therefore,

$$\nabla_\theta q = \left(\frac{\partial f}{\partial \theta}\right)^T \nabla_f \eta_c = J_f^T(\eta_c(e_c - \eta))$$

Since $q = \eta_c$

$$\mathcal{I}_Z(\theta \mid X = x) = J_f^T \frac{1}{\eta_c(1-\eta_c)} (\eta_c(e_c - \eta))(\eta_c(e_c - \eta)^T) J_f$$

$$\mathcal{I}_Z(\theta \mid X = x) = J_f^T \left(\frac{\eta_c}{(1-\eta_c)} (e_c - \eta)(e_c - \eta)^T\right) J_f$$

To relate the two, express the class probabilities as $\eta = [\,\eta_c,\, (1-\eta_c)\pi\,]$, where $\pi_k = \eta_k/(1-\eta_c)$ for $k \neq c$ and $\pi_c = 0$. Substituting this form into the multiclass curvature in logit space gives

$$\begin{aligned} &\mathrm{Diag}(\eta) - \eta\eta^\top \\ &= \frac{\eta_c}{1-\eta_c}(e_c - \eta)(e_c - \eta)^\top + (1-\eta_c)\big(\mathrm{Diag}(\pi) - \pi\pi^\top\big). \end{aligned}$$

Multiplying by $J_f^\top$ and J_f on both sides yields the desired decomposition:

$$\begin{aligned} &\mathcal{I}_Y(\theta \mid X = x) = \mathcal{I}_Z(\theta \mid X = x) \\ &+ (1-\eta_c)\, J_f^\top \big(\mathrm{Diag}(\pi) - \pi\pi^\top\big) J_f \succeq \mathcal{I}_Z(\theta \mid X = x). \end{aligned}$$

The second term is positive semi–definite and quantifes the information lost by collapsing all non–target classes into a single background label.

□

A.6. Limitations for Classification Tasks

While the proposed theory extends naturally to segmentation, it becomes less reliable in standard classification settings due to severe class imbalance. When K classes are mapped into a single "target vs. rest" binary task, the positive class often occupies a vanishing fraction of the data, yielding posterior probabilities $\eta_c(x;\theta)$ that are extremely close to 0 or 1. In this regime, the Bernoulli Fisher term $\eta_c(1-\eta_c)$ degenerates, leading to ill-conditioned or singular information matrices. As a result, the asymptotic efficiency and variance comparisons derived under balanced or well-behaved posteriors no longer hold. In practice, this means that for highly imbalanced classification datasets, coarsening amplifies numerical instability and the theoretical ordering $\mathcal{I}_Y(\theta) \succeq \mathcal{I}_Z(\theta)$ may not provide a useful description of estimator behavior.

A.7. Note on Natural Images vs 3D Medical Images

The observed strength of the proposed framework in 3D medical imaging tasks can be attributed to the high variability and structured heterogeneity of the *background* in volumetric scans. Unlike 2D natural-image segmentation, where the "non-target" regions are often homogeneous and semantically uninformative (e.g., sky, wall, road), medical volumes contain multiple anatomical structures—organs, vessels, and tissues—each contributing distinct non-target gradients. This diversity amplifies the within-rest Fisher term in our decomposition, yielding a substantial gap between the multiclass and collapsed-binary information. Consequently, modeling each anatomical class preserves rich curvature directions in parameter space, improving statistical efficiency and convergence. In contrast, for dichotomous segmentation tasks in natural images or datasets with relatively uniform backgrounds, the variation is minimal, making the binary and multiclass formulations nearly equivalent.

B. Heuristics for selecting auxiliary classes

In practice, the choice of auxiliary support structures plays an important role in maximizing the benefits of BackSplit. We outline several practical heuristics that can guide this selection.

1. Select the organ containing the lesion. The most effective auxiliary structure is typically the organ in which the lesion resides, as it fully encloses the target region. Including this structure allows the model to learn fine-grained variations in local tissue appearance, which directly improves boundary precision and reduces ambiguity.

2. Include adjacent or surrounding organs. Structures that spatially border the lesion serve as anatomical anchors. These surrounding organs provide contextual cues that help the model disambiguate lesion boundaries and reduce drift into neighboring regions. This is particularly useful for lesions located near organ interfaces.

3. Add organs prone to false positives. A third useful strategy is to incorporate organs whose appearance may cause confusion with the target lesion. These structures often contain components or textures that resemble the lesion class, leading to false positives in standard training. By ex-

plicitly modeling these tissues, the network learns discriminative cues that reduce such errors and improve robustness.

These heuristics offer a practical starting point for selecting auxiliary structures and can be adapted based on dataset characteristics, anatomical complexity, and known failure modes of baseline models.

C. Model Weights and Code

Code and pretrained weights will be released following publication. All models are implemented in PyTorch [48], and inference can be performed directly using the nnU-Net [28] library.

D. Training Configuration and Hyperparameters

The following section summarizes the training configurations used for all models.

D.1. U-Nets

All U-Net [50] models were trained using the nnU-Net [28] 3D full-resolution configuration. A batch size of 2 and varying patch sizes were used across experiments, depending on dataset-specific memory requirements. All data were normalized according to imaging modality (CT or MRI) and subsequently resampled to the standardized nnU-Net resolution. The network backbone followed nnU-Net's generic 3D U-Net design, employing InstanceNorm3d and LeakyReLU nonlinearities.

Training used nnU-Net's default Dice + Cross-Entropy loss with deep supervision enabled. Optimization was performed with stochastic gradient descent (momentum 0.99), an initial learning rate of 0.01, weight decay of 3×10^{-5}, and the PolyLR learning rate schedule. Each model was trained for 1000 epochs, with 250 iterations per epoch. All experiments were conducted on an NVIDIA A100 GPU.

D.2. ResEncU-Net

The ResEncU-Net [29] models follow the same training configuration as the U-Nets, differing only in network architecture. We use the Residual Encoder U-Net with the Medium configuration, as implemented in the nnU-Net framework. All optimization settings, normalization procedures, and training schedules remain identical to those described in the U-Net subsection.

D.3. SegResNet

For the SegResNet experiments, we use the MONAI [4] SegResNet [46] implementation wrapped inside the nnU-Net framework, keeping all data preprocessing, normalization, and resampling steps identical to the nnU-Net. The primary differences lie in the network architecture and optimizer settings. Unlike the standard nnU-Net configuration, which uses SGD, SegResNet is optimized using Adam with a learning rate of 1×10^{-4}, a weight decay of 1×10^{-5}, an ε value of 1×10^{-5} and a PolyLR learning rate schedule. Models were trained using the default Dice + Cross-Entropy loss without deep supervision enabled.

Auxiliary Task (Support Structure)	**Dice** ↑ ⊕ 7 clicks	**Dice** ↑ ⊕10 clicks
(a) KiTS23		
Kidney	0.7420	0.7419
Tumor	0.5457	0.5428
(b) PANTHER-MR		
Pancreas	0.6779	0.6823
(c) NSCLC-Rad		
Esophagus	0.7155	0.7256
Heart	0.8408	0.8390
Lung Left	0.9606	0.9687
Lung Right	0.9627	0.9688
Spinal Cord	0.8244	0.8263
(d–e) AutoPET (MSWAL)		
Aorta	0.8421 (0.9335)	0.8506 (0.9335)
Gall Bladder	0.6765 (0.8003)	0.6750 (0.8003)
Kidney Left	0.9333 (0.9585)	0.9334 (0.9585)
Kidney Right	0.9416 (0.9589)	0.9396 (0.9589)
Liver	0.9388 (0.9662)	0.9360 (0.9662)
Pancreas	0.7228 (0.8497)	0.7374 (0.8497)
Postcava	0.7648 (0.7724)	0.7804 (0.7724)
Spleen	0.9379 (0.9655)	0.9353 (0.9655)
Stomach	0.8718 (0.8895)	0.8791 (0.8895)

Table 5. Simulated interactive segmentation results using 7 and 10 user clicks (⊕) across five datasets for generating support structures. For KiTS, PANTHER-MR, and NSCLC-Rad, ground-truth masks were used to randomly sample 7 or 10 seed points, followed by segmentation using nnInteractive. For AutoPET and MSWAL, the seed points were sampled from model-derived organ segmentations.

D.4. SwinUNETR

For the transformer-based experiments (as shown in Sec. G.5), we employ the 3D SwinUNETR [22] architecture from MONAI [4], integrated into the nnU-Net framework while keeping all preprocessing, normalization, and resampling procedures consistent with the standard nnU-Net pipeline. The main differences arise in the network architecture and optimization strategy.

SwinUNETR is trained using the AdamW optimizer with an initial learning rate of 8×10^{-4}, a weight decay of 0.01, and an ε value of 1×10^{-5}. To better suit transformer training dynamics, we replace nnU-Net's default PolyLR schedule with a CosineAnnealingLR scheduler, setting $T_{\max}$ to the total number of training epochs and a minimum learning rate of 1×10^{-6}.

Auxiliary Task (Support Structure)	Dice ↑ TotalSegmentator	Dice ↑ VIBESegmentator
(a–b) AutoPET (MSWAL)		
Aorta	0.9228 (0.9216)	0.7748 (0.8211)
Gall Bladder	0.8307 (0.7612)	0.3875 (0.4609)
Kidney Left	0.9192 (0.9427)	0.9266 (0.8799)
Kidney Right	0.9214 (0.9433)	0.9282 (0.8723)
Liver	0.9761 (0.9719)	0.571 (0.8673)
Pancreas	0.8834 (0.8841)	0.6592 (0.697)
Postcava	0.9062 (0.8614)	0.8275 (0.7092)
Spleen	0.9631 (0.9578)	0.848 (0.8284)
Stomach	0.9473 (0.9377)	0.8728 (0.8036)

Table 6. Auxiliary segmentation performance from large pretrained models on AutoPET and MSWAL. Reported Dice scores reflect organ segmentations generated by TotalSegmentator and VIBESegmentator, with values in parentheses indicating corresponding performance on MSWAL. These automatically derived masks are used as support structures for BackSplit.

All other training settings—including loss function (Dice + Cross-Entropy), batch size, and epoch count—are kept consistent with the U-Net and SegResNet configurations unless otherwise noted.

D.5. Minimal Parameter Overhead of BackSplit

An advantage of BackSplit is its minimal parameter overhead. Even in the most demanding setting—where the largest number of auxiliary classes is added—the increase in model parameters is approximately 0.02%. This negligible overhead ensures that BackSplit remains practical, adding virtually no computational burden while providing strong performance gains.

E. Interactive Segmentation Performance

Table 5 reports the segmentation quality of auxiliary support structures generated using nnInteractive [30] with 7 and 10 positive clicks. Although the resulting masks exhibit only modest accuracy, they are sufficiently informative for BackSplit to produce strong performance gains in downstream lesion segmentation tasks. This demonstrates that BackSplit remains effective even when auxiliary labels are noisy and derived from lightweight interactive segmentation rather than precise manual annotations.

F. Large Models Segmentation Performance

To further assess the robustness of BackSplit, we evaluate auxiliary segmentations produced by large pretrained models—TotalSegmentator [11] and VIBESegmentator [20]—on the AutoPET and MSWAL datasets. These models are widely used for comprehensive whole-body and abdominal organ parsing and provide a strong reference point for high-quality automatic segmentation. We compare their outputs with our model-derived segmentations obtained from a U-Net trained on the AbdomenAtlas-Mini1.0 [49] dataset. As shown in Table 6, the pretrained models achieve strong organ segmentation accuracy on AutoPET and MSWAL.

Method	Patch Size Variation 1 Dice ↑	HD-95 ↓	NSD ↑	Patch Size Variation 2 Dice ↑	HD-95 ↓	NSD ↑
(a) Dataset: KiTS23 with Target Class = {Cyst}						
Patch Size Value	[128,128,128]			[160,128,128]		
Regular Training	0.2033	425.3273	0.1906	0.2117	422.4271	0.1981
BackSplit	0.5297	249.5371	0.6703	0.5295	251.8412	0.6732
(b) Dataset: PANTHER-MR with Target Class = {Tumor}						
Patch Size Value	[48,192,224]			[48,224,224]		
Regular Training	0.3828	157.3470	0.2320	0.4409	152.3861	0.2623
BackSplit	0.4906	54.2010	0.2796	0.4732	52.0249	0.2718
(c) Dataset: NSCLC-Radiomics with Target Class = {GTV}						
Patch Size Value	[64,192,192]			[32,160,192]		
Regular Training	0.5279	140.2698	0.4049	0.4984	164.0937	0.3771
BackSplit	0.5862	124.2557	0.4533	0.5724	123.7844	0.4431

Table 7. Single-fold evaluation of BackSplit vs. regular training using a U-Net backbone across two patch-size configurations (batch size fixed at 2). Results are shown for (a) KiTS23, (b) PANTHER-MR, and (c) NSCLC-Radiomics. Across all datasets and both patch-size settings, BackSplit consistently outperforms regular training on Dice, HD-95, and NSD metrics, demonstrating robustness to changes in patch size.

G. BackSplit Performance during different training settings

In this section, we evaluate the effectiveness of BackSplit across different training conditions and architectural variations to further demonstrate its robustness as a general paradigm.

G.1. Evaluating BackSplit Under Varying Batch Sizes

To assess the robustness of BackSplit, we first evaluate its performance under varying batch sizes. Given typical GPU memory constraints for 3D medical segmentation models, we consider realistic batch sizes of 2 and 4, and conduct experiments on the KiTS23 [25], PANTHER [3], and NSCLC-Radiomics [2] datasets. In all cases, BackSplit consistently outperforms regular training, demonstrating its effectiveness irrespective of batch size (as shown in Tab. 8). All experiments were performed on a single fold using U-Net [50] with the nnU-Net framework [28].

Method	Batch Size = 2			Batch Size = 4		
	Dice ↑	HD-95 ↓	NSD ↑	Dice ↑	HD-95 ↓	NSD ↑
(a) Dataset: KiTS23 with Target Class = {Cyst}						
Regular Training	0.2033	425.3273	0.1906	0.2630	404.6592	0.2624
BackSplit	0.5297	249.5371	0.6703	0.5261	249.6824	0.6716
(b) Dataset: PANTHER-MR with Target Class = {Tumor}						
Regular Training	0.3828	157.3470	0.2320	0.3636	202.4875	0.2129
BackSplit	0.4906	54.2010	0.2796	0.4553	59.1720	0.2696
(c) Dataset: NSCLC-Radiomics with Target Class = {GTV}						
Regular Training	0.5279	140.2698	0.4049	0.5530	136.1963	0.4270
BackSplit	0.5862	124.2557	0.4533	0.5549	127.0088	0.4353

Table 8. Single-fold evaluation of BackSplit vs. regular training using a U-Net backbone across two batch sizes (2 and 4). Results are shown for (a) KiTS23, (b) PANTHER-MR, and (c) NSCLC-Radiomics. In all datasets and for both batch-size settings, BackSplit consistently outperforms regular training across Dice, HD-95, and NSD metrics, demonstrating robustness to changes in batch size.

G.2. Evaluating BackSplit Under Varying Patch Sizes

To further evaluate the robustness of BackSplit, we examine its performance under varying patch sizes for each dataset. Patch size is a critical factor in 3D medical segmentation due to differences in anatomical scale and memory constraints. We conduct experiments on the KiTS23 [25], PANTHER [3], and NSCLC-Radiomics [2] datasets using multiple patch-size configurations. Across all settings, BackSplit consistently outperforms regular training, demonstrating its stability with respect to patch-size variation (as shown in Tab. 7). All experiments were performed on a single fold using U-Net [50] within the nnU-Net framework [28] keeping batch size constant, at a size of two.

G.3. BackSplit with and without Deep Supervision

We also assess the impact of deep supervision [35], a mechanism used in the nnU-Net framework to provide intermediate supervision from the layers and enhance performance in 3D segmentation networks. For each dataset, we compare models trained with and without deep supervision while keeping all other training settings fixed. In both configurations, BackSplit consistently outperforms standard training, indicating that its benefits are complementary to deep supervision rather than dependent on it (as shown in Tab. 9). All experiments were conducted on a single fold using U-Net [50] within the nnU-Net framework [28].

Method	with Deep Supervision [35]			without Deep Supervision [35]		
	Dice ↑	HD-95 ↓	NSD ↑	Dice ↑	HD-95 ↓	NSD ↑
(a) Dataset: KiTS23 with Target Class = {Cyst}						
Regular Training	0.2033	425.3273	0.1906	0.1339	441.7240	0.1300
BackSplit	0.5297	249.5371	0.6703	0.4758	281.8879	0.6137
(b) Dataset: PANTHER-MR with Target Class = {Tumor}						
Regular Training	0.3828	157.3470	0.2320	0.4412	179.4845	0.2590
BackSplit	0.4906	54.2010	0.2796	0.4776	46.5688	0.2905
(c) Dataset: NSCLC-Radiomics with Target Class = {GTV}						
Regular Training	0.5279	140.2698	0.4049	0.5059	158.2856	0.3765
BackSplit	0.5862	124.2557	0.4533	0.5829	115.2792	0.4564

Table 9. Single-fold evaluation of BackSplit with and without Deep Supervision [35] using a U-Net backbone. Results are shown for (a) KiTS23, (b) PANTHER-MR, and (c) NSCLC-Radiomics. Across all datasets and under both training conditions, BackSplit consistently outperforms regular training on Dice, HD-95, and NSD metrics, demonstrating that its benefits hold regardless of whether Deep Supervision is employed.

Method	3D U-Net [28, 50]			2D U-Net [28, 50]		
	Dice ↑	HD-95 ↓	NSD ↑	Dice ↑	HD-95 ↓	NSD ↑
(a) Dataset: KiTS23 with Target Class = {Cyst}						
Regular Training	0.2033	425.3273	0.1906	0.3234	352.3277	0.4231
BackSplit	0.5297	249.5371	0.6703	0.4111	313.9846	0.5126
(b) Dataset: PANTHER-MR with Target Class = {Tumor}						
Regular Training	0.3828	157.3470	0.2320	0.2349	265.2399	0.1312
BackSplit	0.4906	54.2010	0.2796	0.3611	137.2006	0.2031
(c) Dataset: NSCLC-Radiomics with Target Class = {GTV}						
Regular Training	0.5279	140.2698	0.4049	0.4695	102.8473	0.3676
BackSplit	0.5862	124.2557	0.4533	0.4753	125.5158	0.3728

Table 10. Single-fold evaluation of BackSplit using 3D and 2D U-Net architectures on (a) KiTS23, (b) PANTHER-MR, and (c) NSCLC-Radiomics. Across all datasets and for both 3D and 2D models, BackSplit consistently outperforms regular training on Dice, HD-95, and NSD metrics. These results demonstrate that the benefits of the BackSplit paradigm hold irrespective of network dimensionality.

G.4. BackSplit Performance on 2D U-Net Architectures

We additionally evaluate BackSplit using both 3D U-Net and 2D U-Net architectures to assess its effectiveness across different network dimensionalities. While 3D models typically capture richer volumetric context and 2D models offer computational efficiency, BackSplit yields consistent performance improvements in both settings. This demonstrates

Method	3D U-Net [28, 50] Dice ↑	HD-95 ↓	NSD ↑	3D SwinUNETR [22] Dice ↑	HD-95 ↓	NSD ↑
(a) Dataset: KiTS23 with Target Class = {Cyst}						
Regular Training	0.2033	425.3273	0.1906	0.1336	438.8709	0.1035
BackSplit	0.5297	249.5371	0.6703	0.3863	309.6816	0.5010
(b) Dataset: PANTHER-MR with Target Class = {Tumor}						
Regular Training	0.3828	157.3470	0.2320	0.3113	163.1136	0.1480
BackSplit	0.4906	54.2010	0.2796	0.3705	59.1148	0.1926
(c) Dataset: NSCLC-Radiomics with Target Class = {GTV}						
Regular Training	0.5279	140.2698	0.4049	0.4936	154.8597	0.3674
BackSplit	0.5862	124.2557	0.4533	0.5344	131.8715	0.3942

Table 11. Single-fold evaluation of BackSplit on 3D U-Net and transformer-based (3D SwinUNETR) architecture across (a) KiTS23, (b) PANTHER-MR, and (c) NSCLC-Radiomics. In all datasets and for both architectures, BackSplit consistently outperforms regular training on Dice, HD-95, and NSD metrics. These results demonstrate that the BackSplit paradigm generalizes beyond CNNs and remains effective for transformer-based segmentation models.

that the paradigm is not tied to a specific architectural dimensionality and remains effective whether the model processes full 3D volumes or individual 2D slices (as shown in Tab. 10). All experiments were performed on a single fold using implementations with the nnU-Net framework [28] using their "3d_fullres" and "2d" configurations.

G.5. BackSplit Performance on Transformer-based Architectures

We also evaluate BackSplit on transformer-based architectures, despite prior work indicating that such models often underperform compared to convolutional counterparts in medical image segmentation [29]. In particular, we employ a 3D SwinUNETR [22]. Even in this setting, BackSplit yields consistent improvements over standard training (as shown in Tab. 11), demonstrating that its benefits extend beyond convolutional models and remain effective for transformer-based architectures. However, as noted in the literature it does not perform as well as U-Nets.

H. Empirical Analysis of BackSplit with Increasing Auxiliary Classes

In this section, we analyze how performance metrics change when the number of auxiliary classes used in BackSplit training is increased linearly, and compare these results against standard training.

For this experiment, we use the AutoPET [19] dataset and incrementally add auxiliary classes from the AbdomenAtlas1.0Mini [49] dataset (with the left and right kidneys added simultaneously as two separate classes). This

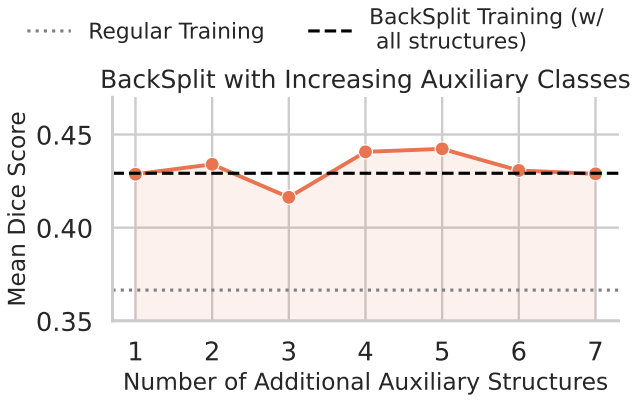

Figure 4. Adding even a single auxiliary structure yields an immediate improvement over regular training, and although incorporating multiple structures introduces mixed behavior with both gains and small drops, performance remains consistently well above the regular-training baseline. This experiment is conducted on the AutoPET dataset by incrementally adding auxiliary classes from AbdomenAtlas1.0Mini using a U-Net backbone.

yields seven distinct models, in addition to the regular training model and the BackSplit model containing all auxiliary structures, as shown in the main paper. All models were trained for a single fold using the U-Net [50] architecture within the nnU-Net framework [28].

We observe that the effect of BackSplit is almost immediate: even adding a single auxiliary structure yields a notable performance gain (as shown in Fig. 4). This is intuitive, as the inclusion of even one anatomical structure provides a meaningful contextual anchor for identifying lesions. Although adding multiple structures shows a mix of positive and negative trends whose underlying causes remain unclear, BackSplit consistently outperforms regular training across all settings. As noted in Sec. 5, identifying which support structures are most beneficial remains an important direction for future work.

I. BackSplit vs. Virtual Classes

We also conduct a brief comparison against the virtual classes paradigm [5], which has been explored in classification tasks by adding an additional class only at the softmax layer during training. This approach is hypothesized to encourage more discriminative feature embeddings and thereby improve performance. To emulate this idea in our setting, we train a U-Net with an added empty class, effectively mimicking the virtual class formulation.

We observe a similar effect in which the virtual class formulation yields a modest increase in performance metrics; however, these gains remain substantially lower than those achieved with BackSplit, as shown in Tab. 12. These results suggest that the virtual class approach primarily encourages

Method	KiTS23 Dice ↑	KiTS23 HD-95 ↓	KiTS23 NSD ↑
Regular Training	0.2033	425.3273	0.1906
Training with Virtual Class	0.2469	414.6015	0.2467
BackSplit	0.5297	249.5371	0.6703

Table 12. Single-fold evaluation on KiTS23 (Target = Cyst) using a U-Net backbone. BackSplit substantially outperforms both regular training and the Virtual Class paradigm across all metrics.

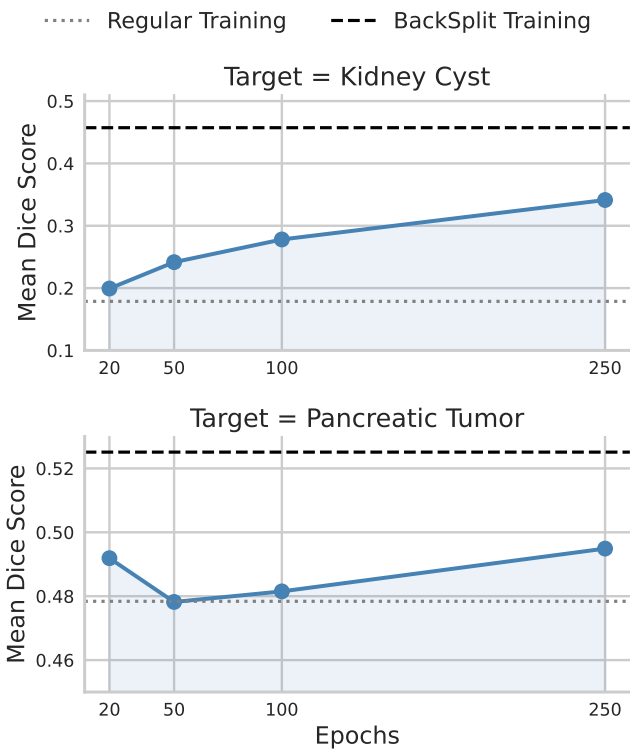

Figure 5. **(Top)** Effect of fine-tuning with BackSplit on a pretrained binary model for kidney cyst segmentation (KiTS23). Mean Dice steadily improves with training epochs, approaching full BackSplit performance. **(Bottom)** Similar trend observed for pancreatic tumor segmentation (PANTHER-MR), where fine-tuning progressively narrows the gap between regular and full BackSplit.

more discriminative features compared to standard training, whereas BackSplit provides improved predictions through reduced variance and richer anatomical context.

From a practical implementation standpoint, we again use U-Net [50] within the nnU-Net framework [28] to demonstrate this effect on the KiTS23 [25] dataset for a single fold. The virtual class is implemented by adding an additional prediction channel without providing corresponding labels in the dataset, and performance is evaluated solely on the cyst label.

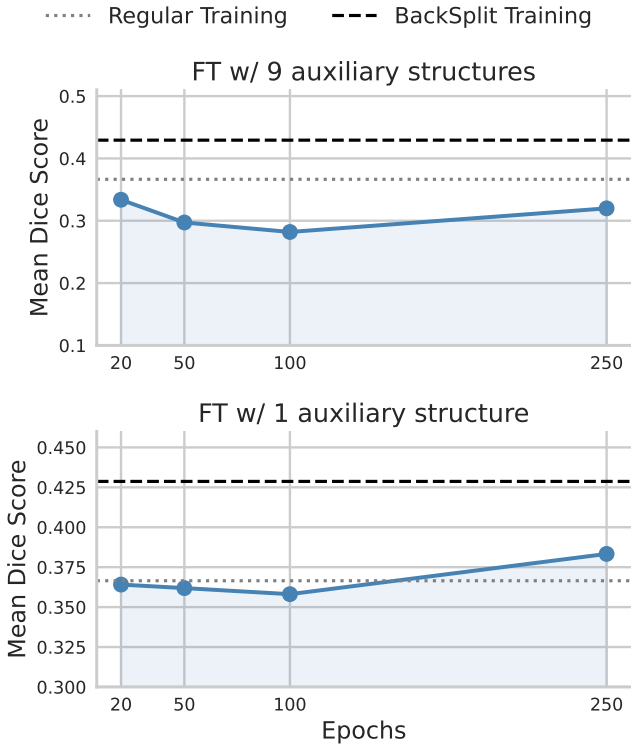

Figure 6. **(Top)** Fine-tuning on AutoPET with 9 auxiliary support structures. When many auxiliary labels are present, the model tends to learn the "easier" auxiliary classes first, delaying progress on the harder lesion target. Under a limited training budget (250 epochs), the model does not sufficiently reach the target class and fails to surpass the regular-training baseline. **(Bottom)** Fine-tuning with only 1 auxiliary structure avoids this and mirrors the behavior seen in KiTS23 and PANTHER, where performance steadily improves with training epochs.

J. Extended Analysis of BackSplit under Fine-Tuning

We further analyze the behavior of BackSplit in the fine-tuning setting, providing additional guidance for readers on how to apply the paradigm effectively for improved performance.

BackSplit is most effective in settings with a limited number of support structures. We demonstrate this empirically through three experiments on the KiTS23 [25], PANTHER [3], and AutoPET [19] datasets (the latter under two configurations). As shown in the main paper, KiTS23 exhibits clear improvements (Fig. 3 Left and Fig. 5 Top), and PANTHER displays a similar trend (as shown in Fig. 5 Bottom). However, when fine-tuning AutoPET with BackSplit, the model struggles to surpass the baseline performance (as shown in Fig. 6 Top).

We attribute this weaker performance to the experimental setup used for AutoPET, where BackSplit incorporated a large number of support structures (nine in total). In such cases, the model tends to learn "easier" auxiliary labels first,

delaying progress on the harder target label. With a training budget of only 250 epochs, the model does not sufficiently progress to learning the target class, which leads to the observed underperformance.

To mitigate this effect, we fine-tune the model using only a subset of support structures rather than all nine throughout training. Under this setting, AutoPET follows the same trend observed for KiTS23 and PANTHER (as shown in Fig. 6 Bottom). This provides a practical way to manage training dynamics when dealing with a large number of auxiliary structures. Here, we just add one additional support structure.

All experiments in this section are conducted using U-Net [50] via the nnU-Net [28] framework. For Fig. 5, we report full 5-fold cross-validation results. In contrast, Fig. 6 presents results from a single fold, as AutoPET is a substantially larger dataset and requires significantly more computational resources. This experiment is included primarily to provide empirical support for the claims made in this section. Additionally, following standard practice, we fine-tune the models using a lower initial learning rate of 1×10^{-3}, compared to the typical 1×10^{-2} used in nnU-Net training.

K. Qualitative Performance

In this section, we present qualitative comparisons illustrating the improvements achieved by BackSplit. In Fig. 7, we compare predictions from standard training and BackSplit across multiple architectures, highlighting clearer lesion boundaries under BackSplit. In Fig. 8, we show qualitative results using auxiliary labels derived from interactive segmentation. Despite the noisier supervision, BackSplit continues to produce stable and accurate lesion segmentations, demonstrating its robustness even under imperfect auxiliary annotations.

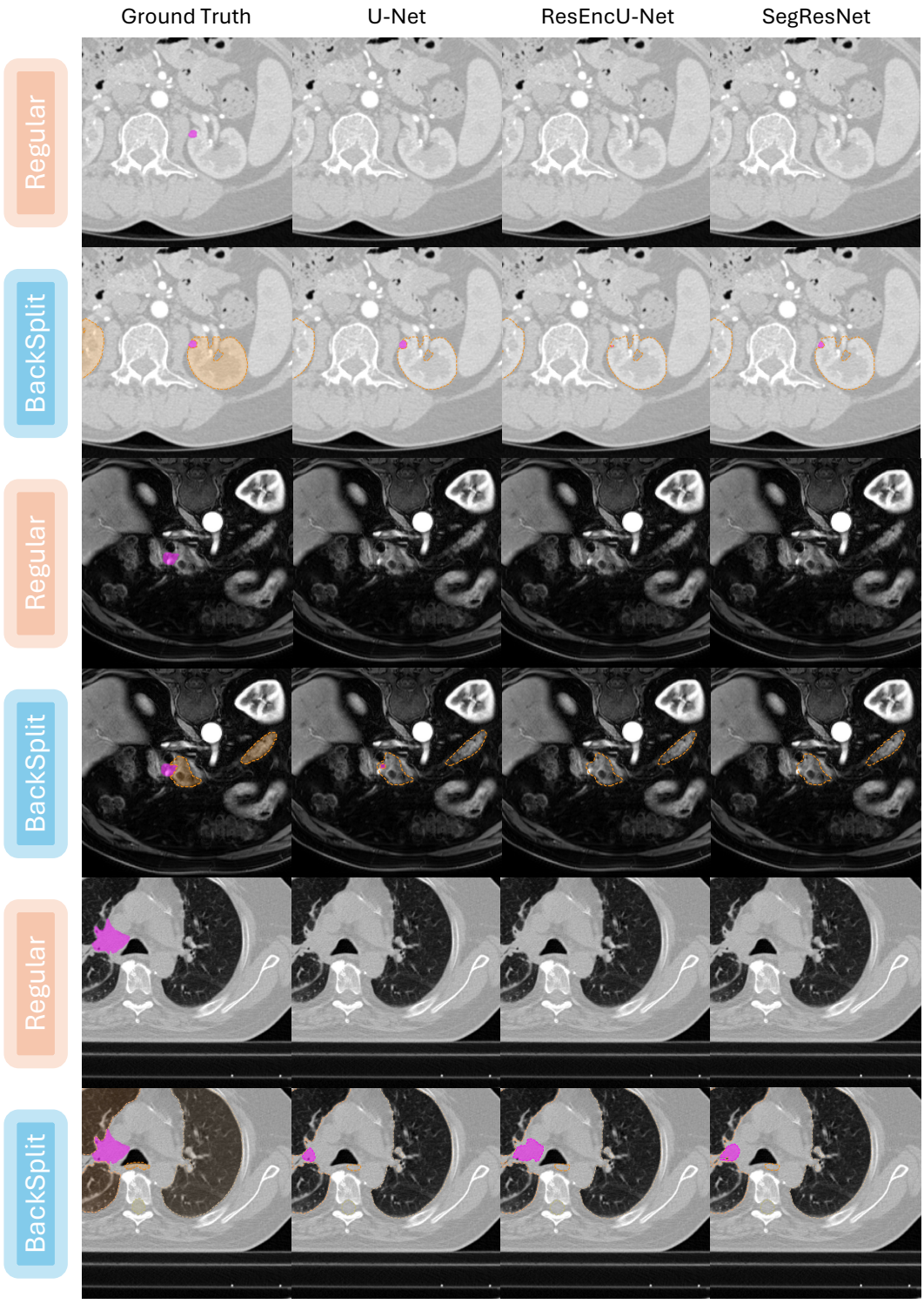

Figure 7. Qualitative comparison on three datasets: KiTS23 (top block), PANTHER-MR (middle block), and NSCLC-Radiomics (bottom block)—across three architectures (U-Net, ResEncU-Net, SegResNet). Each row pair shows regular training (top) and BackSplit (bottom). Regular training frequently misses small or low-contrast lesions, whereas BackSplit, with auxiliary anatomical structures, produces more complete and precise lesion masks. Lesions are shown in pink, while auxiliary support structures predicted under BackSplit are shown in yellow/orange hues, highlighting the additional contextual cues that lead to improved segmentation quality.

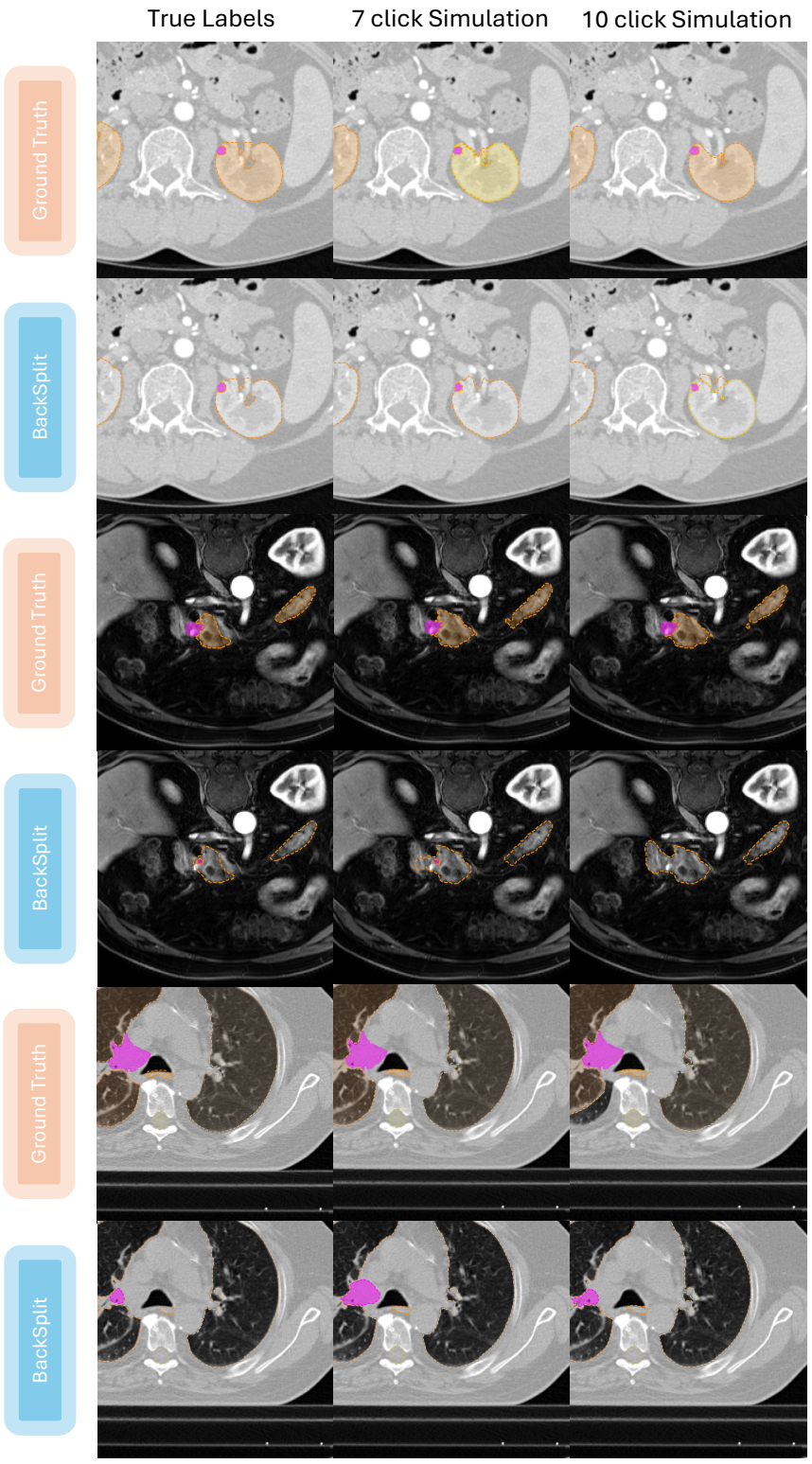

Figure 8. Qualitative comparison of BackSplit using true auxiliary labels versus noisy auxiliary structures generated by nnInteractive with 7 and 10 positive-click simulations. Rows correspond to three datasets: KiTS23, PANTHER-MR, and NSCLC-Radiomics—and each row pair shows Ground Truth auxiliary labels (top) and BackSplit predictions (bottom). U-Net backbone is used for all experiments. Lesions are shown in pink, while auxiliary support structures are shown in yellow/orange hues. Despite substantial noise in the interactively generated auxiliary masks, BackSplit maintains strong lesion segmentation quality, demonstrating robustness to imperfect supervision.