SACP: Spatially-Aware Conformal Prediction in Uncertainty Quantification of Medical Image Segmentation

Jacqueline Isabel Bereska[1,2] J.I.BERESKA@AMSTERDAMUMC.NL
[1] *Department of Radiology and Nuclear Medicine, Department of Biomedical Engineering and Physics, and Department of Surgery, Amsterdam UMC, University of Amsterdam, Amsterdam, Netherlands*
[2] *Cancer Center Amsterdam, Amsterdam, Netherlands*

Hamed Karimi[3] HAMED.KARIMI@TORONTOMU.CA
[3] *Department of Electrical, Computer, and Biomedical Engineering, Toronto Metropolitan University, Toronto, Ontario, Canada*

Reza Samavi[3,4] SAMAVI@TORONTOMU.CA
[4] *Vector Institute, Toronto, Ontario, Canada*

Editors: Under Review for MIDL 2025

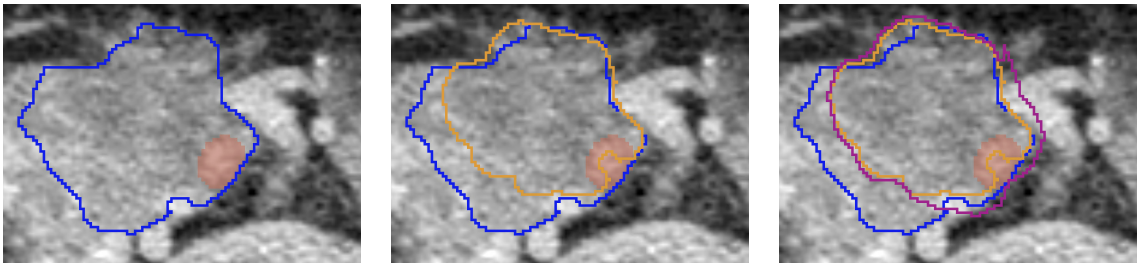
Abstract

Conformal Prediction provides statistical coverage guarantees for uncertainty quantification but fails to account for spatially varying importance of predictive uncertainty in medical image segmentation. This paper introduces a spatially-aware conformal prediction framework that enhances uncertainty quantification by incorporating spatial context near critical anatomical interfaces such as a vessel or critical organ. Our framework consists of three key components: (1) a base nonconformity score derived from segmentation model probabilities, (2) a calibration mechanism that applies structure-specific importance weights based on spatial proximity, and (3) a prediction set construction method that preserves mathematical coverage guarantees while providing targeted uncertainty quantification in critical regions. The calibration mechanism employs a distance-weighted scoring function that exponentially decays with distance from key interfaces, allowing for structure-specific importance factors and adaptive uncertainty estimation. We develop pooled and domain-specific calibration strategies to handle multi-center variability, enabling robust performance across different imaging protocols and populations. We validate our approach on tumor segmentation in pancreatic adenocarcinoma imaging from two medical centers. Results demonstrate that our method achieves the desired coverage levels while generating prediction sets that adaptively expand near critical interfaces.

Keywords: Uncertainty Quantification, Conformal Prediction, Medical Imaging, Image Segmentation, Deep Learning.

1. Introduction

Distribution-free uncertainty quantification in computer vision has seen significant advances through Conformal Prediction (CP), which transforms an algorithm's predictions into prediction sets with robust finite-sample validity (Fontana et al., 2023). While CP has shown promising results in classification and regression tasks (Vazquez and Facelli, 2022; Lu et al., 2022), its application to image segmentation presents several challenges, particularly in scenarios where prediction accuracy requirements vary across the output space (Zhou et al., 2024). Standard CP approaches employ uniform nonconformity scores across prediction regions, implicitly assuming homogeneous uncertainty throughout the prediction space.



(a) (b) (c)

Figure 1: Illustration of spatially-aware conformal prediction for PDAC tumor segmentation: **(a)** the base segmentation of the tumor (blue) adjacent to a major vessel (red), where precise delineation of the tumor-vessel interface is crucial for surgical planning; **(b)** the base segmentation of the tumor (blue) and the standard conformal prediction set (orange) **(c)** a comparison between standard CP sets (orange) and SACP sets (purple), where our approach adapts the prediction bounds based on proximity to critical structures.

However, this assumption breaks down in segmentation tasks where different spatial regions demand distinct levels of certainty measurements. For instance, boundary regions of a segment often require substantially different uncertainty characterization than interior regions. This limitation becomes particularly crucial when medical imaging is used for surgery planning and the boundary regions of a canonical object (e.g., a tumor) and its closeness to some critical masses (e.g., a vessel) requires varying confidence requirements as described in the following clinical setting.

In pancreatic ductal adenocarcinoma (PDAC) diagnosis and treatment planning, accurate tumor segmentation near critical vascular structures can mean the difference between an operable and inoperable assessment. When a tumor interfaces with major blood vessels, surgeons require millimeter-scale accuracy (with high certainty) in boundary delineation to determine resectability and surgical plans. Conversely, the precise boundary definition of the tumor's interior regions, while important, permits more flexibility in uncertainty quantification. As shown in Figure 1(a), the initial segmentation shows a tumor-vessel interface in a CT scan, where a PDAC tumor boundary (blue) is adjacent to a major vessel (red). To consider uncertainty in the prediction, one may apply CP (e.g., with error rate $\alpha = 10\%$) to ensure at least 90% certainty around the delineated boundaries of the tumor (yellow) as shown in Figure 1(b). CP generates homogeneous uncertainty set across all voxels, while for the surgical planner it's crucial to know, with higher precision (e.g., with error rate only $\alpha = 5\%$) around the vessels, due to its critical impact on surgery, and error rate $\alpha = 10\%$ elsewhere. This is the challenge the classical CP lacks addressing.

In this paper, we propose *Spatially-Aware Conformal Prediction* (SACP), to address the challenge of incorporating varying spatial importance of voxels into conformal prediction set while preserving its theoretical guarantees. We introduce locally adaptive nonconformity scores that explicitly account for distance to critical structures, enabling CP to generate prediction sets with higher precision (thus lower uncertainty) near critical boundaries while remaining CP-level uncertain elsewhere. When SACP is applied to our motivating example, as shown in Figure 1(c), the prediction set (shown in purple) expands only around the vessel

(the critical interface). The set now provides more conservative uncertainty estimates near vessel interfaces while maintaining computational efficiency in less critical regions. This varying conservativeness directly impacts surgical decision-making and may lead the surgeon to seek further and more accurate diagnostic imaging for the region. Our framework consists of three key components: (1) a base nonconformity score derived from segmentation model probabilities, (2) a calibration mechanism that applies structure-specific importance weights based on spatial proximity to critical masses, and (3) a prediction set construction method that maintains original CP coverage guarantees.

Related Work. CP has emerged as a promising framework for distribution-free uncertainty quantification (Fontana et al., 2023; Karimi and Samavi, 2023; Zhou et al., 2024), particularly in medical imaging (Angelopoulos et al., 2020; Karimi and Samavi, 2024). Recent work has extended CP to semantic segmentation (Brunekreef et al., 2024), lightweight post-hoc uncertainty quantification (Mossina et al., 2024), and performance range prediction (Wundram et al., 2024). Researchers also considered incorporating spatial aspects to CP in domains like hyperspectral imaging (Liu et al., 2024). However, the challenge of adapting uncertainty estimates based on spatial importance remains open.

Contributions. First, we develop a novel spatially-aware CP method, which is sensitive to the distance to critical structures, allowing a rigorous presentation of nonuniform uncertainty across voxels of a segment in an image. Our approach is generalizable to other fields of image segmentation where nonuniform uncertainty quantification is needed –e.g., safety in robotics and quality control in manufacturing, where a false negative can lead to accidents. Second, we theoretically prove that SACP maintains the original CP coverage, set size and adaptability characteristics. Third, we experimentally demonstrate that our approach is efficient and clinically insightful for a real-world clinical case. The code for experimental evaluation is publicly available at https://github.com/tailabTMU/SACP.

2. Spatially-Aware Conformal Prediction

Conformal Prediction. Conformal prediction is a statistical framework that produces prediction intervals for any underlying pretrained model with a guarantee on the prediction's reliability (Vovk et al., 2005). For a given significance level $\alpha \in \mathbb{R}^{(0,1)}$, CP ensures that for a calibration dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ and a new test point (x_{n+1}, y_{n+1}) drawn from the same distribution, $\mathbb{P}\big(y_{n+1} \in \mathcal{C}(x_{n+1})\big) \geq 1 - \alpha$. CP defines a nonconformity score $S : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ that quantifies how well x_{n+1} *conforms* to the calibration dataset. The prediction set is then computed based on the empirical quantiles of these nonconformity scores: For a chosen confidence level $1-\alpha$, the prediction set $\mathcal{C}(x_{n+1})$ is defined as $\mathcal{C}(x_{n+1}) = \{\hat{y} \in \mathcal{Y} : S(x_{n+1}, \hat{y}) \leq \tau_\alpha\}$, where τ_α is the $(1-\alpha)$-quantile of the nonconformity scores. This guarantee is unconditional and holds for any model and any distribution as long as the underlying exchangeability assumption is satisfied. This assumption implies that for any permutation σ of $1, 2, \ldots, n$, permutations of the dataset have the same joint distribution as $\mathbb{P}\big((x_1, y_1), \ldots, (x_n, y_n)\big) = \mathbb{P}\big((x_{\sigma(1)}, y_{\sigma(1)}), \ldots, (x_{\sigma(n)}, y_{\sigma(n)})\big)$. We refer to (Angelopoulos and Bates, 2021) for a more in-depth introduction to CP.

2.1. Problem Setup

Let $\mathcal{X} \subset \mathbb{R}^3$ represent a discretized volumetric image obtained from axial slices, where $\mathcal{X}$ is subdivided into a finite, structured grid of cuboidal units called voxels. We define each

voxel $x \in \mathcal{X}$ by its indices (x_a, x_c, x_s) along the axial, coronal, and sagittal axes, respectively. Considering a set of K possible labels $\mathcal{Y} = \{1, ..., K\}$, we represent the true label $y \in \mathcal{Y}$ as the indicator of the organ that the voxel x belongs to, and the baseline segmentation model f_Θ obtains predictive probabilities $p(\hat{y}|x)$ associated with each label $\hat{y} \in \mathcal{Y}$.

Definition 1 (Canonical Object) *Label $l \in \mathcal{Y}$ denotes a canonical object, if l represents a primary structure of interest for the downstream task.*

Definition 2 (Critical Masses) *$\mathcal{M}$ is a set of critical masses in the volume, if the proximity of any $m \in \mathcal{M}$ to the canonical object necessitates conservative decision-making for the downstream task.*

For the clinical settings described in Section 1, a tumor is a *canonical object* for the downstream task of surgery planning for the removal of that tumor and the set of vessels in the volume are *critical masses*, as when a tumor has a vessel in its proximity, the surgeon needs a prediction set with lower uncertainty (more conservative prediction).

2.2. Spatially-Aware Non-Conformity Score

To apply CP on voxel-wise tasks (e.g., tumor segmentation for surgery planning), we need to address two challenges: (1) CP uses a single threshold α across all classes; thus the prediction set for rare classes will be overly conservative (i.e. almost all labels are included in the set); this is particularly crucial in tumor segmentation task as tumors are small structures relative to the total CT image volume. (2) The prediction set generated by CP is invariant to the voxels in the volume, while we expect, the prediction set to be more conservative when voxels of the canonical object (e.g., tumor) is closer to one or more critical masses (e.g., vessels).

To address the first challenge, we adopt Class-Conditional Conformal Prediction (CCCP), where CP is refined to use various quantile thresholds across different classes (Angelopoulos et al., 2023). We compute a distinct threshold $\tau_\alpha^{\hat{y}}$ for each label $\hat{y} \in \mathcal{Y}$ independently as the $(1-\alpha)$-quantile of the nonconformity scores as,

$$\tau_\alpha^{\hat{y}} = \text{Quantile}_{1-\alpha}\Big(\{S_{\text{base}}(x_i, y_i) : y_i = \hat{y}\}_{i=1}^{n}\Big) \ . \tag{1}$$

For a new test data point (x_{n+1}, y_{n+1}), the prediction set $\mathcal{C}(x_{n+1})$ is constructed as,

$$\mathcal{C}(x_{n+1}) = \left\{\hat{y} \in \mathcal{Y} : S_{\text{base}}(x_{n+1}, \hat{y}) \leq \tau_\alpha^{\hat{y}}\right\} \ . \tag{2}$$

To address the second challenge, we define a new score function, S_{SACP}, that augments the original CP nonconformity score function, S_{base}, with a parameterized weight $w_v \in \mathbb{R}^{[0.5,1)}$, as a multiplicative factor denoted as,

$$S_{\text{SACP}}(x|\hat{y} = l) = w_v(x, l) \cdot S_{\text{base}}(x|\hat{y} = l) \ , \tag{3}$$

where $v \in \mathcal{M}$ is the nearest critical mass to the voxel x, and $l \in \mathcal{Y}$ is the canonical object. Our intention is to make the impact of the weight irrelevant ($w_v \approx 1$) when voxels of the volume are far from both critical masses and canonical object, therefore generating a prediction set as conservative as the original base function and maximizing the impact of the weight ($w_v = 0.5$) when voxels are very close to critical masses and the canonical object.

The weight has to be also impacted by our confidence in segmenting the canonical object (tumor) as well as the relevancy of the different critical masses, as we may have more than one critical mass, each with a different relevancy factor for the downstream task. Formally, we have four parameters for our weight function:

1. δ_m: The Euclidean distance of each voxel x to the critical mass $m \in \mathcal{M}$.
2. ϕ_l: The Euclidean distance of each voxel x to the canonical object $l \in \mathcal{Y}$.
3. $\mathcal{I}(l)$: The information-theoretical surprisal or unexpectedness of observing the canonical object l with probability $p(\hat{y} = l|x)$ that inversely accounts for our confidence on correct segmentation of the canonical object, where $\mathcal{I}(l) \stackrel{\text{def}}{=} -\log p(\hat{y} = l|x)$.
4. $\gamma_m \in \mathbb{R}^{(0,1]}$: A hyperparameter capturing the relevancy and criticality of each critical mass $m \in \mathcal{M}$.

Putting all together, our spatially-aware weight function for each voxel x is defined as,

$$w_v(x, l) = \sigma\left(\overbrace{\frac{1}{\gamma_v}\Big(\phi_l + \delta_v \mathcal{I}(l)\Big)}^{\tilde{w}_v}\right) \qquad s.t. \qquad v = \arg\min_{m \in \mathcal{M}} \delta_m \ , \tag{4}$$

where $\tilde{w}_v : \mathcal{X} \times \mathcal{Y} \times \mathcal{M} \to \mathbb{R}^+$ is a function that represents the raw weight value and the constraint ensures we consider the nearest critical mass. $\tilde{w}_v$ is then normalized to $w_v(x, l) \in \mathbb{R}^{[0.5,1)}$ using the sigmoid function $\sigma(.)$ (further details in Appendix B).

Since $w_v \in \mathbb{R}^{[0.5,1)}$, when δ_v is small and $p(\hat{y} = l|x)$ is high (i.e. $\mathcal{I}(l)$ is low), w_v yields towards its lower bound, reducing the nonconformity scores of label l, making it more likely to be included in the prediction set (more conservative prediction set). This aligns with the desire to treat voxels around the critical masses and the canonical object more conservatively and expand the prediction set for those areas. Conversely, when δ_v is getting larger and $p(\hat{y} = l|x)$ smaller ($\mathcal{I}(l)$ is higher), w_v moves closer to its upper bound of 1, eliminating the impact of the weight and making it less likely for distant regions to be included in the prediction set. ϕ_l also behaves similarly. Lower ϕ_l (the voxel being closer to the canonical object) yields lower weight, making the set more conservative and vice versa. The relevancy hyperparameter γ_v accepts values between zero (strictly greater than zero), for the least critical mass to 1, for the most critical mass. γ_v has a diminishing impact on $\tilde{w}_v$ and in turn to w_v. For example, if the user sets the relevancy for a critical mass low (e.g., $\gamma_v = 0.5$), the weight computed based on all other factors will be doubled, w_v gets closer to 1 and diminishes its impact on original nonconformity score. In contrast, when relevancy increases, w_v gets closer to its lower bound of 0.5 and increases the prediction set size. Note the value of γ_v needs to be fine-tuned depending on the context of the application.

Theorem 3 (SACP Conservativeness) *If $S_{base}(x, \hat{y})$ denote the base nonconformity score for a voxel $x \in \mathcal{X}$ with the predictive label $\hat{y}$, and $\tau_\alpha^{\hat{y}}$ the $(1-\alpha)$-quantile of S_{base} with the error rate α, then for the canonical object $l \in \mathcal{Y}$, the prediction set produced by the nonconformity function $S_{SACP}(x|\hat{y} = l)$ is at least as conservative as sets produced by $S_{base}(x|\hat{y} = l)$.*

See Appendix C for the proof of Theorem 3 and Appendix D for conservativeness in CP.

Corollary 4 *For an unseen data point x_{new} and $S_{SACP}(x_{new}, \hat{y})$ (Equation 3), the inclusion of canonical object label l in the prediction set $\mathcal{C}(x_{new})$ with error rate α depends on spatial properties near high-risk regions that satisfies:*

$$l \in \mathcal{C}(x_{new}) \iff S_{SACP}(x_{new}|\hat{y} = l) \leq \tau_{\alpha}^{l} \ , \tag{5}$$

where τ_{α}^{l} is class-specific threshold as $(1-\alpha)$-quantile of nonconformity scores S_{base}.

The spatial relationship that Equations (3) and (4) promote is particularly valuable in the clinical setting described in Section 1 for determining resectability. This assessment follows region-specific clinical guidelines - for instance, the NCCN guidelines in the United States and DPCG guidelines in the Netherlands - each defining different criteria for vessel involvement and tumor contact thresholds that determine resectability. Consequently, when deploying pretrained tumor segmentation models across different regions, the calibration of vessel-specific importance factors also needs to be adjusted to align with local clinical guidelines and their specific vessel prioritization. The detailed process of generating a prediction set using SACP is described as Algorithm 1 in Appendix A.

3. Experimental Evaluations

3.1. Data Preparation

Datasets. This retrospective study analyzes 30 contrast-enhanced computed tomography (CT) scans from the PANORAMA Challenge (Alves et al., 2024). The dataset includes portal venous phase CT scans from five European centers.

Ground Truth Segmentations. Ground truth segmentations were established through expert radiologist annotations for PDAC tumors. For surrounding anatomical structures (pancreatic parenchyma, duodenum, liver, gallbladder, kidneys, adrenal glands, and spleen), we employed TotalSegmentator (Wasserthal et al., 2023). These complementary segmentations are integrated using a hierarchical fusion approach that prioritizes radiologists' tumor delineations over automated organ segmentations.

AI-Generated Segmentations. We employ two deep learning models: a primary model for PDAC tumors and surrounding anatomical structures (Bereska et al., 2024), and a vessel-specific model focused on structures critical for PDAC resectability assessment. The latter segments five key vessels: the celiac trunk (CeTr), hepatic artery (HA), portal vein (PV), superior mesenteric artery (SMA), and superior mesenteric vein (SMV). For subsequent nonconformity score computation, we preserve (1) the pre-softmax probability maps for all 11 classes (10 anatomical labels plus background) from the primary segmentation model, and (2) distance maps computed from the vessel-specific model, measuring the distance from each voxel to each of the five resectability-determining vessels. To ensure robustness to outliers, both distance and probability values are clipped at their respective 95th percentiles before being used in the nonconformity score computation. The implementation details are further described in Appendix E.

Cropping. To optimize computational efficiency, we use an adaptive 3D bounding box cropping strategy. We identify the minimal volumetric boundary encompassing all voxels with specified target labels (gallbladder, pancreas, duodenum, and tumor) and apply this crop consistently across all corresponding image modalities and their derivatives.

3.2. Evaluation Metrics

We evaluate our spatially-aware framework through metrics assessing both predictive performance and anatomical sensitivity. For each voxel x with confidence level $1-\alpha$, we compute the empirical coverage rate:

$$\text{cov}(\mathcal{C}, \alpha) = \mathbb{E}\big[y \in \mathcal{C}(x)\big] = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \mathbb{1}_{\{y \in \mathcal{C}(x)\}} \ , \tag{6}$$

where $\mathbb{1}$ is the indicator function over non-empty prediction sets, y is the true label, and $\mathcal{C}(x)$ is the prediction set. Coverage is assessed separately for vessel-adjacent ($\delta_v \leq 5$mm) and non-critical regions ($\delta_v > 5$mm). We also define the Relative Width Ratio (RWR) to quantify adaptation of prediction set sizes based on anatomical criticality as,

$$\rho(r) = \frac{\mu(\mathcal{C}|\delta_v \leq r)}{\mu(\mathcal{C}|\delta_v > r)} \ , \tag{7}$$

where $\mu(\mathcal{C}|\delta_v) = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \big|\mathcal{C}(x)\big|$ represents the average set size to evaluate prediction set efficiency and $\mathcal{X}$ represents voxels at distance δ_v from the nearest vessel v.

3.3. Experimental Setup

We use 10 cases for calibration to determine class-specific nonconformity score thresholds $\tau_\alpha^{\hat{y}}$ for each label $\hat{y} \in \mathcal{Y}$ and evaluate on 20 held-out cases. Statistical comparisons use paired t-tests with Benjamini-Hochberg correction ($p < 0.05$). For the vessel-specific analysis, we incorporate anatomical context through a weighted scoring mechanism. Critical vessels are assigned differential weights (γ) based on the NCCN resectability criteria for PDAC, with arterial vessels (CeTr, HA, SMA) receiving higher weights ($\gamma = 0.8$) compared to venous vessels (PV, SMV: $\gamma = 0.6$). This weighting scheme reflects their relative importance in determining resectability, as arterial involvement beyond 180° renders a tumor unresectable, while venous involvement may permit resection with reconstruction.

To achieve sharp transitions in uncertainty estimates near vessel boundaries, we amplify the sigmoid response using a gain factor ($\beta = 10$), creating more pronounced changes in uncertainty estimates as predictions approach critical vascular structures. This enhanced sigmoid sensitivity provides a clearer delineation of high-risk regions for surgical planning.

3.4. Experimental Results

Coverage Analysis. Our framework achieves strong coverage on the PANORAMA dataset ($n = 20$) with an overall coverage of 0.987 (mean per-case: 0.981 ± 0.005 SEM). The coverage significantly exceeds the target coverage of 0.95 (Wilcoxon signed-rank test, $p = 0.0007$).

Distance-Based Analysis. As shown in Table 1, prediction set size decreases with distance from vessels while maintaining high coverage. RWR ranges from 2.762 ± 0.150 SEM near vessels ($\leq$2mm) to 2.525 ± 0.036 SEM beyond 20mm, with coverage remaining consistently high across all distances (0.981-0.988). This decreasing RWR pattern suggests our method adapts to provide more precise predictions in regions farther from vessels, while maintaining wider prediction sets near critical vascular structures.

Vessel-specific analysis (Table 2) demonstrates robust performance across all major vessels, with excellent coverage in vessel-proximate regions. Notably, we achieve high coverage

Table 1: Coverage and RWR analysis across vessel proximity zones for CCCP and SACP.

	CCCP		SACP	
Distance	Coverage	RWR	Coverage	RWR
≤2mm	0.954 ± 0.027	2.887 ± 0.320	0.981 ± 0.008	2.762 ± 0.150
≤5mm	0.970 ± 0.016	2.702 ± 0.262	0.987 ± 0.004	2.684 ± 0.131
≤10mm	0.977 ± 0.016	2.611 ± 0.263	0.988 ± 0.004	2.621 ± 0.122
≤20mm	0.978 ± 0.003	2.574 ± 0.205	0.987 ± 0.001	2.592 ± 0.090
>20mm	0.982 ± 0.002	2.509 ± 0.078	0.988 ± 0.000	2.525 ± 0.036

Table 2: Vessel-specific coverage rates at different proximity zones for CCCP (C) and SACP (S).

	2mm		5mm		10mm		20mm		>20mm	
Vessel	C	S	C	S	C	S	C	S	C	S
CeTr	0.999	1.000	0.999	1.000	0.998	1.000	0.980	0.987	0.980	0.988
HA	0.959	0.980	0.973	0.986	0.987	0.994	0.981	0.989	0.980	0.987
SMA	0.925	0.975	0.967	0.989	0.982	0.994	0.973	0.985	0.984	0.989
PV	0.927	0.953	0.955	0.974	0.957	0.974	0.978	0.989	0.981	0.987
SMV	0.960	0.997	0.956	0.987	0.958	0.980	0.975	0.987	0.983	0.987

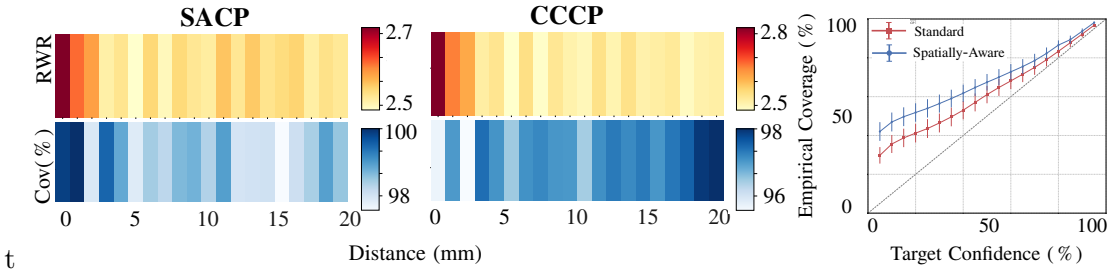


Figure 2: Left: RWR (top) and coverage (bottom) as a function of vessel distance for both datasets. Right: Comparison of empirical coverage at different confidence levels between our method (SACP) and standard Class-Conditional CP (CCCP).

in critical surgical planning zones, particularly near arteries. Visual examples of the prediction sets and their relationship to vessel proximity are provided in Appendix G.

Comparison with Standard Class-Conditional CP. Our spatially-aware approach demonstrates significantly improved coverage (0.981 ± 0.005 SEM vs 0.968 ± 0.038 SEM, paired t-test t=3.366, p=0.003). Near vessels ($\leq 2mm$), we achieve both superior coverage (0.981 vs. 0.954) and reduced RWR (2.762 vs. 2.887). Figure 2 shows consistently better coverage across target confidence levels, particularly in the $40 - 80\%$ range. Our method maintains high coverage while exhibiting decreasing RWR with distance from vessels, from 2.762 ± 0.150 SEM at $\leq 2mm$ to 2.525 ± 0.036 SEM beyond 20mm, demonstrating that our framework effectively adapts prediction sets based on proximity to critical anatomical structures. Results from additional experiments are provided in Appendix F.

4. Conclusion

We presented a spatially-aware conformal prediction framework that provides anatomically informed uncertainty quantification for medical image segmentation. Our method adapts prediction sets based on proximity to critical vascular structures while maintaining theoretical coverage guarantees. Validation on the PANORAMA dataset demonstrates robust performance, with strong coverage in vessel-adjacent regions and efficient adaptation of prediction set sizes based on anatomical criticality. This approach represents an advancement toward clinically reliable AI systems, particularly for applications where precise boundary delineation near critical structures impacts surgical planning and patient care. Future work will explore extending this framework to other anatomical contexts and clinical workflows.

Acknowledgments

R.S. is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery grant RGPIN-2016-06062, the HHS Deep Learning grant, and the TMU start-up grant. J.B. was supported by the Cancer Center Amsterdam's Young Talents Travel Grant for a research visit to Toronto Metropolitan University. We thank Leonard Bereska for his inspiration, proofreading, and contributions to visualization.

References

Nils Alves, Merlijn Schuurmans, Dominik Rutkowski, Derya Yakar, Ingfrid Haldorsen, Marjolein Liedenbaum, Anders Molven, Phillip Vendittelli, Geert Litjens, Jurgen Hermans, and Henkjan Huisman. The PANORAMA study protocol: Pancreatic cancer diagnosis - radiologists meet AI. *Zenodo*, 2024. doi: 10.5281/zenodo.10599559.

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.

Jacqueline I. Bereska, Selina Palic, Leonard F. Bereska, Efstratios Gavves, C. Yung Nio, Marnix P.M. Kop, Femke Struik, Freek Daams, Martijn A. van Dam, Tom Dijkhuis, Marc G. Besselink, Henk A. Marquering, Jaap Stoker, and Inez M. Verpalen. Refining the classroom: The self-supervised professor model for improved segmentation of locally advanced pancreatic ductal adenocarcinoma. 2024. Submitted.

Joren Brunekreef, Eric Marcus, Ray Sheombarsing, Jan-Jakob Sonke, and Jonas Teuwen. Kandinsky conformal prediction: Efficient calibration of image segmentation algorithms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4135–4143, 2024.

Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean-Christophe Fillion-Robin, Sonia Pujol, Christian Bauer, Dominique Jennings, Fiona Fennessy, Milan Sonka, et al. 3d slicer as an image computing platform for the quantitative imaging network. *Magnetic resonance imaging*, 30(9):1323–1341, 2012.

Matteo Fontana, Gianluca Zeni, and Simone Vantini. Conformal prediction: a unified review of theory and new challenges. *Bernoulli*, 29(1):1–23, 2023.

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.

QP Janssen, JL van Dam, BA Bonsing, H Bos, KP Bosscha, PPLO Coene, CHJ van Eijck, IHJT de Hingh, TM Karsten, MB van der Kolk, et al. Total neoadjuvant folfirinox versus neoadjuvant gemcitabine-based chemoradiotherapy and adjuvant gemcitabine for resectable and borderline resectable pancreatic cancer (preopanc-2 trial): study protocol for a nationwide multicenter randomized controlled trial. *BMC cancer*, 21:1–8, 2021.

Hamed Karimi and Reza Samavi. Quantifying deep learning model uncertainty in conformal prediction. In *Proceedings of the AAAI Symposium Series*, volume 1, pages 142–148, 2023.

Hamed Karimi and Reza Samavi. Evidential uncertainty sets in deep classifiers using conformal prediction. *In Proceedings of the Thirteenth Symposium on Conformal and Probabilistic Prediction with Applications*, 230:466–489, 2024.

Kangdao Liu, Tianhao Sun, Hao Zeng, Yongshan Zhang, Chi-Man Pun, and Chi-Man Vong. Spatial-aware conformal prediction for trustworthy hyperspectral image classification. *arXiv preprint arXiv:2409.01236*, 2024.

Charles Lu, Andréanne Lemay, Ken Chang, Katharina Höbel, and Jayashree Kalpathy-Cramer. Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36-11, pages 12008–12016, 2022.

Luca Mossina, Joseba Dalmau, and Léo Andéol. Conformal semantic image segmentation: Post-hoc quantification of predictive uncertainty. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3574–3584, 2024.

Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.

TF Stoop, LW Seelen, F van't Land, DJ Lips, IH de Hingh, S Festen, JH Wijsman, K Bosscha, E van der Harst, F Wit, et al. Surgical outcome after resection of locally advanced pancreatic cancer following systemic treatment: Nationwide retrospective cohort. *HPB*, 24:S310–S311, 2022.

Geertjan Van Tienhoven, Eva Versteijne, Mustafa Suker, Karin BC Groothuis, Olivier R Busch, Bert A Bonsing, Ignace HJT de Hingh, Sebastiaan Festen, Gijs A Patijn, Judith de Vos-Geelen, et al. Preoperative chemoradiotherapy versus immediate surgery for resectable and borderline resectable pancreatic cancer (preopanc-1): A randomized, controlled, multicenter phase iii trial., 2018.

Janette Vazquez and Julio C Facelli. Conformal prediction in clinical medical sciences. *Journal of Healthcare Informatics Research*, 6(3):241–252, 2022.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

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), 2023.

Anna M Wundram, Paul Fischer, Michael Mühlebach, Lisa M Koch, and Christian F Baumgartner. Conformal performance range prediction for segmentation output quality control. In *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pages 81–91. Springer, 2024.

Xiaofan Zhou, Baiting Chen, Yu Gui, and Lu Cheng. Conformal prediction: A data perspective. *arXiv preprint arXiv:2410.06494*, 2024.

Appendix A. Algorithmic Description of SACP

In Algorithm 1, we describe the step-wise procedure and the required computation regarding applying SACP to incorporate spatial context in 3D voxel-wise segmentation and enhance uncertainty quantification.

- In Step 1, a pretrained segmentation model f_Θ generates voxel-wise predictive probabilities in an input volume $\mathcal{X}$ using the softmax function.
- In Step 2, we apply class-conditional calibration to ensure a desired confidence rate of at least $1-\alpha$ for each class $\hat{y} \in \mathcal{Y}$ using S_{base} non-conformity scores of calibration set of voxels. For each class, the $(1-\alpha)$-quantile threshold $\tau_\alpha^{\hat{y}}$ is determined, setting the baseline for prediction set construction.
- In Step 3, we compute spatial properties as the Euclidean distances of each voxel $x \in \mathcal{X}$ to a set of critical masses $m \in \mathcal{M}$ denoted by δ_m and to the canonical object label $l \in \mathcal{Y}$ denoted by ϕ_l.
- In Step 4, we identify the nearest critical mass $v \in \mathcal{M}$ for each voxel, forming a spatial reference. Then, a normalized weight w_v is computed for each voxel x based on its proximity to the canonical object l and the nearest critical mass v, adjusted by a mass-specific relevance factor γ_v. This weight modulates the base non-conformity score S_{base} associated with the canonical object l to refine uncertainty estimation relative to spatial critical structures. Finally, the prediction set $\mathcal{C}(x)$ is constructed by including the canonical object label l in the set if and only if the adjusted score S_{SACP} remains below its respective (class-conditional) quantile threshold τ_α^l.

By integrating spatial information, SACP improves the reliability of conformal prediction in 3D segmentation, particularly in anatomically structured regions where spatial coherence is essential.

Appendix B. Further Details of SACP Parameters

We compute δ_m as the Euclidean distance from voxel x to any of the potential critical masses $m \in \mathcal{M}$ (e.g., major vessels) that is defined by the function $d : \mathcal{X} \times \mathcal{M} \to \mathbb{R}^+$ as,

$$\delta_m = d_{Euc}(x, m) = \min_{x' \in V_m} ||x - x'|| \ , \tag{8}$$

where $m \in \mathcal{M}$ is a critical mass containing a set of voxels $V_m \subset \mathcal{X}$.

We compute ϕ_l as the Euclidean distance from voxel x to the segmentation outcome of a pretrained model f_Θ that is defined by the function $\hat{d} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ as,

$$\phi_l = \hat{d}_{Euc}(x, l) = \min_{x' \in V_l} ||x - x'|| \ , \tag{9}$$

where $l \in \mathcal{Y}$ is the canonical object label and $V_l \subset \mathcal{X}$ contains a set of voxels that are segmented as label l such that:

$$V_l = \left\{ x' \in \mathcal{X} \mid \arg\max_{\hat{y} \in \mathcal{Y}} f_\Theta(x', \hat{y}) = l \right\} \ , \tag{10}$$

Algorithm 1: Spatially-Aware Conformal Prediction (SACP)

Input: 3D input volume $\mathcal{X}$: voxels x with true labels y; Set of all possible labels $\hat{y} \in \mathcal{Y}$; set of critical masses $\mathcal{M}$; canonical object label l; pretrained segmentation model f_Θ; desired error rate α; mass-specific relevance factors $\{\gamma_m\}_{m\in\mathcal{M}}$;

Output: $\mathcal{C}(x)$ as prediction set for each voxel;

// Step 1: Get model predictions

$\forall x \in \mathcal{X}, \hat{y} \in \mathcal{Y}: \; p(\hat{y}|x) \leftarrow \text{softmax}\big(f_\Theta(x, \hat{y})\big)$ // Get predictive probabilities

// Step 2: Class-conditional calibration on n voxels

for *each class* $\hat{y} \in \mathcal{Y}$ **do**

  $\tau_\alpha^{\hat{y}} \leftarrow \text{Quantile}_{1-\alpha}\Big(\big\{S_{\text{base}}(x_i, y_i): \; y_i = \hat{y}\big\}_{i=1}^{n}\Big)$

end

// Step 3: Compute spatial distances

for *each voxel* $x \in \mathcal{X}$ **do**

  $\forall m \in \mathcal{M}: \; \delta_m \leftarrow d_{Euc}(x, m)$ // Distance to the critical masses (Eq. 8)

  $V_l \leftarrow \{x' \in \mathcal{X} | \arg\max_{\{\hat{y}\in\mathcal{Y}\}} f_\Theta(x', \hat{y}) = l\}$ // Set of canonical object voxels

  $\phi_l \leftarrow \hat{d}_{Euc}(x, l)$ // Distance to the canonical object (Eq. 9 and 10)

end

// Step 4: Generate SACP prediction sets

for *each voxel* $x \in \mathcal{X}$ **do**

  $v = \arg\min_{\{m\in\mathcal{M}\}} \delta_m$ // Find the nearest critical mass

  $w_v(x, l) \leftarrow \sigma\Big(\frac{1}{\gamma_v}\big(\phi_l + \delta_v \mathcal{I}(l)\big)\Big)$ // Compute spatial weight (Eq. 4)

  $S_{\text{SACP}}(x|\hat{y} = l) \leftarrow w_v(x, l) \cdot S_{\text{base}}(x|\hat{y} = l)$ // Score for canonical object

  $l \in \mathcal{C}(x) \Leftrightarrow S_{\text{SACP}}(x|\hat{y} = l) \leq \tau_\alpha^l$ // Conservative inclusion of $\hat{y} = l$ (Eq. 3)

end

return $\mathcal{C}(x)$ *for all* $x \in \mathcal{X}$

in which $f_\Theta(x', \hat{y})$ is the outcome of the pretrained segmentation model associated with the label $\hat{y}$ when classifying the voxel x'.

We also compute the confidence of segmentation model defined as the predictive probability associated with the canonical object label l (e.g., a tumor) and denoted by $p(\hat{y} = l|x)$ for each voxel x. High confidence associated with the canonical object label indicates that the model is making reliable predictions that a voxel belongs to that label, which can be valuable in improving reliability in high-risk tasks. The weight function is formulated to represent lower values with higher probabilities (i.e., more confident predictions) and vice versa, emphasizing regions where the model is more confident while discounting less certain regions. To refine the weight computation, we use this segmentation confidence to calculate

the surprisal function $\mathcal{I}(l) \stackrel{\text{def}}{=} -\log p(\hat{y} = l|x)$. This surprisal quantifies the information content or unexpectedness of observing the canonical object l with probability $p(\hat{y} = l|x)$ and accounts for the model's inherent uncertainty during segmentation. By incorporating surprisal, the prediction sets dynamically adapt to the probabilistic confidence of the model.

Following class-conditional CP with the desired confidence level $1 - \alpha$ and according to Equation (1), we independently compute the class-specific quantile $\tau_\alpha^{\hat{y}}$ associated with the canonical object label $\hat{y} = l \in \mathcal{Y}$, based on the S_{base} scores of calibration data. Then, we use S_{SACP} during testing to include the canonical object label l in the voxels' prediction sets as proposed in Corollary 4.

Appendix C. Proof of Theorem 3

Proof Following class-conditional CP, $\tau_\alpha^{\hat{y}}$ denotes the $(1 - \alpha)$-quantile of S_{base} scores associated with calibration data with label $\hat{y}$. Then, for each voxel x, the condition for inclusion the canonical object label $\hat{y} = l$ in the prediction set $\mathcal{C}_{\text{base}}(x)$ generated by S_{base} scores is:

$$S_{\text{base}}(x|\hat{y} = l) \leq \tau_\alpha^l \ . \tag{11}$$

By the definition of in Equation (3), S_{SACP} is computed for each voxel x and the canonical object label $\hat{y} = l$ using the normalized weight w_v as,

$$S_{\text{SACP}}(x|\hat{y} = l) = w_v \cdot S_{\text{base}}(x|\hat{y} = l) \qquad s.t. \qquad w_v = \sigma(\tilde{w}_v) \ , \tag{12}$$

where $\tilde{w}_v \in \mathbb{R}^+$ is the raw weight value defined in Equation (4), and $\sigma(.)$ is the steep sigmoid function (with the gain factor β) defined as $\sigma(\tilde{w}_v) = \frac{1}{1+\exp(-\beta\tilde{w}_v)}$. For other labels $\hat{y} \neq l$, S_{base} is used to include the labels in the sets. As $\tilde{w}_v$ is positive and normalized to be less than 1, so $0.5 \leq w_v < 1$. Then, it follows that:

$$\forall x \in \mathcal{X} : \quad S_{\text{SACP}}(x|\hat{y} = l) < S_{\text{base}}(x|\hat{y} = l) \ . \tag{13}$$

Note that $\lim_{\tilde{w}_v \to +\infty} w_v = 1$, and consequently, $\lim_{\tilde{w}_v \to +\infty} S_{\text{SACP}} = S_{\text{base}}$. According to Equations (11) and (13), the above inequality implies the following condition to include l in the set:

$$S_{\text{SACP}}(x|\hat{y} = l) = w_v \cdot S_{\text{base}}(x|\hat{y} = l) \leq \tau_\alpha^l \ . \tag{14}$$

Therefore, any label $\hat{y} \neq l$ included in $\mathcal{C}_{\text{base}}(x)$ (i.e., $S_{\text{base}}(x|\hat{y}) \leq \tau_\alpha^{\hat{y}}$) is also included in the prediction set $\mathcal{C}_{\text{SACP}}(x)$ generated by SACP, and for the canonical object $\hat{y} = l$, $S_{\text{SACP}}(x|\hat{y} = l) < S_{\text{base}}(x|\hat{y} = l)$ holds. Formally, this means:

$$\mathcal{C}_{\text{base}}(x) \subseteq \mathcal{C}_{\text{SACP}}(x) \ . \tag{15}$$

■

Appendix D. Conservativeness in Conformal Prediction

Conformal prediction constructs set-valued predictions with a user-specified coverage guarantee, ensuring that the empirical coverage of the prediction sets is at least the nominal confidence level. Given a dataset $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$ and a new test point x_{n+1}, CP produces a prediction set $\mathcal{C}_{n,\alpha}(x_{n+1})$ such that

$$\mathbb{P}(y_{n+1} \in \mathcal{C}_{n,\alpha}(x_{n+1})) \geq 1 - \alpha \ . \tag{16}$$

This property, known as *conservativeness*, guarantees that the probability of the true label being included in the prediction set is at least $1 - \alpha$, often making CP slightly over-conservative due to the discrete nature of rank-based p-values in finite samples.

Conservativeness leads to both lower and upper bounds on the empirical coverage. The lower bound is given directly by the validity guarantee, ensuring Equation (16). However, the actual coverage can be higher than $1 - \alpha$ due to the discreteness of conformity scores, leading to an upper bound of the form

$$\mathbb{P}\big(y_{n+1} \in \mathcal{C}_{\alpha}(x_{n+1})\big) \leq 1 - \alpha + \frac{1}{n+1} \ . \tag{17}$$

This small excess coverage diminishes as n grows, ensuring that CP becomes *asymptotically exact*, meaning

$$\lim_{n \to +\infty} \mathbb{P}\big(y_{n+1} \in \mathcal{C}_{\alpha}(x_{n+1})\big) = 1 - \alpha \ . \tag{18}$$

For class-conditional CP, n refers to the number of calibration samples in each class. We encounter stronger conservativeness for rare classes (e.g., tumor label) as classes with small n suffer from higher over-coverage due to the larger impact of discrete rank-based p-values. Due to asymptotic exactness, as $n \to +\infty$, the upper bound tightens, and class-conditional CP approaches exact coverage in Equation (18). Unlike standard CP, class-conditional CP does not enforce a single global coverage level but rather adapts to the structure of the data, ensuring per-class validity.

Thus, conservativeness guarantees *validity* for all sample sizes while maintaining distribution-free coverage guarantees. Class-conditional CP maintains the fundamental conservativeness of standard CP but is more sensitive to class imbalances, making it particularly useful when fairness across classes is a concern.

Appendix E. Experimental Setup Details

E.1. PDAC Segmentation Model Implementation

The PDAC and organ segmentation model utilized a novel tripartite architecture consisting of a teacher, professor, and student model, implemented using 3D UNet cascade architectures. The teacher model was initially trained on 517 contrast-enhanced CT scans from the PREOPANC trials (Amsterdam UMC and Leiden UMC), LAPC registry (Dutch Pancreatic Cancer Group), and control patients who underwent CT prior to transcatheter aortic valve implantation (Van Tienhoven et al., 2018; Janssen et al., 2021; Stoop et al., 2022). Ground truth segmentations were established by three expert radiologists at the Amsterdam University Medial Centers who manually segmented PDAC tumors in 256 LAP-CTs from

120 patients with (borderline) resectable PDAC and 66 LAP-CTs from 66 LAPC patients using 3D Slicer (version 4.11.20210226 (Fedorov et al., 2012)). Additional anatomical context was provided through automated segmentation of surrounding structures (pancreas, duodenum, spleen, kidneys, adrenal glands, liver, and gallbladder) using TotalSegmentator version 1.5.6 (Wasserthal et al., 2023). The professor model, trained on 106 CT scans, was designed to refine the teacher's pseudo-segmentations using an Underestimation Focuser correction matrix that prioritized correctly identified tumors and areas of underestimation. The final student model was trained on an expanded dataset of 1085 CTs from 903 patients, combining manually segmented data with professor-corrected pseudo-segmentations. The model weights are publicly available at https://zenodo.org/records/14782552.

E.2. Vessel Segmentation Model Implementation

The vessel segmentation model was implemented using a 3D nnUNet cascade architecture (low-resolution followed by full-resolution) trained on a dataset of 92 contrast-enhanced CT scans (Isensee et al., 2021). The model was designed to segment nine vascular structures: aorta, celiac trunk, hepatic artery, splenic artery, superior mesenteric artery, inferior vena cava, portal vein, splenic vein, and superior mesenteric vein. Training data was sourced from the PREOPANC trials and control patients, comprising CT scans from patients with varying stages of pancreatic ductal adenocarcinoma (PDAC) and control subjects who underwent CT imaging for transcatheter aortic valve implantation (Van Tienhoven et al., 2018). Ground truth segmentations were established through manual annotation by seven trained observers at the Amsterdam University Medical Centeres using 3D Slicer (version 4.11.20210226) (Fedorov et al., 2012), with particular focus on the five vessels critical for PDAC resectability assessment: celiac trunk, hepatic artery, portal vein, and the superior mesenteric vessels. The model weights are publicly available at https://zenodo.org/records/14782552.

Appendix F. Additional Experimental Results

Dataset Characteristics. We analyze 30 contrast-enhanced computed tomography (CT) scans from the Memorial Sloan Kettering (MSK) Medical Segmentation Decathlon Pancreas dataset (Simpson et al., 2019), comprising portal venous phase CT scans from Memorial Sloan Kettering Cancer Center (New York, USA). Ground truth segmentations were established through expert abdominal radiologist annotations for pancreatic masses (including cysts and tumors), while surrounding anatomical structures were segmented using TotalSegmentator (Wasserthal et al., 2023). These complementary segmentations were integrated using a hierarchical fusion approach that prioritizes radiologists' tumor delineations over automated organ segmentations. This dataset includes a heterogeneous mix of pancreatic masses including resectable PDAC, intraductal papillary mucinous neoplasms (IPMN), and pancreatic neuroendocrine tumors (PNET). This composition notably differs from both the typical clinical presentation of PDAC, where approximately $80 - 85\%$ of patients present with vessel involvement indicating borderline resectable, locally advanced, or metastatic disease, and from our primary dataset which specifically captured the full range of PDAC presentations including locally advanced cases.

Table 3: Vessel-specific coverage rates at different proximity zones for the MSK dataset. The missing values ("-") indicate no tumor voxels were predicted near the celiac trunk and hepatic artery at these distances, consistent with the MSK dataset's focus on resectable PDAC cases.

Vessel	≤2mm	≤5mm	≤10mm	≤20mm	>20mm
CeTr	-	-	-	0.985	0.980
HA	-	0.956	0.753	0.906	0.991
SMA	1.000	0.995	0.997	0.991	0.975
PV	0.821	0.858	0.918	0.967	0.991
SMV	0.973	0.989	0.986	0.988	0.966

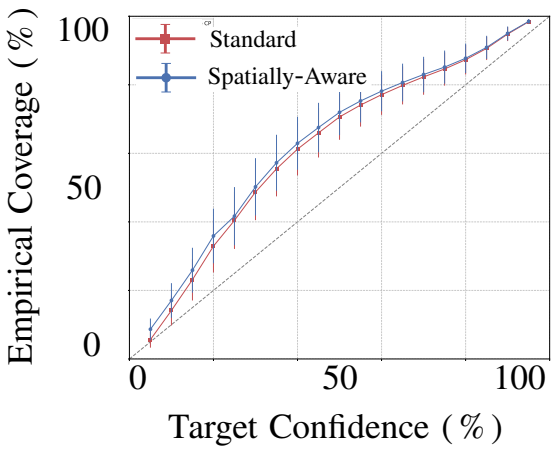
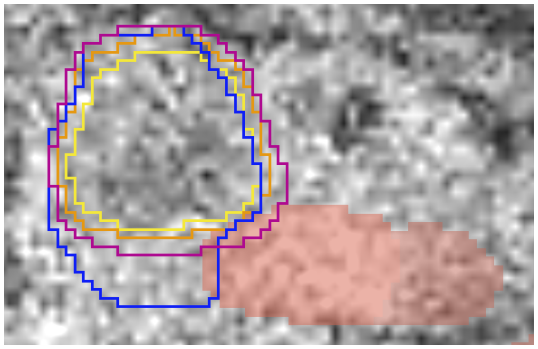


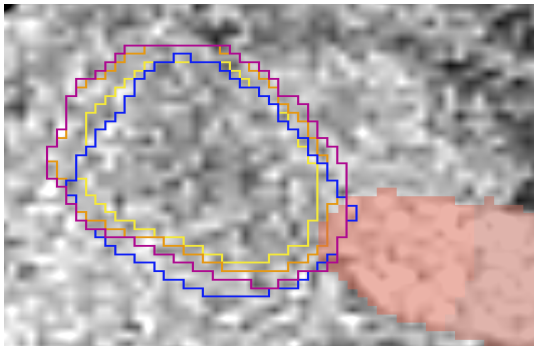
Figure 3: Comparison of empirical coverage at different confidence levels between our method (SACP) and standard Class-Conditional CP (CCCP) on the MSK dataset.

Coverage Analysis. Our framework maintains strong performance on the MSK dataset, achieving an overall coverage of 0.980 (mean per-case: 0.985 ± 0.007 standard error of the mean (SEM)). The coverage significantly exceeds the target coverage of 0.95 (Wilcoxon signed-rank test, $p = 0.0009$).

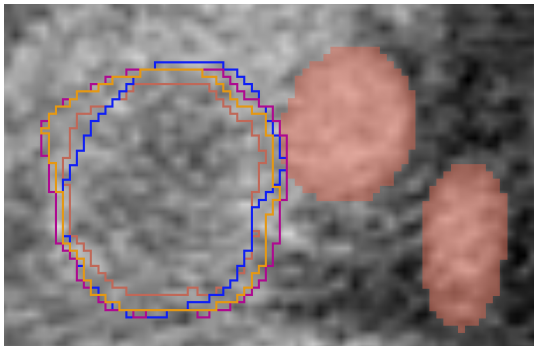
Distance-Based Analysis. Table 3 presents vessel-specific coverage rates across different proximity zones. The coverage patterns reflect the resectable nature of the cases, with notably high coverage rates in regions farther from vessels. Near-vessel regions ($\leq$ 2mm) show more variable coverage (0.821-1.000) when tumor-vessel contact is present. The relative width ratio (RWR) analysis shows a consistent relationship between prediction set size and vessel proximity, though less pronounced than in the primary dataset. Mean RWR values range from 1.141 ± 0.031 SEM in near-vessel regions ($\leq$ 2mm) to 1.655 ± 0.005 SEM beyond 20mm. This pattern of increasing width with vessel proximity persists across all vessels. The results from this dataset complement our primary analysis while highlighting the im-



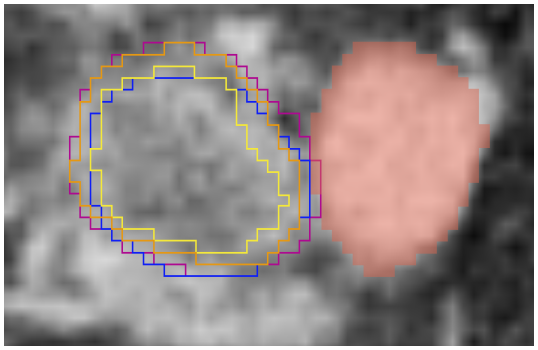
HA contact point: Purple boundary's lateral expansion suggests possible arterial involvement requiring arterial resection planning, while orange CP misses this critical region.



SMA contact point: Spatially-aware expansion identifies possible arterial invasion, a distinction missed by uniform CCCP bounds.



SMV contact point: Purple boundary's circumferential expansion indicates potential venous involvement unlike CCCP's assessment.



Portal-SMV confluence: Focused purple expansion suggests confluence involvement requiring vascular reconstruction planning, which uniform CCCP bounds fail to detect.

Figure 4: Anatomically-aware conformal prediction sets compared to standard CCCP for PDAC cases. Ground truth tumor boundaries (blue), model predictions (yellow), and vessel regions (red) are shown. Our prediction sets (purple) provide adaptive uncertainty bounds based on vessel proximity, unlike the uniform width of standard CCCP (orange), enabling more informed surgical planning in critical regions.

portance of dataset composition in evaluating conformal prediction frameworks for PDAC segmentation. The predominantly resectable cases in the MSK dataset provide insights into framework performance in scenarios with limited vessel involvement, while underscoring the need for diverse datasets that capture the full spectrum of PDAC presentations for comprehensive validation.

Comparison with Standard Class-Conditional CP. As described in Figure 3, our spatially-aware approach yields comparable overall coverage (0.980 vs 0.979) while demonstrating improved stability in anatomically critical regions. Near vessels ($\leq 2mm$), we achieve higher coverage (0.959 vs 0.956) with more efficient prediction sets (RWR 1.141 ± 0.061 SEM vs 1.205 ± 0.095 SEM). The framework shows a more controlled increase in RWR with vessel proximity, ranging from 1.141 ± 0.061 SEM at $\leq 2mm$ to 1.655 ± 0.009 SEM beyond 20mm, demonstrating effective adaptation to anatomical context while maintaining strong coverage guarantees.

Appendix G. Additional Visualization Examples

Figure 4 shows additional examples of our spatially-aware conformal prediction method across different PDAC cases taken from the PANORAMA dataset, demonstrating how the prediction sets adapt to varying tumor-vessel relationships.