Finite-sample Calibration Guarantees for Binary Segmentation

Anonymous authors
Paper under double-blind review

Abstract

Machine learning–based segmentation models are frequently miscalibrated, producing for each pixel a label prediction whose associated confidence does not accurately reflect the true underlying probability. This limitation is particularly critical in high-stakes applications such as medical image segmentation. Although several calibration methods have been proposed, only a small subset of them provides theoretical guarantees on the calibration error, i.e., they do not typically ensure that, with high probability, the model exhibits a low calibration error. In this work, we develop a *post-hoc* calibration method for which formal calibration guarantees can be established. To achieve this, we derive from standard concentration inequalities an improved theoretical bound that explicitly accounts for the spatial dependence inherent in binary segmentation outputs. We empirically demonstrate that the proposed calibrator satisfies the derived guarantee and outperforms existing *post-hoc* calibration approaches.

1 Introduction

The calibration of a segmentation model is a fundamental task in machine learning that needs to be addressed in order to align the predicted confidence of a model with the true reliability of its prediction. It is essential to obtain reliable models for real-world applications. In the context of deep learning, the confidence of a model refers to the probability it assigns to its predicted output. Ideally, this confidence should faithfully reflect the true likelihood of correctness. However, modern neural networks tend to be overconfident, which means that their predicted confidence scores are poorly aligned with the actual accuracy (Guo et al., 2017; Ovadia et al., 2019). This misalignment becomes particularly crucial in critical-domain applications such as medical image segmentation, where erroneous predictions can have severe consequences. In such settings, a miscalibrated model may assign high confidence to incorrect segmentations, leading practitioners to place unwarranted trust in flawed outputs. In contrast, a well-calibrated model ensures that a predicted confidence of, say, 90% corresponds in the long run to a true correctness rate of approximately 90% (Dawid, 1982).

The calibration problem has been extensively studied, and existing approaches can broadly be categorized into three groups (Wang et al., 2023). *Post-hoc* calibration methods operate after training, adjusting the model's output probabilities without modifying its parameters. Classical techniques include Platt scaling (Platt et al., 1999), Temperature Scaling (Guo et al., 2017), Dirichlet calibration (Kull et al., 2019), Histogram Binning (Zadrozny & Elkan, 2001), Meta-cal (Ma & Blaschko, 2021), and Isotonic Regression (Zadrozny & Elkan, 2002). Although originally introduced for classification, these methods can be applied to segmentation by treating each pixel independently. Moreover, there are also spatially aware extensions, such as (Ding et al., 2021; Rousseau et al., 2025; Küppers et al., 2022). The second group, *Train-time* calibration, has also been widely explored. Mixup (Zhang et al., 2018), CutMix (Yun et al., 2019), and label smoothing (Szegedy et al., 2016; Müller et al., 2019) regularize confidence by exposing the model to soft or interpolated targets. This group also includes methods that modify the training loss to penalize miscalibration: representative examples in the classification setting include Focal Loss (Lin et al., 2017), FLSD (Mukhoti et al., 2020), MMCE (Kumar et al., 2018), and DECE (Bohdal et al., 2023), all of which encourage the model to reduce overconfidence while preserving discriminative performance. Several approaches have also been developed specifically for segmentation tasks, such as MDCA (Hebbalaguppe et al., 2022), as well

as methods tailored to the medical imaging domain (Barfoot et al., 2024; 2026; Murugesan et al., 2025; 2023). The third group of calibration approaches are *uncertainty estimation* methods, which includes Monte Carlo dropout techniques (Gal & Ghahramani, 2016; Kendall & Gal, 2017) and Deep Ensemble Networks (Lakshminarayanan et al., 2017; Mehrtash et al., 2020). However, these methods, rather than recalibrating the model, primarily identify and highlight regions of high uncertainty.

In contrast to the works mentioned so far, which are purely empirical, methods that use calibration data to equip a pre-trained machine learning model with theoretical guarantees have seen a strong interest in the past years. Rooted in the work of Vovk et al. (2005), several works leverage statistical learning theory, and in particular conformal prediction, to provide coverage guarantees for segmentation models (Brunekreef et al., 2024; Mossina & Friedrich, 2025; Viti et al., 2026; Bereska et al., 2025; Teng et al., 2023; Liu et al., 2025). An interesting connection between conformal prediction and confidence calibration is developed by Xi et al. (2025). The authors first show that standard calibration approaches fail to improve conformal prediction, then propose a tailored calibration method designed to enhance prediction set efficiency. These works, however, focus primarily on constructing reliable prediction sets, rather than recalibrating pre-trained models with provable guarantees on the calibration error itself.

A notable exception is the work of Gupta et al. (2020); Gupta & Ramdas (2021), which endows histogram-binning calibration with finite-sample statistical guarantees for classification tasks. Following their notation, let the feature and label space be $\mathcal{X}$ and $\mathcal{Y} = \{0, 1\}$. We have at our disposal a pre-trained predictor $f : \mathcal{X} \to [0, 1]$, where $f(X)$ is the (uncalibrated) predicted probability that $Y = 1$, and a calibration dataset $\mathcal{D}_{cal} = \{(X^{(i)}, Y^{(i)})\}_{i=1}^{N_{cal}}$. The pre-trained model f belongs to the space of all measurable functions $\mathcal{F}$ from $\mathcal{X} \to [0, 1]$. We would like to construct a calibrator $H : \mathcal{F} \times \mathcal{D}_{cal} \to \mathcal{F}$, that recalibrates $f \in \mathcal{F}$ given the dataset $\mathcal{D}_{cal}$. We denote the calibrated predictor as $h := H(f, \mathcal{D}_{cal})$. A predictor is perfectly calibrated if $\mathbb{E}[Y|h(X)] = h(X)$, which, in the binary case, corresponds to $\mathbb{P}(Y = 1 \mid h(X)) = h(X)$. While the latter condition is impossible to guarantee in a finite-sample setting, we would like to have a guarantee on the calibration error using the following notions:

Definition 1.1. *(Marginal calibration). Let $\varepsilon > 0$ and $\alpha \in [0, 1]$. A calibrator H is said to be (ε, α)-marginally calibrated if for every predictor $f \in \mathcal{F}$ and distribution $\mathbb{P}$ over $\mathcal{X} \times [0, 1]$*

$$\mathbb{P}(|\mathbb{E}[Y \mid h(X)] - h(X)| \leq \varepsilon) \geq 1 - \alpha.$$

Definition 1.2. *(Conditional calibration). Let $\varepsilon > 0$ and $\alpha \in [0, 1]$. A calibrator H is said to be (ε, α)-conditionally calibrated if for every predictor $f \in \mathcal{F}$ and distribution $\mathbb{P}$ over $\mathcal{X} \times [0, 1]$*

$$\mathbb{P}(\forall r \in Range(h) : |\mathbb{E}[Y \mid h(X) = r] - r| \leq \varepsilon) \geq 1 - \alpha.$$

The fundamental distinction between the two definitions is that in the marginal case the probability is taken over both X and $\mathcal{D}_{cal}$, meaning that the probability holds only on average over $h(X)$. Conditional calibration instead, which implies marginal, requires the bound to hold uniformly in r.

The seminal work of Gupta et al. (2020); Gupta & Ramdas (2021) remains confined to the classification setting. Our contribution builds directly on this line of work and extends it beyond classification. To the best of our knowledge, no existing work provides marginal and conditional guarantees in the segmentation setting. One reason is that standard approaches require large amounts of independent and exchangeable calibration data — an assumption that breaks down in segmentation, where pixel-wise predictions within an image are spatially correlated. This paper fills this gap by leveraging concentration inequalities, specifically adapted to account for spatial pixel correlation in our setting, to derive calibration guarantees for segmentation models.

The main theoretical tool used in (Gupta et al., 2020; Gupta & Ramdas, 2021) to bound the calibration error is Hoeffding's inequality (Hoeffding, 1963), which provides concentration for sums of bounded *i.i.d.* random variables. In segmentation, however, the predictor outputs a probability for each pixel, and the spatial correlation between pixels violates the *i.i.d.* assumption, making such inequalities inapplicable. While extensions to the non-*i.i.d.* setting exist — including martingale-based results such as Azuma (1967) and Freedman

(1975) inequalities, and concentration under β-mixing conditions (Yu, 1994; Mohri & Rostamizadeh, 2010) — these do not capture the structured spatial dependencies inherent to segmentation outputs.

Image segmentation outputs can naturally be modeled as random fields: each pixel label depends not only on its own features but also on the label of its spatial neighbors, inducing dependencies across the output map that are well captured by random field models such as Markov Random Fields (Li, 2009; Zoltan & Josiane, 2012). Bounding the calibration error in this setting therefore requires concentration inequalities that explicitly account for such spatial dependencies. Early approaches to random fields relied on information-theoretic tools: Marton (1998; 2003) and Samson (2000) established concentration results for Φ-mixing processes using transportation and Pinsker inequalities, albeit restricted to exponential bounds. A significant step forward was achieved by Chazottes et al. (2007), who proposed a coupling-based framework for random fields indexed by $\mathbb{Z}^d$ that yields concentration inequalities depending on the strength of spatial dependence. A parallel martingale-based approach was developed in (Kontorovich & Ramanan, 2008), yielding concentration bounds for Markov and hidden Markov chains in terms of mixing coefficients. More recently, Garnier & Langhendries (2022) established a Hoeffding-type inequality specifically for non-causal random fields. Finally, Chazottes et al. (2026) provides sharp Gaussian concentration criteria for Gibbs measures and Markov random fields on $\mathbb{Z}^d$.

Our contribution In this work, we address the problem of calibration for binary segmentation models. We first derive an improved concentration inequality tailored to variables exhibiting the structured spatial dependence of binary segmentation outputs. Building upon the framework of (Chazottes et al., 2007), we model the spatial relationship between pixels and obtain a tighter theoretical bound. We then apply the bound to the calibration setting: we construct a calibrator H following the idea of Gupta & Ramdas (2021), and establish that the calibrator satisfies the theoretical guarantees provided by our bound. This result formally justifies its application in the segmentation setting. A key insight is that treating each individual pixel prediction as a data point vastly increases the effective size of the calibration set. As a result, our method achieves tight statistical guarantees with a comparably small number of calibration images — even with only 50 images — making it practical for real-world applications where calibration data is scarce. Beyond providing theoretical guarantees, we also show that the calibrator H even practically outperforms other *post-hoc* calibration approaches that do not come with such guarantees.

The outline of the paper is the following: In Section 2 we derive the improved theoretical bound and in Section 3 we show how the bound can be exploited to obtain statistical guarantees for calibration errors in the segmentation setting. Finally, in Section 4 we show how to estimate the bound and in Section 5 we empirically show the validity of our approach.

2 Concentration Inequality under Spatial Dependence

For this section we follow the notation of Chazottes et al. (2007). Let A be a finite set, $g : A^n \to \mathbb{R}$ a function of n variables, and $\sigma = (\sigma_1, \dots, \sigma_i, \dots)$ an element of $A^{\mathbb{N}}$. As in (Chazottes et al., 2007), we consider, with an abuse of notation, g as a function on $A^{\mathbb{N}}$, which does not depend on σ_k for $k > n$. We make use of two quantities, both defined later in this section: the coupling matrix $\bar{D}$, which encodes the dependence structure among the n variables and quantifies how well the distribution of each variable is determined by the others; and the sensitivity δg, which measures how much g can change when a single element of σ is perturbed. With this notation, Chazottes et al. (2007) derive concentration inequalities of the form:

$$\mathbb{P}(|g - \mathbb{E}[g]| \geq \varepsilon) \leq 2 \exp\left(\frac{-2\varepsilon^2}{\|\bar{D}\|^2_{\ell^2(\mathbb{N})} \, \|\delta g\|^2_{\ell^2(\mathbb{N})}} \right).$$

We are going to improve the bound through two assumptions that hold true in our setting, namely: $A = \{0, 1\}$ is a Bernoulli alphabet and g is the average of Bernoulli variables.

2.1 Bernoulli alphabet

To establish the concentration inequality, we use a martingale argument which decomposes the quantity of interest $g - \mathbb{E}[g]$ as the sum of martingale differences McDiarmid et al. (1989). For $i = 1, \dots, n$ let $\mathcal{F}_i$ be the sigma-field generated by the random variables $\sigma_1, \dots, \sigma_i$. Then, following Chazottes et al. (2007, Section 2.1), we write

$$g(\sigma) - \mathbb{E}[g] = \sum_{i=1}^{n} V_i(\sigma), \qquad \text{where} \quad V_i(\sigma) := \mathbb{E}[g|\mathcal{F}_i](\sigma) - \mathbb{E}[g|\mathcal{F}_{i-1}](\sigma). \tag{1}$$

Further, we write $\sigma_{<i}$ for the values at sites ordered before i, and define

$$p_i(\sigma) := \mathbb{P}(\sigma_i = 1 \mid \sigma_{<i}), \qquad \Delta_i(\sigma) := \mathbb{E}[g \mid \sigma_{<i},\, \sigma_i = 1] - \mathbb{E}[g \mid \sigma_{<i},\, \sigma_i = 0].$$

For a Bernoulli alphabet $A = \{0, 1\}$, a computation using the law of total expectation (see A.1) yields:

$$V_i(\sigma) = (\sigma_i - p_i)\, \Delta_i. \tag{2}$$

To obtain the concentration inequality, we have to bound the moment generating function (MGF) of each variable V_i. However, Chazottes et al. (2007, Lemma 1) uses only the range of V_i, namely Δ_i, to obtain such a bound. We, instead, make use of the fact that V_i is a scaled centered Bernoulli with conditional variance $\mathrm{Var}(V_i \mid \mathcal{F}_{i-1}) = p_i(1 - p_i)\Delta_i^2$, and exploit the terms $p_i(1 - p_i)$ to provide a tighter MGF bound than the range-only estimate. To do that, we apply Bennett's inequality (Boucheron et al., 2013, Theorem 2.9) and prove it in Appendix A.2:

Lemma 2.1 (Bennett bound)**.** *Let $V_i = (\sigma_i - p_i)\Delta_i$ and $\Delta_i \geq 0$, then $\forall \lambda > 0$:*

$$\log \mathbb{E}\big[e^{\lambda V_i} \mid \mathcal{F}_{i-1}\big] \leq p_i(1 - p_i)\, \phi(\lambda \Delta_i), \qquad \textit{where} \quad \phi(u) := e^u - u - 1. \tag{3}$$

2.2 The coupling matrix $\bar{D}$ and the function g

Now that we have bounded the MGF of V_i by $p_i(1 - p_i)\, \phi(\lambda \Delta_i)$, we need to bound the increment Δ_i. To this aim, for a given configuration σ, we define the coupling matrix D^σ as

$$D^\sigma_{i,j} := \mathbb{P}^{\sigma_{<i}}_{i,0,1}\left(\sigma_j^{(1)} \neq \sigma_j^{(2)}\right),$$

where $\mathbb{P}^{\sigma_{<i}}_{i,0,1}$ denotes the maximal coupling (Lindvall, 2002) of the conditional distributions $\mathbb{P}(\cdot|\sigma_{<i}, \sigma_i = 1)$ and $\mathbb{P}(\cdot|\sigma_{<i}, \sigma_i = 0)$. Intuitively, $D^\sigma_{i,j}$ measures the influence of site i on site j: it is the probability that two coupled copies ($\sigma^{(1)}$ and $\sigma^{(2)}$) of the process, agreeing on the past $\sigma_{<i}$, but initialized with opposite values at site i, still disagree at site j. Notice that for $j < i$ we have $D^\sigma_{i,j} = 0$ and that, in case of independent random variables, the matrix D^σ would be the identity matrix. With D^σ in hand, we can now bound Δ_i via the variation of g. In our setting, g is the empirical mean of Bernoulli variables, and its variation at site i is

$$\delta_i g = \sup_{\substack{\sigma,\, \sigma' \\ \sigma_j = \sigma'_j,\ \forall j \neq i}} |g(\sigma) - g(\sigma')| = \frac{1}{n}, \quad g(\sigma) = \frac{1}{n} \sum_{i=1}^{n} \sigma_i. \tag{4}$$

The last step is using Chazottes et al. (2007, Eq. 5) to bound the increment by $\Delta_i(\sigma) \leq (D^\sigma \delta g)_i$. In our case, since $\delta_j g = 1/n$ for all j, and defining $\bar{D}_{i,j} := \sup_\sigma D^\sigma_{i,j}$ to remove the dependence on σ, this simplifies to (see Appendix A.3):

$$\Delta_i(\sigma) \leq \frac{1}{n} \sum_{j > i}^{n} \bar{D}_{i,j} := \frac{R_i}{n}, \tag{5}$$

where R_i denotes the i-th row sum of $\bar{D}$, i.e., the cumulative forward influence of element i on all subsequent elements. A more detailed discussion and practical estimation of $\bar{D}$ is provided in Section 4.

2.3 Main Theorem

We now state the concentration inequality and prove it in Appendix A.4.

Theorem 2.2 (Concentration for Bernoulli mean)**.** *Let $g = n^{-1}\sum_{i=1}^{n}\sigma_i$. Define the coupling matrix $\bar{D}$ as in Section 2.2, $R_i := \sum_{j>i}^{n}\bar{D}_{i,j}$, and $p_i := \mathbb{P}(\sigma_i = 1 \mid \sigma_{<i})$. Set*

$$S \;:=\; \frac{1}{n^2}\sum_{i=1}^{n} p_i(1-p_i)\,R_i^2, \qquad R_{\max} := \max_{i=1,\dots,n} R_i, \quad c \;:=\; \frac{R_{\max}}{3n}. \tag{6}$$

Then for all $\varepsilon > 0$:

$$\boxed{\mathbb{P}(|g - \mathbb{E}[g]| \geq \varepsilon) \leq 2\exp\left(-\frac{S}{c^2}\eta\Big(\frac{c\varepsilon}{S}\Big)\right)} \qquad \textit{with} \quad \eta(x) = 1 + x - \sqrt{1+2x}. \tag{7}$$

Notice that the notation $\sigma_{<i}$ relies on $\mathbb{N}$ having a natural total order. For variables indexed in $\mathbb{Z}^2$, one fixes an arbitrary enumeration $\Gamma : \mathbb{Z}^2 \to \mathbb{N}$ and writes $\sigma_{<x} := \{\sigma_y : \Gamma(y) < \Gamma(x)\}$. The full extension to $\mathbb{Z}^2$ is given in Appendix A.5 and involves mainly this notational variation.

3 Calibration Guarantees

We are interested in obtaining calibration error bounds in the segmentation setting where the *i.i.d.* assumption does not hold due to the spatial dependence between pixels. Let $x \in \Omega := \{1,\dots,H\} \times \{1,\dots,W\}$ denote pixel positions. Let $\mathcal{X} = [0,1]^{\Omega}$ and $\mathcal{Y} = \{0,1\}^{\Omega}$ be the feature and label spaces of images and their segmentation masks, respectively. Let $f : \mathcal{X} \to [0,1]^{\Omega}$ be a pre-trained predictor, where $f(X)_x$ predicts the probability of $Y_x = 1$. We seek a calibrator H inducing a calibrated predictor h, which maps each pixel-wise confidence $f(X)_x$ to a revised confidence score. We are going to use a binning approach, first introduced in (Zadrozny & Elkan, 2001) and more recently used in (Gupta et al., 2020; Gupta & Ramdas, 2021). Let $\mathcal{D} = \{X^{(i)}, Y^{(i)}\}_{i=1}^{N}$ be a dataset featuring images and binary segmentations. We partition the dataset into two subsets $\mathcal{D}_{\text{bin}}$ and $\mathcal{D}_{\text{cal}}$, using the first one to compute the binning scheme and the latter to construct the calibrator H.

3.1 Binning scheme

The first step is to create a binning scheme that provides a partition $\{\Omega_b\}_{b=1}^{B}$ of the pixel positions. Following the definition of Gupta et al. (2020), a binning scheme $\mathcal{B} = \{I_b\}_{b=1}^{B}$ is any partition of $[0,1]$ into B non-overlapping intervals $I_1,\dots,I_B$ such that $\bigcup_{b=1}^{B} I_b = [0,1]$ and $I_b \cap I_{b'} = \emptyset$ for $b \neq b'$. Therefore, given $\mathcal{B}$ and f, the partition for a generic image i is defined as

$$\Omega_b^{(i)} = \{x \in \Omega : \; f(X^{(i)})_x \in I_b\}.$$

A naive way to construct the binning scheme is to choose fixed-width bins such that $I_j = [\frac{j-1}{B}, \frac{j}{B})$, $j = 1,\dots B-1$ and $I_B = [\frac{B-1}{B}, 1]$. This binning scheme requires no held-out data, but tends to produce highly imbalanced bins, since overconfident predictors typically assign extreme probabilities (close to 0 or 1) to most pixels. Therefore, we empirically find a uniform-mass binning scheme through quantiles (Gupta & Ramdas, 2021). The bins are defined as

$$I_1 = [0, \hat{q}_1),\; I_j = [\hat{q}_{j-1}, \hat{q}_j],\; j = 2,\dots,B-1,\; I_B = (\hat{q}_{B-1}, 1], \quad \hat{q}_j = \mathcal{Q}_{\frac{j}{B}}(\{f(X^{(i)})_x\}_{i=1,x\in\Omega}^{|\mathcal{D}_{bin}|}).$$

where $\hat{q}_j$ is the (j/B)-th quantile, empirically estimated using the images in $\mathcal{D}_{bin}$.

3.2 Calibrator

Now that we have a binning scheme $\mathcal{B}$, we partition the pixels of the N_{cal} images of $\mathcal{D}_{cal}$ using $\mathcal{B}$, and we define the calibrated predictor $h = H(f, \mathcal{D}_{cal})$ as:

$$h(X)_x = \sum_{b=1}^{B} \mathbf{1}_{\{f(X)_x \in I_b\}} \hat{\pi}_b, \quad \text{where} \quad \hat{\pi}_b = \frac{1}{N_b} \sum_{i=1}^{N_{cal}} \sum_{x_b \in \Omega_b^{(i)}} Y_{x_b}^{(i)}, \; N_b = \sum_{i=1}^{N_{cal}} |\Omega_b^{(i)}|. \tag{8}$$

As an example, suppose that on the calibration dataset bin b has an average raw confidence (predicted by f) of 0.8, but only 70% of those predictions have ground-truth equal to 1. Then, for a new test point X_{test}, the calibrator H replaces the raw confidence $f(X_{test})_x$ with the empirical label frequency $\hat{\pi}_b$ observed in bin b on the calibration set, i.e., $h(X_{test})_x = \hat{\pi}_b = 0.7$ if $f(X_{test})_x \in I_b$. We note that $\hat{\pi}_b$ is to be understood as conditioned on the binning scheme $\mathcal{B}$, so that N_b is treated as a constant and the only source of randomness are the Bernoulli labels. The calibration procedure is summarized in Algorithm 1 in the appendix.

3.3 Calibration Guarantees Theorems

We now use the theory derived in Section 2 to bound the calibration error with the following theorems, which are all proved in Appendix A.6. Before stating the calibration theorems, observe that for a fixed bin b, $\hat{\pi}_b$ is the empirical mean of N_b dependent Bernoulli variables. Thus, $\hat{\pi}_b$ plays the role of $g(\sigma)$, with σ being the labels of pixels falling in bin b, i.e., $\sigma = \{Y_{x_b}^{(i)}\}_{i=1, x_b \in \Omega_b^{(i)}}^{N_{cal}}$.

Theorem 3.1 (Per-bin concentration)**.** *Let $\mathcal{B} = \{I_b\}_{b=1}^{B}$ be a fixed binning scheme. Then, given a calibration set $\{(X^{(i)}, Y^{(i)})\}_{i=1}^{N_{cal}}$, let $\{\Omega_b^{(i)}\}_{b=1}^{B}, i = 1, ..., N_{cal}$ be the corresponding partitions and define the observed frequency $\hat{\pi}_b$ in bin b as in equation 8. Denote as $E_{\mathcal{B}}$ a realization of the binning event that fixes the calibration labels $Y_{x_b}^{(i)}$ in the corresponding bins, and fix a bin b. Let $\bar{D}^{(i)}$ be the coupling matrix (as defined in Section 2.2) of image i conditioned on the binning event $E_{\mathcal{B}}$ (i.e. the collection $\{Y_{x_b}^{(i)}\}_{x_b \in \Omega_b^{(i)}}$), and define*

$$R_{x_b}^{(i)} = \sum_{y_b > x_b} \bar{D}_{x_b, y_b}^{(i)}, \qquad p_{x_b}^{(i)} = \mathbb{P}\big(Y_{x_b}^{(i)} = 1 \mid Y_{<x_b}^{(i)}, E_{\mathcal{B}}\big)$$

Set

$$S_b = \frac{1}{N_b^2} \sum_{i=1}^{N_{cal}} \sum_{x_b \in \Omega_b^{(i)}} p_{x_b}^{(i)} (1 - p_{x_b}^{(i)}) \, (R_{x_b}^{(i)})^2, \quad R_{\max,b} = \max_{i=1,\dots,N_{cal}} \max_{x \in \Omega_b^{(i)}} R_{x_b}^{(i)}, \quad c_b = \frac{R_{\max,b}}{3N_b}.$$

Then for all $\varepsilon > 0$:

$$\mathbb{P}(|\hat{\pi}_b - \mathbb{E}[\hat{\pi}_b \mid E_{\mathcal{B}}]| \geq \varepsilon) \leq 2 \exp\left(-\frac{S_b}{c_b^2} \, \eta\left(\frac{c_b \, \varepsilon}{S_b}\right)\right), \quad \textit{with} \quad \eta(x) = 1 + x - \sqrt{1 + 2x}.$$

Theorem 3.2 (Uniform bound over bins)**.** *Let the assumptions of Theorem 3.1 hold. Then:*

$$\mathbb{P}(\exists \, b \in [B] : |\hat{\pi}_b - \mathbb{E}[\hat{\pi}_b \mid E_{\mathcal{B}}]| \geq \varepsilon) \leq 2 \sum_{b=1}^{B} \exp\left(-\frac{S_b}{c_b^2} \, \eta\left(\frac{c_b \varepsilon}{S_b}\right)\right), \quad \textit{with} \quad \eta(x) = 1 + x - \sqrt{1 + 2x}.$$

In the segmentation setting we apply Definitions 1.1-1.2 pixel-wise, replacing Y and $h(X)$ by the per-pixel quantities Y_x and $h(X)_x$, so that the guarantee is required to hold at an arbitrary fixed pixel $x \in \Omega$. This is the natural scope for our binning-based calibrator, which acts on each pixel independently. However, to connect Theorems 3.1–3.2 to this notion, observe that for a variable Y_x, conditioning on the binning event $E_{\mathcal{B}} = \{f(X)_x \in I_b\}$ does not in general coincide with conditioning on $h(X)_x$: the two agree only when $\hat{\pi}_{b'} \neq \hat{\pi}_b$ for every $b' \neq b$. Following Gupta & Ramdas (2021), we resolve this by adding an arbitrarily small uniform perturbation δ to $h(X)_x$, e.g. $\delta = 10^{-20}$, to obtain a randomized version $\tilde{h}(X)_x$. The randomized version is practically indistinguishable from the original calibrated predictor $h(X)_x$ but makes the binning

scheme recoverable from $\tilde{h}(X)_x$. The explicit construction is deferred to the proof of Theorem 3.3 and the randomized calibration procedure is summarized in Algorithm 2. We now state the new calibration guarantees:

Theorem 3.3 (Calibration guarantees)**.** *Let the assumptions of Theorem 3.1 hold. Let $\tilde{H}$ be the randomized calibrator parametrised by δ, such that $\tilde{h} = \tilde{H}(f, \mathcal{D}_{cal})$. Assume moreover that $Y_x \mid \tilde{h}(X)_x \perp\!\!\!\perp \tilde{h}(X)_y,\ y \neq x$, and that, by tanslation invariance, pixels share the same conditional mean $\mathbb{E}[Y_x \mid \tilde{h}(X)_x] = \mathbb{E}[Y_y \mid \tilde{h}(X)_y],\ \forall x, y$. Define*

$$\alpha_{1,b} = 2\exp\left(-\frac{S_b}{c_b^2}\,\eta\left(\frac{c_b\,(\varepsilon-\delta)}{S_b}\right)\right),\ \alpha_2 = 2\sum_{b=1}^{B}\exp\left(-\frac{S_b}{c_b^2}\,\eta\left(\frac{c_b(\varepsilon-\delta)}{S_b}\right)\right).$$

Then, $\tilde{H}$ is $(\varepsilon, \alpha_{1,b})$-marginally calibrated for every $b = 1, \ldots, B$ and every $x_b \in \Omega_b$:

$$\mathbb{P}\left(\left|\mathbb{E}[Y_{x_b} \mid \tilde{h}(X)_{x_b}] - \tilde{h}(X)_{x_b}\right| \leq \varepsilon\right) \geq 1 - \alpha_{1,b}.$$

Moreover $\tilde{H}$ is (ε, α_2)-conditionally calibrated for every $x \in \Omega$:

$$\mathbb{P}\left(\forall\, r \in \mathit{Range}(\tilde{h}) : \left|\mathbb{E}[Y_x \mid \tilde{h}(X)_x = r] - r\right| \leq \varepsilon\right) \geq 1 - \alpha_2.$$

The assumption $Y_x \mid \tilde{h}(X)_x \perp\!\!\!\perp \tilde{h}(X)_y,\ y \neq x$ states that, once the bin assignment at pixel x is known, the bin assignments at all other pixels provide no additional information about the label Y_x. Notice that the pre-trained network f is itself designed to aggregate neighborhood context into the prediction $f(X)_x$, so the bin assignment at x already inherits this spatial information. Crucially, the assumption does not make the problem trivial—the Bernoulli labels Y_x remain dependent across pixels, which is why the spatial concentration inequality is needed for our argument.

4 Estimation of Theoretical Bound

Since the bound is not analytically tractable, due to the matrix $\bar{D}$, we estimate it empirically. While this does not yield a closed-form guarantee, it provides a principled, quantitative basis for evaluating the binning method and guiding its application in the segmentation setting. Let $\Omega \subset \mathbb{Z}^2$ be a finite rectangular image domain with pixel alphabet $A = \{0, 1\}$, so configurations σ are elements of $\{0, 1\}^{\Omega}$. We fix an arbitrary ordering $\Gamma : \Omega \to \{1, \ldots, |\Omega|\}$ of the pixels, and we write $\sigma_{<x}$ for the configuration on $\{y \in \Omega : \Gamma(y) < \Gamma(x)\}$. Following (Chazottes et al., 2007, Eq. 4), and using the total-variation formulation of (Samson, 2000, Eq. 2.1), we define the matrix D^{σ} with entries indexed by pairs $(x, y) \in \Omega \times \Omega$ with $\Gamma(y) > \Gamma(x)$ as

$$D^{\sigma}_{x,y} = \left\|\mathbb{P}(\sigma_y \mid \sigma_{<x}, \sigma_x = 1) - \mathbb{P}(\sigma_y \mid \sigma_{<x}, \sigma_x = 0)\right\|_{\mathrm{TV}}, \tag{9}$$

and $D^{\sigma}_{x,y} = 0$ for all other pairs. This formulation is obtained by choosing the Hamming metric in the definition of D^{σ} (Chazottes et al., 2007, Eq. 4) and is equivalent to the coupling matrices defined in Kontorovich & Ramanan (2008, Eq. 1.2) and Samson (2000, Eq. 2.1). In our application, σ is the vector of labels Y_x restricted to the pixels x_b that are contained in a fixed partition Ω_b, i.e., $\sigma = \{Y_{x_b}\}_{x_b \in \Omega_b}$. Moreover, as detailed in Section 3, the variables Y_x are conditioned on the binning event $E_{\mathcal{B}}$ and therefore on $h(X)_x$. Adapting equation 9 to this conditional process and taking the 'worst' configuration to remove the σ dependence, the coupling matrix that enters Theorem 3.2 becomes

$$\bar{D}_{x_b,y_b} = \sup_{\sigma_{<x_b}} \left|\mathbb{P}(\sigma_{y_b} = 1 \mid \sigma_{<x_b},\, \sigma_{x_b} = 1,\, h(X)_{y_b}) - \mathbb{P}(\sigma_{y_b} = 1 \mid \sigma_{<x_b},\, \sigma_{x_b} = 0,\, h(X)_{y_b})\right|, \tag{10}$$

since for a binary alphabet the total-variation distance equals the absolute difference of the Bernoulli means.

We estimate the conditional probabilities appearing in equation 10 by fitting an autoregressive model that takes as input $h(X)$ and a partition Ω_b for a fixed bin b, and predicts the distribution of the labels

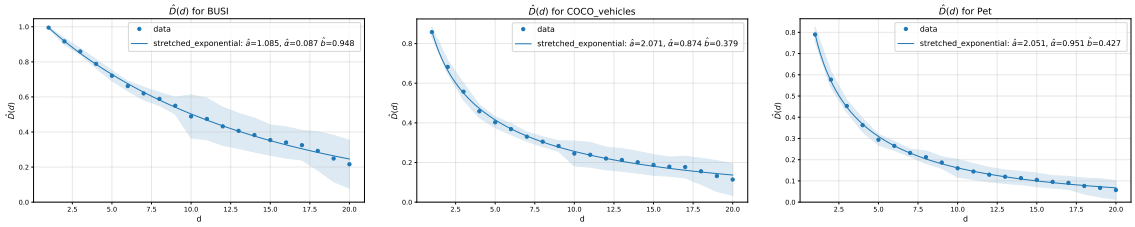

Figure 1: Decay fit $\hat{D}(d)$ for different datasets.

$\sigma = \{Y_{x_b}^{(i)}\}_{x_b \in \Omega_b}$. To this aim, we model the joint distribution of the binary segmentation mask restricted to the partition Ω_b and conditioned to $h(X)$ via the autoregressive factorisation

$$\mathbb{P}_\theta(\sigma \mid h(X)) = \prod_{x_b \in \Omega_b} \mathbb{P}_\theta(\sigma_{x_b} \mid \sigma_{<x_b}, h(X)_{x_b}) \,. \tag{11}$$

Each conditional $\mathbb{P}_\theta(\sigma_{x_b} = 1 \mid \sigma_{<x_b}, h(X)_{x_b})$ is the sigmoid output of a lightweight network trained by maximising the log-likelihood in equation 11 for a set of bins b, partitions Ω_b and $h(X)$. Given the fitted model, the entries of $\bar{D}$, defined in equation 10, are evaluated for each pair $(x_b, y_b) \in \Omega_b \times \Omega_b$ with $\Gamma(y_b) > \Gamma(x_b)$ by computing

$$\hat{D}^{\sigma}_{x_b,y_b} = |\mathbb{P}_\theta(\sigma_{y_b} = 1 \mid \sigma_{<x_b},\, \sigma_{x_b} = 1,\, h(X)_{y_b}) - \mathbb{P}_\theta(\sigma_{y_b} = 1 \mid \sigma_{<x_b},\, \sigma_{x_b} = 0,\, h(X)_{y_b})| \,. \tag{12}$$

Because the conditional $\mathbb{P}_\theta(\sigma_y \mid \sigma_{\leq x}, h(X)_x)$ is a marginal over the intermediate sites $\sigma_{x+1}, \ldots, \sigma_{y-1}$, we estimate it using Monte Carlo iterations. Finally, we estimate $\hat{D}_{x,y}$ by sampling a set of random pasts $\sigma_{<x}$ and taking the configuration correspondent to the largest $\hat{D}^{\sigma}_{x,y}$, obtaining $\hat{D}_{x,y} = \max_{\sigma_{<x}} \hat{D}^{\sigma}_{x,y}$. We compute $\hat{D}_{x,y}$ over a range of positions x and y and empirically observe in Fig. 13 that its entries depend only on the Euclidean distance $d = \|y - x\|_2$. This indicates that $\hat{D}_{x,y}$ is translation-invariant and depends solely on the relative displacement between positions. Therefore, we model the distance dependence using a stretched exponential decay function,

$$\hat{D}(d) = \hat{a}\, \exp(-\hat{\alpha}\, d^{\hat{b}}), \qquad \hat{a},\, \hat{\alpha} > 0, \hat{b} > 0 \tag{13}$$

whose fitted parameters provide a closed-form estimate of the full coupling matrix and can be used to compute the coefficients S and c that appear in Theorem 2.2. Fig. 1 shows the estimated $\hat{D}(d)$ and its exponential fit for a selection of the datasets used in the experiments. See Appendix D for further details.

5 Experiments

We now show that our theoretically derived bound is consistent with the empirical behavior and compare the binning approach with alternative *post-hoc* calibration methods. To do so we compare the true validity curves $V(\varepsilon)$ and the empirical ones $\hat{V}(\varepsilon)$ both for marginal and conditional calibration. Marginal and conditional validity curves are defined as (Gupta & Ramdas, 2021):

$$V_m(\varepsilon) = \mathbb{P}(|\mathbb{E}[Y_x|h(X)_x] - h(X)_x| \leq \varepsilon),$$

$$V_c(\varepsilon) = \mathbb{P}(\forall r \in Range(h) : |\mathbb{E}[Y_x|h(X)_x = r] - r| \leq \varepsilon),$$

and are bounded by Theorem 3.3. Let π_b^{obs} the ground-truth observed frequency,

$$\pi_b^{obs} = \frac{\sum_{i=1}^{N_{test}} \sum_{x \in \Omega} Y_x^{(i)} \mathbf{1}_{\{f(X^{(i)})_x \in I_b\}}}{\sum_{i=1}^{N_{test}} \sum_{x \in \Omega} \mathbf{1}_{\{f(X^{(i)})_x \in I_b\}}}.$$

Then the empirical validity curves can be approximated as

$$\hat{V}_m(\varepsilon) = \frac{1}{N_{test}\,|\Omega|} \sum_{i=1}^{N_{test}} \sum_{b=1}^{B} \sum_{x_b \in \Omega_b^{(i)}} \mathbf{1}_{\{|\pi_b^{obs} - h(X^{(i)})_{x_b}| \leq \varepsilon\}},$$

$$\hat{V}_c(\varepsilon) = \mathbf{1}\Big\{ \max_{i=1,\dots,N_{test}} \max_{b=1,\dots,B} \max_{x_b \in \Omega_b^{(i)}} |\pi_b^{obs} - h(X^{(i)})_{x_b}| \leq \varepsilon \Big\}.$$

Both marginal and conditional calibration are estimated as the mean across 5 random calibration and test splits. The error bars in Figure 2 refer to ± 1 standard deviation across those random splits.

Baselines and Datasets We compare our method with six *post-hoc* calibration baselines. *Temperature Scaling* (Guo et al., 2017) rescales the network logits by a single learned scalar $T > 0$. *Logistic Scaling* (Platt et al., 1999) fits a logistic function on top of the network outputs, learning an affine transformation of the logits to map them to calibrated probabilities. *Meta-Cal* (Ma & Blaschko, 2021) first learns a ranking function to categorize calibration samples in two groups, then applies temperature scaling to one group and sets the other as random predictions. *Selective Scaling* Wang et al. (2023) is an extension of Meta-cal where a binary classifier categorizes pixels in two groups scaled with two different temperatures T_1 and T_2. *Beta Calibration* (Kull et al., 2019) fits a parametric map based on the Beta distribution family, learning three parameters that control the shape and shift of the probability correction. *Isotonic Regression* (Zadrozny & Elkan, 2002) is a non-parametric method that fits a monotone piecewise-constant mapping via the pool-adjacent-violators algorithm. In Appendix B, we provide further details on these baseline approaches, along with a discussion of the effect of *post-hoc* calibration on segmentation accuracy (e.g., DICE score).

To assess the generality of our approach, we evaluate all methods across nine segmentation datasets spanning diverse imaging domains and anatomical structures: ISIC (Gutman et al., 2016), Kvasir (Pogorelov et al., 2017), TN3K (Gong et al., 2023), BUSI (Tasnim & Hasan, 2024; Al-Dhabyani et al., 2020), Cityscapes (Cordts et al., 2015; 2016), FLAIR1 (Garioud et al., 2022), Oxford-IIIT Pet (Parkhi et al., 2012), and two subsets of the COCO dataset (Lin et al., 2014). Where datasets provided multi-class annotations, predictions were binarized by collapsing all foreground classes into a single positive class. The diversity of modalities, domains, and backbone architectures — spanning both medical and non-medical imaging and detailed in Appendix C — tests our method across highly heterogeneous conditions, providing a broad basis for evaluating its generality and robustness.

Results We now compare, and show in Figure 2, the empirical validity curves $\hat{V}(\varepsilon)$ of the binning strategy, and of the other baselines, with our theoretical bound, as well as with the *i.i.d.* bound that would apply under the (hypothetical) assumption that all pixels are independent. Our experiments highlight two main findings. (1) The empirical validity curves of the binning method, evaluated across several datasets, are consistent with our theoretical bound. As expected, marginal calibration is easier to satisfy: the binning curve shows good calibration properties, i.e. it attains small calibration errors even for small ε, and the bound consistently controls the binning curve across the experiments. In some cases, the empirical curve approaches the *i.i.d.* curve, which can be attributed to an intrinsically good calibration of the pre-trained model. Conditional calibration, by contrast, is generally more difficult to achieve. Nevertheless, our derivation yields a meaningful and nontrivial bound that is closely tracked by the binning validity curves. The *i.i.d* curve further confirms that the independence assumption is too strong for calibrating segmentation models, and would yield incorrect guarantees. While the theoretical bound is only available in estimated form, these results reinforce the validity of the binning method in this setting and justify our strategy of considering each pixel as a data point. (2) The binning strategy, supported by our theoretical calibration error bound, outperforms the other *post-hoc* calibration baselines. In the marginal setting, the binning approach achieves superior performance relative to most of the competing approaches. Isotonic regression, beta calibration and logistic scaling exhibit comparable behavior on most datasets. Temperature scaling, selective scaling and meta-cal, by contrast, yields unsatisfactory calibration performance on several datasets. In the conditional setting, the advantage of our method becomes even more pronounced. Specifically, the binning approach is the only method capable of attaining low calibration error (<0.1) with high probability ($>90\%$). Furthermore, we observe that increasing the number of bins B tends to slightly reduce the performance of the approach. This can be attributed to the fact that, as B grows, fewer points fall within each individual bin, leading to less reliable statistics per bin and thus noisier calibration estimates. However, using a larger number of bins also confers the key advantage of more localized and sharper calibration. Ultimately, the optimal number of bins is problem-dependent, and the right balance between statistical reliability and calibration sharpness will vary across applications and datasets.

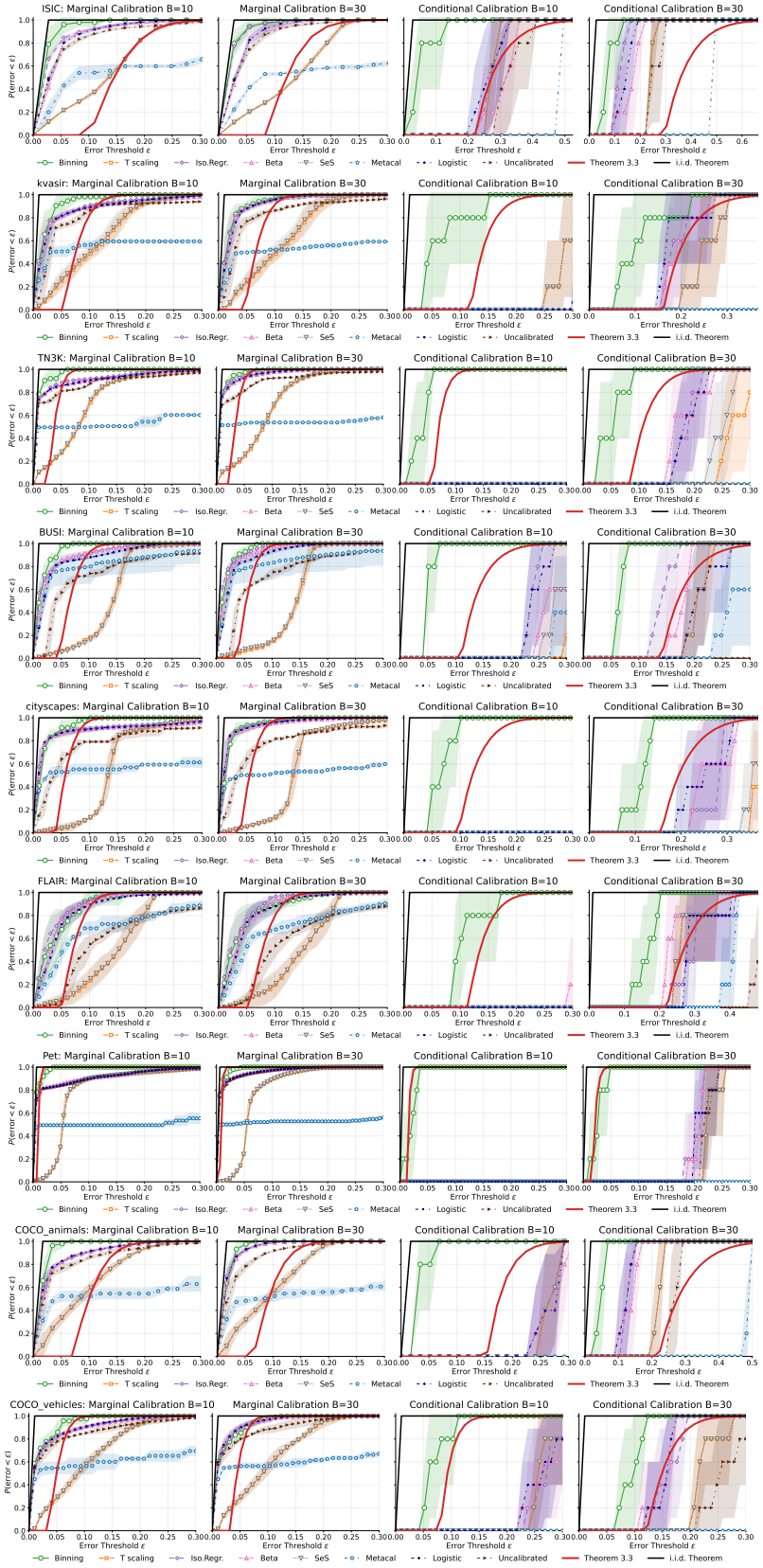

Figure 2: Comparisons of validity curves $V_m(\varepsilon), V_c(\varepsilon)$ (in red) and empirical ones $\hat{V}_m(\varepsilon), \hat{V}_c(\varepsilon)$ across methods and datasets.

6 Conclusions and Limitations

Motivated by the need for calibration guarantees in the context of segmentation, we investigated this problem and obtained the following results. (1) We derived an improved theoretical bound for segmentation-like data using concentration inequalities. Specifically, we model the segmentation task as a Bernoulli random field, explicitly incorporating the variance into the bound, which leads to a tighter guarantee. (2) We revisited the binning approach for binary segmentation calibration, constructing a calibrator that satisfies both marginal and conditional calibration properties, in accordance with the aforementioned theoretical bound. Notably, we formally justified treating each pixel as an individual data point even in the absence of *i.i.d.* assumptions, showing that this choice not only admits rigorous guarantees but also proves effective in practice. In fact, the binning approach consistently achieves superior performance compared to the baselines, particularly in the conditional setting, where most alternative baselines fail to reach comparable levels of calibration across the majority of datasets.

The primary limitations of the present work are as follows. First, our analysis is restricted to binary segmentation; an important direction for future research is the extension of the theoretical framework to multi-class segmentation. Second, while our theoretical justification for a binning approach in the segmentation setting remains valid and supported by the experiments, the bound requires empirical estimation, meaning that the guarantee is evaluated approximately rather than exactly. Finally, we are providing pixel-wise guarantees; a natural question is whether it is possible to bound the calibration error conditioned on specific regions of the image.

References

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020.

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

Theodore Barfoot, Luis C Garcia Peraza Herrera, Ben Glocker, and Tom Vercauteren. Average Calibration Error: A Differentiable Loss for Improved Reliability in Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 139–149. Springer, 2024.

Theodore Barfoot, Luis C Garcia-Peraza-Herrera, Samet Akcay, Ben Glocker, and Tom Vercauteren. Average Calibration Losses for Reliable Uncertainty in Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 2026.

Jacqueline Isabel Bereska, Hamed Karimi, and Reza Samavi. SACP: Spatially-Adaptive Conformal Prediction in Uncertainty Quantification of Medical Image Segmentation. In *Medical Imaging with Deep Learning*, 2025.

Ondrej Bohdal, Yongxin Yang, and Timothy Hospedales. Meta-Calibration: Learning of Model Calibration Using Differentiable Expected Calibration Error. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 02 2013. ISBN 9780199535255.

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*, pp. 4135–4143, 2024.

J-R Chazottes, Pierre Collet, Christof Külske, and Frank Redig. Concentration inequalities for random fields via coupling. *Probability Theory and Related Fields*, 137(1):201–225, 2007.

Jean-René Chazottes et al. Finitary Coding and Gaussian Concentration for Random Fields. *arXiv preprint arXiv:2602.14618*, 2026.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 801–818, 2018.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Scharwächter, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset. In *CVPR Workshop on The Future of Datasets in Vision*, 2015.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

A Philip Dawid. The Well-Calibrated Bayesian. *Journal of the American statistical Association*, 77(379): 605–610, 1982.

Zhipeng Ding, Xu Han, Peirong Liu, and Marc Niethammer. Local Temperature Scaling for Probability Calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6889–6899, 2021.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pp. 100–118, 1975.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *International Conference on Machine Learning*, pp. 1050–1059. PMLR, 2016.

Anatol Garioud, Stéphane Peillet, Eva Bookjans, Sébastien Giordano, and Boris Wattrelos. FLAIR #1: semantic segmentation and domain adaptation dataset. 2022. doi: 10.13140/RG.2.2.30183.73128/1.

Josselin Garnier and Xavier Langhendries. Concentration Inequalities for Non-causal Random Fields. *Electronic Journal of Statistics*, 16:1–44, 2022.

Paul Glasserman. *Monte Carlo Methods in Financial Engineering*, volume 53. Springer, 2004.

Haifan Gong, Jiaxin Chen, Guanqi Chen, Haofeng Li, Guanbin Li, and Fei Chen. Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. *Computers in Biology and Medicine*, 155:106389, 2023. ISSN 0010-4825.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

Chirag Gupta and Aaditya Ramdas. Distribution-free calibration guarantees for histogram binning without sample splitting. In *International Conference on Machine Learning*, pp. 3942–3952. PMLR, 2021.

Chirag Gupta, Aleksandr Podkopaev, and Aaditya Ramdas. Distribution-free binary classification: prediction sets, confidence intervals and calibration. *Advances in Neural Information Processing Systems*, 33: 3711–3723, 2020.

David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC). *arXiv preprint arXiv:1605.01397*, 2016.

Ramya Hebbalaguppe, Jatin Prakash, Neelabh Madan, and Chetan Arora. A Stitch in Time Saves Nine: A Train-Time Regularizing Loss for Improved Neural Network Calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16081–16090, 2022.

Wassily Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. ISSN 01621459, 1537274X.

Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? *Advances in Neural Information Processing Systems*, 30, 2017.

Leonid Kontorovich and Kavita Ramanan. Concentration inequalities for dependent random variables via the martingale method. 2008.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration. *Advances in Neural Information Processing Systems*, 32, 2019.

Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable Calibration Measures For Neural Networks From Kernel Mean Embeddings. In *International Conference on Machine Learning*, pp. 2805–2814. PMLR, 2018.

Fabian Küppers, Anselm Haselhoff, Jan Kronenberger, and Jonas Schneider. Confidence Calibration for Object Detection and Segmentation. In *Deep Neural Networks and Data for Automated Driving: Robustness, Uncertainty Quantification, and Insights Towards Safety*, pp. 225–250. Springer, 2022.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.

Stan Z. Li. *Markov Random Field Modeling in Image Analysis*. Springer, 3rd edition, 2009.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Computer Vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755. Springer, 2014.

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*, pp. 2980–2988, 2017.

Torgny Lindvall. *Lectures on the coupling method*. Courier Corporation, 2002.

Kangdao Liu, Tianhao Sun, Hao Zeng, Yongshan Zhang, Chi-Man Pun, and Chi-Man Vong. Spatial-Aware Conformal Prediction for Trustworthy Hyperspectral Image Classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Xingchen Ma and Matthew B Blaschko. Meta-Cal: Well-controlled post-hoc calibration by ranking. In *International Conference on Machine Learning*, pp. 7235–7245. PMLR, 2021.

Katalin Marton. Measure Concentration for a Class of Random Processes. *Probability Theory and Related Fields*, 110:427–439, 1998.

Katalin Marton. Measure Concentration and Strong Mixing. *Studia Scientiarum Mathematicarum Hungarica*, 40:95–113, 2003.

Colin McDiarmid et al. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

Alireza Mehrtash, William M Wells, Clare M Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 39(12):3868–3878, 2020.

Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte Carlo Gradient Estimation in Machine Learning. *Journal of Machine Learning Research*, 21(132):1–62, 2020.

Mehryar Mohri and Afshin Rostamizadeh. Stability Bounds for Stationary φ-mixing and β-mixing Processes. *Journal of Machine Learning Research*, 11(2), 2010.

Luca Mossina and Corentin Friedrich. Conformal Prediction for Image Segmentation Using Morphological Prediction Sets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 78–88. Springer, 2025.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating Deep Neural Networks using Focal Loss. *Advances in Neural Information Processing Systems*, 33:15288–15299, 2020.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When Does Label Smoothing Help? *Advances in Neural Information Processing Systems*, 32, 2019.

Balamurali Murugesan, Sukesh Adiga Vasudeva, Bingyuan Liu, Herve Lombaert, Ismail Ben Ayed, and Jose Dolz. Trust your neighbours: Penalty-based constraints for model calibration . In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 572–581. Springer, 2023.

Balamurali Murugesan, Sukesh Adiga Vasudeva, Bingyuan Liu, Herve Lombaert, Ismail Ben Ayed, and Jose Dolz. Neighbor-Aware Calibration of Segmentation Networks with Penalty-Based Constraints . *Medical Image Analysis*, 101:103501, 2025.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. *Advances in Neural Information Processing Systems*, 32, 2019.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and Dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

John Platt et al. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in Large Margin Classifiers*, 10(3):61–74, 1999.

Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, MMSys'17, pp. 164–169, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5002-0. doi: 10.1145/3083187.3083212.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.

Axel-Jan Rousseau, Thijs Becker, Simon Appeltans, Matthew Blaschko, and Dirk Valkenborg. Post hoc calibration of medical segmentation models. *Discover Applied Sciences*, 7(3):180, 2025.

Paul-Marie Samson. Concentration of Measure Inequalities for Markov Chains and ϕ-mixing Processes. *The Annals of Probability*, 28:416–461, 2000.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

Jarin Tasnim and Md Kamrul Hasan. CAM-QUS guided self-tuning modular CNNs with multi-loss functions for fully automated breast lesion classification in ultrasound images. *Physics in Medicine & Biology*, 69 (1):015018, 2024.

Jiaye Teng, Chuan Wen, Dinghuai Zhang, Yoshua Bengio, Yang Gao, and Yang Yuan. Predictive Inference with Feature Conformal Prediction. In *The Eleventh International Conference on Learning Representations*, 2023.

Aaron Van Den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional Image Generation with PixelCNN Decoders. *Advances in Neural Information Processing Systems*, 29, 2016a.

Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel Recurrent Neural Networks. In *International Conference on Machine Learning*, pp. 1747–1756. PMLR, 2016b.

Bruno Viti, Elias Karabelas, and Martin Holler. CONSIGN: Conformal Segmentation Informed by Spatial Groupings via Decomposition. In *The Fourteenth International Conference on Learning Representations*, 2026.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Dongdong Wang, Boqing Gong, and Liqiang Wang. On calibrating semantic segmentation models: Analyses and an algorithm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23652–23662, 2023.

HuaJun Xi, Jianguo Huang, Kangdao Liu, Lei Feng, and Hongxin Wei. Does confidence calibration improve conformal prediction? *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pp. 94–116, 1994.

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*, pp. 6023–6032, 2019.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *ICML*, pp. 2001, 2001.

Bianca Zadrozny and Charles Elkan. Transforming Classifier Scores into Accurate Multiclass Probability Estimates. In *Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 694–699, 2002.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*, 2018.

Kato Zoltan and Zerubia Josiane. Markov random fields in image segmentation. *Foundations and Trends in Signal Processing*, 5(1-2):1–155, 2012.

A Proofs

A.1 Derivation of the martingale increment identity

For a Bernoulli alphabet $A = \{0, 1\}$, the martingale increment, using the law of total expectation, reduces to:

$$\begin{aligned} V_i(\sigma) &= \mathbb{E}[g \mid \sigma_{<i}, \sigma_i] - \big[p_i \mathbb{E}[g \mid \sigma_{<i},\, \sigma_i = 1] + (1 - p_i)\mathbb{E}[g \mid \sigma_{<i},\, \sigma_i = 0]\big] \\ &= \begin{cases} (1 - p_i)\Delta_i & \text{if } \sigma_i = 1, \\ -p_i\, \Delta_i & \text{if } \sigma_i = 0. \end{cases} \end{aligned}$$

Both cases are captured by the identity $V_i(\sigma) = (\sigma_i - p_i)\, \Delta_i$ of equation 2.

A.2 Proof of Lemma 2.1

Proof. We first notice that if $\Delta_i = 0$ both sides equal zero and the bound holds trivially. In the general case, fix $\mathcal{F}_{i-1}$ and recall p_i, Δ_i are $\mathcal{F}_{i-1}$-measurable. Given $\mathcal{F}_{i-1}$, $W_i := \frac{V_i}{\Delta_i} = \sigma_i - p_i$ is a centered two-point random variable with:

$$\mathbb{E}[W_i \mid \mathcal{F}_{i-1}] = 0, \quad W_i \leq 1 - p_i \leq 1 =: b, \quad \mathbb{E}[W_i^2 \mid \mathcal{F}_{i-1}] = p_i(1-p_i) =: v.$$

Apply Theorem 2.9 of (Boucheron et al., 2013) with $n = 1$, $b = 1$, $v = p_i(1-p_i)$, and parameter $\lambda\Delta_i$ in place of λ:

$$\log \mathbb{E}\big[e^{\lambda \Delta_i W_i} \mid \mathcal{F}_{i-1}\big] \;\leq\; \frac{v}{b^2}\,\phi(b \cdot \lambda\Delta_i) \;=\; p_i(1-p_i)\,\phi(\lambda\Delta_i).$$

Since $V_i = \Delta_i W_i$, this is exactly equation 3. We use $b = 1$ rather than the tighter $b = 1 - p_i$ to keep the bound symmetric in p_i. □

A.3 Derivation of the Δ_i bound

By Chazottes et al. (2007, Eq. 5), $\Delta_i(\sigma) \leq (D^\sigma \delta g)_i$, where the matrix-vector product is

$$(D^\sigma \delta g)_i := \sum_{j>i}^{n} D^\sigma_{i,j}\,\delta_j g.$$

Since $\delta_j g = 1/n$ for all j equation 4, this simplifies to

$$(D^\sigma\,\delta g)_i = \frac{1}{n}\sum_{j>i}^{n} D^\sigma_{i,j} =: \frac{R^\sigma_i}{n},$$

where R^σ_i denotes the i-th row sum of D^σ. Taking the supremum over all configurations and setting $\bar{D}_{i,j} := \sup_\sigma D^\sigma_{i,j}$ yields

$$\Delta_i \leq (D^\sigma\,\delta g)_i = \frac{1}{n}\sum_{j>i}^{n} D^\sigma_{i,j} \leq \sup_\sigma \frac{1}{n}\sum_{j>i}^{n} D^\sigma_{i,j} = \frac{1}{n}\sum_{j>i}^{n} \bar{D}_{i,j} = \frac{R_i}{n}.$$

A.4 Proof of Theorem 2.2

Proof. From equation 1 and equation 2 we have that:

$$g - \mathbb{E}g = \sum_{i=1}^{n} V_i, \qquad V_i = (\sigma_i - p_i)\,\Delta_i,$$

where both p_i, Δ_i are $\mathcal{F}_{i-1}$-measurable. We apply the exponential Chebyshev inequality, and obtain that for any $\lambda > 0$:

$$\mathbb{P}(g - \mathbb{E}g \geq \varepsilon) \;\leq\; e^{-\lambda\varepsilon}\,\mathbb{E}\Big[e^{\lambda \sum_{i=1}^{n} V_i}\Big]. \tag{14}$$

Focusing on the expectation, we peel off one term using the tower property:

$$\mathbb{E}\Big[e^{\lambda \sum_i V_i}\Big] = \mathbb{E}\Big[\mathbb{E}\big[e^{\lambda V_n} \mid \mathcal{F}_{n-1}\big]\, e^{\lambda \sum_{i=1}^{n-1} V_i}\Big].$$

Now we apply Lemma 2.1 to bound $\mathbb{E}[e^{\lambda V_n} \mid \mathcal{F}_{n-1}]$. Notice that g defined in equation 4 is increasing in σ, therefore $\Delta_i > 0$ and Lemma 2.1 can be applied. We repeat for each i:

$$\log \mathbb{E}\Big[e^{\lambda(g - \mathbb{E}g)}\Big] = \log \mathbb{E}\Big[e^{\lambda \sum_i V_i}\Big] \;\leq\; \sum_{i=1}^{n} p_i(1-p_i)\,\phi(\lambda\Delta_i). \tag{15}$$

For $g = n^{-1}\sum_{i=1}^{n}\sigma_i$, we have $\delta_j g = 1/n$ for all j, therefore, using equation 5 we have $\Delta_i \le (\bar{D}\,\delta g)_i = R_i/n$. Since ϕ is increasing on $[0,\infty)$:

$$\sum_{i=1}^{n} p_i(1-p_i)\,\phi(\lambda\Delta_i) \;\le\; \sum_{i=1}^{n} p_i(1-p_i)\,\phi\Big(\frac{\lambda R_i}{n}\Big).$$

It is possible to show, using Taylor expansions, that for $0 \le u < 3$, the inequality

$$\phi(u) \;\le\; \frac{u^2/2}{1-u/3} \tag{16}$$

holds. Restricting to $\lambda < 3n/R_{\max}$ we ensure $\lambda R_i/n \le \lambda R_{\max}/n < 3$ for all i. We apply equation 16 with $u = \lambda R_i/n$:

$$\sum_{i=1}^{n} p_i(1-p_i)\,\phi\Big(\frac{\lambda R_i}{n}\Big) \;\le\; \frac{\lambda^2}{2}\sum_{i=1}^{n}\frac{p_i(1-p_i)\,R_i^2/n^2}{1-\lambda R_i/(3n)}.$$

Now, we bound the denominator uniformly by its worst case, $1-\lambda R_{\max}/(3n)$, and obtain:

$$\log\mathbb{E}\Big[e^{\lambda(g-\mathbb{E}g)}\Big] \;\le\; \frac{\lambda^2 S/2}{1-\lambda c} \;=:\; \psi(\lambda), \qquad for \quad 0<\lambda<1/c, \tag{17}$$

with S and c as defined in equation 6. (This is the sub-gamma form $\psi(\lambda) = \lambda^2 S/(2(1-c\lambda))$ of Boucheron et al. (2013, Section 2.4).) We plug in equation 17 in equation 14 and obtain:

$$\mathbb{P}(g-\mathbb{E}g \ge \varepsilon) \;\le\; e^{-\lambda\varepsilon}e^{\frac{\lambda^2 S/2}{1-\lambda c}}, \qquad for \quad 0<\lambda<1/c. \tag{18}$$

In order to make the right-hand side minimal, we derive by λ and obtain:

$$\frac{d}{d\lambda}\Big(-\lambda\varepsilon + \frac{\lambda^2 S/2}{1-\lambda c}\Big) = -\varepsilon + \frac{\lambda S(2-\lambda c)}{2(1-\lambda c)^2}$$

Setting the derivative to zero gives a quadratic equation in λ, selecting the λ that satisfies $\lambda < \frac{1}{c}$ gives us:

$$\lambda^* = \frac{1}{c}\Big(1-\sqrt{\frac{S}{S+2\varepsilon c}}\Big)$$

By substituting λ^* in equation 18 we get the following bound:

$$\log\mathbb{P}(g-\mathbb{E}g \ge \varepsilon) \;\le\; -\frac{S}{c^2}\eta\Big(\frac{c\varepsilon}{S}\Big), \quad \eta(x) = 1+x-\sqrt{1+2x}$$

The two-sided bound equation 7 follows selecting $-g$. In this case, $-g$ is decreasing in σ, therefore Δ_i has to be defined as $\Delta_i = \mathbb{E}[g \mid \sigma_{<i}, \sigma_i = 0] - \mathbb{E}[g \mid \sigma_{<i}, \sigma_i = 1]$. Notice that using $-g$ swaps p_i and $1-p_i$ but leaves S and c unchanged since $p_i(1-p_i)$ is symmetric. □

A.5 Extension of Theorem 2.2 to $\mathbb{Z}^2$

To extend Theorem 2.2 to $\mathbb{Z}^2$ we follow the same strategy of (Chazottes et al., 2007, Section 2.4). Let $\Omega \subset \mathbb{Z}^2$ and $\Lambda = \{0,1\}^{\Omega}$. Introduce any enumeration $\Gamma : \mathbb{Z}^2 \to \mathbb{N}$. We use the notation $(\le x) = \{y \in \mathbb{Z}^2 : \Gamma(y) \le \Gamma(x)\}$, which is the $\mathbb{Z}^2$ equivalent of $(\le i)$ in the $\mathbb{N}$ case of Theorem 2.2. Therefore, $\mathcal{F}_{\le x}$ denotes the sigma-field generated by $\sigma_y, y \le x$ and $\mathcal{F}_0$ is the trivial sigma-field. Analogously to equation 1 we have the following decomposition:

$$g(\sigma) - \mathbb{E}[g] = \sum_{x\in\Omega} V_x(\sigma), \qquad \text{where} \quad V_x(\sigma) := \mathbb{E}[g|\mathcal{F}_{\le x}] - \mathbb{E}[g|\mathcal{F}_{<x}]. \tag{19}$$

Finally we introduce the coupling matrices as

$$D^{\sigma}_{x,y} := \mathbb{P}^{\sigma_{<x}}_{x,1,0}\Big(\sigma^{(1)}_y \ne \sigma^{(2)}_y\Big), \quad \bar{D}_{x,y} = \sup_{\sigma} D^{\sigma}_{x,y}, \tag{20}$$

where $\mathbb{P}^{\sigma}_{x,1,0}$ denotes the maximal coupling between the conditional measures $\mathbb{P}(\cdot|\sigma_{<x}, \sigma_x = 1)$ and $\mathbb{P}(\cdot|\sigma_{<x}, \sigma_x = 0)$.

Theorem A.1 (Improved concentration for Bernoulli mean in $\mathbb{Z}^2$)**.** *Let $\Omega \subset \mathbb{Z}^2$ and let $g = |\Omega|^{-1} \sum_{x\in\Omega} \sigma_x$. Define $\bar{D}_{x,y}$ as in equation 20, $R_x := \sum_{y>x} \bar{D}_{x,y}$, and $p_x := \mathbb{P}(\sigma_x = 1 \mid \sigma_{<x})$. Set*

$$S \;:=\; \frac{1}{|\Omega|^2} \sum_{x\in\Omega} p_x(1-p_x)\, R_x^2, \qquad R_{\max} := \max_{x\in\Omega} R_x, \quad c \;:=\; \frac{R_{\max}}{3|\Omega|}. \tag{21}$$

Then for all $\varepsilon > 0$:

$$\boxed{\mathbb{P}(|g - \mathbb{E}[g]| \geq \varepsilon) \;\leq\; 2\exp\!\left(-\frac{S}{c^2}\eta\!\left(\frac{c\varepsilon}{S}\right)\right)} \qquad \textit{with} \quad \eta(x) = 1 + x - \sqrt{1+2x}. \tag{22}$$

Proof. The enumeration Γ maps the indices of $\mathbb{Z}^2$ to $\mathbb{N}$. Therefore, we can apply the proof of Theorem 2.2 and obtain the bound equation 22. □

A.6 Proofs of Calibration Theorems

Proof of Theorem 3.1. Fix a bin $b \in \{1, \ldots, B\}$. The observed frequency can be written as

$$\hat{\pi}_b = \frac{1}{N_b} \sum_{i=1}^{N_{cal}} \sum_{x_b \in \Omega_b^{(i)}} Y_{x_b}^{(i)},$$

which has the form $g = N_b^{-1} \sum_i \sigma_i$, i.e. the empirical mean of N_b Bernoulli variables indexed over the disjoint domains $\Omega_b^{(1)}, \ldots, \Omega_b^{(N_{cal})}$. Moreover, since the binning event $E_{\mathcal{B}}$ is fixed, N_b is deterministic and the randomness is over the Bernoulli variables.

Since images are drawn independently, pixels from different images are independent. Therefore the coupling matrix $\bar{D}$ of the full collection of N_b pixels is block-diagonal, where each block $\bar{D}^{(i)}$ correspond to the coupling matrix of image i conditioned on the binning event $E_{\mathcal{B}}$ (i.e. the collection $\{Y_{x_b}^{(i)}\}_{x_b \in \Omega_b^{(i)}}$). $\bar{D}_{x_b,y_b}$ is defined as in Section 2.2 for pixels x_b, y_b within the same image i, and $\bar{D}_{x_b,y_b} = 0$ for pixels from different images. As a consequence, the row sums of the global coupling matrix reduce to the within-image row sums:

$$R_{x_b}^{(i)} = \sum_{y_b > x_b} \bar{D}_{x_b,y_b}^{(i)},$$

since all cross-image terms vanish. The quantities S_b and c_b are computed by substituting into equation 21:

$$p_{x_b}^{(i)} = \mathbb{P}(Y_{x_b}^{(i)} \mid Y_{<x_b}^{(i)}, E_{\mathcal{B}}), \quad S_b = \frac{1}{N_b^2} \sum_{i=1}^{N_{cal}} \sum_{x_b \in \Omega_b^{(i)}} p_{x_b}^{(i)}(1 - p_{x_b}^{(i)})\,(R_{x_b}^{(i)})^2, \qquad c_b = \frac{R_{\max,b}}{3N_b},$$

with $R_{\max,b} = \max_{i=1}^{N_{cal}} \max_{x\in\Omega_b^{(i)}} R_{x_b}^{(i)}$.

We can now apply Theorem A.1, conditioned on the binning event $E_{\mathcal{B}}$ to fix N_b, with $\sigma = \{Y_{x_b}^{(i)}\}_{i=1, x_b \in \Omega_b^{(i)}}^{N_{cal}}$, $|\Omega| = N_b$, $g = \hat{\pi}_b$:

$$\mathbb{P}(|\hat{\pi}_b - \mathbb{E}[\hat{\pi}_b \mid E_{\mathcal{B}}]| \geq \varepsilon \mid E_{\mathcal{B}}) \leq 2\exp\!\left(-\frac{S_b}{c_b^2}\,\eta\!\left(\frac{c_b\,\varepsilon}{S_b}\right)\right).$$

Marginalising over all the binning events we get

$$\mathbb{P}(|\hat{\pi}_b - \mathbb{E}[\hat{\pi}_b \mid E_{\mathcal{B}}]| \geq \varepsilon) \leq 2\exp\!\left(-\frac{S_b}{c_b^2}\,\eta\!\left(\frac{c_b\,\varepsilon}{S_b}\right)\right).$$ □

Proof of Theorem 3.2. By a union bound over the B bins:

$$\mathbb{P}(\exists\, b \in [B] : |\hat{\pi}_b - \mathbb{E}[\hat{\pi}_b \mid E_{\mathcal{B}}]| \geq \varepsilon) \leq \sum_{b=1}^{B} \mathbb{P}(|\hat{\pi}_b - \mathbb{E}[\hat{\pi}_b \mid E_{\mathcal{B}}]| \geq \varepsilon)$$
$$\leq \sum_{b=1}^{B} 2\exp\left(-\frac{S_b}{c_b^2}\,\eta\left(\frac{c_b\,\varepsilon}{S_b}\right)\right).$$

□

Proof of Theorem 3.3. Marginal calibration. We start by noticing that conditioning on the event, $E_{\mathcal{B}} := \{f(X^{(1)})_x \in I_b, f(X^{(1)})_z \in I_b, \dots, f(X^{(N_{cal})})_x \in I_b, \dots\}$, is not equivalent to conditioning on the event $E_h := \{h(X^{(1)})_x, \dots, h(X^{(N_{cal})})_x, \dots\}$. The two events coincide only if $\hat{\pi}_{b'} \neq \hat{\pi}_b\ \forall b' \neq b$. In that case knowing that pixel x has an associated calibrated predictor $h(X)_x$, and therefore $\hat{\pi}_b$, automatically tells that $f(X)_x \in I_b$, since $\hat{\pi}_b$ is different for each bin b. As introduced in Gupta & Ramdas (2021), a simple trick to not lose the informativeness of h is to randomise it with an arbitrary small value. More formally, define the new calibrated predictor $\tilde{h}(X)_x$ as

$$\tilde{h}(X)_x = \sum_{b=1}^{B} \mathbf{1}_{\{f(X)_x \in I_b\}}(\hat{\pi}_b + \delta U_b), \quad U_b \sim \mathcal{U}[-1, 1],$$

where δ is an arbitrary parameter. If we choose, for example, $\delta = 10^{-20}$, then the new randomized calibrated predictor is in practice indistinguishable from the old one, but, we can now write that $\mathbb{E}[\hat{\pi}_b \mid E_{\mathcal{B}}] = \mathbb{E}[\hat{\pi}_b \mid E_{\tilde{h}}]$. In the following, we denote as $\tilde{h}(X)$ the union of all of the pixel-wise events $\tilde{h}(X)_x\ \forall x \in \Omega$.

We proceed by computing the conditional expected value of $\hat{\pi}_b$ as

$$\begin{aligned}
\mathbb{E}[\hat{\pi}_b \mid E_{\mathcal{B}}] &= \frac{1}{N_b}\mathbb{E}\Bigg[\sum_{i=1}^{N_{cal}}\sum_{x\in\Omega} Y_x^{(i)}\mathbf{1}_{\{f(X^{(i)})_x \in I_b\}} \mid \tilde{h}(X^{(1)}), \dots, \tilde{h}(X^{(N_{cal})})\Bigg] \\
&= \frac{1}{N_b}\sum_{i=1}^{N_{cal}}\sum_{x\in\Omega} \mathbf{1}_{\{f(X^{(i)})_x \in I_b\}}\mathbb{E}[Y_x^{(i)} \mid \tilde{h}(X^{(1)}), \dots, \tilde{h}(X^{(N_{cal})})] \\
&= \frac{1}{N_b}\sum_{i=1}^{N_{cal}}\sum_{x_b\in\Omega_b^{(i)}} \mathbb{E}[Y_{x_b}^{(i)} \mid \tilde{h}(X^{(i)})] && \left(\{X^{(i)}\}_{i=1}^{N_{cal}}\ i.i.d.\right) \\
&= \frac{1}{N_b}\sum_{i=1}^{N_{cal}}\sum_{x_b\in\Omega_b^{(i)}} \mathbb{E}[Y_{x_b}^{(i)} \mid \tilde{h}(X^{(i)})_{x_b}] && \left(Y_{x_b}^{(i)} \mid \tilde{h}(X^{(i)})_{x_b} \perp\!\!\!\perp \tilde{h}(X^{(i)})_{y_b},\ y_b \neq x_b\right) \\
&= \frac{1}{N_b}\sum_{i=1}^{N_{cal}}\sum_{x_b\in\Omega_b^{(i)}} \mathbb{E}[Y_{x_b} \mid \tilde{h}(X)_{x_b}] && \left(\{X^{(i)}\}_{i=1}^{N_{cal}}\ i.i.d.\right) \\
&= \mathbb{E}[Y_{x_b} \mid \tilde{h}(X)_{x_b}]\frac{1}{N_b}\sum_{i=1}^{N_{cal}}\left|\Omega_b^{(i)}\right| && \left(\mathbb{E}[Y_{x_b} \mid \tilde{h}(X)_{x_b}] = \mathbb{E}[Y_{y_b} \mid \tilde{h}(X)_{y_b}]\ \forall x_b, y_b\right) \\
&= \mathbb{E}[Y_{x_b} \mid \tilde{h}(X)_{x_b}] && \left(\sum_{i=1}^{N_{cal}}\left|\Omega_b^{(i)}\right| = N_b\right)
\end{aligned}$$

Since the derivation holds for each bin b, we drop the x_b notation and we continue for a generic pixel x. Substituting $\mathbb{E}[Y_x \mid \tilde{h}(X)_x]$ in Theorem 3.1 we obtain:

$$\mathbb{P}\left(\left|\hat{\pi}_b - \mathbb{E}[Y_x \mid \tilde{h}(X)_x]\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{S_b}{c_b^2}\,\eta\left(\frac{c_b\,\varepsilon}{S_b}\right)\right).$$

We notice that by definition, for a fixed bin b, $\hat{\pi}_b + \delta U_b = \tilde{h}(X)_x$ and therefore

$$\left|\tilde{h}(X)_x - \mathbb{E}[Y_x \mid \tilde{h}(X)_x]\right| = \left|\hat{\pi}_b + \delta U_b - \mathbb{E}[Y_x \mid \tilde{h}(X)_x]\right|.$$

Using the triangle inequality and the definition of U_b we get

$$\left|\tilde{h}(X)_x - \mathbb{E}[Y_x \mid \tilde{h}(X)_x]\right| \leq \left|\hat{\pi}_b - \mathbb{E}[Y_x \mid \tilde{h}(X)_x]\right| + \delta.$$

We proceed by noticing that

$$\left|\tilde{h}(X)_x - \mathbb{E}[Y_x \mid \tilde{h}(X)_x]\right| \geq \varepsilon \implies \left|\hat{\pi}_b - \mathbb{E}[Y_x \mid \tilde{h}(X)_x]\right| \geq \varepsilon - \delta,$$

and therefore

$$\begin{aligned}\mathbb{P}\Big(\left|\tilde{h}(X)_x - \mathbb{E}[Y_x \mid \tilde{h}(X)_x]\right| \geq \varepsilon\Big) &\leq \mathbb{P}\Big(\left|\hat{\pi}_b - \mathbb{E}[Y_x \mid \tilde{h}(X)_x]\right| \geq \varepsilon - \delta\Big) \\ &\leq 2\exp\left(-\frac{S_b}{c_b^2}\,\eta\left(\frac{c_b\,(\varepsilon-\delta)}{S_b}\right)\right).\end{aligned}$$

Taking the complement we obtain the final result

$$\mathbb{P}\Big(\left|\mathbb{E}[Y_x \mid \tilde{h}(X)_x] - \tilde{h}(X)_x\right| \leq \varepsilon\Big) \geq 1 - 2\exp\left(-\frac{S_b}{c_b^2}\,\eta\left(\frac{c_b(\varepsilon-\delta)}{S_b}\right)\right) = 1 - \alpha_{1,b}.$$

Conditional calibration. If we take the complement of Theorem 3.2 we obtain

$$\mathbb{P}(\forall\, b \in [B] : |\hat{\pi}_b - \mathbb{E}[\hat{\pi}_b \mid E_{\mathcal{B}}]| \leq \varepsilon) \geq 1 - 2\sum_{b=1}^{B}\exp\left(-\frac{S_b}{c_b^2}\,\eta\left(\frac{c_b\varepsilon}{S_b}\right)\right).$$

Moreover, by the same argument as in the marginal case

$$\mathbb{P}\Big(\forall\, b \in [B] : \left|\tilde{h}(X)_x - \mathbb{E}[Y_x \mid \tilde{h}(X)_x]\right| \leq \varepsilon\Big) \geq 1 - 2\sum_{b=1}^{B}\exp\left(-\frac{S_b}{c_b^2}\,\eta\left(\frac{c_b(\varepsilon-\delta)}{S_b}\right)\right).$$

After the randomization $\tilde{h}(X)_x = \hat{\pi}_b + \delta U_b$ introduced above, the values $\hat{\pi}_1 + \delta U_1, \ldots, \hat{\pi}_B + \delta U_B$ are pairwise distinct almost surely, so the range of $\tilde{h}(X)_x$ coincides with $\{\hat{\pi}_b + \delta U_b\}_{b=1}^{B}$ and to each r in this range corresponds a unique bin $b(r) \in [B]$ with $\hat{\pi}_{b(r)} + \delta U_{b(r)} = r$. Then

$$\mathbb{P}\Big(\forall\, r \in \mathit{Range}(h) : \left|\mathbb{E}[Y_x \mid \tilde{h}(X)_x = r] - r\right| \leq \varepsilon\Big) \geq 1 - 2\sum_{b=1}^{B}\exp\left(-\frac{S_b}{c_b^2}\,\eta\left(\frac{c_b(\varepsilon-\delta)}{S_b}\right)\right) = 1 - \alpha_2. \qquad \square$$

B Baseline Methods

We now describe the baseline methods. For all the *post-hoc* baselines (and our method), the spatial predictions (e.g. softmax and logits maps of shape $N \times H \times W$) are flattened before fitting, and each pixel is treated independently from the others.

B.1 Temperature Scaling

Temperature Scaling (Guo et al., 2017) is the simplest post-hoc calibration scheme. Given raw logits $\mathbf{z} \in \mathbb{R}^C$ produced by the network, the calibrated probability vector is

$$\hat{\mathbf{p}} = \mathrm{softmax}\left(\frac{\mathbf{z}}{T}\right), \tag{23}$$

where $T > 0$ is a single scalar temperature shared across all classes and all samples. For binary classification, equation 23 reduces to

$$\hat{p} = \sigma\Big(\frac{z}{T}\Big)\,, \tag{24}$$

where σ is the sigmoid function and z is the scalar logit. T is found by minimizing the cross-entropy on the calibration set:

$$T^* = \arg\min_{T>0} \frac{1}{N}\sum_{i=1}^{N} \ell_{\mathrm{CE}}(\mathrm{softmax}(\mathbf{z}_i/T)\,,\, y_i)\,. \tag{25}$$

In practice, over-confident networks yield $T^* > 1$, which softens the output distribution toward uniform. In the experiments, the temperature is log-parametrized as $T = e^{\tau}$ with $\tau \in \mathbb{R}$, ensuring positivity throughout optimization without any explicit constraint. The single parameter τ is optimized with L-BFGS using a strong Wolfe line search.

B.2 Logistic scaling

Logistic scaling (Platt et al., 1999) (also known as Platt scaling) learns an affine transformation of the raw logits followed by a sigmoid, producing calibrated probabilities via the map

$$\hat{p} = \sigma(a \cdot z + b), \tag{26}$$

where $z \in \mathbb{R}$ is the pre-sigmoid logit, $a > 0$ is a scale parameter, and $b \in \mathbb{R}$ is a bias term.

Logistic scaling generalises Temperature Scaling, which corresponds to the special case $a = 1/T$ and $b = 0$. Conversely, it is a special case of Beta Calibration (Kull et al., 2019), which uses the richer map $\hat{p} = \sigma(a\,\log p - b\,\log(1-p) + c)$ with independent parameters for $\log p$ and $\log(1-p)$.

Given calibration pairs $\{(z_i, y_i)\}_{i=1}^{N}$ with $y_i \in \{0, 1\}$, the parameters are found by minimising the binary cross-entropy:

$$a^*,\, b^* = \arg\min_{a>0,\, b\in\mathbb{R}} \frac{1}{N}\sum_{i=1}^{N} \ell_{\mathrm{CE}}\big(\sigma(a\, z_i + b),\, y_i\big). \tag{27}$$

In the experiments, the positivity constraint on a is enforced via a log-parameterisation $a = e^{\alpha}$, $\alpha \in \mathbb{R}$, and the two-parameter problem is solved with L-BFGS.

B.3 Beta Calibration

Beta Calibration (Kull et al., 2019) generalizes logistic (Platt) scaling for binary classifiers by exploiting the Beta family of distributions. The calibration map is

$$\hat{p} = \sigma\big(a \log p - b \log(1-p) + c\big), \tag{28}$$

where $a, b > 0$ are shape parameters and $c \in \mathbb{R}$ is a bias. Equation 28 can be derived as the ratio of two Beta densities evaluated at p. The three-parameter family subsumes temperature scaling ($a = b$, $c = 0$) and Platt scaling ($a = b$, c free) as special cases, providing strictly greater expressiveness. The parameters are optimized by minimizing binary cross-entropy on the calibration set:

$$(a^*, b^*, c^*) = \arg\min_{a,b>0,\, c} \frac{1}{N}\sum_{i=1}^{N} \ell_{\mathrm{BCE}}(\hat{p}_i,\, y_i)\,. \tag{29}$$

In the experiments, the positivity constraints on a and b are handled via log-parametrization ($a = e^{\alpha}$, $b = e^{\beta}$), reducing the constrained problem to an unconstrained three-parameter optimization over (α, β, c). L-BFGS with a strong Wolfe line search is used.

B.4 Isotonic Regression Calibration

Isotonic Regression Calibration (Zadrozny & Elkan, 2002) learns a non-parametric, piecewise-constant monotone mapping

$$f : [0, 1] \to [0, 1], \quad f \text{ non-decreasing}, \tag{30}$$

from raw confidence scores to calibrated probabilities. Given calibration pairs $\{(p_i, y_i)\}_{i=1}^{N}$ with $p_i \in [0, 1]$ and $y_i \in \{0, 1\}$, isotonic regression solves

$$\hat{f} = \arg \min_{f \text{ non-decreasing}} \sum_{i=1}^{N} \big(f(p_i) - y_i\big)^2, \tag{31}$$

via the pool adjacent violators (PAV) algorithm. Unlike Temperature Scaling and Beta Calibration, Isotonic Regression places no parametric assumptions on the shape of the miscalibration, making it more flexible but also more prone to overfitting on small calibration sets.

In the experiments, we use SCIKIT-LEARN's `IsotonicRegression` with `out_of_bounds="clip"`, which fits a piecewise-constant, non-decreasing mapping $f : p \mapsto \hat{p}$ on the flattened validation scores via the pool adjacent violators algorithm. At inference, the fitted mapping is applied element-wise, with input values outside the training range clipped to the nearest boundary.

B.5 Meta-cal

Meta-cal (Ma & Blaschko, 2021) is a post-hoc calibration method that combines temperature scaling with a confidence-aware sample selection strategy. The key idea is to avoid calibrating high-uncertainty predictions—which are unreliable by nature—and instead assign them the uninformative probability 0.5.

Step 1 – Entropy-based partitioning For each pixel with pre-sigmoid logit z, the binary entropy of the uncalibrated prediction $p = \sigma(z)$ is computed as

$$H(z) = -p \log p - (1 - p) \log(1 - p). \tag{32}$$

A threshold τ on this entropy separates the predictions into a *low-entropy* set $\mathcal{L} = \{i : H(z_i) < \tau\}$ and a *high-entropy* set $\mathcal{H} = \{i : H(z_i) \geq \tau\}$.

Step 2 – Threshold selection The threshold is chosen to match a target coverage accuracy a^*. A small held-out (10%) of the calibration data is binned by entropy quantiles, and for each bin the cumulative accuracy is computed. Isotonic regression (decreasing) is then fitted on the (cumulative accuracy, mean entropy) pairs; the threshold is obtained by evaluating the isotonic fit at the target accuracy:

$$\tau = \text{IsoReg}^{-1}(a^*). \tag{33}$$

Step 3 – Temperature scaling on low-entropy samples. On the remaining calibration data, only samples in $\mathcal{L}$ are used to fit a single temperature parameter $T > 0$ by minimising binary cross-entropy:

$$T^* = \arg \min_{T>0} \sum_{i \in \mathcal{L}} \ell_{\text{CE}}\big(\sigma(z_i/T),\, y_i\big). \tag{34}$$

Step 4 – Inference At test time, the calibrated probability is

$$\hat{p} = \begin{cases} \sigma(z/T^*) & \text{if } H(z) < \tau, \\ 0.5 & \text{if } H(z) \geq \tau. \end{cases} \tag{35}$$

By abstaining from calibrating high-entropy pixels, MetaCal avoids distorting predictions where the model is genuinely uncertain, which can improve calibration compared to global temperature scaling. In the experiments $a^* = 0.9$.

B.6 Selective Scaling

Selective Scaling (Wang et al., 2023) extends Meta-cal by learning two temperatures and a binary selector that routes each prediction to the appropriate one. The key observation is that correctly classified and misclassified pixels benefit from different degrees of rescaling.

Step 1 – Selector. A small MLP g_ϕ is trained on 10% of the calibration split to predict whether the network's hard prediction $\hat{y}_i$ matches the ground-truth y_i. Its input is the uncalibrated probability $p_i = \sigma(z_i)$ and its binary target is $c_i = \mathbf{1}[\hat{y}_i = y_i]$. After training, the selector partitions any set of pixels into two groups:

$$\mathcal{C} = \{i : g_\phi(p_i) = 1\}, \qquad \mathcal{M} = \{i : g_\phi(p_i) = 0\}. \tag{36}$$

Step 2 – Two temperatures. On the rest of the calibration split, two scalar temperatures $T_1 > 0$ (for predicted-incorrect pixels $\mathcal{M}$) and $T_2 > 0$ (for predicted-correct pixels $\mathcal{C}$) are jointly optimised by minimising the total cross-entropy:

$$T_1^*, T_2^* = \arg\min_{T_1, T_2 > 0} \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \ell_{\text{CE}}\big(\sigma(z_i/T_1),\, y_i\big) \;+\; \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \ell_{\text{CE}}\big(\sigma(z_i/T_2),\, y_i\big). \tag{37}$$

As in Temperature Scaling, both temperatures are log-parametrised ($T_k = e^{\tau_k}$, $\tau_k \in \mathbb{R}$) to ensure strict positivity, and the two-parameter problem is solved with L-BFGS.

Step 3 – Inference. At test time, the calibrated probability of a new pixel with logit z is

$$\hat{p} = \begin{cases} \sigma(z/T_1^*) & \text{if } g_\phi\big(\sigma(z)\big) = 0 \quad \text{(predicted incorrect)}, \\ \sigma(z/T_2^*) & \text{if } g_\phi\big(\sigma(z)\big) = 1 \quad \text{(predicted correct)}. \end{cases} \tag{38}$$

In practice, over-confident networks typically yield $T_1^* > 1$ (softening wrong predictions) and $T_2^* \leq 1$ (sharpening correct ones), producing a more reliable confidence signal than a single global temperature.

B.7 Effect of Calibration on Accuracy

While the main objective of the paper is to compare the calibration performances of different methods, it is natural to ask what is the effect of these calibration methods on the accuracy of the underlying segmentation model. To do that we are going to evaluate the DICE score difference, before and after calibration, on the datasets used in the experiments. The DICE score is defined as:

$$DICE = \frac{2|Y \bigcap \hat{Y}|}{|Y| + |\hat{Y}|},$$

where Y and $\hat{Y}$ are the ground-truth and predicted segmentation respectively. The mean is evaluated across random calibration/test splits. Tables 1–2 show that *post-hoc* calibration methods yield, on average, a

Table 1: Relative DICE score variation after calibration.

Methods	**Cityscapes**	**COCO-an.**	**COCO-veh.**	**TN3K**
Beta-cal	2.30%	0.22%	0.33%	0.05%
Iso.Regr.	1.55%	0.36%	0.32%	−0.55%
Meta-cal	−33.59%	−14.04%	−8.28%	−53.04%
Logistic	2.6%	0.36%	0.33%	−0.02%
Binning (B=10)	−1.51%	0.05%	−0.07%	−3.43%
Binning (B=30)	1.49%	0.36%	0.22%	−14.2%

Table 2: Relative DICE score variation after calibration.

Methods	**ISIC**	**Kvasir**	**BUSI**	**FLAIR1**	**Pet**
Beta-cal	−0.86%	−0.33%	7.11%	−0.91%	−0.01%
Iso.Regr.	−0.19%	−0.75%	8.31%	−0.56%	−0.01%
Meta-cal	−26.30%	−43.60%	−0.37%	−2.89%	−30.63%
Logistic	−0.41%	−0.07%	4.21%	−3.34%	−0.04%
Binning (B=10)	−0.33%	−9.11%	2.56%	−0.20%	−0.53%
Binning (B=30)	−0.49%	−0.87%	3.84%	−0.67%	−0.38%

marginal decrease in segmentation accuracy on the order of a tenth of a percentage point. There are some exceptions, such as Cityscapes or BUSI, where the calibration brings also an improvement in the accuracy. Meta-cal is the worst method concerning the effect on DICE, often substantially decreasing the accuracy. Moreover, we observe that a larger number of bins is generally preferable: too few bins produce overly coarse intervals, increasing the risk of incorrectly reassigning the confidence values of individual pixels. Temperature scaling and Selective scaling do not change the accuracy of the model, hence are not discussed.

C Datasets and Pre-trained Models

C.1 Datasets

Cityscapes (Cordts et al., 2016; 2015) is a large-scale benchmark for semantic understanding of urban street scenes, collected from a moving vehicle across 50 cities under varying seasonal and lighting conditions. It provides 5,000 images with high-quality pixel-level annotations and an additional 20,000 coarsely annotated frames. The fine-annotation split is pre-divided into 2,975 training, 500 validation, and 1,525 test images at a resolution of 1024×2048 pixels. Annotations span 30 semantic classes organized into 8 higher-level categories, covering both static elements (road, building, vegetation) and dynamic actors (person, car, cyclist). Its combination of high resolution, class diversity, and instance-level masks has made it the de-facto benchmark for segmentation in autonomous driving.

MS COCO (Lin et al., 2014) (Common Objects in Context) is a large-scale, general-purpose dataset for object detection, instance segmentation, and image captioning. It contains over 330,000 images with more than 2.5 million labeled instances drawn from 80 foreground object categories and 91 background 'stuff' categories. Its scale and annotation richness have established COCO as the primary standard for evaluating general-purpose segmentation models. We have created two subsets from the original dataset: COCO- animals and COCO-vehicles. Specifically, we selected images that feature either animals or humans to form the COCO-animals dataset, and images that contain vehicles or humans to create the COCO-vehicles dataset. The COCO-animals dataset includes 314 images with the following labels: background, cat, dog, sheep, cow, horse, bird, and human. In contrast, the COCO-vehicles dataset contains 321 images with these labels: background, train, bus, bicycle, airplane, car, boat, and human. We used those images for the calibration experiments.

TN3K (Gong et al., 2023) The TN3K (Thyroid Nodule Region Segmentation Dataset) comprises 3493 ultrasound images from 2421 patients, captured between January 2016 and August 2020. These images were selected from over 30,000 images provided by different hospitals.

ISIC 2016 (Gutman et al., 2016) is the first public benchmark released by the International Skin Imaging Collaboration (ISIC) for automated skin lesion analysis, hosted as part of the ISBI 2016 challenge. It contains 900 dermoscopic training images and 379 test images, each annotated with a binary pixel-level mask delineating the lesion boundary against the surrounding skin. The classification task is binary (melanoma vs. benign), and the segmentation ground-truth was produced by expert clinicians.

Kvasir-SEG (Pogorelov et al., 2017) is an open-access dataset of 1,000 gastrointestinal polyp images captured during colonoscopy, each paired with a manually delineated binary segmentation mask and a corresponding bounding box annotation. Masks were initially annotated by a medical doctor and then verified by an experienced gastroenterologist, ensuring high ground-truth quality. Accurate polyp segmentation is clinically important because polyps are precursors to colorectal cancer.

BUSI (Tasnim & Hasan, 2024; Al-Dhabyani et al., 2020) is a publicly available collection of breast ultrasound images. It contains 780 images from 600 female patients of age ranging between 25 and 75 years, featuring malignant and benign breast cancers but also normal cases without cancer. The dataset had few anomalies that were later corrected.

FLAIR1 (Garioud et al., 2022) The FLAIR #1 (French Land cover from Aerospace ImageRy) dataset is a large-scale, high-resolution, aerial imagery dataset designed for semantic segmentation of land cover across France. Developed by the French National Institute of Geographical and Forest Information (IGN), it aims to address challenges in domain adaptation, such as environmental heterogeneity and temporal changes. The FLAIR datasets are released under the Open Licence 2.0 from Etalab.

Oxford-IIIT Pet (Parkhi et al., 2012) is a fine-grained recognition and segmentation dataset comprising approximately 7390 images of cats and dogs spanning 37 breeds, with roughly 200 images per class.

C.2 Pre-trained Models and Data Splits

U-Net We train a lightweight U-Net Ronneberger et al. (2015) from scratch for binary foreground segmentation. The encoder follows a standard contracting path with four downsampling stages: starting from a base width $b = 32$, each stage doubles the number of channels ($b \to 2b \to 4b \to 8b \to 16b$, i.e. up to 512 feature maps at the bottleneck) and halves the spatial resolution via 2×2 max-pooling. Every resolution level applies two successive 3×3 convolutions, each followed by batch normalisation and ReLU activation; a spatial dropout layer ($p = 0.2$) is inserted between the two convolutions to regularise training. The decoder mirrors the encoder with four transposed-convolution upsampling steps (2×2, stride 2), each concatenating the corresponding encoder skip connection before applying the same double-convolution block. A final 1×1 convolution projects the b-channel feature map to a single output logit per pixel; the sigmoid activation is applied externally. The model takes 3-channel RGB images of arbitrary spatial resolution as input and produces raw logits of the same spatial size.

This architecture has been used for the following datasets: ISIC 2016, BUSI, Kvasir-SEG and TN3K. Each training has been performed with 200 epochs, batch size of 8 and learning rate of 10^{-3}. Images and masks are resized to 256×256 pixels; images are normalized with ImageNet mean and standard deviation. For ISIC, the dataset is already divided in train/val/test. We use the train and val splits to train the U-Net and the test split, featuring 379 images, for the calibration experiments. For BUSI, we use only the images with foreground (cancer). We train the U-Net with 439 images and use only 150 for the calibration experiments. This dataset is the smallest one in our evaluation and shows that meaningful calibration can be achieved even with few images. For Kvasir, we use 700 images to train the U-Net and the remaining 300 for the calibration experiments. Finally for TN3K, we train the U-Net with 2879 images and used the remaining 614 for the calibration experiments.

Table 3: Comparison of post-hoc calibration baselines.

Dataset	Pre-trained model	# samples for exp. (cal.+test)
Cityscapes	DeepLabv3plus_ResNet101	500
COCO-animals	DeepLabv3plus_ResNet50	314
COCO-vehicles	DeepLabv3plus_MobileNet	321
TN3K	U-Net	614
ISIC	U-Net	379
Kvasir	U-Net	300
BUSI	U-Net	150
FLAIR1	ResNet34-UNet	500
Oxford-Pet	DeepLabv3_resnet50 + fine tuning	739

DeepLabV3 Fine-tuning For the Oxford-IIIT Pet dataset we fine-tune a deeplabv3_resnet50 model from `torchvision.models`. The original classifier head (21-class) is replaced by a 1×1 convolutional layer mapping the 256-channel output to 2 classes (foreground / background).

The dataset consists of JPEG images paired with annotations; the annotations are binarized so that label 1 (pet) maps to the foreground class and all other values map to background. Images and masks are resized to 224×224 pixels; images are normalized with ImageNet mean and standard deviation. A 90/10 random split yields training and test subsets. The model is optimized with Adam (lr = 10^{-4}) using a pixel-wise cross-entropy loss over mini-batches of 32 images for 5 epochs. We use 6651 images for training and obtain the prediction of 739 images for the calibration experiments.

ResNet34-UNet For the FLAIR1 challenge we use a pre-trained model (*FLAIR-INC_rgbie_15cl_resnet34-unet*) provided by the organizer[1]. We use the pre-trained model on the smallest test subsets (*D022_2021*) containing only 500 images, which are later used in our calibration experiments. The predictions contain 19 different classes which are binarized with labels=$0, 8, 10, 11, 12, 17$ as foreground and the rest as background.

DeepLabV3+ Models For the Cityscapes dataset we use a DeepLabV3Plus-ResNet101[2] backbone Chen et al. (2018; 2017) on the validation set featuring 500 images. All of the predictions were randomly cropped to 256×256 and binarized with labels=$11, \ldots, 18$ as foreground and the rest as background. For the COCO-animals experiment we use the DeepLabv3plus_ResNet50 backbone, while for the COCO-vehicles the DeepLabv3plus_MobileNet backbone. All of the predictions were binarized with labels=$1, \ldots, 7$ as foreground and the rest as background. The code is published under the MIT License.

Data Splits During the experiments we perform K random splits of the data. For each random split we divide our dataset $\mathcal{D}$, in three subsets $\mathcal{D}_{bin}$, $\mathcal{D}_{cal}$ and $\mathcal{D}_{test}$. For the *post-hoc* baselines, $\mathcal{D}_{bin}$ and $\mathcal{D}_{cal}$ are used to calibrated the models. For the binning approach, $\mathcal{D}_{bin}$ is used to construct the bins, and $\mathcal{D}_{cal}$ is used to construct the calibrator H.

D Theoretical Bound Estimation

D.1 $\bar{D}$ estimation with autoregressive model

We now go through the implementation of the autoregressive model used to estimate the matrix $\bar{D}$. The coupling matrix $D^{\sigma}_{x,y}$ quantifies how much the distribution of σ_y changes when σ_x is intervened upon, holding the past $\sigma_{<x}$ fixed. Evaluating this quantity requires access to the full conditional distribution $\mathbb{P}(\sigma_y \mid \sigma_{\leq x})$, which in turn depends on the joint distribution over the mask. The autoregressive factorisation,

$$\mathbb{P}_\theta(\sigma) = \prod_{x \in \Omega} \mathbb{P}_\theta(\sigma_x \mid \sigma_{<x}),$$

[1] `https://ignf.github.io/FLAIR/`

[2] `https://github.com/VainF/DeepLabV3Plus-Pytorch`

expresses the conditional $\mathbb{P}(\sigma_y \mid \sigma_{\leq x})$ as a marginal over the intermediate sites $\sigma_{x+1}, \ldots, \sigma_{y-1}$. This marginal is approximated by Monte Carlo. Moreover, since in our application the variable σ_x are conditioned on a realization of the binning scheme and since we need to consider only the labels in a specific bin b, we construct the autoregressive model such that it takes as input a partition Ω_b and calibrated predictions $h(X)$, resulting in the factorisation:

$$\mathbb{P}_\theta(\sigma \mid h(X)) = \prod_{x_b \in \Omega_b} \mathbb{P}_\theta(\sigma_{x_b} \mid \sigma_{<x_b}, h(X)_{x_b}), \tag{39}$$

where $\sigma \in \{0,1\}^{\Omega_b}$ is a binary mask and Ω_b is a subset of a 2-D grid Ω.

Model architecture To ease the notation we are going to drop the b-dependent notation. We model the per-site conditional as a PixelCNN Van Den Oord et al. (2016b;a)

$$\mathbb{P}_\theta(\sigma_x = 1 | \sigma_{<x}, h(X)_x) \tag{40}$$

with sites in raster order, and $\sigma_{<x} = (\sigma_1, \ldots, \sigma_{x-1})$. The network has two parallel streams: a σ-stream of causally masked $k \times k$ convolutions carrying $\sigma_{<x}$, and an h-stream of 1×1 convolutions only, carrying $h(X)_x$. The streams are summed at the readout, so equation 40 depends on $\sigma_{<x}$ and on $h(X)_x$ alone and never on $h(X)_y$ for $y \neq x$.

For a kernel of odd size k with center at $c = (k-1)/2$, the causal masks are

$$M^{(\mathrm{A})}_{x_1,x_2} = \mathbf{1}\big[x_1 < c \text{ or } (x_1 = c,\, x_2 < c)\big], \qquad M^{(\mathrm{B})}_{x_1,x_2} = M^{(\mathrm{A})}_{x_1,x_2} + \mathbf{1}\big[(x_1, x_2) = (c, c)\big]. \tag{41}$$

Mask A excludes the center, mask B includes it. A masked convolution multiplies the weight tensor by $M^{(\tau)}$ before convolving: $\mathrm{MaskedConv}^{(\tau)}_k(w) = (W \odot M^{(\tau)}) * w + b$.

The σ-stream uses mask A in its first layer (so its output at site x does not depend on σ_x) and mask B thereafter (centre allowed because upstream activations are already strictly causal in σ):

$$s^{(0)} = \mathrm{MaskedConv}^{(\mathrm{A})}_k(\sigma), \tag{42}$$

$$s^{(\ell+1)} = s^{(\ell)} + \mathrm{MaskedConv}^{(\mathrm{B})}_k\big(ReLU\big(\mathrm{MaskedConv}^{(\mathrm{B})}_k\big(ReLU(s^{(\ell)})\big)\big)\big), \qquad \ell = 0, \ldots, L{-}2. \tag{43}$$

The h-stream has the same skeleton with $k = 1$ and no masking: $c^{(0)} = Conv_1(h(X)_x)$ and residual blocks formed analogously from two 1×1 convolutions and a $ReLU$.

The readout sums the streams and applies a final pointwise layer:

$$\mathbb{P}_\theta(\sigma_x = 1 \,|\, \sigma_{<x},\, h(X)_x) = \mathrm{sigmoid}\big(Conv_1\big(ReLU\big(s^{(L-1)} + c^{(L-1)}\big)\big)_x\big). \tag{44}$$

Coupled sampling Concretely, we estimate $D^\sigma_{x,y}$ by sampling with a common-random-numbers coupling (Mohamed et al., 2020; Glasserman, 2004). Fix a context $\sigma_{<x}$, then for each Monte Carlo trial we run two chains A and B:

$$\sigma^A_x = 1, \qquad \sigma^B_x = 0, \qquad \sigma^A_{<x} = \sigma^B_{<x} = \sigma_{<x}.$$

For each subsequent site $t = x{+}1, \ldots, HW$ (where $H \times W$ are the image dimensions) we perform one batched forward pass to obtain the two conditional probabilities

$$p^A_t = P_\theta(\sigma_t = 1 \mid \sigma^A_{<t}), \qquad p^B_t = P_\theta(\sigma_t = 1 \mid \sigma^B_{<t}),$$

draw a single threshold $u_t \sim \mathcal{U}[0,1]$ shared between the two chains, and set

$$\sigma^A_t = \mathbf{1}_{[u_t < p^A_t]}, \qquad \sigma^B_t = \mathbf{1}_{[u_t < p^B_t]}.$$

Each chain therefore advances under its own conditional — so neither estimator is biased — while the shared threshold ensures that sites where the two conditionals agree contribute exactly zero to the difference $\sigma^A_t - \sigma^B_t$. Averaged over M trials,

$$\hat{D}^\sigma_{x,t}(\sigma_{<x}) \;=\; \Big| \frac{1}{M} \sum_{m=1}^{M} \big(\sigma^{A,(m)}_t - \sigma^{B,(m)}_t\big) \Big|$$

is an unbiased Monte Carlo estimator of $|D^\sigma_{x,t}|$ conditional on $\sigma_{<x}$, with variance reduced compared to two independent runs.

Full decay curve A key observation is that the autoregressive structure allows us to estimate $D^{\sigma}_{x,y}$ for all $y > x$ simultaneously from a single pair of coupled trajectories. The recorded differences $\sigma^A_t - \sigma^B_t$ at every $t > x$ already provide one Monte Carlo realisation of $D^{\sigma}_{x,t}$ at every target site. Repeating the coupled run M times under the same $\sigma_{<x}$ gives the full empirical decay $\hat{D}^{\sigma}_{x,t}(\sigma_{<x})$ for $t = x+1, \dots, HW$.

Aggregation and the decay fit To reduce dependence on any particular grid location, we repeat the procedure for n_s source sites x drawn from the interior of the grid, and for n_c test contexts providing different prefixes $\sigma_{<x}$. For each source–context pair we compute the per-target estimates $\hat{D}^{\sigma}_{x,y}(\sigma_{<x})$ as described above and bin them by Euclidean distance $d = \|y - x\|_2$. The aggregation across the three nested sources of randomness — trials, contexts, sources — is performed as follows:

1. Within each $(x, \sigma_{<x})$, the Monte Carlo trials are averaged ($\hat{D}^{\sigma}_{x,y}(\sigma_{<x})$).
2. For each context, we average the per-target values over all y falling in the same distance bin d, yielding one number per $(x, \sigma_{<x}, d)$.
3. Across contexts, we take the maximum over $\sigma_{<x}$. This is a worst-case aggregation: the calibration bound of Theorem 2.2 depends on $\sup_{\sigma_{<x}} |D^{\sigma}_{x,y}|$ rather than its average, since the bound must hold for every realisable prefix.
4. Across sources, we average the resulting per-source worst-case values to obtain the empirical decay curve $\hat{D}(d)$.

Fitting the parametric form equation 13 to this curve serves two purposes. First, it provides a closed-form approximation of the coupling matrix for any distance, including distances not directly observed in the finite grid. Second, it allow us to compute the coefficients S and c of the bounds.

Moreover, since we need to estimate the coupling matrix for variables Y_{x_b} in bin b, we have to use the factorisation in equation 39. In practice, we estimate the matrix $\hat{D}(d)$ for different bins b and average the results.

Experiment settings During the experiments we use the calibration dataset ($\mathcal{D}_{cal}$) to estimate the matrix $\hat{D}$. We sample random contexts, of shape 32×32 (64×64 for the FLAIR dataset, due to the larger image dimensions) from the ground-truth segmentation. We use 5000 contexts to train the autoregressive model. We use the Adam optimizer and train the model for 100 epochs with learning rate 10^{-3} and batch size equal to 32.

For the evaluation the hyperparameters are: the number of random source points $n_s = 20$, the number of random contexts $n_c = 20000$ and the number of random Monte Carlo trials $M = 10$. We use masks of size 3×3, with 32 hidden channels and 1 layer.

D.2 Bound estimation

Now we can compute the final bound. The row sums of the global coupling matrix are defined as

$$R_{x_b} \;=\; \sum_{y_b > x_b} \hat{D}\big(\|y_b - x_b\|_2\big),$$

where $\hat{D}(d) = \hat{a}\, \exp(-\hat{\alpha} d^{\hat{b}})$ is the fitted decay, $\|y - x\|_2$ is the Euclidean distance between pixels x and y, and $y > x$ refers to the raster order of the autoregressive scan. Intuitively, R_{x_b} aggregates the total residual dependence between x_b and all in-bin pixels that follow it in the scan. From these row sums we can build the coefficients of Theorem 2.2:

$$S^{(i)}_b \;=\; \frac{N_{\text{cal}}}{N_b^2} \hat{p}\,(1 - \hat{p}) \sum_{x_b \in \Omega^{(i)}_b} (R^{(i)}_{x_b})^2, \qquad c^{(i)}_b \;=\; \frac{1}{3N_b} \max_{x_b \in \Omega^{(i)}_b} R^{(i)}_{x_b},$$

Table 4: Computational time of the calibration methods.

	Binning	**T-scal.**	**Beta-cal**	**Log.-cal**	**Iso-regr**	**Meta-cal**	**Sel.Scal**
Time (s)	$0.6-5$	$6-55$	$12-45$	$4-11$	$3-21$	$6-59$	$200-1200$

where $\hat{p}$ is the base rate of the positive class estimated in D.1 and $N_b = \sum_{i=1}^{N_{cal}} |\Omega_b^{(i)}|$ is the total number of bin-b pixels across $\mathcal{D}_{\text{cal}}$. In this estimation we assume for simplicity that all the images contributes the same. We approximate the bound by averaging over a random subset ($n = 20$) of calibration images,

$$\bar{S}_b \;=\; \tfrac{1}{n}\sum_{i=1}^{n} S_b^{(i)}, \qquad \bar{c}_b \;=\; \tfrac{1}{n}\sum_{i=1}^{n} c_b^{(i)}.$$

Moreover, for the marginal experiments we use an average of the coefficients:

$$\bar{S} \;=\; \tfrac{1}{B}\sum_{b=1}^{B} \bar{S}_b, \qquad \bar{c} \;=\; \tfrac{1}{B}\sum_{b=1}^{B} \bar{c}_b.$$

Final bound Assembling all components, the marginal calibration bound (used in the experiments) at tolerance ε takes the form of Theorem 2.2:

$$\mathbb{P}\Big(\,|h(X)_x - \mathbb{E}[Y_x \mid h(X)_x]| \geq \varepsilon\,\Big) \leq 2\exp\Big(-\frac{\bar{S}}{\bar{c}^2}\eta(u)\Big), \qquad u = \frac{\bar{c}\varepsilon}{\bar{S}}, \qquad \eta(u) = 1 + u - \sqrt{1+2u}.$$

The conditional bound is instead approximated as

$$\mathbb{P}\Big(\forall r \in \mathit{Range}(h) : |r - \mathbb{E}[Y_x \mid h(X)_x = r]| \geq \varepsilon\,\Big) \leq 2\sum_{b=1}^{B}\exp\Big(-\frac{\bar{S}_b}{\bar{c}_b^2}\eta(u_b)\Big).$$

E Implementation details and Computational Expenses

We perform each experiment using a GPU NVIDIA A100-SXM4-40GB. The average computing time of each method, for a calibration split, is given in Table 4.

The calibration algorithm and its randomized version are detailed in Tables 1-2.

Algorithm 1 Calibration algorithm

Input: number of bins B, pre-trained predictor f, dataset $\{(X^{(i)}, Y^{(i)})\}_{i=1}^{N}$
Output: calibrated predictor h

Partition the dataset into $\mathcal{D}_{\text{bin}}$ and $\mathcal{D}_{\text{cal}} = \{(X^{(i)}, Y^{(i)})\}_{i=1}^{N_{cal}}$

// Binning scheme: uniform-mass binning via quantiles on $\mathcal{D}_{\text{bin}}$
Collect the pool of predicted confidences $\mathcal{P} \leftarrow \{f(X^{(i)})_x : \ i \in \mathcal{D}_{\text{bin}}, \ x \in \Omega\}$
for $j = 1, \ldots, B-1$ **do**
    $\hat{q}_j \leftarrow \mathcal{Q}_{j/B}(\mathcal{P})$ ▷ empirical (j/B)-th quantile
$I_1 \leftarrow [0, \hat{q}_1), \quad I_j \leftarrow [\hat{q}_{j-1}, \hat{q}_j]$ for $j = 2, \ldots, B-1, \quad I_B \leftarrow (\hat{q}_{B-1}, 1]$

// Calibrator: empirical label frequency per bin on $\mathcal{D}_{\text{cal}}$
for $i = 1, \ldots, N_{cal}$ and $b = 1, \ldots, B$ **do**
    $\Omega_b^{(i)} \leftarrow \{x \in \Omega : \ f(X^{(i)})_x \in I_b\}$ ▷ assign pixels to bins
for $b = 1, \ldots, B$ **do**
    $N_b \leftarrow \sum_{i=1}^{N_{cal}} |\Omega_b^{(i)}|$
    $\hat{\pi}_b \leftarrow \dfrac{1}{N_b} \sum_{i=1}^{N_{cal}} \sum_{x_b \in \Omega_b^{(i)}} Y_{x_b}^{(i)}$ ▷ empirical positive-class frequency
return $h(X)_x = \displaystyle\sum_{b=1}^{B} \mathbf{1}_{\{f(X)_x \in I_b\}} \, \hat{\pi}_b$

Algorithm 2 Randomized Calibration algorithm

Input: number of bins B, pre-trained predictor f, dataset $\{(X^{(i)}, Y^{(i)})\}_{i=1}^{N}$, randomization parameter δ
Output: calibrated predictor $\tilde{h}$

Partition the dataset into $\mathcal{D}_{\text{bin}}$ and $\mathcal{D}_{\text{cal}} = \{(X^{(i)}, Y^{(i)})\}_{i=1}^{N_{cal}}$

// Binning scheme: uniform-mass binning via quantiles on $\mathcal{D}_{\text{bin}}$
Collect the pool of predicted confidences $\mathcal{P} \leftarrow \{f(X^{(i)})_x : \ i \in \mathcal{D}_{\text{bin}}, \ x \in \Omega\}$
for $j = 1, \ldots, B-1$ **do**
    $\hat{q}_j \leftarrow \mathcal{Q}_{j/B}(\mathcal{P})$ ▷ empirical (j/B)-th quantile
$I_1 \leftarrow [0, \hat{q}_1), \quad I_j \leftarrow [\hat{q}_{j-1}, \hat{q}_j]$ for $j = 2, \ldots, B-1, \quad I_B \leftarrow (\hat{q}_{B-1}, 1]$

// Calibrator: empirical label frequency per bin on $\mathcal{D}_{\text{cal}}$
for $i = 1, \ldots, N_{cal}$ and $b = 1, \ldots, B$ **do**
    $\Omega_b^{(i)} \leftarrow \{x \in \Omega : \ f(X^{(i)})_x \in I_b\}$ ▷ assign pixels to bins
for $b = 1, \ldots, B$ **do**
    $N_b \leftarrow \sum_{i=1}^{N_{cal}} |\Omega_b^{(i)}|$
    $\hat{\pi}_b \leftarrow \dfrac{1}{N_b} \sum_{i=1}^{N_{cal}} \sum_{x_b \in \Omega_b^{(i)}} Y_{x_b}^{(i)}$ ▷ empirical positive-class frequency
    $U_b \sim \mathcal{U}[-1, 1]$
return $\tilde{h}(X)_x = \displaystyle\sum_{b=1}^{B} \mathbf{1}_{\{f(X)_x \in I_b\}} \, (\hat{\pi}_b + \delta U_b)$