TOWARDS CERTIFIED SHORTCUT UNLEARNING IN MEDICAL IMAGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Many medical AI imaging applications rely on segmentation-classification pipelines. In practice, coarse or imprecise segmentation masks often admit *spurious* background cues and the model learns the associations between these cues and the pathological regions of interest (organs, lesions, etc.). Such *shortcuts* tend to severely affect the accuracy of the downstream task. Prior art relies on fine-grained pixel-level annotations to separate background cues from pathology regions. Unfortunately, high-quality annotations are expensive. We aim to ensure high performance on downstream tasks without relying on expensive annotations. We propose using machine unlearning to mitigate shortcut learning. In contrast to relying on heuristic methods, we propose a principled approach that, for the first time, provides guarantees for shortcut unlearning. We develop a definition of certified unlearning for segmentation tasks and formally connect it to its canonical definition. Further, we introduce an information-theoretic objective function, *global information reduction*, and prove that any (ε, δ)-certified machine unlearning operator bounds it relative to the fine-mask reference. Importantly, the bound further limits the accuracy any decision rule can extract from the artefact. We discover existing certified operators collapse at the scale of real medical decoders, and introduce a new machine unlearning algorithm that challenges the SOTA in segmentation at tight (ε, δ) budgets. On ISIC melanoma trap sets and SIIM-ACR pneumothorax with a naturally occurring chest-tube shortcut, we compare unlearning with state-of-the-art refinement methods on segmentation, shortcut robustness, fine-mask budget and compute. Unlearning spans the Pareto front on both datasets, no refinement method reaches it, and the certified operator is ahead of exact retraining, the only other method with a guarantee, at every fine-mask budget.

1 INTRODUCTION

In medical imaging, deep models are encouraged to focus on the pathology of interest (e.g., an organ or lesion) rather than on background pixels before performing classification. In practice, however, coarse or imprecise segmentation masks may admit spurious background cues, leading to *shortcut learning* and models that perform well on correlated data but may fail when those correlations change. For instance, Winkler et al. (2019) demonstrates that surgical ink markings on skin lesions (which are unrelated to melanoma) cause CNNs to falsely raise melanoma scores. Similar problems arise in chest radiography, where chest tubes, the treatment for a pneumothorax, become a proxy for the disease (Olesen et al., 2024). Such shortcuts not only hurt accuracy but can introduce fairness concerns when artefact frequencies differ across patient groups (Schrouff et al., 2022).

Fine annotations are costly, since experts need many more clicks to outline lesions (Saab et al., 2022), which creates a cost-accuracy trade-off. We propose *certified machine unlearning* to resolve it. Suppose a model pre-trained on cheap coarse masks has learned shortcuts from them, and an expert then supplies fine masks for a small set of images (e.g., Jiménez-Sánchez et al., 2023). We want to "unlearn" the coarse-mask pixels on these images, so that the model behaves as if it had been trained on the fine labels only (Figure 1). To our knowledge, this is the first certified unlearning method for segmentation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

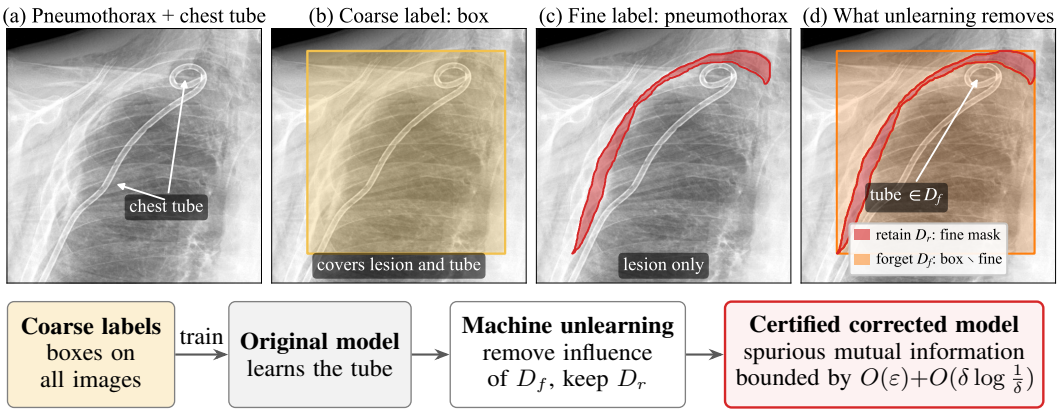


Figure 1: **Shortcut unlearning.** (a) A pneumothorax treated with a chest tube, a strong correlate of the disease. (b) The coarse bounding box covers the lesion and the tube, so a model trained on boxes can learn the tube as a shortcut. (c) The fine mask covers the lesion only. (d) The rest of the box, which contains the tube, is the forget set D_f. Bottom: certified unlearning corrects the deployed model and bounds the spurious information it retains.

Our experiments answer two questions. First, is certification affordable? At privacy budgets down to $\varepsilon = 0.001$, the certified operator matches or exceeds heuristic unlearning in shortcut robustness, at some cost in segmentation on melanoma. Second, is unlearning worthwhile? Compared on segmentation, shortcut robustness, the share of images that need fine masks, and compute, correcting the deployed model by unlearning spans the Pareto front, achieving state of the art performance, which refinement methods that rebuild the model do not reach.

Our Contributions.

- **Segmentation refinement as machine unlearning.** We cast the refinement of coarse segmentation masks as machine unlearning: the pixels a coarse mask adds around the lesion form the forget set, the data to be removed, and the fine mask the retain set, the signal to maintain, which makes certified unlearning well defined for segmentation.
- **Certified bounds on shortcut learning.** We prove that any certified unlearning operator bounds the spurious mutual information between spurious cues and pathology relative to a model trained on fine masks, and hence the accuracy any decision rule can extract from the spurious cues.
- **A state-of-the-art Pareto front.** We introduce a heuristic unlearning algorithm and a new certified unlearning operator for image segmentation, with the latter outperforming existing state-of-the-art certified methods, whose performance collapses in our challenging setting. On two challenging medical settings, ISIC melanoma detection and SIIM-ACR pneumothorax, our unlearning operators achieve the state-of-the-art Pareto front across sample efficiency, time efficiency, downstream classification and segmentation: no state-of-the-art refinement baseline reaches it, and the certified operator is ahead of exact retraining, the only other method with a guarantee, at every fine-mask budget.

2 Theoretical Formulation

We begin by formalising the connection between correcting coarse segmentation masks and *machine unlearning*: correcting a dilated mask is set-theoretically isomorphic to forgetting a specific subset of training constraints. Simply put, this allows us to cast segmentation refinement as machine unlearning.

2.1 Preliminaries

Machine Unlearning. Consider a learning task with a dataset $D = \{(x_i, y_i)\}$. A subset $D_f \subset D$ must be "forgotten," leaving the retained set $D_r = D \setminus D_f$ (Nguyen et al., 2025). Let $M_{\mathcal{R}}$ be the model trained from scratch on D_r. A machine unlearning algorithm takes the original model

M (trained on D) and outputs an unlearned model $M_{\mathcal{U}}$ that is intended to mimic $M_{\mathcal{R}}$ without incurring the cost of full retraining. *Certified unlearning* (Sekhari et al., 2021) formalizes this goal by demanding (ε, δ)-indistinguishability between $M_{\mathcal{U}}$ and $M_{\mathcal{R}}$, analogous to differential privacy (Dwork & Roth, 2014).

Segmentation with Coarse Labels. Extending the framework of Saab et al. (2022), we focus on pixel-wise segmentation tasks. Let $X \in \mathbb{R}^{C\times H\times W}$ be an image. The true segmentation mask is $Y \in \mathcal{Y}^{H\times W}$ where $\mathcal{Y} = \{0, \ldots, K\}$ represents K semantic classes, with $Y_{ij} = 0$ indicating a background pixel. We impose a *single-class* assumption, namely that any given image contains at most one non-background pathology class $k \in \{1, \ldots, K\}$.

In practice, fine-grained masks are expensive to obtain. Instead, we often have access to a *coarse* mask $Y^{(r)}$ (Figure 1 (b)), defined via spatial dilation with radius $\Delta_r = \lfloor \max(H, W)/r \rfloor$. Specifically,

$$Y^{(r)}_{ij} := \max_{(p,q)\in\mathcal{B}^{\ell_\infty}_{\Delta_r}(i,j)} Y_{pq}, \tag{1}$$

where $\mathcal{B}^{\ell_\infty}_{\Delta_r}(i, j) := \{(p, q) : \max(|p - i|, |q - j|) \le \Delta_r\}$ is the ℓ_∞ ball of radius Δ_r centred at pixel (i, j), and Y_{pq} is the true label at pixel (p, q). Thus $Y^{(r)}$ is the morphological dilation of Y: a background pixel is labelled foreground whenever a lesion pixel lies within distance Δ_r. This is the coarse-annotation artefact we later unlearn. A larger r gives a smaller radius and a tighter mask.

2.2 Mask Refinement as Data Removal

The atomic data unit is a *pixel constraint*: an image together with a pixel location. Let $D^{(r)}$ contain the foreground constraints supplied by the coarse mask $Y^{(r)}$. For $r_1 < r_2$, dilation monotonicity gives $D^{(r_2)} \subseteq D^{(r_1)}$. Thus refinement identifies the forget and retain sets as

$$D_f = D^{(r_1)} \setminus D^{(r_2)}, \qquad D_r = D^{(r_2)}. \tag{2}$$

In words, we remove the foreground constraints introduced by coarse annotation while preserving those supported by the finer mask (Figure 1 (d)). Appendix A gives the full construction and its geometric interpretation.

2.3 Derivation of Certified Pixel-Level Unlearning

We derive our criterion directly from the general axioms of *Approximate Unlearning* (Nguyen et al., 2025), which relaxes exact distributional equality to a cryptographic (ε, δ)-indistinguishability bound.

Definition 2.1 (Standard Approximate Unlearning)**.** Let $\mathcal{A}$ be a learning algorithm. An unlearning mechanism $\mathcal{U}$ is (ε, δ)-certified with respect to a forget set D_f if for any measurable subset of hypotheses $\mathcal{T} \subseteq \mathcal{H}$:

$$\mathbb{P}(\mathcal{U}(\mathcal{A}(D)) \in \mathcal{T}) \le e^{\varepsilon}\mathbb{P}(\mathcal{A}(D \setminus D_f) \in \mathcal{T}) + \delta, \tag{3}$$

$$\mathbb{P}(\mathcal{A}(D \setminus D_f) \in \mathcal{T}) \le e^{\varepsilon}\mathbb{P}(\mathcal{U}(\mathcal{A}(D)) \in \mathcal{T}) + \delta. \tag{4}$$

To instantiate Definition 2.1 for dense prediction, we follow *weak unlearning* (Nguyen et al., 2025) and move indistinguishability from parameters to model outputs. Unconditional output matching is inadequate for segmentation, since it admits vacuous solutions (e.g., permutations that preserve frequencies but destroy utility), so we condition on the true label $Y_{ij} = y$.

Definition 2.2 (Certified Pixel-Level Unlearning)**.** Consider two masks $Y^{(r_1)}, Y^{(r_2)}$ for $1 \le r_1 < r_2$, and an unlearning operator $\mathcal{U}$. We say $\mathcal{U}$ is a *pixel-level* (ε, δ)*-certified unlearning operator* if for every pixel (i, j), true label $y \in \mathcal{Y}$, and predicted class $k \in \{1, \ldots, K\}$

$$\mathbb{P}(\hat{Y}_{ij} = k \mid Y_{ij} = y) \le e^{\varepsilon}\mathbb{P}(Y^{(r_2)}_{ij} = k \mid Y_{ij} = y) + \delta, \tag{5}$$

$$\mathbb{P}(Y^{(r_2)}_{ij} = k \mid Y_{ij} = y) \le e^{\varepsilon}\mathbb{P}(\hat{Y}_{ij} = k \mid Y_{ij} = y) + \delta, \tag{6}$$

where $\hat{Y} := \mathcal{U}[Y^{(r_1)}]$ denotes the unlearned mask.

The two inequalities formalise a single statement: the unlearned model behaves as if it had been trained on the fine masks. Taking the fine-mask model as reference, (5) bounds how much *more* often the unlearned model can produce a prediction, which forbids retaining the dilation shortcut (under-forgetting); (6) bounds how much *less* often, which forbids degrading beyond what refinement causes, for instance by collapsing predictions to remove the shortcut (over-forgetting). No test can distinguish the two models beyond odds e^{ε}, up to a failure probability δ.

Remark 2.3 (From certified machine unlearning to certified pixel-level unlearning). We translate between certified machine unlearning (Definition 2.1) and certified pixel-level unlearning (Definition 2.2) through a conditional projection lemma (Lemma D.3, Appendix D.2): any operator certified in the sense of Definition 2.1 is indistinguishable from the fine-mask reference on every event of the generative variables, and (5)–(6) are the instance $E = \{Y_{ij} = y\}$. Certification also preserves the prediction mass of the reference, so the conditioning events of Section 3 cannot collapse unless they are equally rare for it.

3 From Certified Unlearning to Reducing Shortcut Learning

Having defined the formalism for machine unlearning in segmentation, we now ask whether certified unlearning *provably* mitigates shortcut learning. We first introduce an information-theoretic objective function to track the entanglement between spurious cues and the pathology.

3.1 Defining the Shortcut Objective

A naive approach might measure the standard conditional mutual information $I(S; Y \mid \hat{Y})$ between spurious features S and pathology Y given the prediction $\hat{Y}$. However, in segmentation tasks where the vast majority of pixels are background (class 0), this standard global metric is dominated by the trivial independence found in background regions. We are specifically interested in whether the model relies on spurious features *when it flags a region as pathological*. To capture this, we introduce the *Global Spurious Mutual Information*, which explicitly targets the information leakage within the predicted foreground.

Definition 3.1 (Global Spurious Mutual Information)**.** We define the Global Spurious Mutual Information for a random mask variable Z as the maximum over non-background classes of the sum of pixel-wise spurious mutual information terms, weighted by the probability of the mask predicting that class at that location:

$$\mathcal{S}_{\text{global}}(Z) := \max_{k \in \{1,\dots,K\}} \sum_{i,j} \mathbb{P}\left(Z_{ij} = k\right) I\left(S_{ij}, Y_{ij} \mid Z_{ij} = k\right).$$

The maximum over classes captures worst-case leakage, so minimising $\mathcal{S}_{\text{global}}$ formally defines shortcut removal from positive predictions.

3.2 Theoretical Analysis

A Generative Model of Shortcuts. Inspired by Saab et al. (2022), we introduce a formal data model that captures the presence of pixel-level spurious features. Let $S \in \{0, \dots, K\}^{H \times W}$ be a mask indicating the location of such artefacts (e.g., surgical markers or tubes). The generative process for constructing an image X, true pathology Y, coarse annotation $Y^{(r)}$, and spurious mask S proceeds in stages. First, global labels for disease presence $y \in \{0, \dots, K\}$ and artifact presence s are drawn from a joint distribution $\mathcal{P}_{y,s}$. Next, the dense masks Y and S are drawn conditioned on y, s. Crucially, if $y = 0$ then Y is the zero mask. Furthermore, if $y = k$, the spurious mask S is constrained to take values in $\{0, k\}$ (with S being the zero mask if $s = 0$). The observed coarse mask $Y^{(r)}$ is constructed deterministically from Y via Eq. (1). Note that it satisfies the dominance property where $Y_{ij} = k \implies Y^{(r)}_{ij} = k$ for any $k \in \{1, \dots, K\}$. Finally, the observed image X is drawn from a conditional distribution $\mathcal{P}(X \mid Y, S)$.

The Shortcut Mechanism. A coarse mask $Y^{(r)}$ with small r covers the lesion and surrounding background, so it may overlap S, and a model fitting $Y^{(r)}$ can use S as a shortcut. Our model allows arbitrary statistical dependence between S and Y and enforces only *disjoint support*, $Y_{ij} =$

$k > 0 \implies S_{ij} = 0$: artefacts do not physically overlap the lesion. We find this is a reasonable assumption in practice (see Figure 1 (d)).

3.3 Certified Shortcut Reduction

We first relate spatial specificity to spurious dependence. Saab et al. (2022) state a result of this form under $S_{ij} \perp Y_{ij}^{(r)} \mid Y_{ij}$, which cannot hold under their own data-generating process: the global disease label is a common cause of the dilated mask and the artefact (counterexample in Appendix E). We prove it under disjoint support, which permits arbitrary dependence between disease and artefact, via a new closed-form representation $\mathcal{J}(\alpha, m)$ of the conditional spurious information (8) (Appendix D.3), which the main theorem also uses.

Proposition 3.2 (Spatial Specificity Reduces Spurious Dependence)**.** *Under the proposed data model, there exists a minimal $r_0 \in [1, n)$ such that $I(S_{ij}, Y_{ij} | Y_{ij}^{(r)} = k)$ is decreasing in r for $r \geq r_0$ and any $k > 0$.*

The monotonicity concerns this data model and refinement regime, not arbitrary conditioning; Appendix G gives a numerical diagnostic.

Our main per-pixel result compares the unlearned mask with the fine-mask reference $Y^{(r_2)}$, with an error controlled by the certification parameters (proof in Appendix D.4).

Theorem 3.3 (Certified Pixel-Level Reduction of Spurious MI)**.** *Assume $\mathcal{U}$ satisfies the conclusion of Lemma D.3, in particular any $\mathcal{U}$ certified in the sense of Definition 2.1. Then under the proposed data model, for $r_0 \leq r_1 < r_2$ and any class $k > 0$, assuming the non-degeneracy condition $\mathbb{P}(S_{ij} = 1 \mid Y_{ij} = 0, Y_{ij}^{(r_2)} = 1) < 1$ and δ sufficiently small (both quantified in the proof), we have*

$$\left| I\left(S_{ij}, Y_{ij} \mid \mathcal{U}\left[Y^{(r_1)}\right]_{ij} = k\right) - I\left(S_{ij}, Y_{ij} \mid Y_{ij}^{(r_2)} = k\right) \right| \leq 2\varepsilon I\left(S_{ij}, Y_{ij} \mid Y_{ij}^{(r_2)} = k\right) + \mathcal{O}(\varepsilon) + \mathcal{O}(\delta \log(1/\delta)).$$

Remark 3.4 (From certified pixel-level unlearning to global spurious mutual information). Aggregating the per-pixel bounds gives our main global guarantee on $\mathcal{S}_{\text{global}}$ (proof in Appendix D.5). Appendix B discusses the scope of the disjoint-support assumption, the role of the hypothesis and what distinguishes Theorem 3.3 from standard indistinguishability guarantees.

3.4 Main Result: Global Spurious Information Reduction Guarantees

Corollary 3.5 (Certified Reduction of Global Spurious MI)**.** *Under the assumptions of Theorem 3.3, for $r_0 \leq r_1 < r_2$, we have*

$$\left| \mathcal{S}_{global}\left(\mathcal{U}\left[Y^{(r_1)}\right]\right) - \mathcal{S}_{global}\left(Y^{(r_2)}\right) \right| \leq \mathcal{O}(\varepsilon) + \mathcal{O}(\delta \log(1/\delta)).$$

By certifying the precise forget set (the dilation artefacts), the total spurious information of the unlearned model tracks that of the ideal fine-mask model up to the certification error (Figure 1 (Bottom)).

From information to reliance. A bound on information becomes a bound on reliance through the accuracy the artefact can confer on any decision rule, including the trained classifier. Fix a pixel and a class k, let $E = \{\hat{Y}_{ij} = k\}$, and let $R_\emptyset^*$ and R_S^* be the Bayes errors of predicting Y_{ij} on E without and with access to S_{ij}.

Proposition 3.6 (Bounded shortcut advantage)**.** *With natural logarithms, $R_\emptyset^* - R_S^* \leq \sqrt{\frac{1}{2} I(S_{ij}; Y_{ij} \mid E)}$.*

The proof (Appendix D.7) bounds the gap by the total variation between the joint law of (S_{ij}, Y_{ij}) on E and the product of its marginals, and applies Pinsker's inequality. Combined with Theorem 3.3, the shortcut's Bayes-optimal advantage after certified unlearning is controlled by that of the fine-mask reference: where the reference carries no spurious information, no decision rule gains more than $\mathcal{O}(\sqrt{\varepsilon + \delta \log(1/\delta)})$ accuracy from the artefact.

4 EXPERIMENTAL METHODOLOGY

Hypotheses. Our experiments test two hypotheses. **H1 (certification is affordable):** a certified unlearning operator is competitive with heuristic unlearning, so the guarantee costs little in segmentation or shortcut robustness. **H2 (machine unlearning is Pareto-efficient):** we compare methods on four axes, namely segmentation (Dice), downstream classification (robust AUROC), sample efficiency (the share of training images that need fine masks) and time efficiency (compute relative to retraining from scratch); unlearning the original model lies on the Pareto front of these axes, and refinement methods that rebuild it do not.

Three families of methods. All methods start from the same coarse annotations and receive the same fine-mask budget. *Refinement* rebuilds the segmenter: *Retrain* trains it from random weights on the fine masks, and the semi-supervised *Mean Teacher* (Tarvainen & Valpola, 2017) and *Cross-Pseudo Supervision* (CPS) (Chen et al., 2021) also use the remaining images unlabelled. *Heuristic unlearning* (ours) corrects the deployed original model: *NegGrad-Seg*, our segmentation version of NegGrad+ (Kurmanji et al., 2023), fine-tunes on the retain set D_r and ascends the loss on the forget set D_f with weight w_{D_f}; $w_{D_f} = 1$ is plain *fine-tuning*. *Certified unlearning* (ours) adds an (ε, δ) guarantee. The state of the art is per-step noisy fine-tuning (Koloskova et al., 2025; Dolgova & Stich, 2026), which clips the optimisation and injects noise at every step; noisy fine-tuning collapse to chance on our decoders, because clipping the model to the radius of prior configurations removes nearly all of its signal (Appendix I.3).

One-shot certified unlearning. Algorithm 1 injects Gaussian noise *once*, without an initial model-clipping step. Following the proximity framework of Dolgova & Stich (2026), we assume a bound Δ on the distance between coupled outputs of full-data training and retain-only retraining. Sharing training randomness aligns the two runs; the bound describes their remaining discrepancy. This replaces the worst-case clipping diameter in the Gaussian calibration. Rescaling to data-independent initialisation norms and recovery on D_r preserve the guarantee by post-processing. Appendix C specifies the coupling assumption, noise calibration, and reference distribution needed to apply Theorem 3.3.

Algorithm 1 One-shot certified instantiation

Require: decoder $\hat{x} = \mathcal{A}(D)$; data-independent initialisation x_{init}; retain set D_r; valid fixed discrepancy bound $\Delta > 0$ as specified in Appendix C; budget (ε, δ)

1: Set $\sigma = \Delta\sqrt{q^*/(2\varepsilon_r^*)}$ using Eq. (7)
2: Draw $\xi \sim \mathcal{N}(0, \sigma^2 I_d)$ independently and set $x_0 \leftarrow \hat{x} + \xi$
3: Rescale each parameter tensor of x_0 to the norm of its counterpart in x_{init}
4: Reset the optimiser state and **return** x_0 fine-tuned on D_r

Setup. A frozen ResNet-50 (He et al., 2016) encoder (pneumothorax: ImageNet weights, 256×256 input replicated to three channels, $8 \times 8 \times 2048$ stage-4 features) feeds a decoder with approximately $18{,}000$ parameters; features are cached, so the same code path runs unmodified on both datasets. Fine labels, whose pixel-level supervision costs about $50\times$ more than bounding boxes (Saab et al., 2022), cover 50% of the training images in the main experiments (911 ISIC and 1,500 SIIM-ACR images); Section 5.2 varies this budget. All recovery-based methods share the recovery schedule of each dataset (Appendix I), and every method is evaluated on the same data splits and seeds. We report the certified operator at $\varepsilon \in \{0.1, 0.01, 0.001\}$ with $\delta = 10^{-5}$. We validate the theory on a controlled synthetic benchmark in which the spurious information bounded by Theorem 3.3 can be measured exactly (Appendix F).

5 RESULTS

ISIC melanoma. We use ISIC 2018 (Codella et al., 2019) (2,594 dermoscopic images with pixel-level masks) and the trap sets of Bissoto et al. (2020), which make artefacts and diagnosis disagree at test time; coarse annotations are minimum bounding boxes (Figure 4). Segmentations are turned

Table 1: **ISIC melanoma trap sets** (50% fine annotations; mean ± std over 15 runs, 5 trap sets × 3 seeds). Robust AUROC: trap-set AUROC averaged over artefact categories. Certified rows: $\delta = 10^{-5}$. Bold: best non-reference method. $^{\blacktriangle}$/$^{\blacktriangledown}$: significantly higher/lower than the original model (paired t-test over the matched runs, $p < 0.05$).

Method	Dice ↑	Robust AUROC ↑	Test AUROC ↑
Refinement			
Original model (bounding box)	0.161 ± 0.028	0.546 ± 0.035	0.552 ± 0.034
Retrain from scratch	$0.094 \pm 0.094^{\blacktriangledown}$	0.564 ± 0.073	0.567 ± 0.102
Mean Teacher	$0.116 \pm 0.072^{\blacktriangledown}$	$\mathbf{0.606 \pm 0.040}^{\blacktriangle}$	$\mathbf{0.624 \pm 0.030}^{\blacktriangle}$
CPS	$0.044 \pm 0.056^{\blacktriangledown}$	$0.590 \pm 0.059^{\blacktriangle}$	$0.607 \pm 0.050^{\blacktriangle}$
Heuristic unlearning (ours)			
Fine-tuning ($w_{D_f} = 1$)	$\mathbf{0.178 \pm 0.016}^{\blacktriangle}$	0.548 ± 0.043	0.563 ± 0.032
NegGrad-Seg ($w_{D_f} = 2$)	0.171 ± 0.017	0.551 ± 0.043	0.568 ± 0.031
Certified unlearning (ours)			
One-shot, $\varepsilon = 0.1$	0.125 ± 0.084	$\mathbf{0.604 \pm 0.063}^{\blacktriangle}$	$\mathbf{0.621 \pm 0.061}^{\blacktriangle}$
One-shot, $\varepsilon = 0.01$	0.162 ± 0.073	$0.586 \pm 0.050^{\blacktriangle}$	$0.609 \pm 0.034^{\blacktriangle}$
One-shot, $\varepsilon = 0.001$	0.142 ± 0.082	0.592 ± 0.095	0.603 ± 0.099

Table 2: **SIIM-ACR pneumothorax** (50% fine annotations; mean ± std over 12 runs, 3 data splits). Robust AUROC: worst-group AUROC over images with and without a chest tube. Certified rows: $\delta = 10^{-5}$. Bold: best non-reference method. $^{\blacktriangle}$/$^{\blacktriangledown}$: significantly higher/lower than the original model (paired t-test over the matched runs, $p < 0.05$).

Method	Dice ↑	Robust AUROC ↑	AUROC ↑
Refinement			
Original model (bounding box)	0.057 ± 0.024	0.598 ± 0.022	0.744 ± 0.021
Retrain from scratch	0.057 ± 0.023	0.588 ± 0.036	0.746 ± 0.080
Mean Teacher	0.067 ± 0.023	0.605 ± 0.023	0.761 ± 0.029
CPS	$0.075 \pm 0.020^{\blacktriangle}$	0.607 ± 0.024	$0.768 \pm 0.028^{\blacktriangle}$
Heuristic unlearning (ours)			
Fine-tuning ($w_{D_f} = 1$)	$0.078 \pm 0.013^{\blacktriangle}$	$\mathbf{0.610 \pm 0.018}^{\blacktriangle}$	$\mathbf{0.780 \pm 0.014}^{\blacktriangle}$
NegGrad-Seg ($w_{D_f} = 2$)	$\mathbf{0.081 \pm 0.021}^{\blacktriangle}$	$\mathbf{0.610 \pm 0.023}$	$\mathbf{0.779 \pm 0.017}^{\blacktriangle}$
Certified unlearning (ours)			
One-shot, $\varepsilon = 0.1$	0.078 ± 0.031	0.592 ± 0.051	0.745 ± 0.075
One-shot, $\varepsilon = 0.01$	0.075 ± 0.035	0.588 ± 0.050	0.740 ± 0.079
One-shot, $\varepsilon = 0.001$	$\mathbf{0.079 \pm 0.031}$	$\mathbf{0.608 \pm 0.024}$	$\mathbf{0.772 \pm 0.015}^{\blacktriangle}$

into image-level scores by average pixel pooling (Hooper et al., 2023), and *robust AUROC* is trap-set AUROC averaged over artefact categories. Robust AUROC is the empirical counterpart of Proposition 3.6: within a trap subgroup the artefact is constant and cannot separate the classes, so a model scores only by looking at the lesion. Because the trap sets are adversarial, the original model (trained on coarse masks) sits near chance. Dice measures thresholded localisation of malignant tissue, whereas the image score averages soft probabilities, so the two need not agree across methods; within a method they are positively correlated (Spearman $\rho = 0.35$–0.59).

SIIM-ACR pneumothorax. We run the complete pipeline on SIIM-ACR (Society for Imaging Informatics in Medicine, 2019), where the shortcut is a chest tube: the treatment for a pneumothorax, hence strongly dependent on it ($\mathbb{P}(\text{tube} \mid \text{pneumothorax}) = 0.74$ vs. 0.13 without), yet in a distinct anatomical region, a naturally occurring instance of disjoint support that conditional independence excludes. Pathology occupies 0.32% of pixels; *robust AUROC* is the worst-group AUROC over images with and without a chest tube, which again holds the artefact fixed within each group, using the 1,951 tube labels of Saab et al. (2022).

5.1 H1: Certified versus Heuristic Machine Unlearning

Certification costs little. On ISIC, the certified operator is the most shortcut-robust of our operators (Table 1): at every privacy budget it raises robust AUROC above heuristic unlearning (at $\varepsilon = 0.1$: 0.604 vs. 0.551, $p = 0.007$), whereas heuristic unlearning barely moves the original model off its shortcut. The price is segmentation: heuristic unlearning keeps the sharper masks of the deployed model (Dice 0.171 vs. 0.125), because the certified noise erases most of what the original model learned. On SIIM-ACR, where heuristic unlearning is the strongest family, certification is statistically free: at $\varepsilon = 0.001$ the certified operator is within 0.007 AUROC and 0.002 Dice of NegGrad-Seg, and no difference is significant at any privacy budget ($p \geq 0.68$; Table 2). Tightening the privacy budget from $\varepsilon = 0.1$ to 0.001 changes neither picture, so the strongest guarantee comes at no additional cost. **H1 therefore holds on both datasets.**

5.2 H2: Machine Unlearning Spans the Pareto Front

We compare the families on four axes: segmentation (Dice), downstream classification (robust AUROC), sample efficiency, varying the share of training images that carry fine masks from 10 to 50% while every method sees the same images and bounding boxes, and time efficiency, the wall-clock cost relative to retraining from scratch on all images, which we approximate by the cost of training the original model (Tables 1, 2 and 6, Figure 2; Dice at every budget in Appendix I.2).

ISIC. The two unlearning families span the Pareto front. Heuristic unlearning holds its segmentation end: it keeps the original model's masks (Dice 0.147–0.178) at 0.24–0.31 of the cost of retraining, but barely moves the shortcut. The certified operator holds its robustness end: with 25 and 50% fine masks it raises robust AUROC above the original model and heuristic unlearning ($p \leq 0.007$), at 0.22 of the cost. The refinement baselines reach at most the same robustness (Mean Teacher at 50%: 0.606 vs. 0.604, $p = 0.86$) with no better segmentation, at 2.3–3.3 times the certified operator's cost (Table 6); retraining costs as little but is significantly less robust at every budget. With 10% fine masks, every method except retraining stays close to the original model (robust AUROC 0.537–0.554).

SIIM-ACR. Heuristic unlearning spans the front on its own: at every fine-mask budget it is the most shortcut-robust family or statistically tied with it, segments best (Dice 0.074–0.081, against 0.051–0.067 for Mean Teacher), and costs 0.47–0.50 of retraining, against 0.63 for CPS and 1.17 for Mean Teacher. Across all runs 50% study, NegGrad-Seg raises the original model's AUROC from 0.760 to 0.784 ($p = 2.3 \times 10^{-6}$), in images both with and without a chest tube.

When a guarantee is required. The only refinement method with a guarantee is exact retraining, and the certified operator is ahead of it at every fine-mask budget on both datasets, significantly so on ISIC ($p < 0.001$, $p = 0.003$ and $p = 0.03$ at 10, 25 and 50%). On SIIM-ACR it matches Mean Teacher's robustness ($p \geq 0.37$) at about 0.65 of the cost of retraining, but segments less well than heuristic unlearning at low budgets.

H2 therefore holds on both datasets: unlearning spans the Pareto front, and no refinement method reaches it.

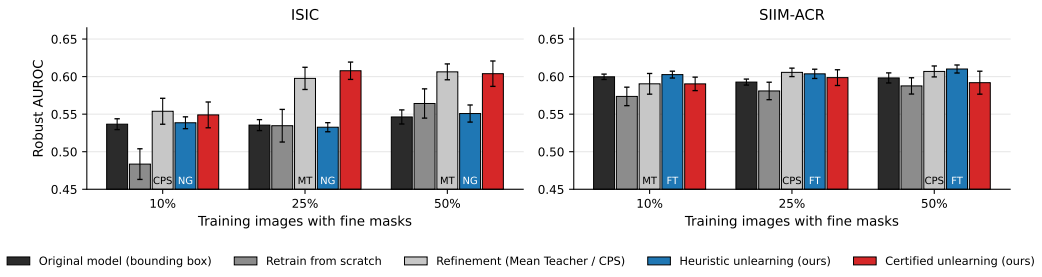


Figure 2: **Sample efficiency.** Robust AUROC of the best method of each family at each share of training images with fine masks (inside the bars: MT Mean Teacher, CPS Cross-Pseudo Supervision, FT fine-tuning, NG NegGrad-Seg); error bars: standard error. All methods in Table 6 (Appendix I.1).

6 Related Work

Machine unlearning. Machine unlearning removes the influence of a forget set from a trained model without retraining it from scratch. Exact approaches restructure training so that the affected components can be retrained cheaply (Bourtoule et al., 2021). Approximate approaches edit the trained weights, for instance by scrubbing the information a model holds about the forget set (Golatkar et al., 2020) or by combining descent on the retain set with ascent on the forget set, as in NegGrad+ and SCRUB (Kurmanji et al., 2023), on which our heuristic NegGrad-Seg builds. Certified unlearning instead requires the unlearned model to be (ε, δ)-indistinguishable from one retrained without the forget set (Sekhari et al., 2021), and has been obtained through randomised gradient smoothing (Zhang et al., 2022), second-order updates with calibrated noise (Zhang et al., 2024), and noisy gradient descent and Langevin dynamics (Chien et al., 2024a;b). The state of the art for deep networks is the noisy fine-tuning framework of Koloskova et al. (2025), which certifies unlearning without convexity assumptions through model clipping, gradient clipping and noise injected at every step, together with its refinement by Dolgova & Stich (2026), which spreads the noise over orthogonal parameter subspaces to prevent the accuracy collapse caused by noise injection. All of these methods remove training examples in classification. **We define forget sets over pixel constraints, show that per-step operators collapse on segmentation decoders, and give a certified algorithm that remains usable at tight budgets**.

Refinement from coarse or partial annotation. Learning segmentation when only part of the data carries fine labels is the domain of semi-supervised segmentation. Mean Teacher (Tarvainen & Valpola, 2017) encourages consistency between student predictions and those of a teacher whose weights are an exponential moving average of the student's weights; this framework also applies to pixelwise segmentation (French et al., 2019). Cross Pseudo Supervision (CPS) (Chen et al., 2021) trains two differently initialised segmentation networks using ground-truth masks on labelled images and each other's pseudo-labels. We use these established semi-supervised approaches as refinement baselines under a limited fine-mask budget.

Shortcuts in medical imaging. Shortcuts are documented in dermoscopy, where surgical markings, rulers and other artefacts raise melanoma scores (Winkler et al., 2019; Bissoto et al., 2020; Nauta et al., 2021), and in chest radiography, where chest drains correlate with pneumothorax (Jiménez-Sánchez et al., 2023; Olesen et al., 2024). Segmentation-for-classification restricts the classifier to the lesion (Hooper et al., 2023), and Saab et al. (2022) show that finer masks reduce spurious information, which they measure after training. **We give a worst-case guarantee on this information as a function of the privacy budget, for a model that is already deployed**.

7 Limitations

We assume the availability a retain set of fine annotations from expert annotators. The forget set is defined by the systematic difference between coarse and fine protocols, so biases of the fine protocol are inherited. The noise required for certified unlearning grows with the trainable dimension d (7). This is one of the factors for the collapse in noisy fine tuning methods. We alleviate this concern via transfer learning in this work. The guarantee covers spatially separate cues only; occluding artefacts are outside its scope. Our metrics diagnose shortcut reliance, not clinical validity.

8 Conclusion

We connected segmentation refinement to certified unlearning and proved that any certified unlearning algorithm bounds the spurious information of a segmenter relative to a fine-mask reference, and hence the accuracy any decision rule can extract from spurious cues. Empirically, certification is affordable: our certified unlearning operator holds its guarantee down to $\varepsilon = 0.001$, and matches or exceeds heuristic unlearning in shortcut robustness. Across segmentation, shortcut robustness, fine-mask budget and compute, unlearning spans the Pareto front on both datasets, while on melanoma refinement methods reach the certified operator's robustness only at more than twice its compute. Because unlearning corrects a deployed model rather than rebuilding it, its cost advantage should grow with model size, where retraining is prohibitively expensive.

AI USE STATEMENT

We used generative AI tools to support writing, rephrasing, and manuscript organisation. Additionally, we have used AI tooling to support the code implementation. We have reviewed all AI-assisted work, including code development work. We developed the experimental design, conducted the experiments, and constructed and verified the mathematical proofs without AI involvement. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our work aims to improve the trustworthiness of medical AI by reducing reliance on spurious cues. In positive terms, this can enhance diagnostic fairness. By focusing models on pathology rather than background, we help ensure performance does not vary unfairly across patient subgroups (e.g., if artefacts appear more in one group). Moreover, our certified unlearning framework aligns with data privacy principles, since it systematically removes information about specified data points (the coarse masks), which is conceptually similar to the "right to be forgotten" under GDPR. The present experiments evaluate shortcut reduction rather than patient-level deletion or clinical deployment; those uses require separate validation.

REPRODUCIBILITY STATEMENT

Section 2 defines the unlearning setting and Section 4 describes the evaluated operators, including the certified one-shot instantiation (Algorithm 1), and model architectures. Appendix G.1 specifies the synthetic data generator. Mathematical proofs appear in Appendix D. Appendix I lists the hyperparameters and Appendix I.3 the per-step operators; Section 5 gives the SIIM-ACR protocol. Every comparative claim reports its seed and split structure, and paired tests start all arms from an identical original model.

REFERENCES

Alceu Bissoto, Eduardo Valle, and Sandra Avila. Debiasing skin lesion datasets and models? not so fast. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3192–3201. IEEE, 2020.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pp. 141–159. IEEE, 2021.

Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2613–2622. IEEE, 2021.

Eli Chien, Haoyu Wang, Ziang Chen, and Pan Li. Certified machine unlearning via noisy stochastic gradient descent. *Advances in Neural Information Processing Systems*, 37:38852–38887, 2024a.

Eli Chien, Haoyu Wang, Ziang Chen, and Pan Li. Langevin unlearning: A new perspective of noisy gradient descent for machine unlearning. *Advances in neural information processing systems*, 37: 79666–79703, 2024b.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

Polina Dolgova and Sebastian U Stich. Sequential subspace noise injection prevents accuracy collapse in certified unlearning. *arXiv preprint arXiv:2601.05134*, 2026.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3-4):211–487, 2014.

Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*, 2019.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9301–9309. IEEE, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Sarah Hooper, Mayee Chen, Khaled Saab, Kush Bhatia, Curtis Langlotz, and Christopher Ré. A case for reframing automated medical image classification as segmentation. *Advances in Neural Information Processing Systems*, 36:55415–55441, 2023.

Amelia Jiménez-Sánchez, Dovile Juodelyte, Bethany Chamberlain, and Veronika Cheplygina. Detecting shortcuts in medical images-a case study in chest x-rays. In *2023 IEEE 20th international symposium on biomedical imaging (ISBI)*, pp. 1–5. IEEE, 2023.

Anastasia Koloskova, Youssef Allouah, Animesh Jha, Rachid Guerraoui, and Sanmi Koyejo. Certified unlearning for neural networks. *arXiv preprint arXiv:2506.06985*, 2025.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. *Advances in neural information processing systems*, 36:1957–1987, 2023.

Meike Nauta, Ricky Walsh, Adam Dubowski, and Christin Seifert. Uncovering and correcting shortcut learning in machine learning models for skin cancer diagnosis. *Diagnostics*, 12(1):40, 2021.

Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–46, 2025.

Vincent Olesen, Nina Weng, Aasa Feragen, and Eike Petersen. Slicing through bias: explaining performance gaps in medical image analysis using slice discovery methods. In *MICCAI Workshop on Fairness of AI in Medical Imaging*, pp. 3–13. Springer, 2024.

Khaled Saab, Sarah Hooper, Mayee Chen, Michael Zhang, Daniel Rubin, and Christopher Ré. Reducing reliance on spurious features in medical image classification with spatial specificity. In *Machine Learning for Healthcare Conference*, pp. 760–784. PMLR, 2022.

Jessica Schrouff, Natalie Harris, Sanmi Koyejo, Ibrahim M Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alexander Brown, Subhrajit Roy, Diana Mincu, Christina Chen, et al. Diagnosing failures of fairness transfer across distribution shift in real-world medical settings. *Advances in Neural Information Processing Systems*, 35:19304–19318, 2022.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34:18075–18086, 2021.

Society for Imaging Informatics in Medicine. SIIM-ACR pneumothorax segmentation. Kaggle, 2019. URL `https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation`.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Julia K Winkler, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm Stolz, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA dermatology*, 155(10):1135–1141, 2019.

Binchi Zhang, Yushun Dong, Tianhao Wang, and Jundong Li. Towards certified unlearning for deep neural networks. *arXiv preprint arXiv:2408.00920*, 2024.

Zijie Zhang, Yang Zhou, Xin Zhao, Tianshi Che, and Lingjuan Lyu. Prompt certified machine unlearning with randomized gradient smoothing and quantization. *Advances in Neural Information Processing Systems*, 35:13433–13455, 2022.

A Mask Refinement as Data Removal

A.1 The Atomic Data Space

To bridge segmentation and unlearning, we define the atomic data unit not as an image, but as a *pixel constraint*. Let the atomic data space be $\mathcal{Z} = \mathcal{X} \times \Lambda$, where $\Lambda = \{1, \dots, H\} \times \{1, \dots, W\}$ is the pixel grid.

Definition A.1 (The Empirical Universe)**.** The *empirical universe* $\mathcal{E}$ is the set of all possible atomic constraints derived from the N training images:

$$\mathcal{E} := \{z_{n,i,j} = (X_n, (i,j)) \mid n \in \{1, \dots, N\}, (i,j) \in \Lambda\} \subset \mathcal{Z}.$$

Remark A.2. When $r = 1$, the dilation corresponds to a global image-level label. While our theoretical results often utilise the limit $r \to \infty$, in the discrete pixel grid setting, $r > \max\{H, W\}$ is sufficient to recover exact pixel-precision.

A.2 The Unlearning Isomorphism

We now map the properties of the dilation operator to the set-theoretic operations of machine unlearning. Consider the transition from a coarse mask $Y^{(r_1)}$ to a finer mask $Y^{(r_2)}$ (where $r_1 < r_2$). Due to the monotonicity of dilation, the set of annotated pixels in the finer mask is a subset of the coarse mask. This nesting property allows us to identify the "correction" of a mask with the "forgetting" of dilation artefacts.

The Full Dataset ($D \equiv D^{(r_1)}$): The set of positive constraints imposed by the coarse annotations $Y^{(r_1)}$.

The Forget Set (D_f): The set of "spurious" positive constraints, namely pixels labelled as foreground solely due to dilation artefacts.

$$D_f := D \setminus D^{(r_2)} = \left\{ z_{n,i,j} \in D \mid \left(Y_n^{(r_2)}\right)_{ij} = 0 \right\}$$

Mechanically, removing these elements corresponds to "flipping" the label of pixel (i, j) from foreground to background.

The Retained Set (D_r): The set of constraints imposed by the ideal fine annotations $Y^{(r_2)}$:

$$D_r := D \setminus D_f = D^{(r_2)}.$$

Remark A.3. The change in dilation radius, $\Delta_{r_1} - \Delta_{r_2}$, controls the geometry and size of the forget region. Since $\Delta_r = \lfloor \max(H, W)/r \rfloor$, this change is not equal to $r_2 - r_1$. The forget set consists of the foreground constraints removed between these two radii. In practice, the refinement is determined by the annotation protocol (e.g. bounding box vs. expert mask) rather than being a tuneable hyperparameter.

B Further Discussion of the Theory

Preserved prediction mass.

Corollary B.1 (Preserved prediction mass)**.** *Under Definition 2.2, for every pixel (i, j) and class $k \in \{1, \dots, K\}$,*

$$\left|\mathbb{P}(\hat{Y}_{ij} = k) - \mathbb{P}(Y_{ij}^{(r_2)} = k)\right| \leq (e^{\varepsilon} - 1)\, \mathbb{P}(Y_{ij}^{(r_2)} = k) + \delta.$$

The proof (Appendix D.6) uses the law of total probability over Y_{ij}; this preserved mass is the weight each pixel carries in $\mathcal{S}_{\text{global}}$ below.

Scope of the assumption. Disjoint support restricts geometry, not statistics. It holds exactly for most catalogued shortcuts, where cue and lesion are spatially separate (surgical ink, rulers, dark corners, gel borders, chest tubes); a chest tube is the treatment for a pneumothorax, hence strongly dependent on it, yet lies in a distinct anatomical region (Section 5). It fails for occluding cues such as dense hair over a melanoma, which lie outside the scope of our guarantee.

Remark B.2. This is a stability guarantee around the fine-mask reference: a low-dependence reference yields a low-dependence unlearned mask when the certification error is small, but not exact independence at nonzero error.

Remark B.3 (On the hypothesis). Definition 2.2 alone does not suffice: it constrains how often $\hat{Y}$ fires on background, not where, so an operator placing its false positives on artefact pixels satisfies it with $\varepsilon = \delta = 0$ yet relies on the shortcut. The proof uses Lemma D.3 on $\{Y_{ij} = 0, S_{ij} = 1\}$, which every operator certified per Definition 2.1 satisfies.

What is new in Theorem 3.3. The theorem is not an instance of an existing indistinguishability guarantee: it transfers a guarantee of one kind (an (ε, δ) privacy budget on outputs) into a bound of another kind (on conditional spurious information), and to our knowledge no prior result bounds spurious information as a function of a privacy budget. Indistinguishability controls probabilities multiplicatively, whereas mutual information is a nonlinear functional of them that is singular near zero. The proof therefore decomposes both the background posterior and the artefact mass of the unlearned mask against the reference (sandwiching $m^{\mathcal{U}}$ rather than $\rho^{\mathcal{U}}$, which keeps constants uniform in r_2), and then controls the representation $\mathcal{J}$ through a stability estimate (Lemma D.2) that yields the $\mathcal{O}(\delta \log(1/\delta))$ rate.

C Certification of the One-Shot Operator

Coupling. Fix the training procedure and its data-independent initialisation. Let (X, X') be a coupling with marginals $X \sim \mathcal{A}(D)$ and $X' \sim \mathcal{A}(D_r)$. The runs may share random seeds and sampling randomness, provided each run retains the prescribed training distribution on its own dataset. Following Dolgova & Stich (2026), our certificate requires a fixed discrepancy bound satisfying

$$\Pr[\|X - X'\|_2 \le \Delta] = 1.$$

The bound must hold for every dataset and deletion request covered by the claim, and its selection must not depend on the forget data. Both worlds use the same Δ and noise variance.

Noise calibration. For $\varepsilon > 0$ and $\delta \in (0, 1)$, draw $\xi \sim \mathcal{N}(0, \sigma^2 I_d)$ independently of training. Optimising the standard Rényi-to-(ε, δ) conversion over the Rényi order gives (Dolgova & Stich, 2026, Appendix H.1)

$$\begin{aligned} a &= \log(1/\delta), \\ q^* &= 1 + \frac{a + \sqrt{a(a+\varepsilon)}}{\varepsilon}, \\ \varepsilon_r^* &= \varepsilon + a - \sqrt{a(a+\varepsilon)}, \\ \sigma &= \Delta \sqrt{\frac{q^*}{2\varepsilon_r^*}}. \end{aligned} \tag{7}$$

Privacy accounting. For any coupled pair (x, x') with $\|x - x'\|_2 \le \Delta$, the Gaussian Rényi divergence satisfies

$$D_{q^*}\big(\mathcal{N}(x, \sigma^2 I_d) \,\big\|\, \mathcal{N}(x', \sigma^2 I_d)\big) = \frac{q^* \|x - x'\|_2^2}{2\sigma^2} \le \varepsilon_r^*,$$

with the same bound in the reverse direction. Since

$$\varepsilon_r^* + \frac{\log(1/\delta)}{q^* - 1} = \varepsilon,$$

the two Gaussian releases satisfy the (ε, δ) indistinguishability inequalities. Averaging these inequalities over the coupling gives the same guarantee for the random training outputs.

Let R_{D_r} denote the rescaling and recovery procedure in Algorithm 1. Its target norms are fixed by x_{init}, its optimiser state is reset, and it uses only retained or public validation data and fresh randomness. The certified comparison is between

$$Y = R_{D_r}(X + \xi) \quad \text{and} \quad Y' = R_{D_r}(X' + \xi'), \qquad \xi, \xi' \sim \mathcal{N}(0, \sigma^2 I_d),$$

where each Gaussian is independent of training. By post-processing, for every measurable set O,

$$\Pr[Y \in O] \leq e^{\varepsilon} \Pr[Y' \in O] + \delta,$$
$$\Pr[Y' \in O] \leq e^{\varepsilon} \Pr[Y \in O] + \delta.$$

Thus the certifying reference is retain-only retraining followed by the same perturbation, rescaling, and recovery; it is not unperturbed $\mathcal{A}(D_r)$.

If the proximity bound instead holds with probability $1 - \rho$, the same calibration yields $(\varepsilon, \delta + \rho)$-unlearning. The main text uses $\delta = 10^{-5}$ and $\varepsilon \in \{0.1, 0.01, 0.001\}$; these are the complete certified budgets only under the almost-sure proximity assumption.

D PROOFS

D.1 PRELIMINARIES

Lemma D.1. *Let $p \in (0,1)$. Then the entropy $H_b(p)$ of a* Bernoulli(p) *random variable satisfies*

$$H_b(p) = -p\log(p) + \mathcal{O}(p),$$

as $p \downarrow 0$.

Proof. We have

$$\begin{aligned} H_b(p) &= -p\log(p) - (1-p)\log(1-p) \\ &= -p\log(p) + (1-p)\sum_{k=1}^{\infty}\frac{p^k}{k} \\ &= -p\log(p) + \mathcal{O}(p). \end{aligned}$$

□

The following stability estimate is used in the proof of Theorem 3.3.

Lemma D.2. *For $0 \le \mu \le a < 1$ let*

$$\mathcal{J}(a,\mu) := H_b(\mu) - aH_b(\mu/a) = \varphi(a) - \varphi(a-\mu) + \varphi(1-\mu), \qquad \varphi(x) := -x\log x. \tag{8}$$

Fix $\gamma \in (0,1)$ and $C > 0$. Suppose $0 \le m \le \alpha \le \gamma$ with $m/\alpha \le \gamma$, and

$$\alpha' = \alpha(1+\eta_a) + c_a\delta, \qquad m' = m(1+\eta_m) + c_m\delta,$$

with $|\eta_a|$ and $|\eta_m|$ at most $2\varepsilon + \mathcal{O}(\varepsilon^2)$, with $|c_a|$ and $|c_m|$ at most C, and with (α', m') in the same domain as (α, m). Then, for ε and δ sufficiently small,

$$|\mathcal{J}(\alpha', m') - \mathcal{J}(\alpha, m)| \le \mathcal{O}(\varepsilon) + \mathcal{O}(\delta\log(1/\delta)),$$

with constants depending only on γ and C.

Proof. Set $\tilde{\alpha} := \alpha(1+\eta_a)$ and $\tilde{m} := m(1+\eta_m)$, so that $\alpha' = \tilde{\alpha} + c_a\delta$ and $m' = \tilde{m} + c_m\delta$, and split

$$\mathcal{J}(\alpha', m') - \mathcal{J}(\alpha, m) = \big[\mathcal{J}(\tilde{\alpha}, \tilde{m}) - \mathcal{J}(\alpha, m)\big] + \big[\mathcal{J}(\alpha', m') - \mathcal{J}(\tilde{\alpha}, \tilde{m})\big]. \tag{9}$$

The first bracket is the effect of the factors $1+\eta_a$ and $1+\eta_m$, the second the effect of the shifts $c_a\delta$ and $c_m\delta$. We bound them in turn, using (8) to write each bracket as a sum of three differences of φ. Throughout, ε and δ are small enough that $|\eta_a|, |\eta_m| \le 1/2$, that $|\tilde{\eta}| \le 1/2$ for $\tilde{\eta}$ defined in (14) below, and that every argument of φ appearing lies in $[0,1]$, with $1-m$, $1-\tilde{m}$ and $1-m'$ in $[(1-\gamma)/2, 1]$.

Two elementary facts about φ are used. For $c > 0$ and $x \ge 0$,

$$\varphi(cx) = c\,\varphi(x) - cx\log c. \tag{10}$$

For $|\eta| \le 1/2$, the bound $\log t \le t - 1$ applied to $t = 1+\eta$ when $\eta \ge 0$, and to $t = (1+\eta)^{-1}$ when $\eta < 0$, gives $(1+\eta)|\log(1+\eta)| \le 2|\eta|$. Hence (10) with $c = 1+\eta$ and $x \in [0,1]$ yields

$$\big|\varphi((1+\eta)x) - \varphi(x)\big| = \big|\eta\,\varphi(x) - (1+\eta)x\log(1+\eta)\big| \le |\eta|\,\varphi(x) + 2|\eta|x \le |\eta|\,(e^{-1}+2). \tag{11}$$

For $x, x+t \in [0,1]$ with $t \ge 0$, concavity of φ with $\varphi(0) = 0$ gives $\varphi(x+t) \le \varphi(x) + \varphi(t)$, while $\varphi(x) - \varphi(x+t) = \int_x^{x+t}(1+\log\xi)\,\mathrm{d}\xi \le t$ because $\log\xi \le 0$ on $(0,1]$. Exchanging the roles of x and $x+t$ when $t < 0$,

$$\big|\varphi(x+t) - \varphi(x)\big| \le \varphi(|t|) + |t|. \tag{12}$$

For the first bracket in (9), the definition (8) gives

$$\begin{aligned} \mathcal{J}(\tilde{\alpha}, \tilde{m}) - \mathcal{J}(\alpha, m) = &\big[\varphi(\tilde{\alpha}) - \varphi(\alpha)\big] - \big[\varphi(\tilde{\alpha} - \tilde{m}) - \varphi(\alpha - m)\big] \\ &+ \big[\varphi(1-\tilde{m}) - \varphi(1-m)\big]. \end{aligned} \tag{13}$$

Since $\tilde{\alpha} = (1+\eta_a)\alpha$, the first difference is bounded by (11) with $\eta = \eta_a$. For the second,

$$\tilde{\alpha} - \tilde{m} = (\alpha - m)(1+\tilde{\eta}), \qquad \tilde{\eta} := \frac{\eta_a \alpha - \eta_m m}{\alpha - m}, \tag{14}$$

and $\alpha - m \geq \alpha(1-\gamma)$ together with $m \leq \alpha$ gives $|\tilde{\eta}| \leq (|\eta_a| + |\eta_m|)/(1-\gamma)$, so (11) applies with $\eta = \tilde{\eta}$. For the third, $1 - \tilde{m} = (1-m) - m\eta_m$, and the mean value theorem gives $|\varphi(1-\tilde{m}) - \varphi(1-m)| \leq m|\eta_m| \sup|\varphi'| \leq |\eta_m|\big(1 + \log\frac{2}{1-\gamma}\big)$, the supremum being over $[(1-\gamma)/2, 1]$, where $|\varphi'(\xi)| = |1 + \log\xi| \leq 1 + \log\frac{2}{1-\gamma}$. Adding the three bounds,

$$\begin{aligned}\big|\mathcal{J}(\tilde{\alpha},\tilde{m}) - \mathcal{J}(\alpha,m)\big| &\leq (e^{-1}+2)\Big(|\eta_a| + \frac{|\eta_a| + |\eta_m|}{1-\gamma}\Big) + |\eta_m|\Big(1 + \log\frac{2}{1-\gamma}\Big) \\ &= \mathcal{O}(\varepsilon).\end{aligned} \tag{15}$$

For the second bracket in (9), the definition (8) gives

$$\begin{aligned}\mathcal{J}(\alpha',m') - \mathcal{J}(\tilde{\alpha},\tilde{m}) = \big[\varphi(\tilde{\alpha} + c_a\delta) - \varphi(\tilde{\alpha})\big] - \big[\varphi(\tilde{\alpha} - \tilde{m} + (c_a - c_m)\delta) - \varphi(\tilde{\alpha} - \tilde{m})\big] \\ + \big[\varphi(1 - \tilde{m} - c_m\delta) - \varphi(1-\tilde{m})\big].\end{aligned} \tag{16}$$

Each difference is of the form (12) with $|t| \leq 2C\delta$, and $t \mapsto \varphi(t) + t$ is increasing on $[0, e^{-1}]$, so

$$\big|\mathcal{J}(\alpha',m') - \mathcal{J}(\tilde{\alpha},\tilde{m})\big| \leq 3\big(\varphi(2C\delta) + 2C\delta\big) = \mathcal{O}(\delta\log(1/\delta)). \tag{17}$$

Combining (9), (15) and (17) proves the lemma. □

D.2 Conditional Pixel-Level Projection

The following lemma translates certified machine unlearning into certified pixel-level unlearning: any certified operator is indistinguishable from the fine-mask reference on every slice of the generative variables, of which (5)–(6) are the instance $E = \{Y_{ij} = y\}$.

Lemma D.3 (Conditional pixel-level projection)**.** *Assume Definition 2.1 holds for every fixed dataset. Then for every pixel (i,j), every class k, and every event E measurable with respect to (Y,S) with $\mathbb{P}(E) > 0$,*

$$\mathbb{P}(\hat{Y}_{ij} = k \mid E) \leq e^{\varepsilon}\,\mathbb{P}(Y_{ij}^{(r_2)} = k \mid E) + \delta, \tag{18}$$

$$\mathbb{P}(Y_{ij}^{(r_2)} = k \mid E) \leq e^{\varepsilon}\,\mathbb{P}(\hat{Y}_{ij} = k \mid E) + \delta. \tag{19}$$

Proof. Write $V := (X,Y,S)$ and μ for its law. A hypothesis $h \in \mathcal{H}$ is a model mapping an image x to a mask $h(x) \in \{0,\dots,K\}^{H\times W}$, with $h(x)_{ij}$ its predicted class at pixel (i,j). Both $\mathcal{U}$ and $\mathcal{A}$ return models rather than masks, and the masks in the lemma are their outputs on the observed image, $\hat{Y} = \mathcal{U}(D)(X)$ and $Y^{(r_2)} = \mathcal{A}(D \setminus D_f)(X)$, where D and D_f are built from the annotations in V.

Let $g_{\mathcal{U}}$ and $g_{\mathcal{A}}$ be the measurable functions of v that represent the conditional probabilities given V,

$$g_{\mathcal{U}}(v) := \mathbb{P}\big(\hat{Y}_{ij} = k \mid V = v\big), \qquad g_{\mathcal{A}}(v) := \mathbb{P}\big(Y_{ij}^{(r_2)} = k \mid V = v\big),$$

so that $\mathbb{P}(\hat{Y}_{ij} = k \mid V) = g_{\mathcal{U}}(V)$ and $\mathbb{P}(Y_{ij}^{(r_2)} = k \mid V) = g_{\mathcal{A}}(V)$ almost surely. For $E \in \sigma(Y,S) \subseteq \sigma(V)$ the indicator $\mathbf{1}_E$ is $\sigma(V)$-measurable, so the tower property gives

$$\mathbb{P}\big(\{\hat{Y}_{ij} = k\} \cap E\big) = \mathbb{E}\big[\mathbf{1}_E\,\mathbf{1}_{\{\hat{Y}_{ij}=k\}}\big] = \mathbb{E}\Big[\mathbf{1}_E\,\mathbb{E}\big[\mathbf{1}_{\{\hat{Y}_{ij}=k\}} \mid V\big]\Big] = \mathbb{E}\big[\mathbf{1}_E\,g_{\mathcal{U}}(V)\big], \tag{20}$$

and likewise $\mathbb{P}(\{Y_{ij}^{(r_2)} = k\} \cap E) = \mathbb{E}[\mathbf{1}_E\,g_{\mathcal{A}}(V)]$.

The hypothesis enters through the values of $g_{\mathcal{U}}$ and $g_{\mathcal{A}}$. Fix $v = (x,y,s)$ and let $\mathcal{T}_k(v) := \{h \in \mathcal{H} : h(x)_{ij} = k\}$ be the measurable set of models labelling pixel (i,j) of x as k, so that $\{\hat{Y}_{ij} = k\} = \{\mathcal{U}(D) \in \mathcal{T}_k(V)\}$. Both $\mathcal{U}$ and $\mathcal{A}$ read V only through their dataset argument, and their internal randomness is generated independently of V. Conditionally on $V = v$ the model $\mathcal{U}(D)$ therefore has the law of $\mathcal{U}$ run on the fixed dataset $D(v)$, and likewise for $\mathcal{A}$. Hence, for μ-almost every v,

$$g_{\mathcal{U}}(v) = \mathbb{P}\big(\mathcal{U}(D(v)) \in \mathcal{T}_k(v)\big), \qquad g_{\mathcal{A}}(v) = \mathbb{P}\big(\mathcal{A}(D(v) \setminus D_f(v)) \in \mathcal{T}_k(v)\big). \tag{21}$$

Applying the hypothesis of the lemma at the dataset $D(v)$ and the set $\mathcal{T}_k(v)$ gives, for μ-almost every v,

$$g_{\mathcal{U}}(v) \le e^{\varepsilon} g_{\mathcal{A}}(v) + \delta. \tag{22}$$

Dividing (20) by $\mathbb{P}(E) > 0$ and using (22),

$$\mathbb{P}\big(\hat{Y}_{ij} = k \mid E\big) = \mathbb{E}\big[g_{\mathcal{U}}(V) \mid E\big] \le \mathbb{E}\big[e^{\varepsilon} g_{\mathcal{A}}(V) + \delta \mid E\big] = e^{\varepsilon}\, \mathbb{P}\big(Y_{ij}^{(r_2)} = k \mid E\big) + \delta,$$

which is (18). The reverse hypothesis gives (19) identically. □

D.3 Proof for Proposition 3.2

Proof. Without loss of generality, let $k = 1$. Furthermore, let $\alpha^{(r)}$ denote the posterior probability that a pixel is background given it is predicted as abnormal, and let $\rho^{(r)}$ denote the conditional probability that a spurious feature is present given a false positive prediction,

$$\alpha^{(r)} := \mathbb{P}\left(Y_{ij} = 0 \mid Y_{ij}^{(r)} = 1\right),$$
$$\rho^{(r)} := \mathbb{P}\left(S_{ij} = 1 \mid Y_{ij} = 0, Y_{ij}^{(r)} = 1\right).$$

Let r_0 be minimal such that for $r \ge r_0$ the quantities $\alpha^{(r)}$ and $\rho^{(r)}$ are both decreasing in r and $\alpha^{(r)} \le e^{-1}$. Such an r_0 exists because $\alpha^{(r)} \to 0$ as $r \to \infty$. On this range $I(S_{ij}, Y_{ij} \mid Y_{ij}^{(r)} = 1) = \mathcal{J}(\alpha^{(r)}, \rho^{(r)}\alpha^{(r)}) = H_b(\rho^{(r)}\alpha^{(r)}) - \alpha^{(r)} H_b(\rho^{(r)})$ is nondecreasing in each of $\alpha^{(r)}$ and $\rho^{(r)}$ separately, since $\partial_\rho\big[H_b(\rho\alpha) - \alpha H_b(\rho)\big] = \alpha\left[\log\frac{1-\alpha\rho}{\alpha\rho} - \log\frac{1-\rho}{\rho}\right] \ge 0$ for $\alpha \le 1$ and $\partial_\alpha\big[H_b(\rho\alpha) - \alpha H_b(\rho)\big] = \rho \log\frac{1-\alpha\rho}{\alpha\rho} - H_b(\rho) \ge 0$ for $\alpha \le e^{-1}$, so the conditional mutual information is decreasing in r. It remains to establish the asymptotic expansion

$$I\left(S_{ij}, Y_{ij} \mid Y_{ij}^{(r)} = 1\right) = -\rho^{(r)}\alpha^{(r)} \log\left(\rho^{(r)}\alpha^{(r)}\right) + \mathcal{O}\left(\alpha^{(r)}\right), \tag{23}$$

as $r \to \infty$.

By standard properties of conditional mutual information, we decompose the term of interest into the difference of marginal and conditional entropies,

$$I\left(S_{ij}, Y_{ij} \mid Y_{ij}^{(r)} = 1\right) = H\left(S_{ij} \mid Y_{ij}^{(r)} = 1\right) - H\left(S_{ij} \mid Y_{ij}, Y_{ij}^{(r)} = 1\right). \tag{24}$$

We first analyze the marginal entropy term, $H(S_{ij} \mid Y_{ij}^{(r)} = 1)$. Using the law of total probability and the constraint that spurious features and abnormalities are non-overlapping (i.e., $S_{ij} = 0$ whenever $Y_{ij} = 1$), we derive the probability of a spurious feature being present within the mask,

$$\begin{aligned}
\mathbb{P}\left(S_{ij} = 1 \mid Y_{ij}^{(r)} = 1\right) &= \sum_{y \in \{0,\dots,K\}} \mathbb{P}\left(S_{ij} = 1 \mid Y_{ij}^{(r)} = 1, Y_{ij} = y\right) \mathbb{P}\left(Y_{ij} = y \mid Y_{ij}^{(r)} = 1\right) \\
&= \mathbb{P}\left(S_{ij} = 1 \mid Y_{ij}^{(r)} = 1, Y_{ij} = 0\right) \alpha^{(r)} \\
&= \rho^{(r)}\alpha^{(r)}.
\end{aligned}$$

Consequently, the first term is simply the entropy of a Bernoulli variable with parameter $\rho^{(r)}\alpha^{(r)}$,

$$H\left(S_{ij} \mid Y_{ij}^{(r)} = 1\right) = H_b\left(\rho^{(r)}\alpha^{(r)}\right),$$

where $H_b(p)$ denotes the binary entropy function.

Similarly, for the second term in (24), we expand the conditional entropy over the values of Y_{ij}. Since S_{ij} is deterministic when $Y_{ij} = 1$, the corresponding entropy term vanishes, giving

$$\begin{aligned}
H\left(S_{ij} \mid Y_{ij}, Y_{ij}^{(r)} = 1\right) &= \sum_{y \in \{0,\dots,K\}} H\left(S_{ij} \mid Y_{ij}^{(r)} = 1, Y_{ij} = y\right) \mathbb{P}\left(Y_{ij} = y \mid Y_{ij}^{(r)} = 1\right) \\
&= H\left(S_{ij} \mid Y_{ij}^{(r)} = 1, Y_{ij} = 0\right) \alpha^{(r)} \\
&= \alpha^{(r)} H_b(\rho^{(r)}).
\end{aligned}$$

Combining the above, the fact that $\rho^{(r)}$ is non-negative and bounded, and using Lemma D.1 we can re-write (24) as

$$I(S_{ij}, Y_{ij} \mid Y_{ij}^{(r)} = 1) = H_b\left(\rho^{(r)}\alpha^{(r)}\right) - \alpha^{(r)} H_b\left(\rho^{(r)}\right) = -\rho^{(r)}\alpha^{(r)} \log\left(\rho^{(r)}\alpha^{(r)}\right) + \mathcal{O}\left(\alpha^{(r)}\right), \tag{25}$$

as $r \to \infty$, which completes the proof. □

D.4 Proof for Theorem 3.3

Proof. We focus on the binary case ($K = 1$). The multi-class setting reduces to this case under the single-class assumption of Section 2. For $r \geq 1$ let $\alpha^{(r)}$ be defined as in the proof of Proposition 3.2. Similarly, let $\alpha^{\mathcal{U}}$ denote the posterior probability that a pixel is background given it is predicted as abnormal after unlearning,

$$\alpha^{\mathcal{U}} := \mathbb{P}\left(Y_{ij} = 0 \mid \mathcal{U}\left[Y^{(r_1)}\right]_{ij} = 1\right).$$

For ease of exposition, let

$$\begin{aligned}
\pi_y &:= \mathbb{P}\left(Y_{ij} = y\right), \\
p_y^{(r)} &:= \mathbb{P}\left(Y_{ij}^{(r)} = 1 \mid Y_{ij} = y\right), \\
q_y &:= \mathbb{P}\left(\mathcal{U}\left[Y^{(r_1)}\right]_{ij} = 1 \mid Y_{ij} = y\right),
\end{aligned}$$

for $y \in \{0, 1\}$. Then using Bayes' formula, the definition of (ε, δ)-unlearning, and the fact that $D := p_0^{(r)}\pi_0 + p_1^{(r)}\pi_1$ is uniformly bounded away from zero by some $C_0 > 0$ for $r \geq r_0$, we have

$$\begin{aligned}
\alpha^{\mathcal{U}} &= \frac{q_0\pi_0}{q_0\pi_0 + q_1\pi_1} \\
&\leq \frac{e^{\varepsilon}\left(e^{\varepsilon}p_0^{(r_2)} + \delta\right)\pi_0}{p_0^{(r_2)}\pi_0 + p_1^{(r_2)}\pi_1 - \delta} \\
&= e^{\varepsilon}\left(e^{\varepsilon}p_0^{(r_2)} + \delta\right)\pi_0 \sum_{i=0}^{\infty} \frac{\delta^i}{D^{i+1}} \\
&\leq e^{2\varepsilon}\left[\frac{p_0^{(r_2)}\pi_0}{p_0^{(r_2)}\pi_0 + p_1^{(r_2)}\pi_1}\right] + \delta\left[\left(\frac{e^{\varepsilon}}{C_0} + 1\right)\sum_{i=0}^{\infty}\frac{e^{\varepsilon}\pi_0\delta^i}{C_0^{i+1}}\right].
\end{aligned}$$

The same chain of inequalities gives the lower bound

$$\alpha^{\mathcal{U}} \geq e^{-2\varepsilon}\left(\alpha^{(r_2)} - \mathcal{O}(\delta)\right).$$

Thus, we may express $\alpha^{\mathcal{U}}$ as

$$\alpha^{\mathcal{U}} = \alpha^{(r_2)}(1 + \eta) + C_1\delta$$

for some coefficients satisfying $|C_1| \leq C_2 + \mathcal{O}(\delta) + \mathcal{O}(\varepsilon)$ and $|\eta| \leq 2\varepsilon + \mathcal{O}(\varepsilon^2)$, where $C_2 > 0$, and C_1 is independent of r_2. Note this decomposition is exact, so all terms implicitly ensure that $\alpha^{\mathcal{U}}$ remains in $[0, 1]$.

By the computation leading to (25), applied verbatim with $\hat{Y} := \mathcal{U}[Y^{(r_1)}]$ in place of $Y^{(r)}$ (only the disjoint support property and the law of total probability are used there),

$$I\left(S_{ij}, Y_{ij} \mid \hat{Y}_{ij} = 1\right) = H_b\left(m^{\mathcal{U}}\right) - \alpha^{\mathcal{U}} H_b\left(m^{\mathcal{U}}/\alpha^{\mathcal{U}}\right), \qquad m^{\mathcal{U}} := \mathbb{P}\left(S_{ij} = 1 \mid \hat{Y}_{ij} = 1\right). \tag{26}$$

The same computation with $Y^{(r_2)}$ in place of $\hat{Y}$ gives

$$I\left(S_{ij}, Y_{ij} \mid Y_{ij}^{(r_2)} = 1\right) = H_b(m) - \alpha H_b(m/\alpha),$$

where $\alpha := \alpha^{(r_2)}$ and $m := \rho^{(r_2)}\alpha^{(r_2)} = \mathbb{P}(S_{ij} = 1 \mid Y_{ij}^{(r_2)} = 1)$. In the notation of Lemma D.2, the two mutual informations are $\mathcal{J}(\alpha^{\mathcal{U}}, m^{\mathcal{U}})$ and $\mathcal{J}(\alpha, m)$, and the theorem bounds their difference. If $\alpha^{\mathcal{U}} = 0$ then $m^{\mathcal{U}} = 0$ and the left side of (26) vanishes, so we may assume $\alpha^{\mathcal{U}} > 0$.

The decomposition of $\alpha^{\mathcal{U}}$ alone does not control this difference. The unlearned mutual information involves its own conditional probability $\rho^{\mathcal{U}} := \mathbb{P}(S_{ij} = 1 \mid Y_{ij} = 0, \hat{Y}_{ij} = 1) = m^{\mathcal{U}}/\alpha^{\mathcal{U}}$, so $m^{\mathcal{U}}$ must be decomposed against m in the same way. Let

$$\begin{aligned} \pi_S &:= \mathbb{P}(S_{ij} = 1, Y_{ij} = 0), \\ b &:= \mathbb{P}\left(Y_{ij}^{(r_2)} = 1 \mid Y_{ij} = 0, S_{ij} = 1\right), \\ b^{\mathcal{U}} &:= \mathbb{P}\left(\hat{Y}_{ij} = 1 \mid Y_{ij} = 0, S_{ij} = 1\right). \end{aligned}$$

Since $S_{ij} = 1$ forces $Y_{ij} = 0$, Bayes' formula gives

$$m = \frac{b\pi_S}{D}, \qquad m^{\mathcal{U}} = \frac{b^{\mathcal{U}}\pi_S}{\mathbb{P}(\hat{Y}_{ij} = 1)}.$$

Both numerators and both denominators are sandwiched by Lemma D.3. With $E = \{Y_{ij} = 0, S_{ij} = 1\}$, the inequalities (18)–(19) read

$$e^{-\varepsilon}(b - \delta) \le b^{\mathcal{U}} \le e^{\varepsilon} b + \delta,$$

and averaging (5)–(6) over y with weights π_y gives

$$e^{-\varepsilon}(D - \delta) \le \mathbb{P}(\hat{Y}_{ij} = 1) \le e^{\varepsilon} D + \delta.$$

Since $D \ge C_0$ and $\delta \le C_0/2$ by hypothesis, the denominator $\mathbb{P}(\hat{Y}_{ij} = 1)$ is bounded below by $e^{-\varepsilon}C_0/2$. Repeating the chain of inequalities used for $\alpha^{\mathcal{U}}$ with these bounds yields

$$m^{\mathcal{U}} = m(1 + \eta_m) + C_1'\delta, \qquad |\eta_m| \le 2\varepsilon + \mathcal{O}(\varepsilon^2), \qquad |C_1'| \le C_2' + \mathcal{O}(\varepsilon) + \mathcal{O}(\delta), \tag{27}$$

for some $C_2' > 0$, with C_1' independent of r_2. Sandwiching $m^{\mathcal{U}}$ rather than $\rho^{\mathcal{U}}$ is what keeps the additive constant uniform in r_2, since the corresponding decomposition of $\rho^{\mathcal{U}}$ carries a factor $1/p_0^{(r_2)}$.

It remains to apply Lemma D.2. Fix $\gamma \in (0, 1)$ with $\alpha^{(r_2)} \le \gamma$ and $\rho^{(r_2)} \le \gamma$. The first bound holds for $r_2 \ge r_0$ and every $\gamma \ge e^{-1}$, by the choice of r_0 in the proof of Proposition 3.2, and the second is the non-degeneracy hypothesis of the theorem. Thus (α, m) lies in the domain of the lemma, and for ε and δ sufficiently small so does $(\alpha^{\mathcal{U}}, m^{\mathcal{U}})$. The hypotheses of Lemma D.2 are therefore met with $(\alpha', m') = (\alpha^{\mathcal{U}}, m^{\mathcal{U}})$, with $\eta_a = \eta$ and $c_a = C_1$ from the decomposition of $\alpha^{\mathcal{U}}$, and with η_m and $c_m = C_1'$ from (27). The lemma gives

$$\begin{aligned} \left| I\left(S_{ij}, Y_{ij} \mid \hat{Y}_{ij} = 1\right) - I\left(S_{ij}, Y_{ij} \mid Y_{ij}^{(r_2)} = 1\right) \right| &= \left| \mathcal{J}\left(\alpha^{\mathcal{U}}, m^{\mathcal{U}}\right) - \mathcal{J}\left(\alpha, m\right) \right| \\ &\le \mathcal{O}(\varepsilon) + \mathcal{O}(\delta \log(1/\delta)). \end{aligned}$$

The claimed bound follows, its first term $2\varepsilon I(S_{ij}, Y_{ij} \mid Y_{ij}^{(r_2)} = 1)$ being nonnegative. The multiplicative form is retained in the statement for comparability with Remark B.2. All $\mathcal{O}(\cdot)$ constants depend only on the fixed data model, through C_0, γ and $(1 - \rho^{(r_2)})^{-1}$, as the constants in the decomposition of $\alpha^{\mathcal{U}}$ already do. □

D.5 Proof for Corollary 3.5

Proof. Let $\mathcal{S}_k(Z) := \sum_{i,j} \mathbb{P}\left(Z_{ij} = k\right) I\left(S_{ij}, Y_{ij} \mid Z_{ij} = k\right)$. Fix a class $k \in \{1, \ldots, K\}$. Using the definition of (ε, δ)-certified unlearning and Theorem 3.3, we analyze the difference term-wise for

$\mathcal{S}_k$ and obtain

$$\begin{aligned}
\mathcal{S}_k\left(\mathcal{U}\left[Y^{(r_1)}\right]\right) &= \sum_{i,j} \mathbb{P}\left(\mathcal{U}\left[Y^{(r_1)}\right]_{ij} = k\right) I\left(S_{ij}, Y_{ij} \mid \mathcal{U}\left[Y^{(r_1)}\right]_{ij} = k\right) \\
&= \sum_{i,j} \Bigg\{ \mathbb{P}\left(\mathcal{U}\left[Y^{(r_1)}\right]_{ij} = k \mid Y_{ij} = k\right) \mathbb{P}\left(Y_{ij} = k\right) \\
&\qquad + \mathbb{P}\left(\mathcal{U}\left[Y^{(r_1)}\right]_{ij} = k \mid Y_{ij} \neq k\right) \mathbb{P}\left(Y_{ij} \neq k\right) \Bigg\} \\
&\qquad\qquad \times I\left(S_{ij}, Y_{ij} \mid \mathcal{U}\left[Y^{(r_1)}\right]_{ij} = k\right) \\
&\leq \sum_{i,j} \left(e^{\varepsilon} \mathbb{P}\left(Y_{ij}^{(r_2)} = k\right) + \delta\right) \\
&\qquad\qquad \times \left(I\left(S_{ij}, Y_{ij} \mid Y_{ij}^{(r_2)} = k\right) + \mathcal{O}(\varepsilon) + \mathcal{O}(\delta \log(1/\delta))\right) \\
&= \mathcal{S}_k\left(Y^{(r_2)}\right) + \mathcal{O}(\varepsilon) + \mathcal{O}(\delta \log(1/\delta)).
\end{aligned}$$

Analogously, we obtain the lower bound, establishing that $|\mathcal{S}_k(\mathcal{U}[Y^{(r_1)}]) - \mathcal{S}_k(Y^{(r_2)})| \leq \mathcal{O}(\varepsilon) + \mathcal{O}(\delta \log(1/\delta))$ for all k. Finally, using the property that $|\max_k f(k) - \max_k g(k)| \leq \max_k |f(k) - g(k)|$, we conclude that

$$\begin{aligned}
\left|\mathcal{S}_{\text{global}}\left(\mathcal{U}\left[Y^{(r_1)}\right]\right) - \mathcal{S}_{\text{global}}\left(Y^{(r_2)}\right)\right| &= \left|\max_k \mathcal{S}_k\left(\mathcal{U}\left[Y^{(r_1)}\right]\right) - \max_k \mathcal{S}_k\left(Y^{(r_2)}\right)\right| \\
&\leq \max_k \left|\mathcal{S}_k\left(\mathcal{U}\left[Y^{(r_1)}\right]\right) - \mathcal{S}_k\left(Y^{(r_2)}\right)\right| \\
&\leq \mathcal{O}(\varepsilon) + \mathcal{O}(\delta \log(1/\delta)).
\end{aligned}$$

□

D.6 Proof for Corollary B.1

Proof. Write $\pi_y := \mathbb{P}(Y_{ij} = y)$. By the law of total probability and (5),

$$\begin{aligned}
\mathbb{P}(\hat{Y}_{ij} = k) &= \sum_y \mathbb{P}(\hat{Y}_{ij} = k \mid Y_{ij} = y)\, \pi_y \leq \sum_y \left(e^{\varepsilon}\, \mathbb{P}(Y_{ij}^{(r_2)} = k \mid Y_{ij} = y) + \delta\right) \pi_y \\
&= e^{\varepsilon}\, \mathbb{P}(Y_{ij}^{(r_2)} = k) + \delta,
\end{aligned}$$

so $\mathbb{P}(\hat{Y}_{ij} = k) - \mathbb{P}(Y_{ij}^{(r_2)} = k) \leq (e^{\varepsilon} - 1)\mathbb{P}(Y_{ij}^{(r_2)} = k) + \delta$. Averaging (6) in the same way gives $\mathbb{P}(Y_{ij}^{(r_2)} = k) \leq e^{\varepsilon}\mathbb{P}(\hat{Y}_{ij} = k) + \delta$, hence

$$\mathbb{P}(Y_{ij}^{(r_2)} = k) - \mathbb{P}(\hat{Y}_{ij} = k) \leq (1 - e^{-\varepsilon})\, \mathbb{P}(Y_{ij}^{(r_2)} = k) + e^{-\varepsilon}\delta \leq (e^{\varepsilon} - 1)\, \mathbb{P}(Y_{ij}^{(r_2)} = k) + \delta,$$

using $1 - e^{-\varepsilon} \leq e^{\varepsilon} - 1$ and $e^{-\varepsilon} \leq 1$. Combining the two directions proves the claim. □

D.7 Proof for Proposition 3.6

Proof. Write P for the joint law of (S_{ij}, Y_{ij}) conditional on E, P_S and P_Y for its marginals, and $Q := P_S \otimes P_Y$. Both variables take finitely many values. The Bayes errors are

$$R^*_{\emptyset} = 1 - \max_y P_Y(y), \qquad R^*_S = 1 - \sum_s \max_y P(s, y).$$

Since $\sum_s \max_y Q(s, y) = \sum_s P_S(s) \max_y P_Y(y) = \max_y P_Y(y)$,

$$R^*_{\emptyset} - R^*_S = \sum_s \Big[\max_y P(s, y) - \max_y Q(s, y)\Big] \leq \sum_s \max_y \big(P(s, y) - Q(s, y)\big),$$

where the inequality evaluates $Q(s,\cdot)$ at a maximiser of $P(s,\cdot)$. For each s, $\sum_y (P-Q)(s,y) = P_S(s) - P_S(s) = 0$, so the positive and negative parts of $(P-Q)(s,\cdot)$ have equal mass and $\max_y (P-Q)(s,y) \le \frac{1}{2}\sum_y |P-Q|(s,y)$. Summing over s gives $R^*_\emptyset - R^*_S \le \mathrm{TV}(P,Q)$. Pinsker's inequality gives $\mathrm{TV}(P,Q) \le \sqrt{\frac{1}{2}\mathrm{KL}(P\,\|\,Q)}$, and $\mathrm{KL}(P\,\|\,Q) = I(S_{ij}; Y_{ij} \mid E)$. □

Remark D.4 (Converse). Conversely, Fano's inequality lower-bounds the error P_e of any estimator of Y_{ij} from S_{ij} on E: with $Y_{ij} \in \{0,\dots,K\}$, $H_b(P_e) + P_e \log K \ge H(Y_{ij} \mid E) - I(S_{ij}; Y_{ij} \mid E)$. Small spurious information therefore leaves the artefact-based error close to the prior uncertainty of Y_{ij} on E.

E VIOLATION OF CONDITIONAL INDEPENDENCE IN PRIOR WORK

We demonstrate that the conditional independence assumption $S_{ij} \perp Y_{ij}^{(r)} \mid Y_{ij}$ employed by Saab et al. (2022) does not hold under their synthetic data generation process.

Causal Mechanism. The violation arises because the global label y acts as a common cause for both Y (and thus the dilated mask $Y^{(r)}$) and the spurious feature map S. Observing $Y_{ij}^{(r)} = 1$ while $Y_{ij} = 0$ reveals that an abnormality exists in a neighboring pixel, implying $y = 1$. Since $y = 1$ increases the probability of spurious features throughout the image (via correlation ρ), we obtain the dependency path $Y_{ij}^{(r)} \leftarrow Y_{\text{neigh}} \leftarrow y \rightarrow S \rightarrow S_{ij}$. Conditioning on Y_{ij} alone fails to block this path.

Counterexample. Consider a 3×3 grid where both Y and S have support 1×1. The data generation proceeds in two steps. Global labels $y, s \in \{0, 1\}$ are drawn independently. If $y = 1$ (resp. $s = 1$), the abnormality (resp. spurious feature) is placed uniformly at random on the grid with probability μ (resp. ρ) per pixel. For pixel $(1, 1)$ conditioned on $Y_{11} = 0$, independence would require $\mathbb{P}[Y_{11}^{(r)} = 1, S_{11} = 1 \mid Y_{11} = 0] = \mathbb{P}[Y_{11}^{(r)} = 1 \mid Y_{11} = 0] \cdot \mathbb{P}[S_{11} = 1 \mid Y_{11} = 0]$. However, direct computation yields

$$\mathbb{P}[S_{11} = 1 \mid Y_{11} = 0] = \frac{8\mu\rho}{9(9-\mu)}, \quad \mathbb{P}[Y_{11}^{(r)} = 1 \mid Y_{11} = 0] = \frac{3\mu}{9-\mu},$$
$$\mathbb{P}[Y_{11}^{(r)} = 1, S_{11} = 1 \mid Y_{11} = 0] = \frac{\mu\rho}{3(9-\mu)}.$$

The joint probability exceeds the product of marginals by

$$\Delta = \frac{\mu\rho}{3(9-\mu)} - \frac{8\mu\rho}{9(9-\mu)} \cdot \frac{3\mu}{9-\mu} = \frac{3\mu\rho(1-\mu)}{(9-\mu)^2},$$

which is strictly positive for every $\mu \in (0, 1)$ and $\rho > 0$, so the conditional independence assumption cannot be satisfied.

F SYNTHETIC VALIDATION OF THE THEORY

To verify the theory of Section 3 under full control, we generate 2,900 images (128×128, downsampled to 64×64 for training; 2,000/200/700 train/validation/test) in which pathology and a brighter artefact occupy disjoint squares, with $\mathbb{P}(s = 1 \mid y = 1) = 0.8$ and $\mathbb{P}(s = 1 \mid y = 0) = 0.2$ (Appendix G.1). The model is a UNet-inspired network (Saab et al., 2022) with approximately 15,000 parameters.

Table 3: **Synthetic benchmark.** Test Dice and spurious information $I(Z; S \mid Y)$ (bits). One-shot uses rescaling and $\delta = 10^{-5}$.

Method	Certified	Dice ↑	$I(Z; S \mid Y)$ ↓
Retrain from scratch	yes	0.873 ± 0.050	1.18×10^{-3}
Mean Teacher	no	0.841 ± 0.069	1.36×10^{-3}
CPS	no	0.886 ± 0.031	9.58×10^{-5}
Fine-tuning ($w_{D_f} = 1$)	no	0.881 ± 0.023	$\mathbf{4.34 \times 10^{-5}}$
NegGrad-Seg ($w_{D_f} = 2$)	no	$\mathbf{0.910 \pm 0.026}$	2.64×10^{-4}
One-shot, $\varepsilon = 0.1$	yes	0.855 ± 0.093	1.97×10^{-3}
One-shot, $\varepsilon = 0.01$	yes	0.888 ± 0.023	1.11×10^{-4}

Spurious information and reliance. Every refinement method keeps the measure the theory bounds at the 10^{-5}–10^{-3} level; the lowest value is attained by fine-tuning (4.34×10^{-5}). Along certified unlearning it tracks reliance: task accuracy rises as it falls across 6 seeds and 54 trajectory points (Spearman $\rho = -0.62$, $p = 6.6 \times 10^{-7}$; per-seed Pearson r negative on 5 of 6 seeds, down to -0.997).

G Synthetic Data and Additional Diagnostics

G.1 Synthetic Dataset Generation

Let $\mathcal{X} = \mathbb{R}^{n \times n}$ be the image space with $n = 128$. We simulate a binary classification task where labels $(y, s) \in \{0, 1\}^2$ are drawn with marginal disease prevalence $\mathbb{P}(y = 1) = 0.5$. To introduce strong spurious correlation, we set $\mathbb{P}(s = 1 \mid y = 1) = 0.8$ and $\mathbb{P}(s = 1 \mid y = 0) = 0.2$.

For each sample, we generate binary masks Y, S containing disjoint square regions of size 20×20 and 24×24, respectively, subject to the disjoint support assumption. The image X is synthesized via Gaussian sampling $X_{ij} \sim \mathcal{N}(\mu, \sigma^2)$, where the mean intensity is set to $\mu_y = 1.0$ for pathology pixels ($Y_{ij} = 1$), $\mu_s = 2.0$ for spurious pixels ($S_{ij} = 1$), and 0 otherwise. With noise $\sigma = 0.3$, the higher salience of artefacts ($\mu_s > \mu_y$) encourages the model to learn statistical shortcuts. To simulate varying degrees of spatial annotation granularity, we generate coarse masks $Y^{(r)}$ via the dilation operator defined in Eq. (1).

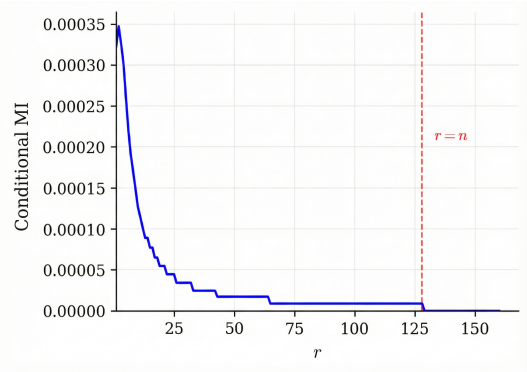


Figure 3: Numerical verification of Proposition 3.2 on the synthetic dataset. Increasing spatial specificity reduces the conditional mutual information between spurious features and the target label.

H ISIC Visualisations

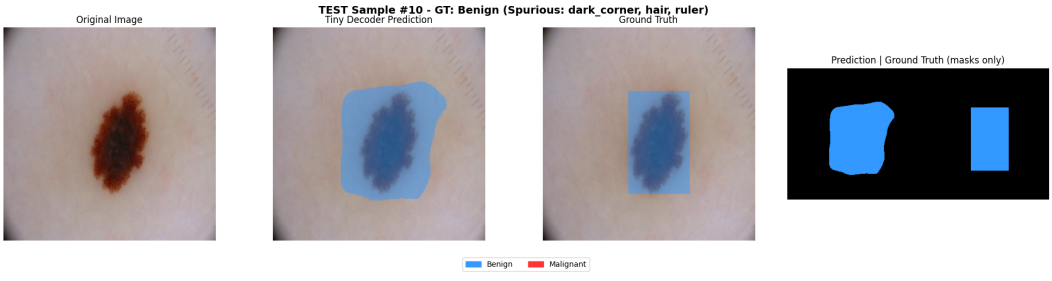


Figure 4: **Representative ISIC image and coarse annotation.** The original image, decoder prediction, bounding-box target, and masks are shown from left to right. This is an example from the melanoma experiment, not a diagram of the unlearning pipeline.

I Hyperparameters

I.1 Sample and Computational Efficiency

Table 6 lists every method of Figure 2 with its wall-clock cost.

Table 4: **Hyperparameters, ISIC.** Shared by all methods.

Encoder	frozen ImageNet ResNet-50
Decoder	17,722 parameters
Optimiser	SGD, one-cycle lr (peak 0.06), weight decay 5×10^{-4}
Batch size	16
Max. epochs / patience	1,000 / 50
NegGrad-Seg	$w_{D_f} = 2$
Mean Teacher	EMA 0.999, consistency weight 1
CPS	weight 1
One-shot	$\delta = 10^{-5}$, $\varepsilon \in \{0.1, 0.01, 0.001\}$

Table 5: **Hyperparameters, SIIM-ACR.** Shared by all methods.

Encoder	frozen ImageNet ResNet-50
Decoder	17,722 parameters
Optimiser	SGD, one-cycle lr (peak 0.06), weight decay 5×10^{-4}
Batch size	16
Max. epochs / patience	150 / 25
NegGrad-Seg	$w_{D_f} = 2$
Mean Teacher	EMA 0.999, consistency weight 1
CPS	weight 1
One-shot	$\delta = 10^{-5}$, $\varepsilon \in \{0.1, 0.01, 0.001\}$

Table 6: **Sample and computational efficiency.** Robust AUROC by share of training images with fine masks (mean $\pm$ std). Certified rows: $\epsilon = 0.1$ and $\delta = 10^{-5}$. Bold: highest mean; $^{\blacktriangle}$/$^{\blacktriangledown}$: significantly higher/lower than the original model, paired t-test, $p < 0.05$. Time: median wall-clock at 50% relative to training the original model on all images.

	Robust AUROC ↑, ISIC			Robust AUROC ↑, SIIM-ACR			Time ↓	
Number of samples	10%	25%	50%	10%	25%	50%	ISIC	SIIM
Original model (bounding box)	0.537 ± 0.035	0.535 ± 0.036	0.546 ± 0.035	0.600 ± 0.013	0.593 ± 0.014	0.598 ± 0.022	1.00×	1.00×
Retrain from scratch	$0.484 \pm 0.100^{\blacktriangledown}$	0.535 ± 0.106	0.564 ± 0.073	0.574 ± 0.046	0.581 ± 0.042	0.588 ± 0.036	0.21×	0.54×
Mean Teacher	0.540 ± 0.074	$0.598 \pm 0.072^{\blacktriangle}$	$\mathbf{0.606 \pm 0.040}^{\blacktriangle}$	0.590 ± 0.052	0.605 ± 0.036	0.605 ± 0.023	0.72×	1.17×
CPS	$\mathbf{0.554 \pm 0.085}$	$0.596 \pm 0.062^{\blacktriangle}$	$0.590 \pm 0.059^{\blacktriangle}$	0.582 ± 0.061	$\mathbf{0.606 \pm 0.020}$	0.607 ± 0.024	0.50×	0.63×
Fine-tuning ($w_{D_f} = 1$)	0.537 ± 0.035	0.530 ± 0.029	0.548 ± 0.043	$\mathbf{0.603 \pm 0.017}$	0.604 ± 0.022	$\mathbf{0.610 \pm 0.018}$	0.24×	0.47×
NegGrad-Seg ($w_{D_f} = 2$)	0.539 ± 0.039	0.533 ± 0.030	0.551 ± 0.043	0.601 ± 0.015	0.595 ± 0.016	0.610 ± 0.023	0.31×	0.50×
One-shot, $\varepsilon = 0.1$	0.549 ± 0.084	$\mathbf{0.608 \pm 0.057}^{\blacktriangle}$	$0.604 \pm 0.063^{\blacktriangle}$	0.590 ± 0.034	0.599 ± 0.038	0.592 ± 0.051	0.22×	0.65×

I.2 Sample efficiency: segmentation

Table 7 reports Dice for the runs of Table 6.

Table 7: **Sample efficiency: segmentation.** Dice by share of training images with fine masks, for the runs of Table 6 (mean $\pm$ std over 25, 25 and 15 runs on ISIC and 15, 14 and 12 on SIIM-ACR; NegGrad-Seg on SIIM-ACR over 11, 10 and 12). Certified rows: $\varepsilon = 0.1$ and $\delta = 10^{-5}$. Bold: highest mean; $^{\blacktriangle}$/$^{\blacktriangledown}$: significantly higher/lower than the original model, paired t-test, $p < 0.05$.

	Dice ↑, ISIC			Dice ↑, SIIM-ACR		
Method	10%	25%	50%	10%	25%	50%
Original model (bounding box)	$\mathbf{0.158 \pm 0.039}$	0.162 ± 0.036	0.161 ± 0.028	0.063 ± 0.018	0.056 ± 0.021	0.057 ± 0.024
Retrain from scratch	$0.004 \pm 0.017^{\blacktriangledown}$	$0.024 \pm 0.066^{\blacktriangledown}$	$0.094 \pm 0.094^{\blacktriangledown}$	$0.024 \pm 0.010^{\blacktriangledown}$	0.043 ± 0.016	0.057 ± 0.023
Mean Teacher	$0.002 \pm 0.006^{\blacktriangledown}$	$0.028 \pm 0.051^{\blacktriangledown}$	$0.116 \pm 0.072^{\blacktriangledown}$	0.051 ± 0.027	0.067 ± 0.025	0.067 ± 0.023
CPS	$0.007 \pm 0.025^{\blacktriangledown}$	$0.000 \pm 0.000^{\blacktriangledown}$	$0.044 \pm 0.056^{\blacktriangledown}$	$0.040 \pm 0.019^{\blacktriangledown}$	0.066 ± 0.019	$0.075 \pm 0.020^{\blacktriangle}$
Fine-tuning ($w_{D_f} = 1$)	0.152 ± 0.024	$\mathbf{0.164 \pm 0.020}$	$\mathbf{0.178 \pm 0.016}^{\blacktriangle}$	0.069 ± 0.021	$0.073 \pm 0.016^{\blacktriangle}$	$0.078 \pm 0.013^{\blacktriangle}$
NegGrad-Seg ($w_{D_f} = 2$)	0.147 ± 0.025	0.157 ± 0.023	0.171 ± 0.017	$\mathbf{0.076 \pm 0.018}$	$\mathbf{0.074 \pm 0.015}^{\blacktriangle}$	$\mathbf{0.081 \pm 0.021}^{\blacktriangle}$
One-shot, $\varepsilon = 0.1$	$0.016 \pm 0.046^{\blacktriangledown}$	$0.032 \pm 0.064^{\blacktriangledown}$	0.125 ± 0.084	$0.040 \pm 0.023^{\blacktriangledown}$	0.052 ± 0.026	0.078 ± 0.031

I.3 PER-STEP CERTIFIED OPERATORS

The per-step operators of Koloskova et al. (2025) clip the model to a radius κ, clip every gradient and inject noise at every step. With the clipping radius of prior configurations ($\kappa = 0.156$) they collapse on ISIC: in all 50 paired runs the recovered decoder predicts no lesion (Dice 0.000) and sits at chance (robust AUROC 0.500). The trained decoders have norms between 7.7 and 10.4, so clipping them to 0.156 removes nearly all of their signal before unlearning begins. This motivates Algorithm 1, which never clips the model.