SELECTIVE PREDICTION FORGETTING FOR ADAPTATION OF WEAKLY SUPERVISED LOCALIZATION OF HISTOLOGY IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Under strong cross-domain shifts, weakly supervised object localization (WSOL) models often lose their capacity to discriminate and develop a strong prediction bias toward a subset of classes. As a result, the majority of target samples are assigned to these dominant classes. Since source-free (unsupervised) domain adaptation (SFDA) methods typically rely on pseudo-labels, inaccurate pseudo-label estimation causes suboptimal adaptation and poor classification and localization performance. This issue is particularly challenging in computational histopathology, where shifts in organ-specific morphology, staining protocols, and acquisition conditions can significantly impact class-discriminative visual cues. This paper introduces SFDA with selective prediction forgetting (`SFDA-SPF`), a method designed to mitigate excessive concentration of predictions for dominant classes during target-domain adaptation. `SFDA-SPF` identifies prediction-dominant classes based on the model's target predictions and selectively revises uncertain samples assigned to them. Importantly, this procedure does not require knowledge of the true target class distribution. In contrast with previous SFDA methods, `SFDA-SPF` introduces a "forget set" composed of uncertain samples from dominant classes near their decision boundaries, enabling the classification boundaries to shift away from biased predictions. A second "retain set" is used to anchor certain and accurate samples. To forget the predictions of dominant classes, `SFDA-SPF` introduces a complementary optimization objective that reduces their probability while allowing alternative class assignments to emerge. This iterative optimization process reassesses target images for forget and retain sets at regular intervals. In addition to improving image-level pseudo-label estimation, the pixel-level classifier is jointly optimized, further enhancing localization. Experiments[1] on cross-organ and cross-center histopathology benchmarks (`GlaS`, `CAMELYON16`, and `CAMELYON17`) show that `SFDA-SPF` consistently outperforms state-of-the-art SFDA methods in both classification and localization, including across different levels of target-domain class imbalance.

1 INTRODUCTION

Computer-aided diagnostic systems based on deep learning are increasingly explored in digital pathology to support large-scale cancer screening and grading (Komura & Ishikawa, 2018; Gurcan et al., 2009; Rony et al., 2023; Guichemerre et al., 2025; Karimian et al., 2026). However, most state-of-the-art approaches rely on densely annotated datasets, where pixel-level annotations of tumor regions remain labor-intensive and require expert pathologists (Rony et al., 2023; Guichemerre et al., 2024; 2025). Deep WSOL models address this limitation by jointly predicting image-level diagnoses and localizing discriminative tissue regions using only image-level class labels for training (Zhou et al., 2016; Murtaza et al., 2023; Rony et al., 2023; Guichemerre et al., 2025).

Despite this advantage, the performance of WSOL models degrades under domain shifts, which are common in histopathology due to variations in staining protocols, scanner characteristics, tissue processing, inter-center acquisition, and organ-specific tissue architecture (Guichemerre et al.,

[1] Our code is included in the suppl. materials and will be made public.

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.

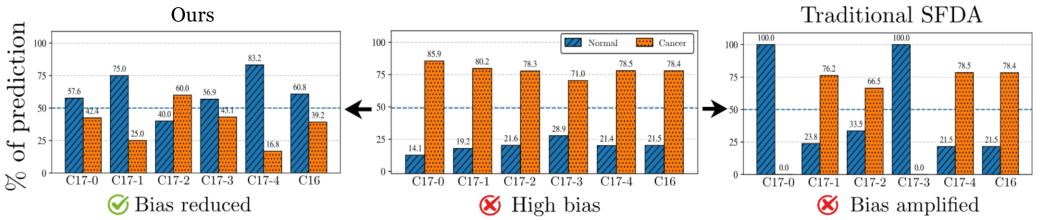


Figure 1: (a) SFDA-SPF reduces predictive bias across centers. (b) Source-only inference shows a strong predictive bias toward the cancer class when transferring from GlaS (source) to CAMELYON16 and CAMELYON17 (targets) with PixelCAM. (c) In contrast, traditional SFDA methods, such as SFDA-DE, amplify the bias and often collapse toward a dominant class.

2024; 2025; Chen et al., 2025). Empirical analysis (Fig. 1) shows that WSOL performance drops substantially under cross-organ shifts and dataset-scale imbalance (Guichemerre et al., 2024; 2025). When a WSOL model trained on a small source dataset (possibly from a different organ) is deployed on a large target dataset collected at another center, both classification accuracy and spatial localization quality degrade (Guichemerre et al., 2024). In contrast, models trained on large datasets generalize more consistently across domains. This transfer pattern suggests that WSOL degradation under severe shifts can be influenced by differences in organ-specific tissue structure and dataset size between the source and target domains. Additionally, appearance variations remain a major source of domain shift. In digital pathology, spatial drift makes predictions less reliable and harder to interpret (Guichemerre et al., 2024).

SFDA provides a practical framework that adapts a source-trained model using only unlabeled target data (Sharafi et al., 2025; Zeeshan et al., 2026), relying mainly on pseudo-labels generated from source-model predictions (Liang et al., 2020; Ding et al., 2022a; Guichemerre et al., 2024). Under strong target shift and the high visual inter-class similarity, WSOL models can lose their discriminative power, leading them to assign most samples to the same dominant classes (Guichemerre et al., 2024; 2025). Since standard SFDA methods do not explicitly detect and correct this prediction bias (Fang et al., 2024), the resulting skewed pseudo-label distribution tends to reinforce dominant-class over-prediction during adaptation (Guichemerre et al., 2024), leading to poor performance. A common approach to reduce bias in SFDA is to diversify target predictions (Liang et al., 2020; Yang et al., 2021; 2022). Nevertheless, this mechanism is applied globally across all classes and samples, without identifying the individual samples responsible for the dominant classes, which limits its effectiveness. In a WSOL setting for histopathology data, these limitations of SFDA affect both classification and localization, as localization is derived from class-discriminative evidence learned through image-level classification supervision (Rony et al., 2023; Guichemerre et al., 2024; 2025).

To overcome these limitations, we propose a novel SFDA method, SFDA-SPF, that identifies regions of the target prediction space where erroneous assignments are more likely to occur. Under severe domain shift, prediction errors can accumulate in classes that receive a disproportionately large share of target samples. SFDA-SPF first identifies these prediction-dominant classes and then uses predictive uncertainty to find the samples within them that are most likely to be unreliable. These samples form a *forget set*, where the probability assigned to their current predicted class is reduced, allowing alternative class assignments to emerge. The remaining samples from all classes form a *retain set*, which preserves stable target predictions and limits excessive model drift.

The key contributions of our work are: (1) We identify and empirically characterize dominant-class prediction bias as a critical obstacle to SFDA of WSOL models for histopathology. In this case, strong shifts in staining, acquisition, tissue preparation, and organ-specific morphology can obscure subtle class-discriminative cues, causing the source model to assign most target samples to a small subset of dominant classes. **(2)** SFDA-SPF, a novel SFDA strategy that acts directly on the uncertain target samples that predict dominant classes. A *forget set* identifies uncertain samples currently assigned to dominant classes, and a new loss reduces the probability assigned to their current dominant-class prediction, allowing alternative class predictions to emerge. In parallel, a *retain set* preserves the remaining target predictions to anchor adaptation and limit excessive model drift. Both sets are periodically reconstructed as the target decision boundary evolves. A jointly

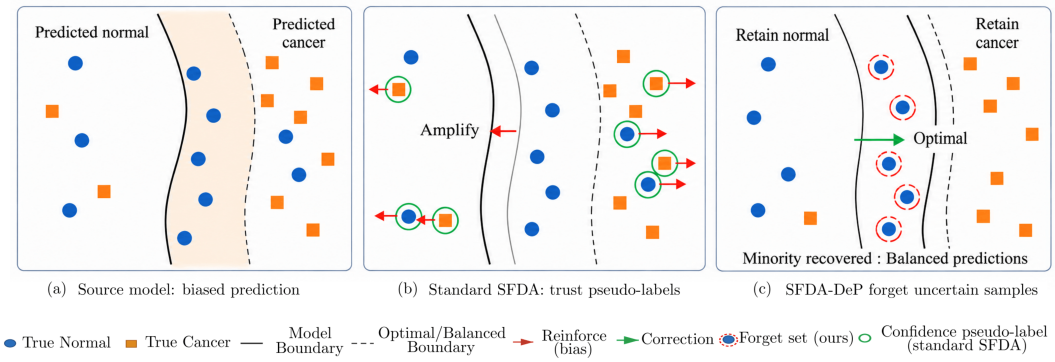


Figure 2: **Conceptual overview of `SFDA-SPF`.** In source-free adaptation for histological WSOL, the source model is biased toward the dominant class. **(a) Source model: biased prediction.** Many samples lie near the decision boundary, where high-entropy predictions are informative but error-prone. **(b) Standard SFDA: trust pseudo-labels.** Conventional methods reinforce confident pseudo-labels. However, in histology, confidence does not necessarily imply correctness, and erroneous pseudo-labels amplify the initial bias and attract uncertain samples toward the dominant class. **(c) `SFDA-SPF` forget uncertain samples (ours).** We instead identify uncertain dominant-class samples as a *forget set*. The forgetting objective reduces the probability assigned to the current dominant-class prediction, allowing probability mass to be redistributed toward alternative classes.

optimized pixel-level classifier using CAM-derived partial pseudo-labels further improves model localization. **(3)** An extensive evaluation of `SFDA-SPF` under cross-organ and cross-center shifts involving `GlaS`, `CAMELYON16`, and the five `CAMELYON17` centers, using CNN- and ViT-based WSOL models. For the six target domains, `SFDA-SPF` achieves the highest average localization performance for all three methods, outperforming the strongest competing SFDA baseline by over 10 `PxAP` points on average.

2 Related Work

(a) Weakly Supervised Object Localization: WSOL models predict spatial localization from image-level supervision through the class activation mapping (CAM) mechanism (Zhou et al., 2016; Selvaraju et al., 2020; Chattopadhyay et al., 2018; Jiang et al., 2021; Belharbi et al., 2022b; Gao et al., 2021), where localization maps are induced by classifier response. Representative WSOL approaches can be categorized into bottom-up (forward-pass information) and top-down strategies (exploit both forward and backward information) (Zhang et al., 2021; Shao et al., 2021; Rony et al., 2023). Bottom-up approaches include `DeepMIL` (Ilse et al., 2018), which uses instance-level attention weights to identify diagnostically relevant patches. Transformer-based methods (Chen et al., 2022; Gupta et al., 2022) such as `SAT` (Wu et al., 2023) instead exploit interactions between image tokens to refine class-specific attention maps. In histopathology, methods such as `NEGEV` (Belharbi et al., 2022a) and `PixelCAM` (Guichemerre et al., 2025) introduce additional spatial-refinement mechanisms to improve pixel-level attribution. Robustness of WSOL models under cross-domain conditions remains relatively underexplored in computational histopathology applications. Recent studies (Guichemerre et al., 2024; 2025) report significant degradation in both classification and localization when models are transferred across organs or centers.

(b) Source-Free Domain Adaptation: SFDA addresses domain shift (Li et al., 2020; Kurmi et al., 2021; Zhang et al., 2023) by adapting a pretrained source model to a new target domain using only unlabeled target data, avoiding annotation costs and the need for source data. A major family of SFDA methods relies on using reliable target predictions and features (Chen et al., 2021; Ding et al., 2022b; Fang et al., 2024; Guichemerre et al., 2024; Li et al., 2024). For example, SHOT (Liang et al., 2020) and CDCL (Wang et al., 2022) refine target pseudo-labels using class centroids. SFDA-DE (Ding et al., 2022a) instead estimates class-conditioned source feature distributions from high-confidence target pseudo-labels. Such strategies reduce the influence of noisy supervision, but their effectiveness strongly depends on the correctness of the selected pseudo-labels (Guichemerre et al., 2024). Moreover, discarding less reliable samples limits the use of a substantial portion of

samples from the target domain, including ambiguous samples located near the decision boundary. To exploit a broader set of target samples, uncertainty-aware approaches assign sample-specific weights according to prediction confidence or entropy (Lee et al., 2022; Roy et al., 2022; Litrico et al., 2023; Xu et al., 2025). However, weighting modifies the impact of pseudo-labels without reconsidering the resulting class distribution. Under strongly concentrated target predictions, uncertain samples assigned to dominant classes may still influence adaptation, albeit with reduced weight. Consequently, these methods preserve or amplify bias rather than correct the underlying class assignment. Another common strategy is to encourage confident sample-level predictions and diverse class assignments (Liang et al., 2020; Yang et al., 2021; 2022). However, those objectives act as global regularizers and do not identify which samples assigned to over-predicted classes should be reconsidered.

(c) Uncertainty in SFDA: In the absence of source data and target labels, estimating the reliability of target predictions is a central and key challenge in SFDA. Several approaches exploit predictive or representation uncertainty to reduce the impact of unreliable images by reweighting target samples or prioritizing confident predictions during adaptation (Roy et al., 2022; Lee et al., 2022; Litrico et al., 2023). These approaches use uncertainty to determine how strongly a target prediction should contribute to optimization. More recently, (Xu et al., 2025) studied uncertainty in contrastive-learning-based SFDA from an instance-dependent perspective and proposed controlling prediction variance during adaptation. While this improves the utilization of uncertain target data, the resulting supervision remains derived from the original model's prediction history. Consequently, under severe domain shifts as in histopathology where predictions become skewed toward a dominant class, such supervision may still preserve candidate assignments induced by the biased source model.

Although these global regularizers can reduce excessively imbalanced predictions, they do not explicitly determine which classes are over-predicted or which samples assigned to these classes should be reconsidered. As highlighted in (Guichemerre et al., 2024), conventional SFDA methods face substantial difficulties when deployed with WSOL models for computational histopathology. Since localization maps are derived from class-specific activations, their quality is closely tied to the model's ability to correctly predict the image class. When SFDA reinforces erroneous pseudo-labels, the backbone is optimized using incorrect class-level signals, which degrades the spatial feature maps, leading to poor performance in both classification and localization tasks.

3 SFDA WITH SELECTIVE PREDICTION FORGETTING

Notation. Let us denote by $\mathbb{T} = \{x_i\}_{i=1}^{N}$ an unlabeled target training set composed of N images $x_i : \Omega \subset \mathbb{R}^2$, where Ω is a discrete image domain. We consider a classifier $f_\theta(x) \in [0,1]^K$ that was pre-trained on source data $\mathbb{S} = \{(x,y)_i\}_{i=1}^{M}$ labeled only with image classes $y \in \{1, \cdots, K\}$, where K is the total number of classes. The feature extractor of f is denoted by $g(x) = Z \in \mathbb{R}^{|\Omega| \times d}$, where d is the feature dimension. We denote by $z_s \in \mathbb{R}^d$ the spatial embedding at location s in Ω. During adaptation, the source dataset $\mathbb{S}$ is no longer accessible. For each target image x_i, we denote its predicted class by $\hat{y}_i = \arg\max_{k \in \{1,\ldots,K\}} p_{i,k}$.

We first measure the empirical prediction distribution of the current model. The fraction of target samples currently assigned to class k is denoted by:

$$\widehat{\pi}(k) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\hat{y}_i = k]$$

Using $1/K$ as a class-symmetric reference, we define the set of prediction-dominant classes as: $\mathcal{C}_{\text{dom}} = \left\{k \in \{1, \ldots, K\} : \widehat{\pi}(k) > \frac{1}{K}\right\}$. Since patch-level class proportions can naturally vary across histopathology cohorts, the uniform reference is not intended to model the underlying target class prevalence. Instead, $(1/K)$ provides a class-symmetric reference point for determining how the current model distributes its predictions across classes. A prediction frequency above this reference indicates that a class attracts a large fraction of the model predictions. Under domain shift, when the source model exhibits degraded target performance, such prediction concentration can cause prediction-dominant classes to accumulate an important fraction of the misclassified target samples. We therefore use prediction dominance as a dataset-level cue to identify where potentially

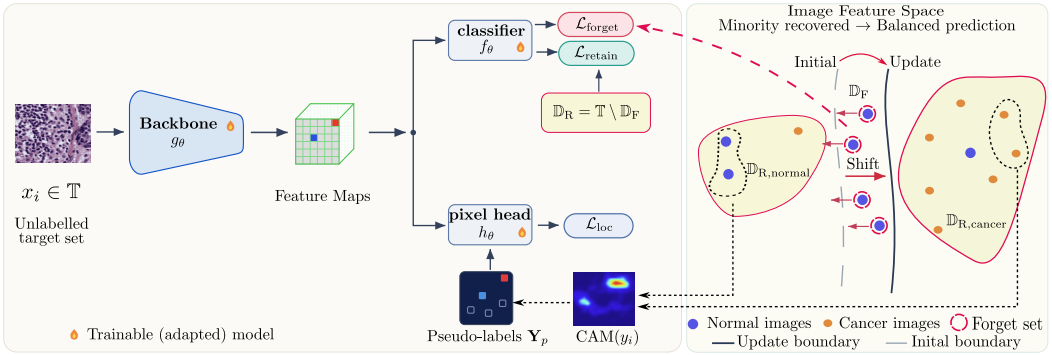


Figure 3: **Overview of the proposed `SFDA-SPF` pipeline.** Starting from a source-pretrained WSOL model and an unlabeled target set, `SFDA-SPF` first identifies dominant classes from the target prediction distribution. For each dominant class, the fraction ρ of samples with the highest normalized entropy is assigned to the forget set, while the remaining target samples form the retain set. During adaptation, $\mathcal{L}_{\text{F}}$ reduces the probability assigned to the current dominant-class predictions of forget samples, whereas $\mathcal{L}_{\text{R}}$ preserves the current assignments of the remaining samples to limit excessive model drift. An auxiliary pixel-level loss, $\mathcal{L}_{\text{L}}$, uses CAM-based partial pseudo-labels from the frozen source model to preserve spatial representations. Target predictions and the dominant, forget, and retain sets are periodically recomputed as the decision boundary evolves.

erroneous assignments are likely to concentrate. `SFDA-SPF` therefore combines this global signal with instance-level predictive uncertainty. Among samples assigned to prediction-dominant classes, only highly uncertain predictions are selected for forgetting, as these samples are more likely to lie near unreliable target decision boundaries, while more confident assignments are retained. In this way, the uniform reference identifies where prediction errors may accumulate, whereas uncertainty identifies which individual assignments should be explicitly reconsidered.

Forget and Retain Sets. `SFDA-SPF` identifies target samples whose current assignments may contribute to the observed prediction bias and progressively revises these assignments during adaptation. This process shifts the decision boundaries away from excessive dominant-class assignments, thereby reducing the skew in the target prediction distribution. To do this, we sample uncertain samples near the boundaries from dominant classes by using the entropy defined by:

$$H(x_i) = -\frac{1}{\log K} \sum_{c=1}^{K} p_\theta(y = c \mid x_i) \log p_\theta(y = c \mid x_i). \tag{1}$$

For each dominant class $k \in \mathcal{C}_{\text{dom}}$, we define its class-specific candidate set:

$$\mathbb{B}_k = \{(x_i, \hat{y}_i) \mid x_i \in \mathbb{T},\ \hat{y}_i = k\}, \qquad \forall k \in \mathcal{C}_{\text{dom}}. \tag{2}$$

Within each $\mathbb{B}_k$, we select the top $\rho\%$ most uncertain samples, which are the closest to the boundary:

$$\mathbb{D}_{\text{F}}^k = \text{Top}_\rho\left(\mathbb{B}_k, H\right). \tag{3}$$

The complete forget set, $\mathbb{D}_{\text{F}}$, is obtained by aggregating the selected samples from all dominant classes:

$$\mathbb{D}_{\text{F}} = \bigcup_{k \in \mathcal{C}_{\text{dom}}} \mathbb{D}_{\text{F}}^k. \tag{4}$$

To prevent excessive model drift and preserve the overall target prediction structure, we define the retain set, $\mathbb{D}_{\text{R}}$, as the complement of the forget set:

$$\mathbb{D}_{\text{R}} = \{(x_i, \hat{y}_i) \mid x_i \in \mathbb{T},\ (x_i, \hat{y}_i) \notin \mathbb{D}_{\text{F}}\}. \tag{5}$$

Importantly, it contains samples from all classes, including confident samples assigned to dominant classes, which helps maintain the model's ability to predict every class.

Forget and Retain Losses. For samples selected in the forget set, we explicitly reduce the probability assigned to their current predicted class. Since these samples are uncertain and currently assigned to prediction-dominant classes, preserving their pseudo-labels may reinforce potentially erroneous target assignments. We therefore define:

$$\mathcal{L}_{\mathrm{F}} = -\frac{1}{|\mathbb{D}_{\mathrm{F}}|} \sum_{x_i \in \mathbb{D}_{\mathrm{F}}} \log\left(1 - p_\theta(y = \hat{y}_i \mid x_i)\right), \tag{6}$$

where $\hat{y}_i$ denotes the current prediction of x_i. By minimizing this objective over the forget set, we discourage the model from preserving these uncertain dominant-class assignments, thereby allowing alternative class predictions to emerge as the target decision boundary evolves. For samples in $\mathbb{D}_{\mathrm{R}}$, we seek to preserve their current pseudo-labels $\hat{y}_i$ assigned by the model using a standard cross-entropy objective:

$$\mathcal{L}_{\mathrm{R}} = -\frac{1}{|\mathbb{D}_{\mathrm{R}}|} \sum_{x_i \in \mathbb{D}_{\mathrm{R}}} \log p_\theta(y = \hat{y}_i \mid x_i). \tag{7}$$

This objective maintains the prediction structure of samples that were not selected for the forgetting process. This loss complements $\mathcal{L}_{\mathrm{F}}$ by preserving the remaining class assignments while the decision boundaries are adjusted around ambiguous samples from dominant classes.

Our *forget-retain* approach progressively reduces the skew in the target class-prediction distribution. This gradually adjusts the decision boundaries away from dominant classes. As adaptation proceeds, some samples initially included in $\mathbb{D}_{\mathrm{F}}$ no longer provide relevant corrective signals once their assignments have changed or moved away from the updated decision boundary. Conversely, other samples become ambiguous with respect to the evolving boundary while remaining assigned to a dominant class. We periodically reassess the prediction bias and reconstruct both sets every m epochs. At each reconstruction step, we recompute the empirical target prediction distribution $\widehat{\pi}$, update the set of dominant classes $\mathcal{C}_{\mathrm{dom}}$, and assign the currently most ambiguous samples predicted as these classes to $\mathbb{D}_{\mathrm{F}}$. The remaining target samples are assigned to $\mathbb{D}_{\mathrm{retain}}$.

Localization Task: In WSOL, localization is obtained by class activation maps conditioned on the predicted image-level class. Therefore, improving the accuracy of target pseudo-labels during adaptation provides more reliable class-specific training signals and encourages the model to focus on regions associated with the correct semantic class. As our forget/retain method improves image-level predictions, they can now be used to provide more reliable class-specific localization cues during adaptation. To further preserve discriminative local representations, we follow PixelCAM (Guichemerre et al., 2025) and use an auxiliary pixel-level classifier $h : \mathbb{R}^d \to [0, 1]^2$, which predicts whether a spatial embedding z_s belongs to the foreground or background. The locations of the known pseudo-labels are encoded in Ω'. To this end, for each target sample, we define a mask of pixel-wise partial pseudo-labels denoted $\boldsymbol{Y} \in \{0, 1\}^{|\Omega| \times 2}$, $\boldsymbol{Y}_s \in \{0, 1\}^2$ at location s with value 0 for background, 1 for foreground. To reduce overfitting to potentially noisy pixel-level pseudo-labels, only a subset of pixels is randomly sampled from each image at every training iteration. This stochastic sampling encourages the model to explore different spatial regions. For each $x_i \in \mathbb{D}_{\mathrm{loc}}$, the current image-level pseudo-label $\hat{y}_i$ determines the class used for localization. We then extract the CAM associated with $\hat{y}_i$ from the frozen source-pretrained model and convert it into partial foreground and background pseudo-labels following PixelCAM (Guichemerre et al., 2025). The pixel-level classifier is trained using those pseudo-labels with standard binary cross-entropy loss.

$$\mathcal{L}_{\mathrm{loc}} = -\frac{1}{|\Omega'|} \sum_{s \in \Omega'} \left[(1 - \boldsymbol{Y}_s) \log(h(z_s)_0) + \boldsymbol{Y}_s \log(h(z_s)_1)\right]. \tag{8}$$

The total adaptation loss consists of three jointly optimized terms. The first two terms reduce class prediction bias while preserving reliable predictions. The third improves localization by exploring the image and expanding to fully cover ROIs. The loss can be written as

$$\mathcal{L} = \lambda_{\mathrm{R}} \mathcal{L}_{\mathrm{R}} + \lambda_{\mathrm{F}} \mathcal{L}_{\mathrm{F}} + \lambda_{\mathrm{L}} \mathcal{L}_{\mathrm{loc}}, \tag{9}$$

where λ_{R}, λ_{F} and λ_{L} are weighting coefficients.

4 RESULTS AND DISCUSSION

4.1 EXPERIMENTAL METHODOLOGY

Datasets: Our approach was evaluated on three publicly available histopathology benchmarks that enable both cross-organ and cross-center domain shifts. `GlaS` provides gland-level annotations for colon cancer (Sirinukunwattana et al., 2015; Rony et al., 2023), while `CAMELYON16` and `CAMELYON17` focus on breast lymph-node metastasis detection (Ehteshami Bejnordi et al., 2017; Litjens et al., 2018; Rony et al., 2023). We use `CAMELYON17` (Litjens et al., 2018) and construct a WSOL-compliant patch-level benchmark that supports classification and localization evaluation with the provided tumor segmentation masks. Following (Rony et al., 2023), we perform a *center-wise* WSI split that simulates cross-center shifts. Each patch is labeled with an image-level class and a localization mask. Additional details are provided in the supplementary material.

Backbone and WSOL Models: CNN-based models use ResNet-50 (Guichemerre et al., 2025) and transformer-based models use DeiT-Tiny (Dosovitskiy et al., 2021). We evaluate three WSOL models: `DeepMIL` (Ilse et al., 2018), `SAT` (Wu et al., 2023), `PixelCAM` (Guichemerre et al., 2025). For fair comparison, we follow (Guichemerre et al., 2024; 2025) for source pretraining.

`SFDA-SPF` Hyperparameters: The retain and forget loss weights are set as $\lambda_{\mathrm{R}}, \lambda_{\mathrm{F}} \in \{0.5, 1.0\}$. The localization loss weight is chosen from $\lambda_{\mathrm{L}} \in \{0.5, 1.0, 5.0\}$. The forgetting ratio is explored within $\rho \in \{5\%, 15\%, 25\%\}$. Following prior SFDA protocols (Guichemerre et al., 2024; Xu et al., 2025), hyperparameters are selected on a separate validation set, Additional details are provided in Appendix B.

Implementation Details: We compare our method against recent source-free domain adaptation (SFDA) approaches. We evaluate `SHOT` (Liang et al., 2020), `SFDA-DE` (Ding et al., 2022a), `CDCL` (Wang et al., 2022), `ERL` (Yi et al., 2023), and `RGV` (Zhu et al., 2025). Additional information regarding the hyper-parameters for each method is provided in Appendix B.

Table 1: **Localization (`PxAP`) and classification (`CL`) accuracy.** Results shown for `GlaS` → `C16`, `C17-0`, `C17-1`, `C17-2`, `C17-3`, `C17-4`.

	Method	C16		C17-0		C17-1		C17-2		C17-3		C17-4		Avg	
		PxAP	CL	PxAP	CL	PxAP	CL	PxAP	CL	PxAP	CL	PxAP	CL	PxAP	CL
PixelCAM (MIDL'25)	Source only	37.9	51.8	37.2	41.2	27.2	58.7	44.9	64.9	43.1	39.1	31.1	39.9	36.9	49.3
	SHOT (ICML'20)	14.9	67.1	12.1	71.7	19.2	**70.9**	24.4	71.6	15.1	78.9	13.9	**56.4**	16.6	**69.4**
	SFDA-DE (CVPR'22)	37.9	51.8	14.5	50.0	**36.7**	65.7	28.9	70.5	18.9	49.9	31.1	39.9	28.0	54.6
	CDCL (TMM'22)	37.9	51.8	25.3	50.0	21.2	59.3	25.5	69.9	19.3	49.9	22.6	40.9	25.3	53.6
	ERL (ICLR'23)	29.9	59.7	13.2	59.5	14.6	67.4	43.0	69.4	19.3	49.9	32.4	54.0	25.4	60.0
	RGV (CVPR'25)	30.7	52.9	31.0	50.0	29.4	53.5	44.9	64.9	40.2	49.9	32.2	41.6	34.7	52.1
	Ours	**49.5**	**67.2**	**49.9**	**86.2**	27.2	41.3	**50.2**	**76.1**	**53.8**	**80.6**	**33.9**	51.3	**44.1**	67.1
SAT (ICCV'23)	Source only	24.9	54.0	16.8	54.9	17.4	50.6	28.8	52.5	21.0	53.2	18.9	47.3	21.3	52.1
	SHOT (ICML'20)	27.2	65.1	16.9	68.7	18.9	50.0	33.7	**73.6**	19.7	76.3	17.9	55.0	22.4	64.8
	SFDA-DE (CVPR'22)	26.8	62.9	14.9	**76.3**	17.8	**61.0**	32.8	69.9	14.1	76.3	**22.9**	65.8	21.6	68.7
	CDCL (TMM'22)	32.3	66.6	14.4	72.5	18.3	56.4	33.5	72.8	16.1	77.4	17.4	55.4	22.0	66.9
	ERL (ICLR'23)	32.6	66.6	14.4	69.8	15.1	55.2	32.2	73.2	22.4	**79.3**	16.3	**69.1**	22.2	68.9
	RGV (CVPR'25)	24.9	54.0	16.8	54.9	17.4	50.6	28.0	52.4	20.9	49.9	18.9	48.3	21.2	51.7
	Ours	**36.4**	**66.8**	**24.9**	74.8	**26.9**	57.6	**43.7**	73.2	**27.8**	77.4	22.3	65.4	**30.3**	**69.2**
DeepMIL (ICML'18)	Source only	24.1	54.0	14.1	45.0	18.4	56.4	30.5	61.8	18.1	38.8	19.1	42.6	20.7	49.8
	SHOT (ICML'20)	25.2	61.4	12.8	70.6	17.3	**68.0**	27.5	74.1	19.2	75.0	14.8	67.8	19.5	69.5
	SFDA-DE (CVPR'22)	21.6	51.9	14.5	50.0	18.4	56.4	32.5	72.5	16.6	49.9	19.1	42.6	20.5	53.9
	CDCL (TMM'22)	24.0	54.1	27.2	50.0	18.4	56.4	27.8	72.3	37.6	49.9	28.6	50.0	27.3	55.5
	ERL (ICLR'23)	25.7	53.8	12.1	50.0	10.9	65.1	19.4	71.8	16.1	59.3	12.9	46.6	16.2	57.8
	RGV (CVPR'25)	24.7	52.9	27.1	50.0	18.4	56.9	31.1	62.1	21.0	49.9	21.7	43.9	24.0	52.6
	Ours	**53.5**	**72.1**	**34.0**	**82.8**	**38.8**	54.1	**45.4**	**79.2**	**37.7**	**77.9**	**34.5**	**74.5**	**40.7**	**73.4**

Table 2: **Localization (`PxAP`) and classification (`CL`) accuracy.** Inter-center SFDA on `CAMELYON17`. The source model is trained on center 0 and independently adapted to centers 1–4.

	Methods	C17-1		C17-2		C17-3		C17-4	
		PxAP	CL	PxAP	CL	PxAP	CL	PxAP	CL
PixelCAM (MIDL'25)	SHOT (ICML'20)	22.9	52.9	31.6	77.0	46.8	88.5	51.4	74.1
	SFDA-DE (CVPR'22)	23.9	49.4	18.7	50.0	50.7	92.0	51.4	74.1
	CDCL (TMM'22)	40.1	73.8	30.6	74.7	59.4	88.3	50.2	**83.2**
	ERL (ICLR'23)	17.4	62.2	45.6	58.4	50.7	92.0	51.4	74.1
	RGV (CVPR'25)	36.1	56.9	41.2	50.00	50.7	92.0	**54.3**	70.4
	Ours	**48.5**	**74.4**	**58.1**	**81.6**	**60.4**	**93.8**	46.1	**83.2**

Table 3: **Robustness to class imbalance in the target adaptation set on CAMELYON17.** Starting from the balanced setting (50:50), we progressively increase the proportion of Normal samples up to 90:10. The test set remains balanced across all settings; only the target training set is imbalanced.

	Methods	50:50		60:40		70:30		80:20		90:10	
		PxAP	CL	PxAP	CL	PxAP	CL	PxAP	CL	PxAP	CL
PixelCAM (MIDL'25)	SHOT (ICML'20)	19.2	70.0	22.9	41.6	37.2	41.2	24.2	41.6	30.5	73.6
	CDCL (TMM'22)	25.3	50.0	25.5	45.4	22.6	44.6	22.4	44.6	22.8	44.6
	ERL (ICLR'23)	13.2	59.5	21.8	44.2	20.4	43.1	21.5	43.5	21.0	43.5
	RGV (CVPR'25)	31.0	50.0	33.7	50.0	33.3	50.0	33.1	50.0	32.8	50.0
	Ours	**49.9**	**86.2**	**43.7**	**79.0**	**55.1**	**82.8**	**46.0**	**79.4**	**58.8**	**82.8**

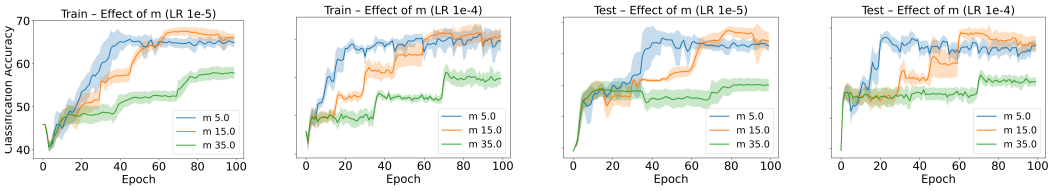


Figure 4: Effect of dynamic resampling frequency during adaptation with PixelCAM.

4.2 Impact on Classification and Localization

Table 1 reports localization (PxAP) and classification (CL) accuracy for the three WSOL methods across the cross-organ (GlaS → C16, C17) and Table 2–3 cross-center C17 settings. It shows that SFDA-SPF improves WSOL adaptation on most target domains and across architectures. The largest gains appear on challenging domains where several SFDA baselines remain near 50% CL and show strong dominant-class prediction bias. On PixelCAM, SFDA-SPF reaches 44.1 PxAP and 67.1 CL on average (+16.1 PxAP/ +12.5 CL over SFDA-DE), with the largest gains on the centers where the source model is most biased and the baselines remain around 50% CL. On C17-0 and C17-3, CL rises from $\sim 50\%$ to 86.2% and 80.6%, while localization also improves, confirming that correcting the classification bias directly benefits the spatial cues that drive the CAMs.

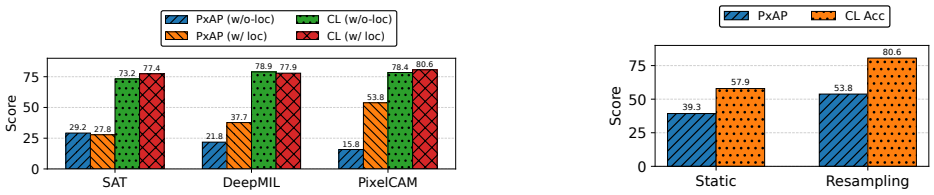
4.3 Ablation Studies

Contribution of Loss Components. Table 4 highlights the role of our three objectives. Using $\mathcal{L}_{\mathrm{R}}$ alone better preserves localization (27.6 PxAP) but fails to correct the classification bias (41.2 CL), whereas $\mathcal{L}_{\mathrm{F}}$ improves classification (85.1 CL) but provides poor localization (10.4 PxAP). Combining both image-level objectives maintains high classification accuracy (84.7 CL), but remains insufficient for spatial adaptation (13.1 PxAP). Including $\mathcal{L}_{\mathrm{L}}$ raises PxAP to 49.9 while preserving 86.1 CL, showing that prediction debiasing and explicit localization guidance are required to jointly improve both tasks.

Table 4: Impact of loss components during SFDA-SPF on GlaS → C17-0 with PixelCAM.

$\mathcal{L}_{\mathrm{R}}$	$\mathcal{L}_{\mathrm{F}}$	$\mathcal{L}_{\mathrm{L}}$	PxAP ↑	CL ↑
✓			27.6	41.2
	✓		10.4	85.1
		✓	9.5	48.1
✓	✓		13.1	84.7
✓	✓	✓	**49.9**	**86.1**

Localization Task. As shown in Fig. 5a, adding $\mathcal{L}_{\mathrm{L}}$ improves PxAP for PixelCAM from 15.8 to 53.8 and for DeepMIL from 21.8 to 37.7, while producing smaller changes in CL. This confirms that the localization objective primarily preserves spatial discrimination. For SAT, PxAP instead slightly decreases from 29.2 to 27.8, possibly because the lower-resolution representations provide less reliable pixel-level pseudo-labels. Nevertheless, SFDA-SPF reaches an average PxAP of 30.3 with SAT, compared with 21.3 for source-only inference and 22.4 for the strongest SFDA baseline, highlighting the effectiveness of the forget-and-retain mechanism for improving transformer-based WSOL adaptation. We analyze the quality of these source-derived localization cues in Appendix D. **Impact of dynamic resampling.** As illustrated in Fig. 4, more frequent resampling enables SFDA-SPF to improve classification more rapidly. This observation supports our hypothesis that uncertain samples located near the current decision boundary provide useful signals to correct dominant-class prediction bias. As shown in Fig. 5b, the static



(a) (b)

Figure 5: Ablation analysis of SFDA-SPF on GlaS → C17-3 with PixelCAM. **(a)** Impact of the localization loss during adaptation. **(b)** Comparison between *static sets*, where forget and retain sets are constructed once before adaptation, and *dynamic resampling* are periodically reconstructed.

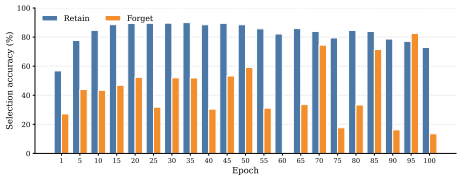


Figure 6: Evolution of retain- and forget-set selection accuracy during GlaS → C17-4 adaptation with DeepMIL. The forget and retain sets are dynamically reconstructed every five epochs.

strategy limits both classification and localization performance. Indeed, as adaptation progressively modifies the target decision boundary, the initial set assignments become increasingly irrelevant.

Quality of forget-set selection. Because SFDA-SPF relies on identifying samples whose current dominant-class assignments should be reconsidered, Fig. 6 shows the evolution of retain and forget selection for DeepMIL on CAMELYON17-4 with GlaS as the source. At each reconstruction step, selection accuracy measures the agreement between the dynamically selected retain/forget samples and their ground-truth status. The retain set remains consistently more reliable, while the forget set exhibits variations throughout adaptation, with low accuracy during training.

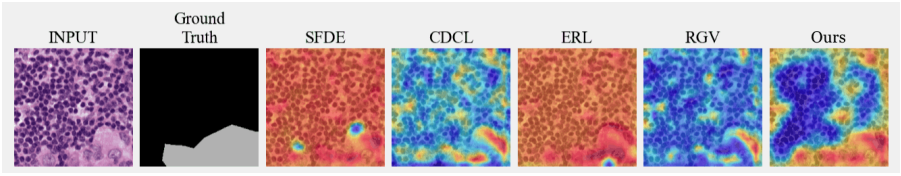


Figure 7: Localization produced by the PixelCAM WSOL model after SFDA-SPF and the state-of-the-art SFDA methods (GlaS → CAMELYON17-3).

5 Conclusion

This paper addresses dominant-class prediction bias, a critical challenge in source-free adaptation for WSOL models in histopathology. We introduced SFDA-SPF, a new SFDA method that corrects this bias through complementary forget and retain objectives. The forget set contains uncertain samples currently assigned to dominant classes and likely to lie near the target decision boundary. The proposed forgetting objective reduces the probability assigned to their current dominant class, allowing under-predicted classes to emerge and progressively reducing the skew in the target prediction distribution. In parallel, the retain objective preserves the predictions of remaining target samples through cross-entropy supervision, thereby maintaining useful target structure and limiting excessive model drift. Experiments show that SFDA-SPF achieves the best average localization performance across all evaluated WSOL architectures while improving average classification accuracy.

Supplementary Material: It provides additional information on datasets, implementation, and quantitative results. It also includes additional analyses of target-class imbalance, hyperparameter sensitivity, source-CAM reliability, the multi-class setting, and entropy-based prediction behavior.

AI USE STATEMENT

AI-based tools were used during the preparation of this manuscript to support language refinement, proofreading, and improvements in clarity and presentation. These tools were not used to generate datasets, produce or alter experimental results, or determine the scientific conclusions of the study. Any AI-assisted content, including text, code, figures, and interpretations, was carefully inspected and validated by the authors. The authors remain fully responsible for the accuracy, integrity, and final content of the manuscript.

REFERENCES

S. Belharbi, M. Pedersoli, I. Ben Ayed, L. McCaffrey, and E. Granger. Negative evidence matters in interpretable histology image classification. In *MIDL*, 2022a.

S. Belharbi, A. Sarraf, M. Pedersoli, I. Ben Ayed, L. McCaffrey, and E. Granger. F-CAM: Full resolution class activation maps via guided parametric upscaling. In *WACV*, 2022b.

Malik Boudiaf, Tom Denton, Bart Van Merriënboer, Vincent Dumoulin, and Eleni Triantafillou. In search for a generalizable method for source free domain adaptation. In *International Conference on Machine Learning*, pp. 2914–2931. PMLR, 2023.

A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *WACV*, pp. 839–847, 2018.

Hsi-Ling Chen, Chun-Shien Lu, and Pau-Choo Chung. Sgcd: Stain-guided cyclediffusion for unsupervised domain adaptation of histopathology image classification. In *NeurIPS*, 2025.

Weijie Chen, Luojun Lin, Shicai Yang, Di Xie, Shiliang Pu, Yueting Zhuang, and Wenqi Ren. Self-supervised noisy label learning for source-free unsupervised domain adaptation. *CoRR*, abs/2102.11614, 2021.

Z. Chen, C. Wang, Y. Wang, G. Jiang, Y. Shen, Y. Tai, C. Wang, W. Zhang, and L. Cao. LCTR: On awakening the local continuity of transformer for weakly supervised object localization. In *AAAI*, volume 36, pp. 410–418, 2022.

Ning Ding, Yixing Xu, Yehui Tang, Chao Xu, Yunhe Wang, and Dacheng Tao. Source-free domain adaptation via distribution estimation. In *CVPR*, 2022a.

Yuhe Ding, Lijun Sheng, Jian Liang, Aihua Zheng, and Ran He. Proxymix: Proxy-based mixup training with label refinery for source-free domain adaptation. *CoRR*, abs/2205.14566, 2022b.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

B. Ehteshami Bejnordi, M. Veta, P. Johannes van Diest, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer . *JAMA*, 2017.

Linus Ericsson, Da Li, and Timothy Hospedales. Better practices for domain adaptation. In Aleksandra Faust, Roman Garnett, Colin White, Frank Hutter, and Jacob R. Gardner (eds.), *Proceedings of the Second International Conference on Automated Machine Learning*, volume 224 of *Proceedings of Machine Learning Research*, pp. 4/1–25. PMLR, 12–15 Nov 2023.

Yuqi Fang, Pew-Thian Yap, Weili Lin, Hongtu Zhu, and Mingxia Liu. Source-free unsupervised domain adaptation: A survey. *Neural Network*, 2024.

W. Gao, F. Wan, X. Pan, Z. Peng, Q. Tian, Z. Han, B. Zhou, and Q. Ye. TS-CAM: token semantic coupled attention map for weakly supervised object localization. In *ICCV*, 2021.

A. Guichemerre, S. Belharbi, T. Mayet, S. Murtaza, P. Shamsolmoali, L. McCaffrey, and E. Granger. Source-free domain adaptation of weakly-supervised object localization models for histology. In *CVPRw*, 2024.

A. Guichemerre, S. Belharbi, M. Shateri, L. McCaffrey, and E. Granger. Pixelcam: Pixel class activation mapping for histology image classification and roi localization. In *MIDL*, 2025.

S. Gupta, S. Lakhotia, A. Rawat, and R. Tallamraju. ViTOL: Vision transformer for weakly supervised object localization. In *CVPRw*, 2022.

Metin N. Gurcan, Laura E. Boucheron, Ali Can, Anant Madabhushi, Nasir M. Rajpoot, and Bulent Yener. Histopathological image analysis: A review. *IEEE Reviews in Biomedical Engineering*, 2: 147–171, 2009.

M. Ilse, J. M. Tomczak, and M. Welling. Attention-based deep multiple instance learning. In *ICML*, 2018.

P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei. Layercam: Exploring hierarchical class activation maps for localization. *TMI*, 30:5875–5888, 2021.

Banafsheh Karimian, Giulia Avanzato, Soufiane Belharbi, Alexis Guichemerre, Luke McCaffrey, Mohammadhadi Shateri, and Eric Granger. Clip-it: Clip-based pairing for histology images classification. In *WACV*, 2026.

Daisuke Komura and Shumpei Ishikawa. Machine learning methods for histopathological image analysis. *CSBJ*, 16:34–42, 2018.

Vinod K Kurmi, Venkatesh K Subramanian, and Vinay P Namboodiri. Domain impression: A source data free domain adaptation method. In *WACV*, 2021.

Jonghyun Lee, Dahuin Jung, Junho Yim, and Sungroh Yoon. Confidence score for source-free unsupervised domain adaptation. In *International conference on machine learning*, pp. 12365–12377. PMLR, 2022.

Jingjing Li, Zhiqi Yu, Zhekai Du, Lei Zhu, and Heng Tao Shen. A comprehensive survey on source-free domain adaptation. *IEEE TPAMI*, 2024.

Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *CVPR*, pp. 9638–9647, 2020.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. *ICML*, 2020.

Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob Van de Loo, Rob Vogels, et al. 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. *GigaScience*, 7(6):giy065, 2018.

Mattia Litrico, Alessio Del Bue, and Pietro Morerio. Guiding pseudo-labels with uncertainty estimation for source-free unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7640–7650, 2023.

S. Murtaza, S. Belharbi, M. Pedersoli, A. Sarraf, and E. Granger. Dips: Discriminative pseudo-label sampling with self-supervised transformers for weakly supervised object localization. *IMAVIS*, 2023.

Mehrdad Noori, Gustavo A Vargas Hakim, David Osowiechi, Fereshteh Shakeri, Ali Bahri, Moslem Yazdanpanah, Sahar Dastani, Ismail Ben Ayed, and Christian Desrosiers. Histopath-c: Towards realistic domain shifts for histopathology vision-language adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4890–4900, 2026.

J. Rony, S. Belharbi, J. Dolz, I. Ben Ayed, L. McCaffrey, and E. Granger. Deep weakly-supervised learning methods for classification and localization in histology images: A survey. *MLBI*, 2, 2023.

Subhankar Roy, Martin Trapp, Andrea Pilzer, Juho Kannala, Nicu Sebe, Elisa Ricci, and Arno Solin. Uncertainty-guided source-free domain adaptation. In *European conference on computer vision*, pp. 537–555. Springer, 2022.

R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *IJCV*, 128(2):336–359, 2020.

Feifei Shao, Long Chen, Jian Shao, Wei Ji, Shaoning Xiao, Lu Ye, Yueting Zhuang, and Jun Xiao. Deep learning for weakly-supervised object detection and object localization: A survey, 2021.

Masoumeh Sharafi, Emma Ollivier, Osama Zeeshan, Soufiane Belharbi, Alessandro Koerich, Marco Pedersoli, Simon Bacon, and Eric Granger. Disentangled source-free personalization for facial expression recognition with neutral target data. In *FG*, 2025.

Liyu Shi, Xiaoyan Li, Weiming Hu, Haoyuan Chen, Jing Chen, Zizhen Fan, Minghe Gao, Yujie Jing, Guotao Lu, Deguo Ma, Zhiyu Ma, Qingtao Meng, Dechao Tang, Hongzan Sun, Marcin Grzegorzek, Shouliang Qi, Yueyang Teng, and Chen Li. Ebhi-seg: A novel enteroscope biopsy histopathological hematoxylin and eosin image dataset for image segmentation tasks. *Frontiers in Medicine*, 2023.

K. Sirinukunwattana, D.RJ Snead, and N.M. Rajpoot. A stochastic polygons model for glandular structures in colon histology images. *TMI*, 2015.

Rui Wang, Zuxuan Wu, Zejia Weng, Jingjing Chen, Guo-Jun Qi, and Yu-Gang Jiang. Cross-domain contrastive learning for unsupervised domain adaptation. *IEEE TMM*, 2022.

P. Wu, W. Zhai, Y. Cao, J. Luo, and Z.-J. Zha. Spatial-aware token for weakly supervised object localization. In *ICCV*, 2023.

Gezheng Xu, Hui Guo, Li Yi, Charles Ling, Boyu Wang, and Grace Yi. Revisiting source-free domain adaptation: a new perspective via uncertainty control. In *ICLR*, 2025.

Shiqi Yang, Joost van de Weijer, Luis Herranz, Shangling Jui, et al. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. *NeurIPS*, 34:29393–29405, 2021.

Shiqi Yang, Shangling Jui, Joost Van De Weijer, et al. Attracting and dispersing: A simple approach for source-free domain adaptation. *Advances in Neural Information Processing Systems*, 35: 5802–5815, 2022.

Li Yi, Gezheng Xu, Pengcheng Xu, Jiaqi Li, Ruizhi Pu, Charles Ling, A. Ian McLeod, and Boyu Wang. When source-free domain adaptation meets learning with noisy labels. In *ICLR*, 2023.

Muhammad Osama Zeeshan, Natacha Gillet, Alessandro Lameiras Koerich, Marco Pedersoli, Francois Bremond, and Eric Granger. Musaco: Multimodal subject-specific selection and adaptation for expression recognition with co-training. In *WACV*, 2026.

Dingwen Zhang, Junwei Han, Gong Cheng, and Ming-Hsuan Yang. Weakly supervised object localization and detection: A survey, 2021.

Ningyuan Zhang, Jie Lu, Keqiuyin Li, Zhen Fang, and Guangquan Zhang. Source-free unsupervised domain adaptation: Current research and future directions. *NERUCOMPU*, pp. 126921, 2023.

B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.

Ronghang Zhu, Mengxuan Hu, Weiming Zhuang, Lingjuan Lyu, Xiang Yu, and Sheng Li. Revisiting source-free domain adaptation: Insights into representativeness, generalization, and variety. In *CVPR*, 2025.

SUPPLEMENTARY MATERIAL

Table of Contents

A Dataset Construction and Experimental Protocol 14
A.1 CAMELYON17 Dataset Construction 14
A.2 Visualization of CAMELYON17 15
A.3 EBHI Corruption Benchmark 17
A.4 Visualization of EBHI 18

B Implementation and Reproducibility Details 20
B.1 Training and Adaptation Hyperparameters 20
B.2 Experimental Settings 21
B.3 `SFDA-SPF` Training Algorithm 22

C Additional Quantitative Results 22
C.1 Cross-Organ Results 22
C.2 Robustness to Target-Class Imbalance 23
C.3 Sensitivity to the Forgetting Proportion 23
C.4 Sensitivity to the Localization Keep Ratio 24
C.5 Backbone and Classifier Adaptation 24
C.6 Extension to the Multi-Class Setting 24
C.7 CNNs versus Transformer-Based Models 24

D Reliability of Source CAMs for Pixel Pseudo-Labeling 26
D.1 Reliability of High-Confidence Foreground Regions 26
D.2 Reliability of Low-Activation Background Regions 29

E Additional Visualizations 31
E.1 Entropy Distributions 31
E.2 Entropy–Accuracy Relationships 31
E.3 Additional Visualizations 33

A Dataset Construction and Experimental Protocol

A.1 CAMELYON17 Dataset Construction

In this work, we do not use the standard `CAMELYON17` (Litjens et al., 2018) classification benchmark. Instead, we construct a patch-level version of `CAMELYON17` specifically designed for Weakly Supervised Object Localization (WSOL). Each extracted patch is associated with an image-level normal/tumor label used for classification and WSOL training, while the corresponding pixel-level tumor annotation is retained exclusively for localization evaluation. Our construction protocol closely follows the patch-sampling strategy introduced in (Rony et al., 2023) for `CAMELYON16` (Rony et al., 2023; Guichemerre et al., 2025). Importantly, patches originating from the five acquisition centers of `CAMELYON17` are kept as separate domains, resulting in five domain-specific datasets, denoted `C17-0` to `C17-4`. This construction enables controlled source-to-target experiments across acquisition centers while preserving the same tissue type and prediction task.

A.1.1 Patch Extraction Procedure

Sampling of Metastatic Patches Metastatic patches are sampled exclusively from metastatic whole-slide images (WSIs) and only in regions surrounding cancerous tissue. Each sampled patch is associated with both an image-level label and a pixel-level annotation mask derived from the WSI annotations.

Algorithm 1 Sampling of Metastatic Patches for WSOL

Require: Metastatic WSI with pixel-level annotation, patch size (H, W), and thresholds p_t and p_0
Ensure: Patch x with image-level label and pixel-level mask Y

```
Select a metastatic WSI
Sample a patch x of size (H, W)
Binarize x into a tissue mask x_b using Otsu thresholding
Compute tissue ratio p_x^t in x_b
if p_x^t < p_t then
    discard x
end if
Extract metastatic binary mask Y from the WSI annotation
Compute metastatic ratio p_x^m in Y
if p_x^m < p_0 then
    discard x
else
    Assign image-level label metastatic
    return (x, Y)
end if
```

Calibration of Metastatic Patches Naively sampling all eligible metastatic patches results in a strong bias toward patches fully occupied by cancerous tissue, typically located at the center of metastatic regions. Such patches produce binary masks almost filled with positive labels, which artificially inflates localization performance and undermines a fair evaluation of WSOL methods. To mitigate this issue, we adopt the calibration strategy proposed in (Rony et al., 2023), summarized in Algorithm 2.

Algorithm 2 Calibration of Metastatic Patches

Require: Set of metastatic patches $\mathcal{P}$ with metastatic ratios p_x^m, thresholds p_0 and p_1, sampling ratio p_n, and number of bins b
Ensure: Calibrated set of metastatic patches $\mathcal{P}_{\text{cal}}$

```
Partition P into:
    P_1 = {x | p_0 ≤ p_x^m ≤ p_1}
    P_2 = {x | p_x^m > p_1}
Initialize P_cal ← P_1
n ← |P_1|
k ← n p_n
Compute histogram H of p_x^m for patches in P_2 using b bins
while |P_cal| < n + k do
    Uniformly sample a bin from H with p_x^m > p_1
    Uniformly sample a patch x from the selected bin
    Add x to P_cal
end while
return P_cal
```

Sampling of Normal Patches Normal patches are sampled exclusively from normal WSIs. Patches are extracted randomly and uniformly without overlap. A patch is accepted if its tissue ratio p_x^t satisfies $p_x^t \geq p_t$.

To ensure class balance at the patch level, we sample the same number of normal patches as metastatic patches within each data split (training, validation, or test). Given h normal WSIs and n_m metastatic patches, we sample n_m/h normal patches per WSI. Tissue mass estimation is performed at resolution level 6 using Otsu thresholding and verified at level 0.

Final Configuration All patches are extracted without overlap and resized to 512×512 pixels. We adopt the same hyperparameter configuration as in (Rony et al., 2023). This construction yields a `CAMELYON17` WSOL benchmark with realistic spatial ambiguity, making it well suited for evaluating weakly supervised localization and source-free domain adaptation methods.

Table 5: Dataset splits used in our experiments, showing the number of patches per split for each dataset and center.

	GlaS	C16	C17-0	C17-1	C17-2	C17-3	C17-4
Train	67	24348	634	1066	498	816	940
Val (CL)	18	8850	144	38	110	102	146
Val (`PxAP`)	6	10	10	10	10	10	10
Test	80	15664	262	172	448	376	298

Dataset Statistics Table 5 reports the resulting number of patches for each center and experimental split. Following the same protocol across centers, we construct separate training, classification-validation, localization-validation, and test sets. The localization-validation images and their pixel-level annotations are never used to optimize the adaptation objective and are only employed for model selection/evaluation of localization performance, whereas the classification-validation split is used for classification model selection. We provide additional visualization in the following section. For cancer images, ground-truth tumor regions are overlaid in red, whereas normal images contain no annotated tumor regions.

A.2 Visualization of `CAMELYON17`

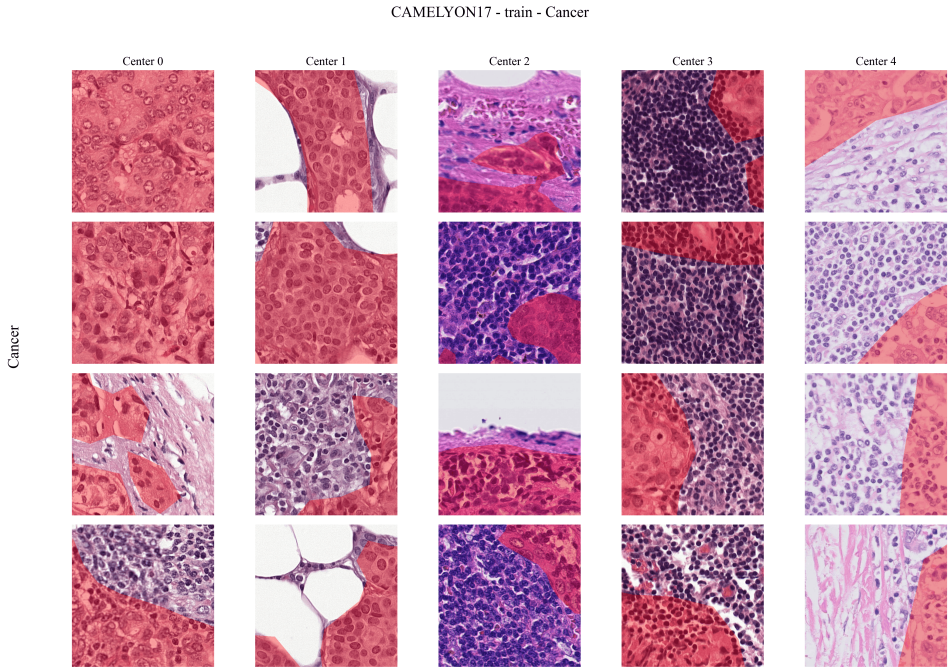


Figure 8: Representative cancer examples from the **training split** of the constructed CAMELYON17 WSOL benchmark across the five centers. Ground-truth tumor regions are overlaid in red.

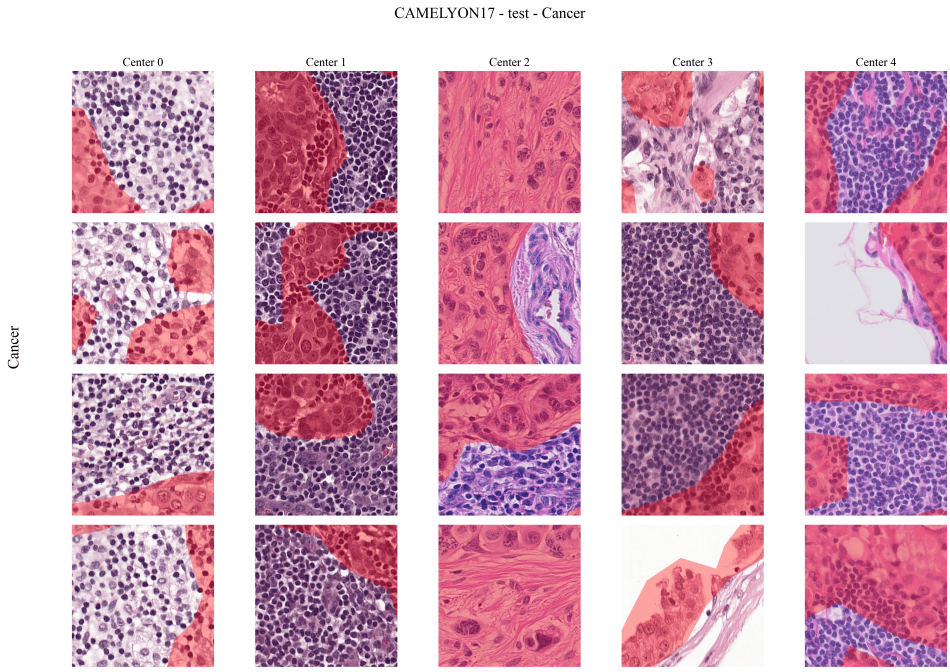


Figure 9: Representative cancer examples from the **testing split** of the constructed CAMELYON17 WSOL benchmark across the five centers. Ground-truth tumor regions are overlaid in red.

A.3 EBHI CORRUPTION BENCHMARK

To extend our evaluation to a challenging multi-class setting, we construct a new benchmark for SFDA and WSOL in colorectal histopathology based on EBHI Shi et al. (2023). EBHI comprises six classes and has been previously used for colorectal cancer classification and segmentation. We split the dataset into disjoint source and target subsets. The source subset contains clean histopathology images and is used exclusively for source-model training, while the target subset is held out from source training and used to construct different target-domain shifts. We further divide each subset into training, validation, and test sets while preserving the original class distribution. Table 6 summarizes the resulting class-wise partition.

Corruption Types : To construct controlled target-domain shifts, we follow the Histopath-C benchmark (Noori et al., 2026), which introduces histopathology-specific synthetic corruptions designed to mimic distribution shifts encountered in digital pathology. Histopath-C groups ten corruption types into five categories: *staining*, *contamination*, *blurring*, *noise*, and *illumination*.

Specifically, we consider two staining perturbations (*Stain-Light* and *Stain-Heavy*), two contamination artifacts (*Dust* and *Air Bubble*), two blur perturbations (*Defocus Blur* and *Motion Blur*), two noise perturbations (*Gaussian Noise* and *Shot Noise*), and two illumination perturbations (*Brightness* and *Contrast*).

We apply each corruption independently to the held-out EBHI−T subset, resulting in 11 distinct target domains. The source domain remains composed essentially of clean images, ensuring that the observed distribution shift arises only from the applied corruption.

Table 6: Class distribution of the EBHI dataset and the constructed source/target splits.

Class	**Dataset**	**Source/Target**	**Train**	**Val**	**Test**
Normal	152	76	**46**	**15**	**15**
Polyp	948	474	**284**	**95**	**95**
Low-grade IN	1,276	638	**382**	**128**	**128**
High-grade IN	372	186	**112**	**37**	**37**
Serrated adenoma	116	58	**34**	**12**	**12**
Adenocarcinoma	1,590	795	**477**	**159**	**159**
Total	**4,454**	**2,227**	**1,335**	**446**	**446**

Table 7: Parameters used to construct the corrupted EBHI-Seg target domains. We follow the official Histopath-C implementation (Noori et al., 2026).

Category	**Corruption**	**Parameters**
Staining	Stain-Light	$\theta = 0.05$
	Stain-Heavy	$\theta = 0.10$
Contamination	Dust	intensity $= 0.7$, $N_{\text{smudges}} \in [3, 10]$
	Air Bubble	$N_{\text{bubbles}} \in [10, 20]$, $\alpha = 0.3$, blur severity $= 2$
Blurring	Defocus Blur	severity $= 5$, radius $= 10$, alias blur $= 0.5$
	Motion Blur	severity $= 5$, radius $= 20$, $\sigma = 15$
Noise	Gaussian Noise	severity $= 3$, $\sigma = 0.18$
	Shot Noise	severity $= 3$, $c = 12$
Illumination	Brightness	severity $= 5$, offset $= 0.5$
	Contrast	severity $= 5$, scale $= 0.05$

A.4 VISUALIZATION OF EBHI

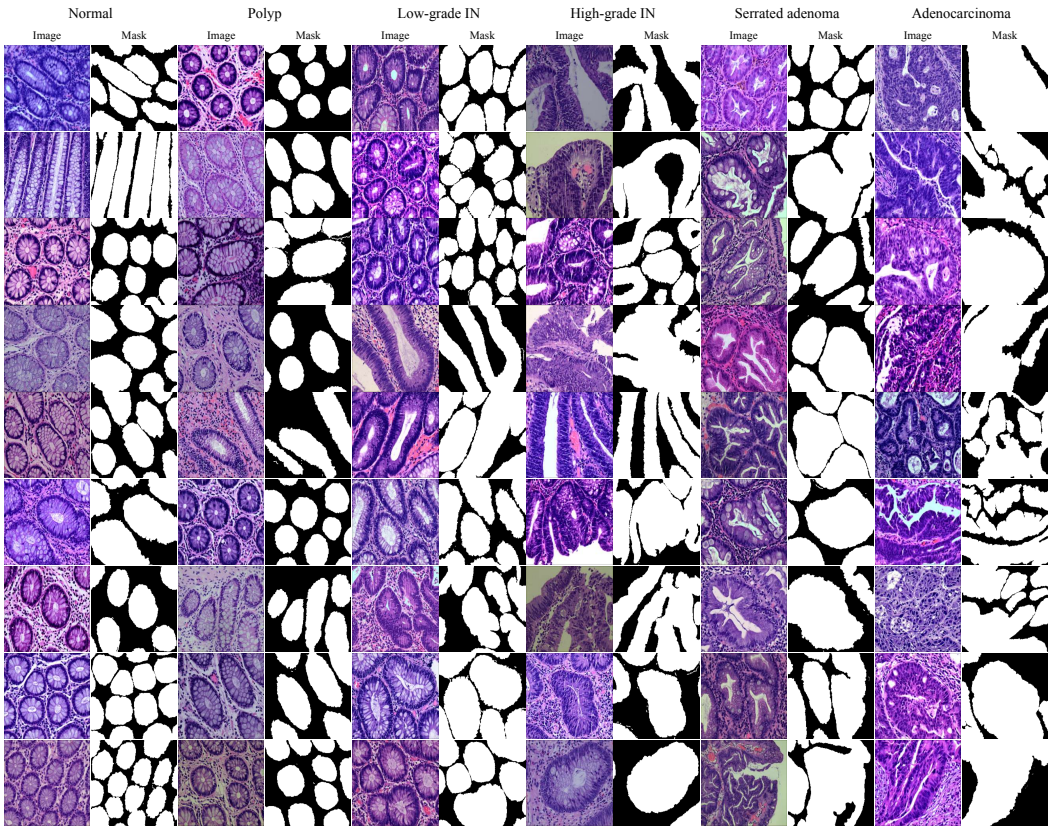


Figure 10: Examples from the **training split** of the constructed EBHI−S WSOL benchmark.

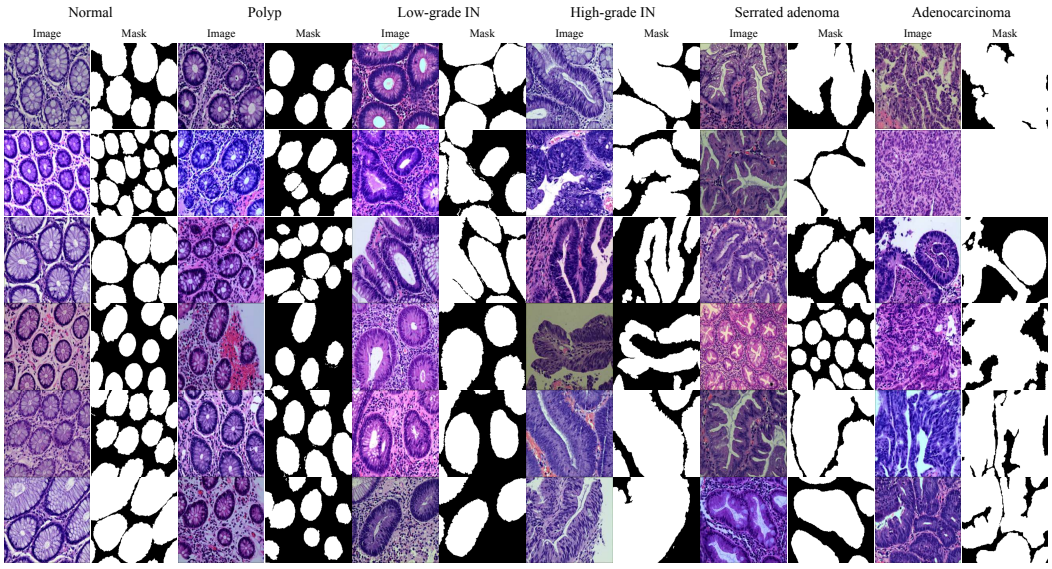


Figure 11: Examples from the **testing split** of the constructed EBHI−S WSOL benchmark.

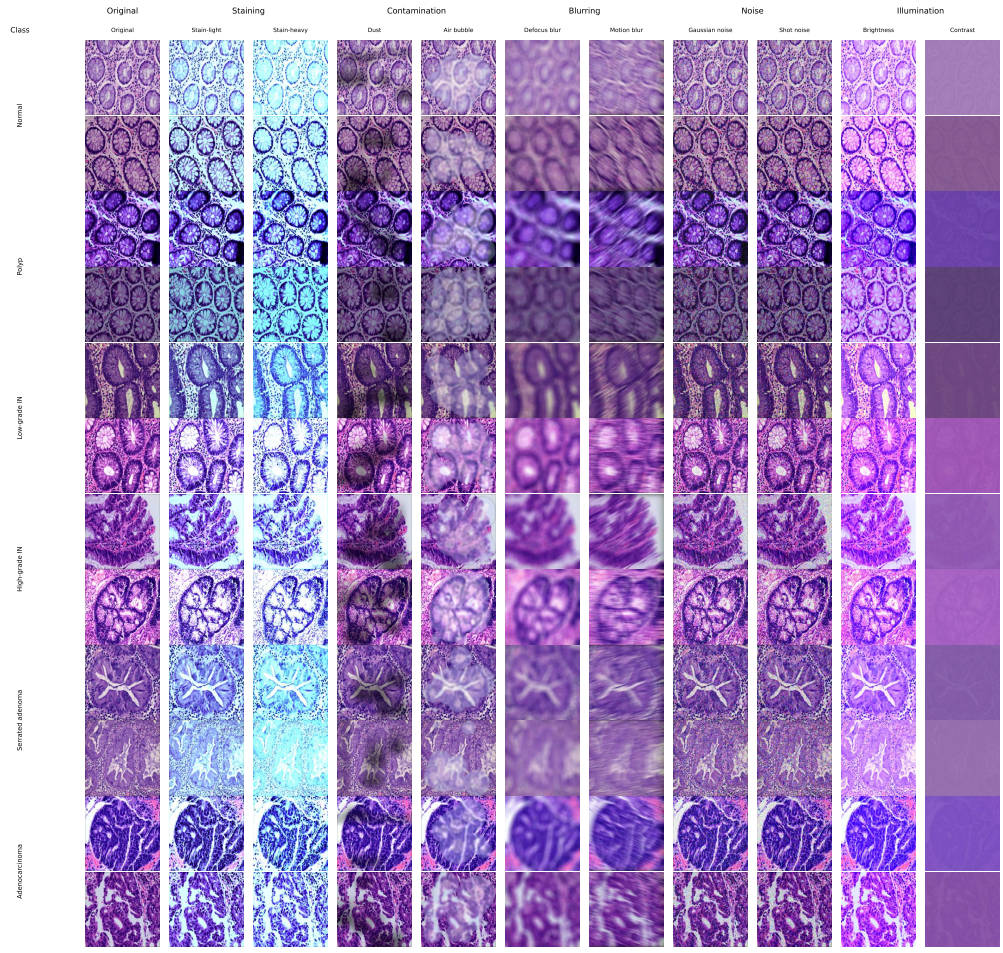


Figure 12: Examples from the **testing split** of the constructed `CAMELYON17` WSOL benchmark.

B Implementation and Reproducibility Details

B.1 Training and Adaptation Hyperparameters

This section provides the complete list of hyperparameters used throughout our experiments. We distinguish between general training parameters, WSOL-specific parameters, and method-specific hyperparameters for SFDA baselines.

Table 8: General hyperparameters used during source model training.

Hyperparameter	Value
Backbones	ResNet50 / DeiT
Optimizer	SGD
Nesterov acceleration	True
Momentum	$\in \{0.1, 0.4, 0.9\}$ for source models
Weight decay	0.0001
Learning rate	$\in \{0.0001, 0.001, 0.002, 0.01, 0.02, 0.1\}$ for source
Learning rate decay	`GlaS`: $0.1 \in \{150, 250, 350e\}$ epochs for source
Mini-batch size	32
Random flip	Horizontal/vertical random flip
Random color jittering	$\mathrm{Brightness}$, $\mathrm{contrast}$, and $\mathrm{saturation}$ at 0.5 and hue at 0.05
Image size	Resize image to 225×225. Then, crop random patches of 224×224
Learning epochs	`GlaS`: 1000

During the SFDA, all architectural choices and data augmentations are kept identical to source model training. Only optimization-related parameters and specific hyperparameters are varied. All SFDA methods share the same backbone, optimizer, batch size, and training schedule. Only method-specific loss weights and thresholds are varied.

Table 9: General hyperparameters used during SFDA training.

Hyperparameter	Value
Momentum	0.9
Weight decay	0.0001
Learning rate	$\in \{0.000001, 0.00001, 0.0001, 0.001\}$
Learning rate decay	`CAMELYON17`: $0.1 \in \{20\}$ epochs `CAMELYON16`: $0.1 \in 5$ epochs
Flip ratio (forgotten samples)	$\in \{0.05, 0.15\}$
`SFDA-SPF` period (epochs)	`CAMELYON17`: 5,15,35 epochs `CAMELYON16`: 3,5 epochs
Learning epochs	`CAMELYON17`: 100, `CAMELYON16`: 15
`SFDA-SPF`	$\lambda_F \in \{0.5, 1.0\}$ $\lambda_R \in \{0.5, 1.0\}$ $\lambda_L \in \{0.5, 1.0, 5.0\}$

Table 10: Per-method hyperparameters for SFDA methods.

Hyperparameter	Value
SHOT	$\lambda_{ent} \in \{0.5, 1.0\}$ $\lambda_{div} \in \{0.5, 1.0\}$ $\lambda_{peu} \in \{0.5, 1.0\}$
CDCL	$\tau \in \{0.04, 0.08, 0.5\}$ $\lambda_{cdcl} \in \{0.5, 1.0\}$
SFDE	$\tau \in \{0.8, 1.0\}$ $\lambda_{sfde} \in \{0.5, 1.0\}$
ERL	$\tau \in \{0.04, 0.08, 0.5\}$ $\lambda_{cdcl} \in \{0.5, 1.0\}$ $\lambda_{erl} \in \{0.5, 1.0\}$

The hyperparameter search spaces reported in Tables 8, 9, and 10 correspond to the primary experimental setting reported in the main paper, i.e., adaptation from `GlaS` to `CAMELYON16` and the five `CAMELYON17` centers. For the additional experimental scenarios reported in the supplementary material and cross-center adaptation, we use a fixed learning rate of 10^{-4}, a learning-rate decay factor $\gamma = 0.9$, and train the adaptation model for 100 epochs. To limit the computational cost of the extended experiments while maintaining a consistent model-selection protocol across methods, all method-specific loss weights are selected from the search space $\{0.5, 1.0\}$, independently of the considered SFDA method and loss term. Unless otherwise stated, all remaining architectural choices, optimization settings, and data augmentations are kept unchanged across scenarios. All experiments were conducted using NVIDIA H100 GPUs.

B.2 Experimental Settings

B.2.1 SFDA Methods

We compare our proposed method against five representative SFDA baselines covering different adaptation strategies.

- **SHOT** (Liang et al., 2020) exploits the structure of the target domain through pseudo-label refinement, with a diversity objective (information maximization loss) that discourages predictions from collapsing toward a subset of classes.
- **SFDA-DE** (Ding et al., 2022a) estimates source feature distributions from the source model classifier and reliable target features, and uses the surrogate source representations to reduce the source–target distribution discrepancy.
- **CDCL** (Wang et al., 2022) relies essentially on confident pseudo-labels and performs cross-domain contrastive learning using class prototypes (from the source classifier), encouraging features from the same pseudo-labels to remain close while separating features associated with different classes.
- **ERL** (Yi et al., 2023) exploits prediction uncertainty through entropy-based regularization to progressively refine the target model without accessing source samples.
- **RGV** (Zhu et al., 2025) leverages a memory-based mechanism to exploit relationships among target samples during adaptation.

B.2.2 Hyperparameter Selection and Validation Protocol

Hyperparameter selection constitutes a key challenge in SFDA. While traditionally SFDA is performed without any source or target annotations, benchmarking additionally requires a model-selection protocol for choosing hyperparameters such as the learning rate, number of epochs, and method-specific parameters.

Previous work has highlighted that model selection remains an unresolved issue in domain adaptation. Different protocols are used across the literature (Ericsson et al., 2023). Historically, oracle selection based on target performance has been used in several adaptation methods, while other

works rely on source validation data or dedicated validation sets (Guichemerre et al., 2024; Sharafi et al., 2025; Xu et al., 2025). In this work, we adopt an explicit validation-assisted protocol. The target training set used by all the SFDA methods remains entirely unlabeled, and no target annotation is involved in the adaptation objective. Hyperparameters and checkpoints are instead selected using a separate labeled validation split, while the test set remains strictly unseen and is accessed only for the final evaluation.

This protocol is consistent with validation strategies considered in recent SFDA methods. For example, (Boudiaf et al., 2023) explicitly use a validation set for hyperparameter and model selection when evaluating SFDA approaches. More recently, (Xu et al., 2025) investigate a similar validation-assisted setting by manually annotating 300 samples from the target domain to form a small validation set for hyperparameter selection.
Our protocol follows such previous work and our adaptation remains source-free and is performed only on unlabeled target training data, whereas a separate validation set is used only to determine the experimental configuration.

B.3 SFDA−SPF TRAINING ALGORITHM

Algorithm 3 SFDA-SPF Adaptation

Require: Unlabeled target set $\mathbb{T}$, source-pretrained WSOL model f^s, forgetting ratio ρ, reconstruction period m, number of epochs E, and loss weights $\lambda_{\mathrm{F}}, \lambda_{\mathrm{R}}, \lambda_{\mathrm{L}}$
Ensure: Target-adapted WSOL model f

Initialize f from the source-pretrained model f^s
Keep a frozen copy of f^s for CAM pseudo-label generation
for $e = 1$ to E **do**
  if $e = 1$ **or** $(e-1) \bmod m = 0$ **then**
    for all $x_i \in \mathbb{T}$ **do**
      Compute prediction p_i, pseudo-label $\hat{y}_i$, and normalized entropy $H(x_i)$
    end for
    Estimate the target prediction distribution $\hat{\pi}$
    Identify the dominant classes $\mathcal{C}_{\mathrm{dom}}$
    for all $k \in \mathcal{C}_{\mathrm{dom}}$ **do**
      Construct the class-specific candidate set $\mathbb{B}_k$
      Select its top-ρ highest-entropy samples as $\mathbb{D}_{\mathrm{F}}^k$
    end for
    Construct the forget set $\mathbb{D}_{\mathrm{F}} = \bigcup_{k \in \mathcal{C}_{\mathrm{dom}}} \mathbb{D}_{\mathrm{F}}^k$
    Construct the retain set $\mathbb{D}_{\mathrm{R}} = \mathbb{T} \setminus \mathbb{D}_{\mathrm{F}}$
  end if
  for all target mini-batches $\mathcal{B} \subset \mathbb{T}$ **do**
    Obtain forget and retain subsets $\mathcal{B}_{\mathrm{F}}$ and $\mathcal{B}_{\mathrm{R}}$
    Compute the forgetting loss $\mathcal{L}_{\mathrm{F}}$ on $\mathcal{B}_{\mathrm{F}}$
    Compute the retaining loss $\mathcal{L}_{\mathrm{R}}$ on $\mathcal{B}_{\mathrm{R}}$
    for all $x_i \in \mathcal{B}$ **do**
      Generate CAM-based partial pseudo-labels using the frozen source model f^s
      Randomly sample supervised pixel locations Ω_i'
    end for
    Compute the localization loss $\mathcal{L}_{\mathrm{L}}$
    Compute the total adaptation loss $\mathcal{L}$
    Update the backbone, image-level classifier, and pixel-level classifier
  end for
end for
return f

C ADDITIONAL QUANTITATIVE RESULTS

C.1 CROSS-ORGAN RESULTS

In Tables 1, 2, and 3, we evaluate SFDA−SPF under both intra-organ and inter-organ domain shifts. The intra-organ setting corresponds to breast-to-breast adaptation across different CAMELYON17 centers, whereas the inter-organ setting considers adaptation from colon to breast tissue. In this

section, we further evaluate the reverse inter-organ scenario (breast-to-colon adaptation). Inter-organ adaptation represents a particularly challenging setting because the target domain differs from the source not only in acquisition conditions and staining characteristics, but also in tissue morphology and the histological manifestations of malignancy. For instance, the cellular and tissue structures associated with normal and cancerous regions in colorectal tissue strongly differ from those observed in breast tissue.

Table 11: Reverse cross-dataset adaptation from each CAMELYON17 center to GlaS. Each entry reports classification accuracy/localization PxAP. The last column reports the average across the five CAMELYON17 source centers.

Method	C17-0 →GlaS	C17-1 →GlaS	C17-2 →GlaS	C17-3 →GlaS	C17-4 →GlaS	Average
SHOT	52.50 / 62.14	55.00 / 59.33	48.75 / 56.84	**53.75** / 66.37	53.75 / 56.83	52.75 / 60.30
SFDA-DE	**53.75** / 63.54	50.00 / 56.22	55.00 / 55.90	**53.75** / **66.37**	38.75 / 59.71	50.25 / **60.35**
CDCL	**53.75** / 63.64	46.25 / 52.55	47.50 / 55.60	**53.75** / 64.25	46.25 / 52.70	49.50 / 57.75
ERL	**53.75** / 56.92	70.00 / **59.40**	56.25 / 56.21	**53.75** / 43.52	51.25 / 60.24	57.00 / 55.26
RGV	**53.75** / **64.84**	53.75 / 50.40	55.00 / 55.90	**53.75** / 63.72	53.75 / **60.62**	54.00 / 59.10
SFDA-SPF	**53.75** / 53.11	**77.50** / 56.62	**61.25** / **59.83**	**53.75** / 62.04	**55.00** / 59.87	**60.25** / 58.29

As observed in the breast-to-colon adaptation scenario, most competing methods experience a substantial performance degradation. In contrast, SFDA−SPF consistently achieves better classification performance when adapting from each CAMELYON17 center to GlaS.

C.2 Robustness to Target-Class Imbalance (Cancer scenario)

In Table 3, we analyze the robustness of SFDA−SPF under different levels of target-domain class imbalance. We first consider a Normal-majority setting, which reflects a realistic histopathology scenario in which tumor regions occupy only a limited fraction of a whole-slide image (WSI), while large portions of the surrounding tissue remain non-tumorous. Consequently, patch-based sampling from such slides may yield substantially more normal than cancer samples. We also investigate the complementary Cancer-majority setting in Table 12, which can arise when tissue sampling is concentrated within large tumor regions, resulting in only a small proportion of normal tissue patches.

Table 12: Robustness to cancer-majority class imbalance in the target adaptation (train) set on CAMELYON17. Starting from the balanced setting (50:50), we progressively increase the Cancer/Normal ratio up to 90:10. Results are reported as classification accuracy (CL) / localization PxAP (%). Best results are shown in **bold**.

Method	50:50 **CL / PxAP**	60:40 **CL / PxAP**	70:30 **CL / PxAP**	80:20 **CL / PxAP**	90:10 **CL / PxAP**	**Average** **CL / PxAP**
CDCL	50.0 / 25.3	44.6 / 23.4	48.1 / 25.3	46.6 / 25.2	46.6 / 25.5	47.2 / 24.9
SHOT	70.0 / 19.2	41.2 / 37.2	41.2 / 37.2	41.2 / 37.2	41.2 / 37.2	47.0 / 33.6
ERL	59.5 / 13.2	43.1 / 28.2	44.3 / 21.9	43.9 / 22.1	43.9 / 21.7	46.9 / 21.4
RGV	50.0 / 31.0	50.0 / 34.2	50.0 / 34.3	50.0 / 34.9	50.0 / 35.3	50.0 / 33.9
SFDA-SPF	**86.2 / 49.9**	**82.1 / 47.6**	**83.2 / 50.9**	**83.9 / 49.7**	**84.7 / 50.2**	**84.0 / 49.7**

Table 13: Sensitivity to the forgetting proportion p under inter-center adaptation on CAMELYON17. The source model is trained on center 0 and independently adapted to centers 1–4. Results are reported as classification accuracy (CL) / localization PxAP (%). Best results for each target are shown in **bold**.

Target	$p = 0.01$ **CL / PxAP**	$p = 0.05$ **CL / PxAP**	$p = 0.10$ **CL / PxAP**
C17-0→1	67.44 / 25.37	74.42 / 34.69	**75.58** / **40.59**
C17-0→2	78.35 / 48.17	**79.24** / 54.34	79.02 / **59.12**
C17-0→3	92.29 / **63.45**	**93.88** / 60.44	93.09 / 62.39
C17-0→4	**83.22** / 46.13	79.87 / **48.88**	77.85 / 47.43

Table 14: Sensitivity to the localization keep ratio on CAMELYON17 inter-center adaptation using center 0 as source. Results are reported as classification accuracy (CL) / localization PxAP (%).

Target	Ratio of Images for Localization		
	0.2 **CL / PxAP**	0.5 **CL / PxAP**	1.0 **CL / PxAP**
C17-1	68.02 / 37.95	**75.58 / 40.59**	74.42 / 34.69
C17-2	79.24 / **59.38**	79.02 / 53.47	**79.24** / 54.34
C17-3	92.29 / **63.45**	93.35 / 60.61	**93.88** / 60.44
C17-4	**83.22** / 46.13	78.52 / 38.03	79.87 / **48.88**

Table 15: Impact of jointly adapting the backbone and image-level classifier. Results are reported on GlaS $\rightarrow$ C17-0 with PixelCAM. The source model used for CAM-based pixel pseudo-label generation remains frozen in all configurations.

Backbone	Classifier	PxAP ↑	CL ↑
Frozen	Trainable	<u>31.4</u>	47.7
Trainable	Frozen	<u>31.4</u>	<u>85.8</u>
Trainable	Trainable	**49.9**	**86.1**

Many SFDA approaches freeze the source classifier and adapt only the feature extractor, based on the assumption that the classifier preserves a class-discriminative structure learned from the source domain that remains transferable to the target domain. We examine whether this assumption remains valid under a strong cross-organ shift. Table 15 reports the effect of adapting the backbone and image-level classifier independently or jointly. Updating only the classifier while keeping the backbone frozen yields 47.7 CL and 31.4 PxAP, indicating that adjusting the decision layer alone cannot compensate for the mismatch between source and target image representations. In contrast, adapting the backbone while freezing the classifier improves classification accuracy to 85.8 CL, showing that feature adaptation plays an important role in recovering target-domain class discrimination. However, localization performance remains limited to 31.4 PxAP. Jointly adapting the backbone and classifier increases PxAP to 49.9, while improving classification accuracy of 86.1 CL.

C.3 CNNs VERSUS TRANSFORMER-BASED MODELS.

Figure 13 provides an explanation for the more modest gains observed with transformer-based WSOL models compared with CNN-based models. Unlike PixelCAM, for which uncertain dominant-class predictions are enriched in misclassification errors, SAT produces a highly concentrated low-entropy distribution.

Consequently, many erroneous predictions remain highly confident and are assigned to the retain set, making entropy less effective at separating reliable from incorrect predictions. Despite this less favorable confidence structure, SFDA-SPF improves SAT from 21.3 to 30.3 PxAP and from 52.1 to 69.2 CL on average. Its benefit is particularly pronounced for localization, exceeding the strongest competing SFDA method by 7.9 PxAP points.

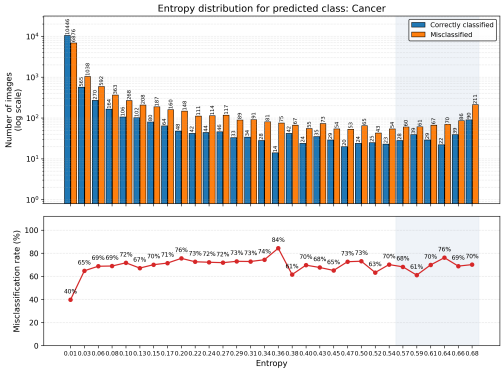


Figure 13: Entropy distributions of correctly and incorrectly classified samples predicted as cancer by SAT.

C.4 Extension to the Multi-Class Setting

Table 16: Localization (PxAP) and classification (F1) performance on EBHI−T under controlled domain shifts. The source model is trained on the clean source domain and independently adapted to each target condition. For all methods, results are reported at the checkpoint selected according to the best F1 score. Best results are shown in **bold** and second-best results are underlined.

	SHOT (ICML'20)		CDCL (TMM'22)		ERL (ICLR'23)		RGV (CVPR'25)		**Ours**	
Target condition	PxAP	F1	PxAP	F1	PxAP	F1	PxAP	F1	PxAP	F1
Target clean	83.7	77.2	83.7	77.2	83.7	77.2	83.7	77.2	**89.3**	**82.3**
Stain-light	83.5	**75.1**	83.8	67.1	83.7	69.9	83.8	67.1	**90.9**	**75.1**
Stain-heavy	**86.7**	**70.3**	81.3	55.9	70.5	56.7	80.8	43.9	82.7	57.9
Dust	82.9	**79.4**	83.1	71.0	83.1	71.0	83.1	71.0	**86.7**	72.4
Air bubble	77.8	**78.1**	76.8	45.3	76.8	45.3	76.8	45.3	**78.4**	66.5
Defocus blur	**73.3**	**42.2**	71.9	17.4	71.9	17.4	71.9	17.4	76.5	21.6
Motion blur	70.7	21.1	67.9	19.4	67.9	19.4	67.9	19.4	**72.9**	**29.3**
Gaussian noise	82.9	**60.9**	84.1	45.5	77.1	53.1	84.1	45.5	**88.2**	57.2
Shot noise	84.3	46.5	79.1	50.5	75.3	**54.5**	84.3	46.5	**86.8**	52.6
Brightness	81.1	70.9	81.1	70.9	78.0	72.5	81.1	70.9	**84.6**	**79.4**
Contrast	**71.1**	**30.8**	63.4	9.1	63.4	9.1	63.4	9.1	62.6	12.1
Average	79.8	**59.3**	77.8	48.1	75.6	49.7	78.3	46.7	**82.3**	58.5

Table 17: Per-class source-model predictions under representative target shifts on the EBHI training set. For each class, we report the ground-truth distribution (GT), per-class recall (Rec.), precision (Prec.), F1-score, and predicted class distribution (Pred.). Target Clean and Brightness represent conditions where the source model preserves predictions across classes, whereas Defocus Blur and Contrast illustrate severe prediction collapse.

			Target Clean (Shift 1)				**Brightness (Shift 10)**				**Defocus Blur (Shift 6)**				**Contrast (Shift 11)**			
Class	**GT**	**GT (%)**	**Rec.**	**Prec.**	**F1**	**Pred.**	**Rec.**	**Prec.**	**F1**	**Pred.**	**Rec.**	**Prec.**	**F1**	**Pred.**	**Rec.**	**Prec.**	**F1**	**Pred.**
Normal	22	3.3	13.6	100.0	24.0	3 (0.5%)	18.2	80.0	29.6	5 (0.8%)	0.0	0.0	0.0	0 (0.0%)	0.0	0.0	0.0	0 (0.0%)
Polyp	143	21.4	94.4	83.9	88.8	161 (24.1%)	95.1	81.0	87.5	168 (25.2%)	10.5	65.2	18.1	23 (3.4%)	6.3	81.8	11.7	11 (1.6%)
Low-grade IN	191	28.6	91.1	91.1	91.1	191 (28.6%)	81.2	92.8	86.6	167 (25.0%)	0.0	0.0	0.0	0 (0.0%)	0.0	0.0	0.0	0 (0.0%)
High-grade IN	55	8.2	69.1	73.1	71.0	52 (7.8%)	69.1	64.4	66.7	59 (8.8%)	90.9	9.4	17.0	535 (**80.2%**)	98.2	8.8	16.1	614 (**92.1%**)
Serrated adenoma	17	2.5	94.1	94.1	94.1	17 (2.5%)	94.1	94.1	94.1	17 (2.5%)	5.9	100.0	11.1	1 (0.1%)	0.0	0.0	0.0	0 (0.0%)
Adenocarcinoma	239	35.8	96.2	94.7	95.4	243 (36.4%)	96.7	92.0	94.3	251 (37.6%)	25.1	55.6	34.6	108 (16.2%)	14.2	81.0	24.2	42 (6.3%)

In the 2-class setting, SFDA−SPF achieves the best localization and classification performance in the majority of domain shifts, supporting its effectiveness. In a more challenging multi-class setting with class imbalance, SFDA−SPF remains competitive in classification, outperforming CDCL, ERL, and RGV on most shifts. It achieves particularly large gains on Motion Blur and Brightness cases. Nevertheless, all SFDA methods experience degradation under some important shifts, such as Dust, Air Bubble, Defocus Blur, Noise, and Contrast, where SHOT can outperform our approach. This limitation is explained by the combination of severe class imbalance and multiclass similarity.

Table 17 provides insight into the shift-dependent behavior observed in Table 16. For Target Clean and Brightness, where SFDA−SPF performs particularly well, the source model preserves meaningful predictions across all six classes. In contrast, Defocus Blur and Contrast induce severe prediction collapse: although High-grade IN represents only 8.2% of the target samples, it accounts for 80.2% and 92.1% of the source-model predictions, respectively, while some classes receive no predictions at all. Under such extreme collapse, identifying samples to forget is only part of the problem. Unlike in the 2-class setting, reconsidering a dominant prediction leaves several possible alternative classes, some of which are absent from the initial predictions. This extreme prediction collapse leaves very limited information for recovering classes that are entirely absent from the initial predictions, making adaptation particularly difficult for all evaluated SFDA methods. In this specific scenario, SHOT can occasionally benefit from its diversity objective, which encourages predictions across multiple classes and may help recover initially unrepresented classes. This advantage is, however, specific to some severe multi-class collapse cases and is not consistently observed across the evaluated shifts.

D RELIABILITY OF SOURCE CAMS FOR PIXEL PSEUDO-LABELING

D.1 RELIABILITY OF HIGH-CONFIDENCE FOREGROUND REGIONS

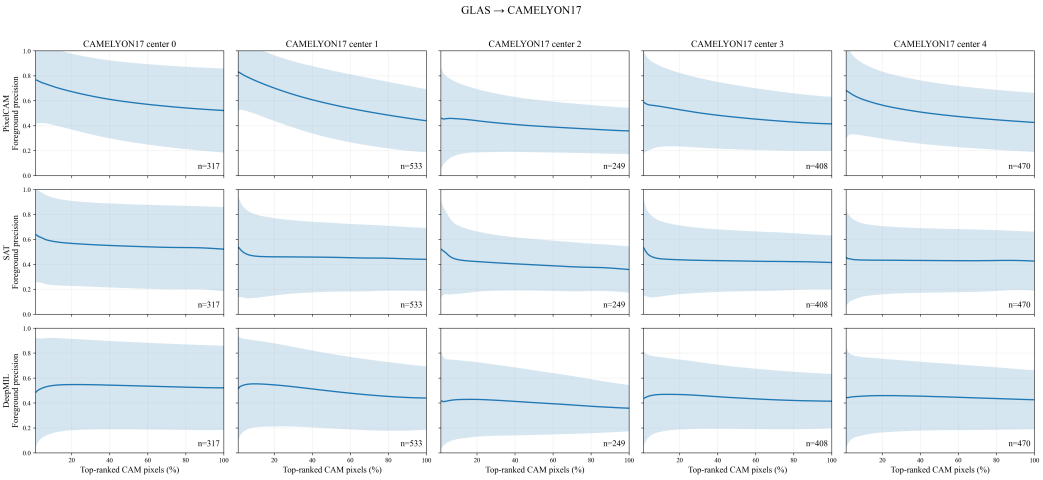


Figure 14: Reliability of high-confidence foreground regions for GlaS → CAMELYON17. For each positive target image, CAM pixels are ranked according to their activation score, and foreground precision is computed as the fraction of top-ranked pixels overlapping the ground-truth foreground mask. Results are reported for PixelCAM, SAT, and DeepMIL across the five CAMELYON17 centers.

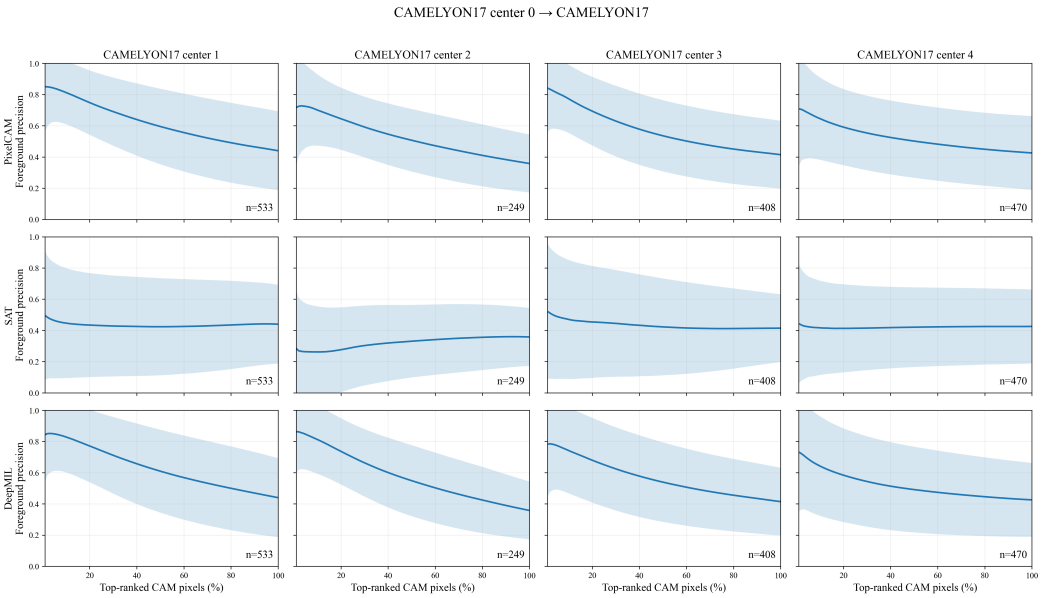


Figure 15: Reliability of high-confidence foreground regions under CAMELYON17 inter-center domain shifts using center 0 as the source. The source models are evaluated on centers 1–4. Results are reported for PixelCAM, SAT, and DeepMIL.

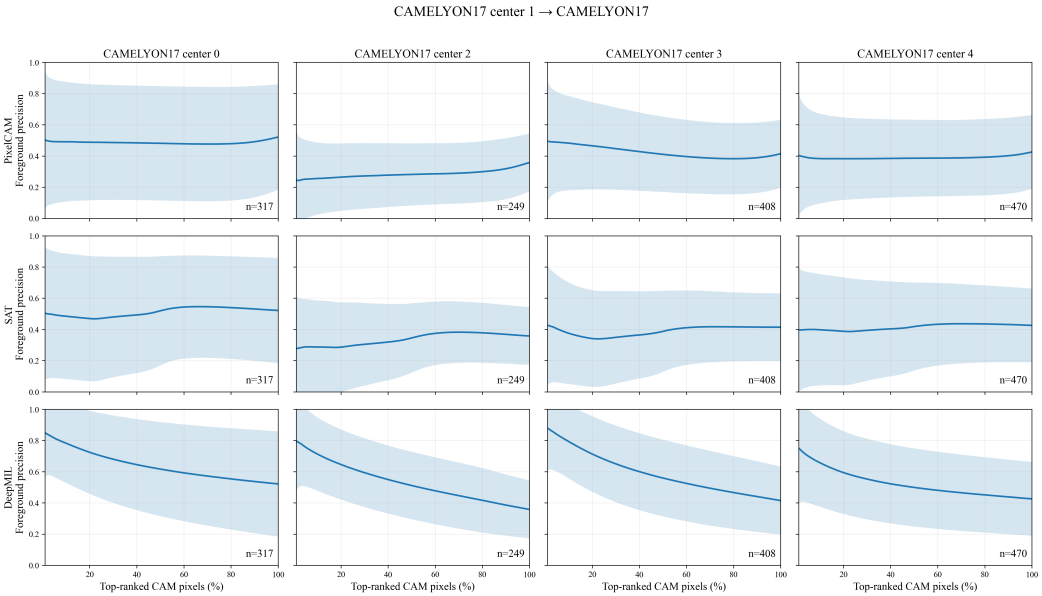


Figure 16: Reliability of high-confidence foreground regions under CAMELYON17 inter-center domain shifts using center 1 as the source. The source models are evaluated on centers 0, 2, 3, and 4. Results are reported for PixelCAM, SAT, and DeepMIL.

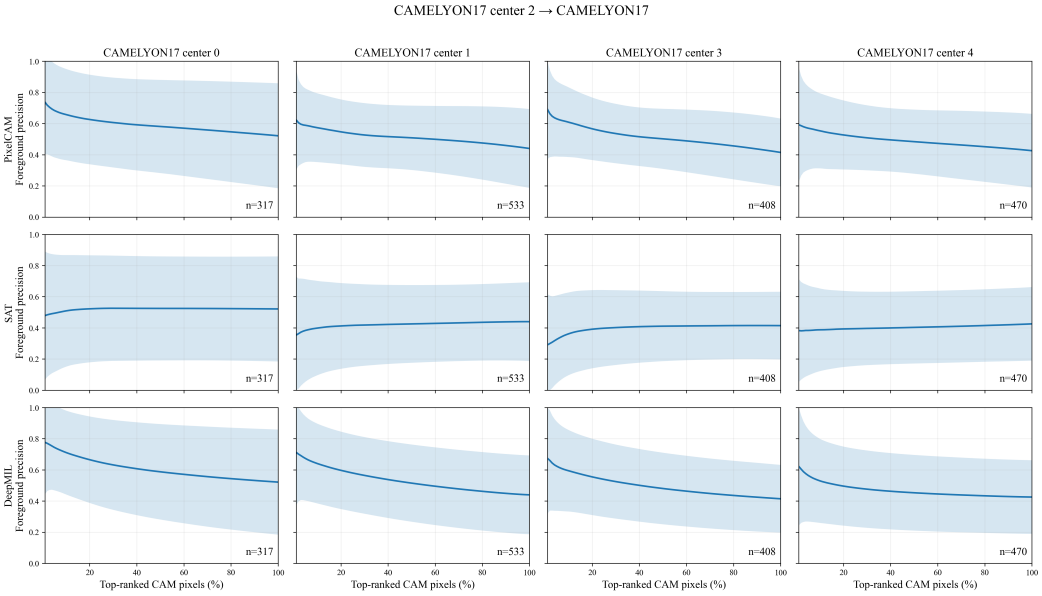


Figure 17: Reliability of high-confidence foreground regions under CAMELYON17 inter-center domain shifts using center 2 as the source. The source models are evaluated on centers 0, 1, 3, and 4. Results are reported for PixelCAM, SAT, and DeepMIL.

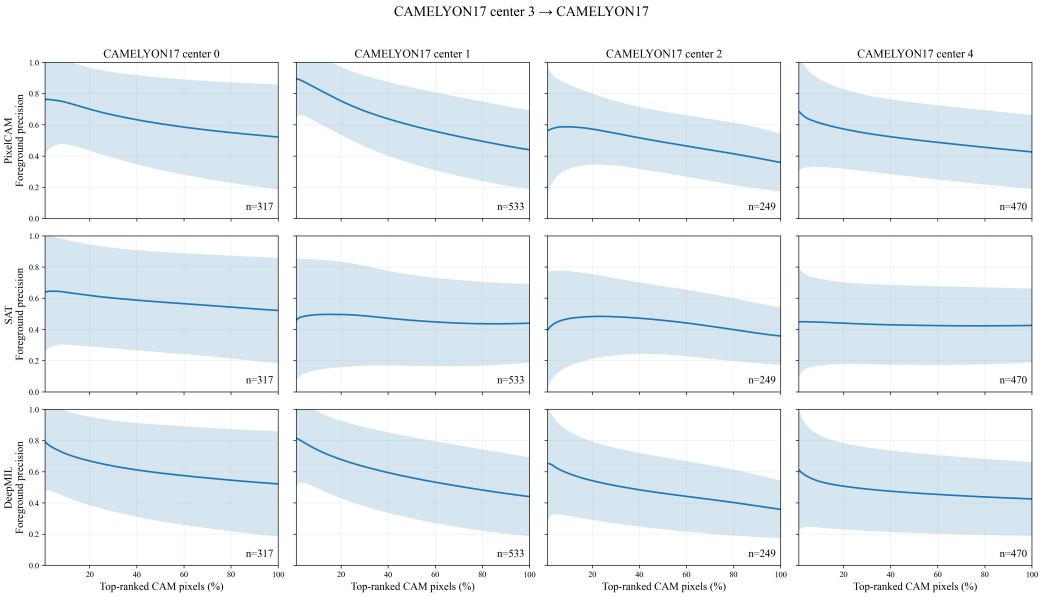


Figure 18: Reliability of high-confidence foreground regions under CAMELYON17 inter-center domain shifts using center 3 as the source. The source models are evaluated on centers 0, 1, 2, and 4. Results are reported for PixelCAM, SAT, and DeepMIL.

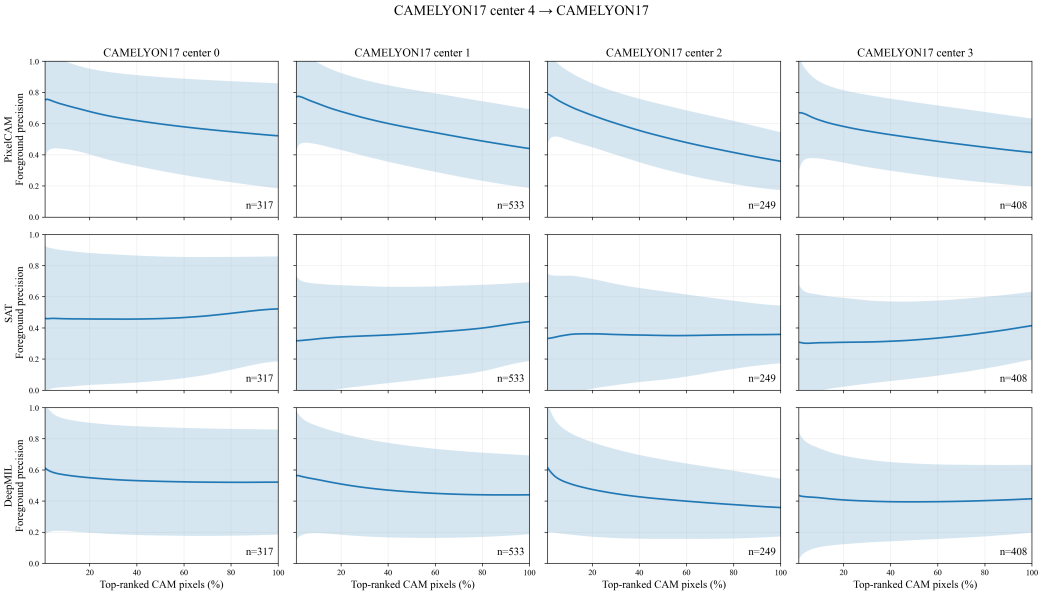


Figure 19: Reliability of high-confidence foreground regions under CAMELYON17 inter-center domain shifts using center 4 as the source. The source models are evaluated on centers 0–3. Results are reported for PixelCAM, SAT, and DeepMIL.

D.2 Reliability of Low-Activation Background Regions

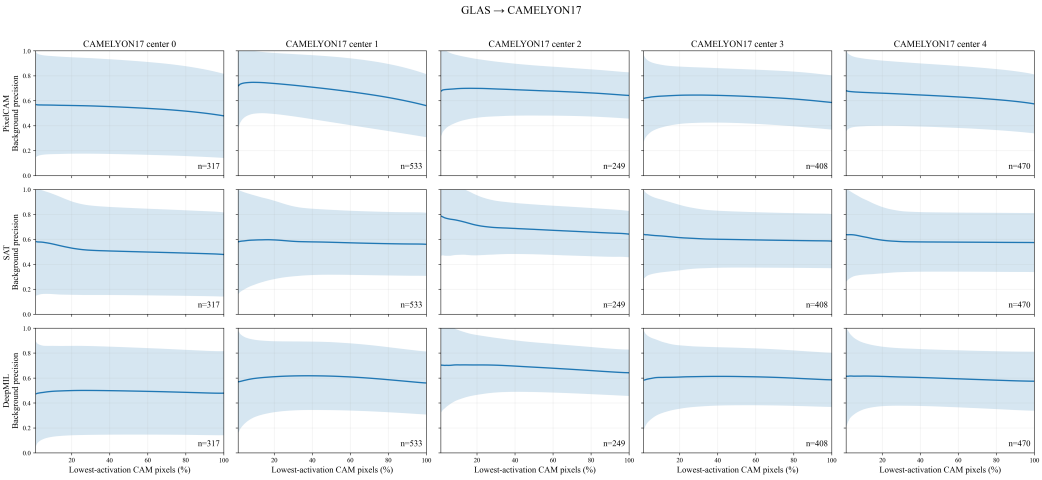


Figure 20: Reliability of low-activation background regions for GlaS → CAMELYON17. For each positive target image, CAM pixels are ranked according to their activation score, and background precision is computed as the fraction of the lowest-activation pixels overlapping the ground-truth background region. Results are reported for PixelCAM, SAT, and DeepMIL across the five CAMELYON17 centers.

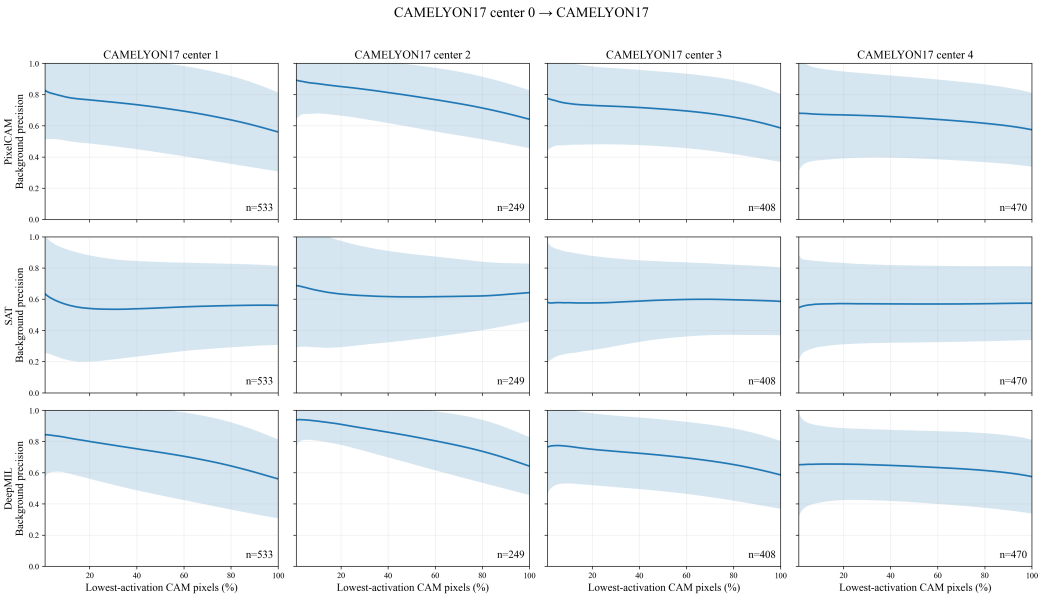


Figure 21: Reliability of low-activation background regions under CAMELYON17 inter-center domain shifts using center 0 as the source. The source models are evaluated on centers 1–4. Results are reported for PixelCAM, SAT, and DeepMIL.

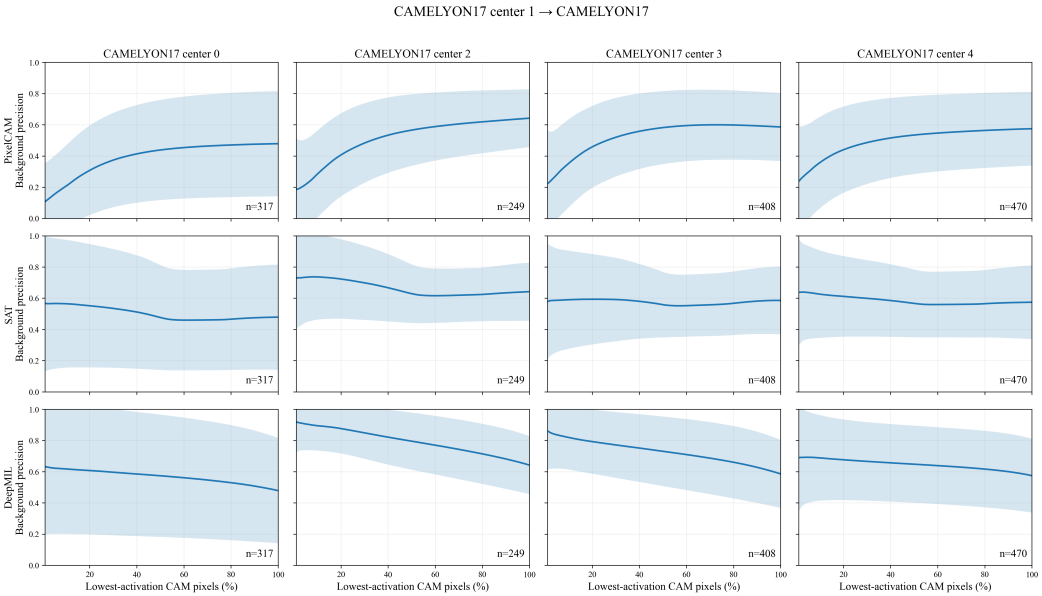


Figure 22: Reliability of low-activation background regions under CAMELYON17 inter-center domain shifts using center 1 as the source. The source models are evaluated on centers 0, 2, 3, and 4. Results are reported for PixelCAM, SAT, and DeepMIL.

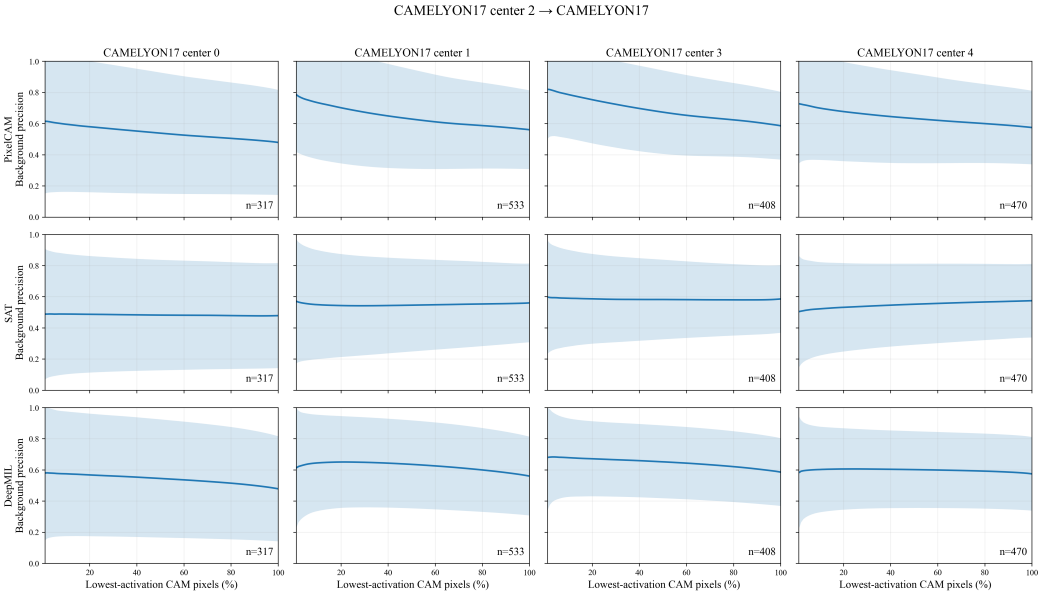


Figure 23: Reliability of low-activation background regions under CAMELYON17 inter-center domain shifts using center 2 as the source. The source models are evaluated on centers 0, 1, 3, and 4. Results are reported for PixelCAM, SAT, and DeepMIL.

E ADDITIONAL VISUALIZATIONS

SFDA−SPF motivation is that uncertain predictions within over-predicted classes are particularly informative for identifying potentially erroneous target predictions. Rather than relying exclusively on confident pseudo-labels, SFDA−SPF explicitly identifies uncertain samples from over-predicted classes and encourages the model to forget these predictions. This mechanism allows the decision boundary to progressively move away from source predictions and better separate normal and cancer samples in the target domain. To empirically support this motivation, Figure 24 analyzes the distribution of correct and incorrect predictions across predictive-entropy levels under different domain shifts and target class priors.

E.1 ENTROPY DISTRIBUTIONS



Figure 24: Distribution of correct and incorrect predictions across predictive-entropy bins under different domain shifts and target class priors. Results are shown for PixelCAM. Bars report the percentage of target samples within each normalized entropy bin, decomposed into correctly (blue) and incorrectly (orange) classified samples.

E.2 ENTROPY–ACCURACY RELATIONSHIPS

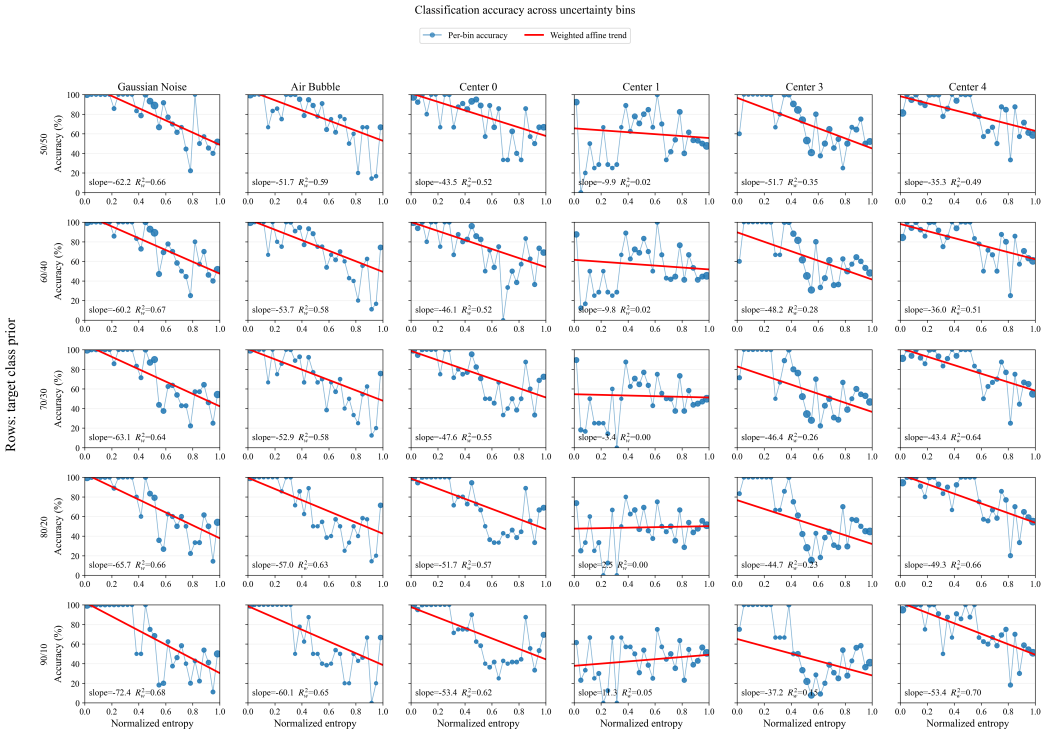


Figure 25: Relationship between prediction entropy and classification accuracy under CAMELYON17 inter-center domain shifts using center 2 as the source. Results are reported for PixelCAM.

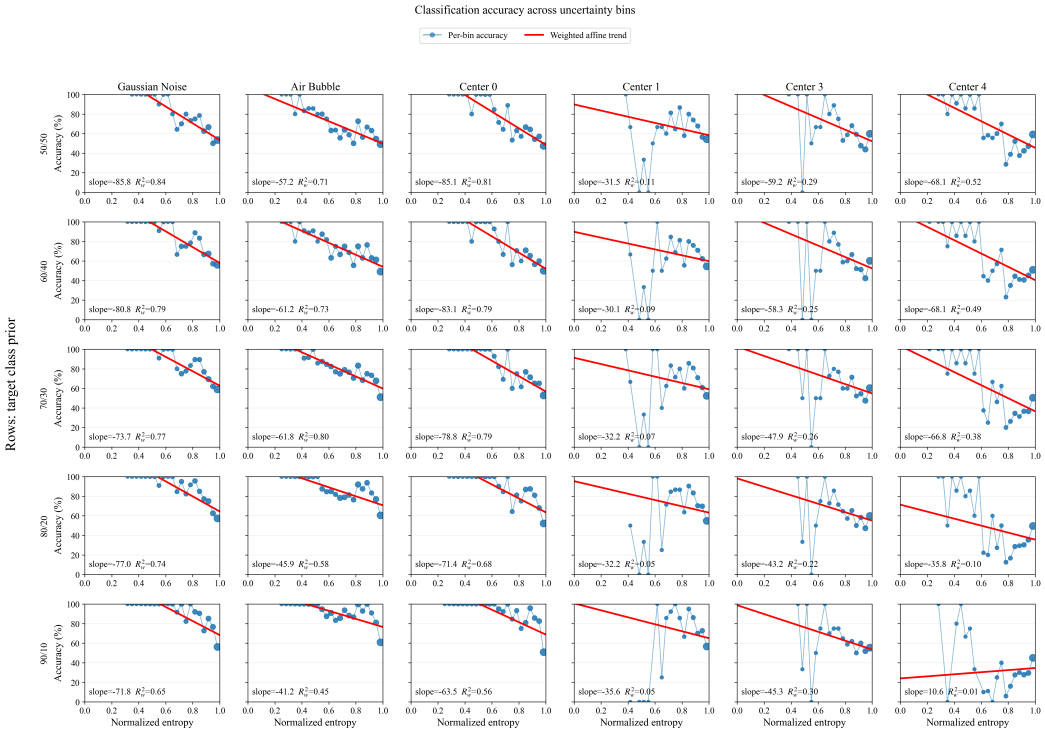


Figure 26: Relationship between prediction entropy and classification accuracy under CAMELYON17 inter-center domain shifts using center 2 as the source. Results are reported for SAT.

E.3 Additional Visualizations

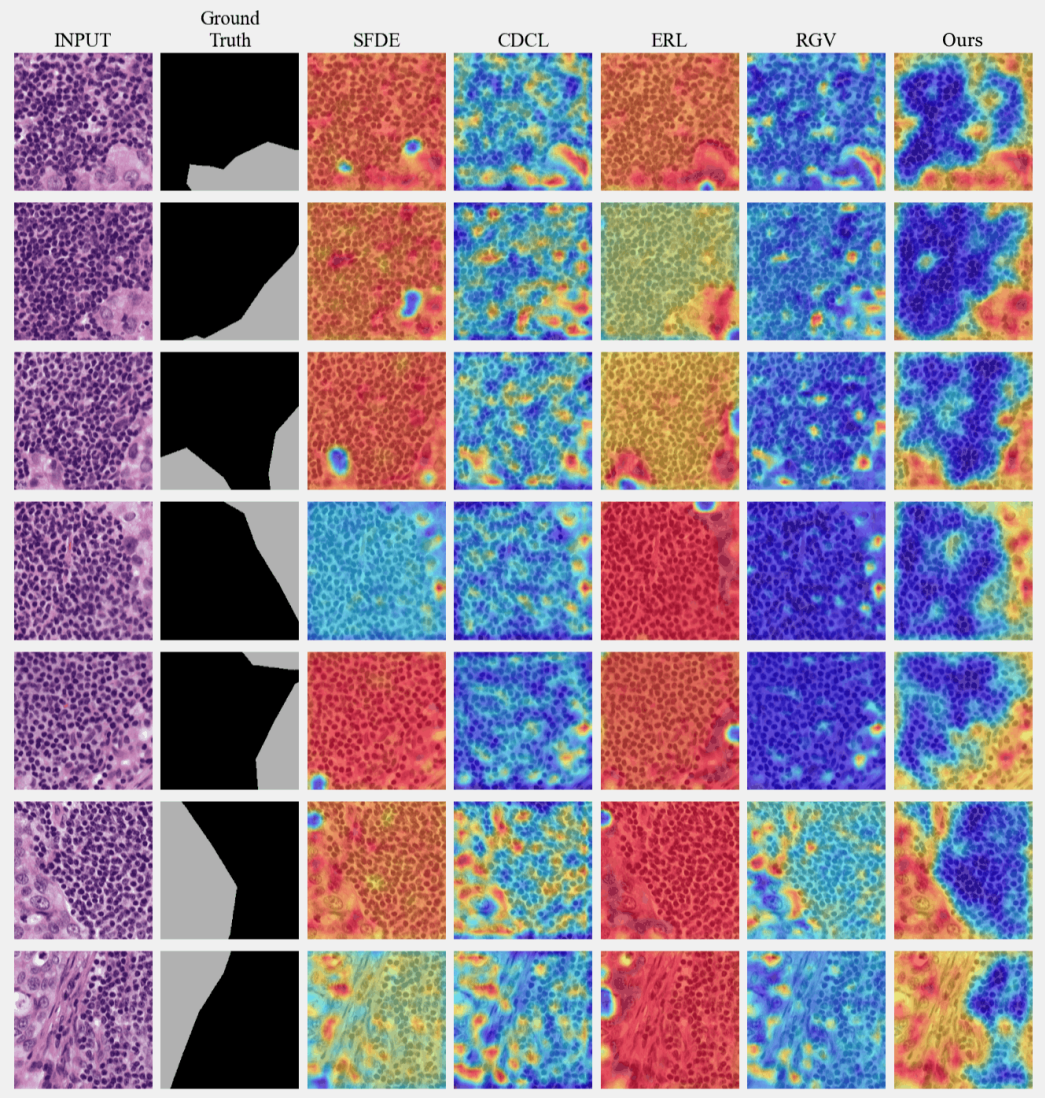


Figure 27: Localization produced by the `PixelCAM` WSOL model after `SFDA-SPF` and the state-of-the-art SFDA methods (`GlaS` → `CAMELYON17-3`).

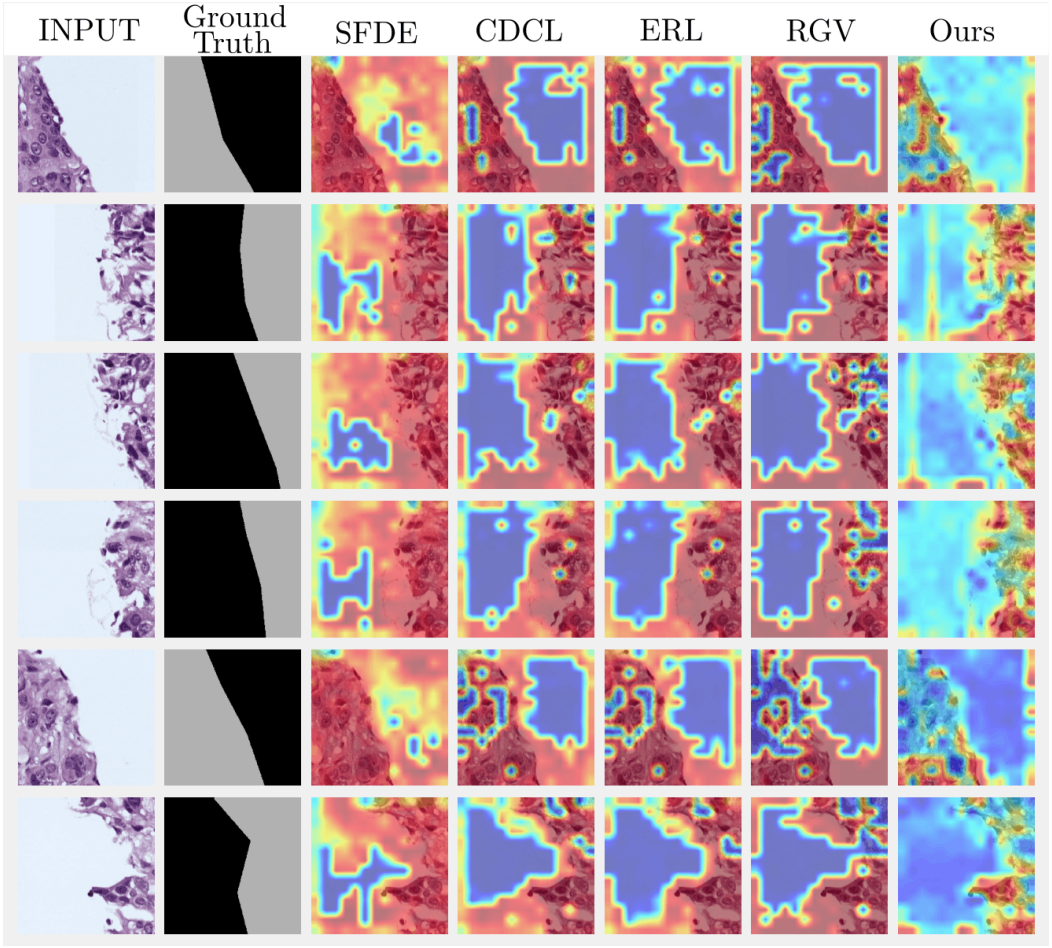


Figure 28: Localization produced by the SAT WSOL model after SFDA-SPF and the state-of-the-art SFDA methods (GlaS → CAMELYON17-3).

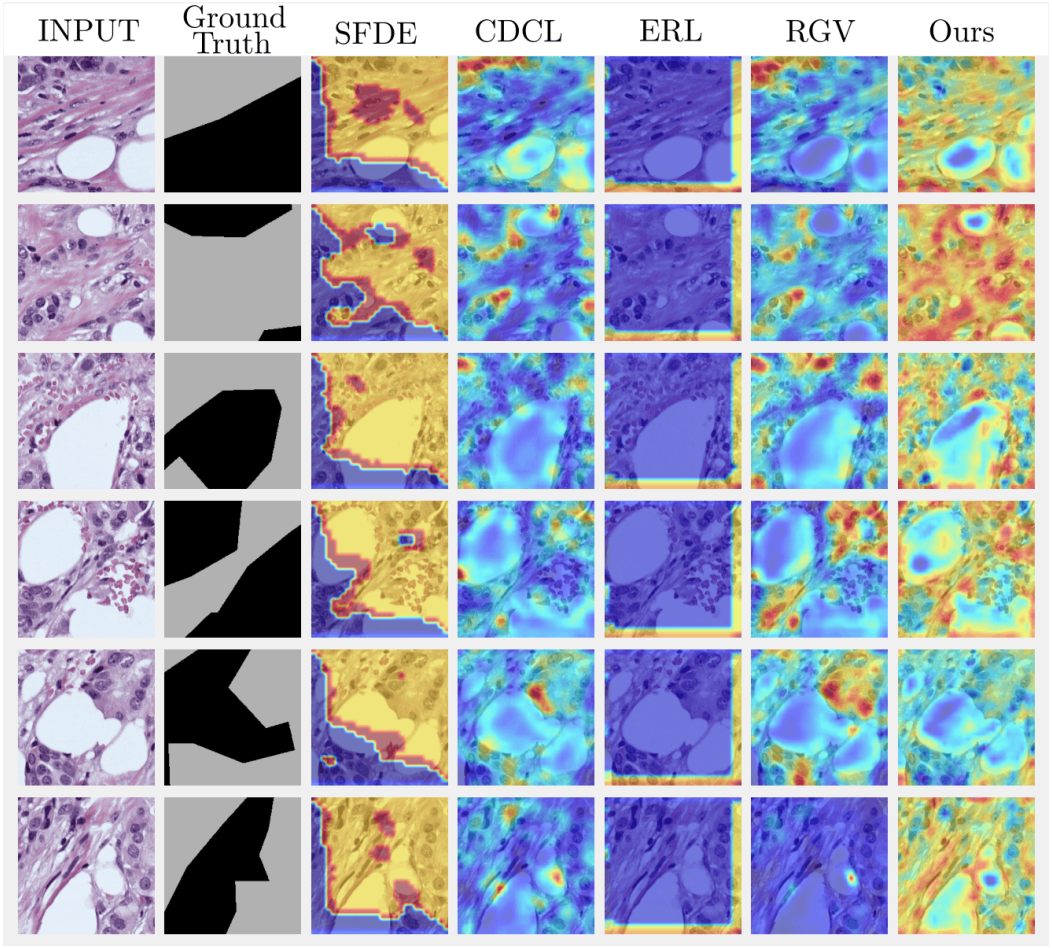


Figure 29: Localization produced by the `DeepMIL` WSOL model after `SFDA-SPF` and the state-of-the-art SFDA methods (`GlaS` → `CAMELYON17-3`).