Are nuclear masks all you need for improved out-of-domain generalisation? A closer look at cancer classification in histopathology

Dhananjay Tomar
University of Oslo
dhananjt@ifi.uio.no

Alexander Binder
Otto-von-Guericke University Magdeburg,
Singapore Institute of Technology
alexander.binder@ovgu.de

Andreas Kleppe*
Oslo University Hospital, University of Oslo,
UiT The Arctic University of Norway
andrekle@ifi.uio.no

Abstract

Domain generalisation in computational histopathology is challenging because the images are substantially affected by differences among hospitals due to factors like fixation and staining of tissue and imaging equipment. We hypothesise that focusing on nuclei can improve the out-of-domain (OOD) generalisation in cancer detection. We propose a simple approach to improve OOD generalisation for cancer detection by focusing on nuclear morphology and organisation, as these are domain-invariant features critical in cancer detection. Our approach integrates original images with nuclear segmentation masks during training, encouraging the model to prioritise nuclei and their spatial arrangement. Going beyond mere data augmentation, we introduce a regularisation technique that aligns the representations of masks and original images. We show, using multiple datasets, that our method improves OOD generalisation and also leads to increased robustness to image corruptions and adversarial attacks. The source code is available at https://github.com/undercutspiky/SFL/

1 Introduction

Domain generalisation in histopathology is a crucial challenge because domain shifts naturally occur among hospitals and even within a single hospital or laboratory, e.g., temporally or among human operators and observers such as pathologists. Non-biological factors that substantially alter the images include differences in scanners, staining protocols, fixation of tissue, and even minor aspects like the manufacturer and storage conditions of stains [1].

Collecting data from numerous hospitals to address these domain shifts is often impractical and may not adequately reflect the full variability present in routine clinical practice, thus making it difficult to build computational histopathology models that generalise well. This leads us to focus on single-domain generalisation (S-DG) in this paper, specifically on how to train a model using data from only one hospital (considered a domain here) that generalises well to data from other hospitals. Popular S-GD methods in histopathology apply data augmentation and stain normalisation [2, 3]. The effectiveness of S-GD methods developed for natural images remains underexplored in histopathology [3]. Here, we compare these methods to a new, simple approach that we propose.

*Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Research has shown that Convolutional Neural Networks (CNNs) tend to focus on texture over shape [4, 5]. However, in histopathology, the texture and colour of cell nuclei vary much more across domains than the shape and organisation of cell nuclei. As a result, focusing on shape features could improve a computational histopathology model's ability to generalise to unseen data because it may rely less on domain-specific features that vary across hospitals and more.

Nuclei in cancerous tissue exhibit distinct changes in shape, size, and overall organisation compared to nuclei in normal tissue [6–8]. Pathologists rely on these and other visual cues [9] for cancer diagnosis and grading, underscoring the biological importance and the consistency of nuclear morphology and organisation across domains. We hypothesise that focusing on nuclear morphology and organisation may be sufficient for cancer detection and that exploiting this during training could result in models with good generalisation.

We propose a method that encourages CNNs to focus more on nuclear morphology and organisation by using additional loss terms that prioritise shape-based features. Specifically, our method leverages nuclear segmentation masks during training to steer the learning towards nuclei. Through extensive experimentation, we demonstrate that this method improves performance on out-of-domain data without requiring nuclear segmentation masks at inference time, thus offering a promising and attractive solution for addressing domain generalisation in histopathology. Our contributions include:

- We propose a novel training method that incentivises the model to focus on nuclei.
- We evaluate our method on three datasets comprising hundreds of WSIs in total from various hospitals and organs. Our results show accuracy improvements over all other approaches.
- We evaluate the sensitivity of our method to image corruptions and adversarial attacks. Our results show performance improvements over the baseline.
- We conduct extensive ablation studies to show that models trained with our method focus on nuclei.

2 Related work

The prediction of various properties such as malignancy, grading, and HER2 expression using segmented nuclei has been a well-studied topic for many years [10–16]. Researchers have employed techniques such as watershed segmentation [17], thresholding, level sets [18], and snakes [19], often followed by extracting explicit morphometric features from the segmentations. For example, early work by Hasegawa et al. [20] focused on counting segmented regions, while Lee and Street [21] applied neural networks to the segmentation outputs. In contrast to these approaches, our method does not rely on segmentation during inference. Instead, we adjust the training process to encourage the extraction of nuclear features.

Stain normalisation methods convert the colours of a source image to match those of a target image. These methods were typically designed specifically for the most common type of histopathology images, which are images of tissue stained with haematoxylin and eosin (H&E). One of the earlier methods, Macenko normalisation [1], estimates stain vectors for source and target images and uses them to normalise the source image. Vahadane et al. [22] proposed a method that decouples stain "density maps" from "colour appearances", allowing the combination of the source image's density maps with the target image's colour appearances. Reinhard et al. [23] pioneered colour transfer by adjusting the global statistics of images in a different colour space, effectively transferring the colour characteristics of the target to the source image. Random Stain Normalization and Augmentation (RandStainNA) [24] combines stain normalisation and augmentation. Unlike traditional approaches that normalise using a fixed template, RandStainNA generates random virtual templates in the LAB [23] colour space and uses them to normalise the images during training. The templates are drawn from Gaussian distributions whose means and variances are derived from the training data. For a more comprehensive review, we refer the readers to [25]. In summary, the stain normalisation methods primarily focus on manipulating colour information to remove stain variability. On the other hand, our approach shifts the focus from colour manipulation to nuclear features.

Data augmentation is a common way to facilitate domain generalisation. Tellez et al. [2] evaluated several stain colour augmentation and stain normalisation methods and found that colour augmentation was crucial for good performance on external test sets in histopathology. Faryna et al. [26] extended RandAugment [27] by including certain histopathology-specific augmentations and excluding the ones that produce unrealistic-looking images. Tellez et al. [28] developed a data augmentation method

specific to H&E-stained images and used it for domain generalisation in mitosis detection. Pohjonen et al. [29] developed StrongAugment, where varying numbers of transformations are applied to an image to improve domain generalisation. Marini et al. [30] proposed Data-driven colour augmentation (DDCA), which evaluates an augmented image as acceptable or not for training based on its distance from other images in a database. Faryna et al. [31] evaluated different data augmentation strategies in histopathology, including manually selected augmentations, and found them all to be competitive.

Single-domain generalisation (S-DG) methods do not require data from multiple domains during training. Representation Self-Challenging (RSC) [32] works by discarding the features with relatively high gradients, making the model predict with the remaining features during training. Adversarial Domain Augmentation (ADA) [33] generates adversarial examples iteratively to augment the source domain and creates an ensemble of models. Meta-Learning-based ADA (M-ADA) [34] uses Wasserstein Auto-Encoder [35] to generate new samples and uses adversarial training on top along with a meta-learning scheme. Progressive domain expansion network (PDEN) [36] uses multiple autoencoders to generate new samples to expand the training set. Learning to Diversify (L2D) [37] introduces a learnable style-complement module that generates augmented images. The style-complement module is trained to diversify the images as much as possible but still keep the semantic information intact.

Domain adaptation, unlike S-DG, requires having access to some samples from the target domain. In histopathology, domain adaptation methods commonly make use of GAN [38] and CycleGAN [39]. StainGAN [40] uses CycleGAN to make images in the source domain look like the target domain. Residual CycleGAN [41] modifies the CycleGAN objective to have the generator produce the residual between domains instead of recreating the input image. In [42], authors augment a generator in CycleGAN with a stain colour matrix as an auxiliary input to stabilise the training. NST_AD_HRNet [43] uses Neural Style Transfer [44, 45] and GAN to preserve the content of the source image while combining it with the style of the target image. In some earlier works [46, 47], the input image is converted to greyscale and then coloured using a generator network which is based on a target image. While domain adaptation is not S-DG and thus a bit tangential to the focus of this paper, it is worth noting that domain adaptation is impractical in many clinical settings and may result in worse generalisation than stain normalisation and colour augmentation [48].

3 Proposed Method

Our approach aims to enhance S-DG by incentivising the model to focus on shaped-based features of nuclei in histopathological images and thereby reduce overfitting to irrelevant features that may carry higher label noise.

The first step involves generating segmentation masks that highlight specific areas of interest in the image. This step is applied only during training, while test-time evaluation relies solely on H&E-stained images. As we hypothesised that nuclear morphology and organisation contain sufficient information for cancer detection, our segmentation masks are binary images with nuclear pixels as foreground and other pixels as background.

One possible approach to using the segmentation mask is to include it as a fourth channel in the input image. Alternatively, the mask can be used as the sole input to the model. However, both methods necessitate running the segmentation model during inference, which increases computational demands and slows down processing.

Our method circumvents the need for a nuclear segmentation network at inference time by incorporating additional loss terms during training. For a given input image x and its corresponding segmentation mask x', our method involves the following steps:

1. Execute a forward propagation through the neural network model on both the H&E-stained image x and its nuclear mask x', saving the embeddings generated by the network as z and z' for x and x', respectively.
2. Compute the Binary Cross-Entropy (BCE) loss for both x and x'.
3. Compute the ℓ_2-distance between the embeddings z and z'.
4. Minimise the sum of the two CE losses and the ℓ_2-distance.

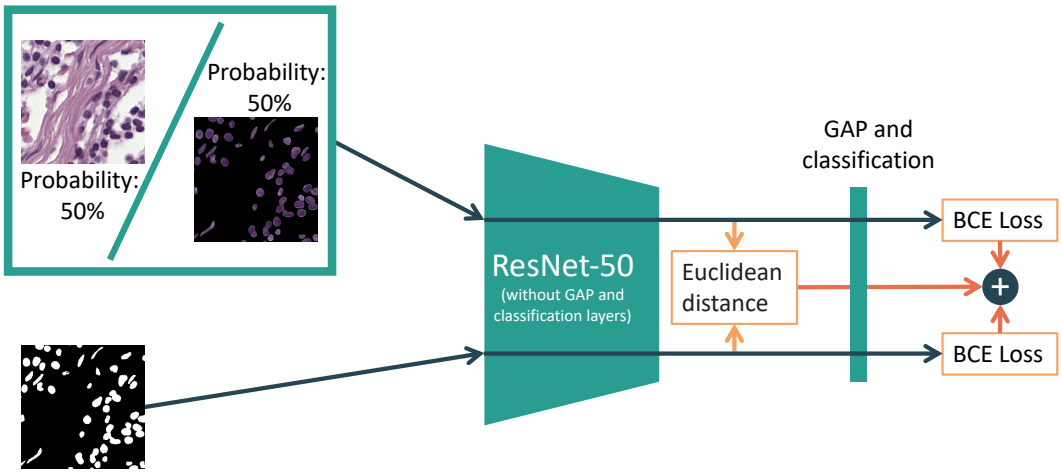


Figure 1: We pass the input image (or, with 0.5 probability, input image multiplied with its nuclear segmentation mask) and its nuclear segmentation mask through the network and minimise the Binary Cross-Entropy (BCE) loss for both the input image and its mask. Additionally, we minimise the ℓ_2-distance between the input image's embedding vector and the mask's embedding vector just before the Global Average Pooling (GAP) layer. The embedding vector is ResNet-50's penultimate layer's feature map, i.e., stage 4's last feature map.

Our approach is illustrated in Figure 1. We employ a ResNet-50 [49] from Torchvision [50] as the base model. We next discuss some details of our approach.

ℓ_2-regularisation: To encourage the network to focus on nuclei, we minimise the distance between the feature map of the original image and that of its nuclear segmentation mask. We use the flattened feature map from ResNet-50's penultimate layer, just before Global Average Pooling, to obtain the embeddings. The regularisation term consists of the ℓ_2-distance between the embeddings of the original image z and its mask z', which is added to the BCE losses for both the image and the mask. Let $\hat{y}$ be the model's prediction for x, $\hat{y}'$ for x', and y be the ground truth. Then, the total loss L is:

$$L = \lambda \|z - z'\|_2^2 + BCE(y, \hat{y}) + BCE(y, \hat{y}') \tag{1}$$

where BCE is the Binary Cross-Entropy loss function for labels in $\{0, 1\}$:

$$BCE(y, p) = -(y \log(p) + (1 - y) \log(1 - p)) \tag{2}$$

Original image times mask: Since the embeddings of the original image and its binary segmentation mask may differ significantly, minimizing their ℓ_2-distance can be challenging for the model. To address this, with a probability of 0.5, we multiply the original image by its segmentation mask, i.e., the network receives $x * x'$ as input half the time instead of x. By multiplying with the segmentation mask, everything in the original image except the nuclei is set to 0. Figure 1 shows what the output looks like. By simplifying the task, the network can more easily reduce the distance between the embeddings of the nuclei-only image and the mask and gradually improve alignment between the embeddings of the original image and the mask. We found this augmentation to help stabilise training.

4 Experiments

4.1 Datasets

CAMELYON17 [51] dataset consists of 1000 H&E-stained Whole Slide Images (WSIs) of breast cancer metastases in lymph node sections from five medical centres in the Netherlands. It contains pixel-level annotations of tumours for 10 WSIs from each medical centre, giving us 50 WSIs to work with. WSIs from centres 0, 3 and 4 were scanned using the same scanner, while the other two centres used a different scanner each. All slides were scanned at $40\times$ resolution. We treat each centre as a different domain.

BCSS [52] dataset consists of 151 H&E-stained WSIs of histologically-confirmed primary breast cancer cases from The Cancer Genome Atlas (TCGA) with triple-negative status determined from

clinical data files. All WSIs have a resolution of $40\times$. The WSIs were annotated at the pixel level using crowdsourcing. Each pixel can have one of the many labels. We consider the label "tumor" to define pixels with a tumour and all other labels except "outside_roi", "exclude", or "undetermined" to define pixels without a tumour.

Ocelot [53] dataset consists of pixel-level annotations of tumour vs non-tumour pixels for 303 WSIs from TCGA. It consists of WSIs of primary tumour from six different organs: Bladder, Endometrium, Head-and-neck, Kidney, Prostate, and Stomach. The annotations in the dataset are at a low resolution, so we upscale the annotations to $40\times$. We exclude two WSIs that have only $20\times$ resolution; all other WSIs have $40\times$ resolution.

4.2 Dataset preparation

We use code from WILDS [54, 55] to prepare the CAMELYON17 dataset with modifications. Tiles are sized (270×270) at $40\times$ resolution. For each domain (medical centre), data is split by patient, ensuring all tiles from a patient are in a single subset. Since the number of tiles varies drastically across patients, we shuffle patients so that the validation subset contains 20%-25% of all tiles per domain. Our processed version of CAMELYON17 is available at [56].

4.3 Experiment setup

We train models using the CAMELYON17 dataset, treating each medical centre as a distinct domain. We use the BCSS and Ocelot datasets as external test datasets. To avoid multiple comparisons and overly optimistic performance estimates, we use the external test datasets only once during the entire project, solely to evaluate the final models [57].

For each combination of medical centre and method, we train ten models using the train subset of that centre. Thus, we train 50 models in total for each method. We use the loss on the validation dataset (of the training domain) to select the best model for each training. All models are trained for 50 epochs using the Adam optimiser with a learning rate $4\mathrm{e}{-5}$ and a weight decay of $1\mathrm{e}{-4}$. We use exponential learning rate decay with a decay rate of 0.955. For our method, we set the parameter λ in equation (1) to $\lambda = 0$ for the first five epochs, effectively training without ℓ_2-distance loss in these epochs, and then use $\lambda = 1$ for the rest of the training. We start saving models for selection of the one with lowest validation loss after ten epochs for our method to allow the network to stabilise while from the first epoch for other methods. For all experiments, unless stated otherwise, we use a ResNet-50 [49] model pre-trained on ImageNet [58]. We use HoVer-Net [59] trained on the CoNSeP [59] dataset to generate nuclear segmentation masks.

While domain generalisation encompasses a wide variety of methods, we have selected several exemplary baselines for comparison: Macenko normalisation [1], RSC [32], L2D [37], RandStainNA (RandSNA in result tables) [24], and DDCA [30]. We also include a baseline where we initialise ResNet-50 with pre-trained weights from HoVer-Net [59]. These methods represent different approaches, including stain normalisation (Macenko, RandStainNA) and generating augmented images (L2D). By selecting these diverse techniques, we ensure a comprehensive evaluation of our method's performance across various S-DG strategies.

It is important to note that our method can be integrated with many existing S-DG approaches, making it a flexible plug-in solution rather than a direct competitor. We evaluate most methods with and without the photometric augmentations selected for ERM. After testing various augmentation strategies available in Torchvision [50], we identified the most effective combination to be: ColorJitter(brightness=[0.5, 1.5], contrast=[0.5, 1.5], saturation=[0.5, 1.5], hue=[-0.3, 0.3]) and GaussianBlur(kernel_size=3). Results using these augmentations are marked as '-Aug' in the results tables. In all experiments, including those without photometric augmentations, we apply the basic geometric augmentations: random horizontal and vertical flips. For all ViT-Tiny [60] experiments, we also add affine augmentations: random rotation (up to 90°) and translation (up to 45 pixels).

We ran the experiments on two clusters with GPUs with 64 GB (AMD MI250X) and 24 GB (Nvidia RTX 3090) GPU RAM each. Each job consumed about 21 to 31 GB of GPU RAM. The proposed method took 5 to 20 hours to train, depending on the train data size while ERM took 2.5 to 11 hours.

We report tile-level accuracy for tumour vs non-tumour tile classification for all datasets. Additionally, we measure robustness to image noise by measuring the accuracy drop on CAMELYON17 for image

Table 1: Out-of-domain accuracy on CAMELYON17. The column name indicates the centre used to train models. The best accuracy for each column is in **bold face** and the second best in *italics*. Method "Ours-no-ℓ_2-A" is shorthand for "Ours-no-ℓ_2-Aug" and refers to our approach without ℓ_2-regularisation. Method "Ours-MO-Aug" refers to our approach with masks only, that is, neither using ℓ_2-regularisation nor using mask-times-input augmentation of H&E images with 50% probability during training. A paired t-test for "L2D-Aug" versus "Ours-Aug" yields a p-value of $2 \cdot 10^{-5}$.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
ERM	72.8 ± 2.3	65.9 ± 3.7	64.1 ± 3.0	55.0 ± 1.3	53.8 ± 3.0	62.4 ± 2.7
Macenko	79.3 ± 2.1	62.4 ± 1.4	73.3 ± 5.0	65.8 ± 2.3	85.9 ± 4.2	73.3 ± 3.0
HoVerNet	72.5 ± 2.4	71.0 ± 2.5	61.3 ± 3.9	55.1 ± 1.9	49.6 ± 8.4	61.9 ± 3.8
RandSNA	75.7 ± 3.1	70.9 ± 4.9	62.4 ± 2.5	57.2 ± 2.8	51.8 ± 2.6	63.6 ± 3.2
RSC	77.1 ± 3.2	64.5 ± 3.1	61.9 ± 3.8	56.8 ± 2.3	51.1 ± 2.2	62.3 ± 2.9
L2D	*93.6 ± 1.0*	72.9 ± 2.5	64.4 ± 13.0	73.6 ± 4.3	84.4 ± 3.7	77.8 ± 4.9
Ours	90.4 ± 1.5	$\mathbf{92.5 \pm 0.3}$	90.1 ± 1.3	82.1 ± 2.7	90.8 ± 1.0	*89.2 ± 1.3*
ERM-Aug	93.1 ± 1.0	78.9 ± 2.1	89.3 ± 2.8	74.8 ± 1.5	91.3 ± 1.6	85.5 ± 1.8
Macenko-Aug	86.3 ± 1.9	78.7 ± 1.5	86.2 ± 4.4	70.0 ± 2.8	90.8 ± 1.2	82.4 ± 2.3
HoVerNet-Aug	93.0 ± 0.6	80.8 ± 2.8	*91.3 ± 1.2*	82.2 ± 1.6	89.6 ± 2.2	87.4 ± 1.7
RandSNA-Aug	92.7 ± 1.1	83.1 ± 2.1	91.0 ± 2.0	78.9 ± 3.0	91.1 ± 1.5	87.4 ± 1.9
DDCA-Aug	92.5 ± 2.4	79.4 ± 1.9	89.4 ± 2.9	78.2 ± 3.1	90.2 ± 2.1	86.0 ± 2.5
RSC-Aug	93.1 ± 0.8	78.2 ± 2.0	89.3 ± 3.4	77.9 ± 2.2	91.0 ± 1.7	85.9 ± 2.0
L2D-Aug	$\mathbf{94.3 \pm 0.1}$	87.6 ± 0.6	87.7 ± 1.4	*83.4 ± 2.6*	$\mathbf{92.3 \pm 0.9}$	89.1 ± 1.1
Ours-Aug	91.8 ± 0.7	*92.2 ± 1.6*	$\mathbf{92.9 \pm 0.7}$	$\mathbf{90.4 \pm 1.1}$	*91.7 ± 0.5*	$\mathbf{91.8 \pm 0.9}$
Ours-no-ℓ_2-A	92.1 ± 0.8	81.4 ± 2.9	91.8 ± 2.0	83.3 ± 1.0	90.5 ± 1.7	87.8 ± 1.7
Ours-MO-Aug	91.8 ± 2.2	88.2 ± 2.1	85.0 ± 2.2	79.7 ± 4.2	77.9 ± 2.8	84.5 ± 2.7

corruptions introduced in [61]. This includes Gaussian-, shot-, impulse- and snow-noise, and two blur types, elastic transform and JPEG compression.

Results on CAMELYON17 (lymph node sections) We test models on their respective out-of-domain data. E.g., a model trained on Centre-3 is tested on all the data from Centre-0,1,2,4. Our method attains 10% higher accuracy than the next best method (L2D) when none used photometric augmentations and was also superior when photometric augmentations were used (Table 1).

Results on BCSS (primary breast cancer) The accuracy of the models trained on a centre in CAMELYON17 drops substantially (12% to 14%) for all the methods when tested on BCSS (Table 2) compared to when tested on other centres in CAMELYON17 (Table 1). This could be due to a mismatch between pathologists' annotations on CAMELYON17 and BCSS but also due to the biological differences between these tissue types. In particular, epithelial cells in lymph nodes would almost certainly be tumour cells, while they could be benign cells in ordinary breast tissue. These results show that the relative performance on CAMELYON17 for different methods is indicative of relative performance on an external test set, as the performance drop is similar for all methods.

Results on Ocelot (primary non-breast cancer) We test our model on the Ocelot dataset to evaluate if our method helps to train models that generalise to other organs as well. Ocelot does not have any data from breast tissue nor does it include lymph node sections (which all the models have been trained on). We report the results of these experiments in Table 3. While our method achieves the highest accuracy also in this case, the difference between our method and L2D is not as big as it is for CAMELYON17 and BCSS. Taking a closer look into the performance for separate organs (Tables 4, 5, 6, 7, 8, and 9 in the Supplement), we can see that our method performs worse than L2D in Endometrium and Kidney, where accuracies are generally lower, and better in the four other organs. This indicates that models trained with our method generalise worse to organs where the transferability from breast tissue is generally low. This is at least the case for Kidney which has by far the lowest accuracies across all methods. Generalising to different cancer types is an emerging experimental topic; see, for example, [62].

In summary, our method yields better accuracy than the baselines, including other S-DG approaches.

Table 2: Out-of-domain accuracy on BCSS. The column name indicates the centre used to train models. The best accuracy for each column is in **bold face** and the second best in *italics*. A paired t-test for "L2D-Aug" versus "Ours-Aug" yields a p-value of $4 \cdot 10^{-5}$.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
ERM	58.0 ± 2.9	69.5 ± 2.8	52.0 ± 1.0	50.0 ± 0.1	52.1 ± 3.9	56.3 ± 2.1
Macenko	68.7 ± 2.4	56.5 ± 1.9	65.5 ± 2.6	56.8 ± 2.3	71.1 ± 3.4	63.7 ± 2.5
HoVerNet	55.7 ± 2.0	65.3 ± 2.2	53.8 ± 0.8	49.8 ± 1.9	47.5 ± 2.9	54.4 ± 2.0
RandSNA	60.5 ± 2.8	67.4 ± 3.8	51.0 ± 0.8	50.1 ± 0.7	50.2 ± 1.6	55.8 ± 1.9
RSC	63.3 ± 3.7	65.7 ± 2.4	50.8 ± 1.0	50.1 ± 0.1	50.2 ± 0.5	56.0 ± 1.5
L2D	79.9 ± 1.2	65.2 ± 0.7	67.4 ± 2.3	63.2 ± 4.5	66.2 ± 2.2	68.4 ± 2.2
Ours	74.2 ± 3.6	*78.3 ± 2.2*	73.4 ± 1.9	63.8 ± 2.8	71.8 ± 2.3	72.3 ± 2.6
ERM-Aug	80.1 ± 1.6	70.7 ± 2.8	73.7 ± 2.6	60.9 ± 1.8	73.3 ± 2.6	71.7 ± 2.3
Macenko-Aug	75.8 ± 2.9	67.6 ± 2.4	72.6 ± 2.6	57.8 ± 2.9	*75.3 ± 1.8*	69.8 ± 2.5
HoVerNet-Aug	79.8 ± 1.4	65.1 ± 1.6	71.2 ± 2.2	64.0 ± 2.2	69.2 ± 6.1	69.9 ± 2.7
RandSNA-Aug	78.5 ± 2.7	73.2 ± 3.1	72.8 ± 3.2	64.5 ± 3.4	75.1 ± 2.5	72.8 ± 3.0
DDCA-Aug	79.1 ± 1.9	71.7 ± 3.5	70.1 ± 2.8	61.4 ± 3.2	71.4 ± 7.3	70.7 ± 3.8
RSC-Aug	79.6 ± 1.5	72.1 ± 2.9	71.6 ± 1.7	63.2 ± 1.6	74.1 ± 4.3	72.1 ± 2.4
L2D-Aug	*81.9 ± 0.3*	74.7 ± 0.6	*74.2 ± 0.6*	*67.5 ± 2.9*	$\mathbf{77.2 \pm 2.2}$	*75.1 ± 1.3*
Ours-Aug	$\mathbf{82.3 \pm 0.8}$	$\mathbf{81.9 \pm 2.3}$	$\mathbf{75.7 \pm 2.2}$	$\mathbf{79.6 \pm 1.0}$	74.8 ± 1.9	$\mathbf{78.8 \pm 1.6}$

Table 3: Out-of-domain accuracy on Ocelot. The column name indicates the centre used to train models. The best accuracy for each column is in **bold face** and the second best in *italics*. A paired t-test for "L2D-Aug" versus "Ours-Aug" yields a p-value of 0.044.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
ERM	65.7 ± 2.4	55.8 ± 1.7	51.7 ± 1.3	45.6 ± 1.4	53.5 ± 5.0	54.5 ± 2.4
Macenko	64.3 ± 1.9	54.4 ± 0.9	63.8 ± 1.4	54.8 ± 2.0	65.3 ± 3.7	60.5 ± 2.0
HoVerNet	62.0 ± 1.2	53.7 ± 2.2	52.1 ± 0.7	48.0 ± 1.9	54.3 ± 2.9	54.0 ± 1.8
RandSNA	67.0 ± 2.5	56.1 ± 1.9	51.0 ± 0.3	46.3 ± 1.9	51.2 ± 2.4	54.3 ± 1.8
RSC	68.1 ± 2.4	55.6 ± 1.0	50.9 ± 0.6	47.2 ± 1.5	50.3 ± 0.8	54.4 ± 1.3
L2D	68.2 ± 1.3	57.3 ± 0.5	56.4 ± 2.4	55.9 ± 3.3	60.1 ± 4.0	59.6 ± 2.3
Ours	67.9 ± 1.4	*70.7 ± 1.0*	66.7 ± 1.2	62.1 ± 1.8	69.2 ± 0.9	67.3 ± 1.3
ERM-Aug	*74.0 ± 1.4*	62.8 ± 1.9	67.6 ± 2.5	56.0 ± 1.6	67.6 ± 3.1	65.6 ± 2.1
Macenko-Aug	68.8 ± 2.0	60.9 ± 1.1	$\mathbf{70.3 \pm 1.5}$	57.6 ± 3.2	*72.5 ± 1.7*	66.0 ± 1.9
HoVerNet-Aug	70.7 ± 1.0	54.2 ± 1.9	*69.4 ± 2.3*	61.2 ± 2.4	69.2 ± 2.9	65.0 ± 2.1
RandSNA-Aug	70.8 ± 3.2	66.8 ± 2.5	68.4 ± 2.5	61.3 ± 2.5	71.7 ± 2.7	67.8 ± 2.7
DDCA-Aug	73.1 ± 2.4	64.9 ± 2.5	68.9 ± 2.7	57.4 ± 2.9	66.9 ± 4.5	66.2 ± 3.0
RSC-Aug	71.9 ± 2.2	61.7 ± 2.4	69.2 ± 2.9	58.4 ± 1.4	70.1 ± 3.6	66.3 ± 2.5
L2D-Aug	$\mathbf{74.7 \pm 0.6}$	68.3 ± 0.5	65.6 ± 0.7	*62.5 ± 2.7*	$\mathbf{74.4 \pm 1.2}$	*69.1 ± 1.1*
Ours-Aug	70.8 ± 0.5	$\mathbf{72.0 \pm 1.3}$	68.9 ± 1.3	$\mathbf{70.4 \pm 0.7}$	70.7 ± 1.3	$\mathbf{70.6 \pm 1.0}$

5 Ablation Study and Discussion

Impact of data augmentation Tables 1, 2, and 3 shows that data augmentation benefits all methods substantially, which is consistent with well-established knowledge.

Impact of ℓ_2-regularisation The result labelled *Ours-no-ℓ_2-A* in Table 1 shows that using data augmentation with nuclear masks alone is insufficient to achieve high accuracy. Without ℓ_2-regularisation (i.e., setting $\lambda = 0$ in Equation (1)), our method only slightly outperforms most baselines that also uses data augmentation. The key factor for effective cross-domain generalisation is the ability to align the feature representation of input images with corresponding mask images, which lack colour and texture. Further evidence supporting this alignment effect is presented in the next paragraph.

Impact of mask-times-input-augmentation The result *Ours-MO-Aug* in Table 1 demonstrates an ablation with two changes: the absence of ℓ_2-regulation and the removal of the 50% probability

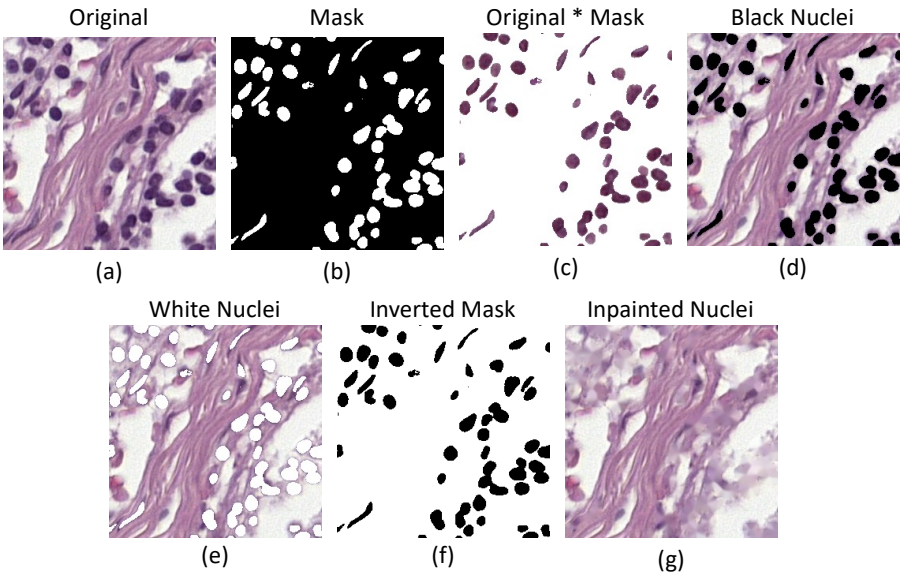


Figure 2: Exemplary image ablations used in this study.

mask-times-input augmentation of H&E images during training (Figure 1), but still having two CE loss terms, one of them over nuclear masks. We see a further decline of accuracies below most baselines with photometric augmentations.

Impact of learned features without nuclear-mask-like features Here, we bring further evidence for the effect of ℓ_2-regularisation about pulling the features towards the representation of nuclear masks. Note that nuclear masks are not used during inference in standard evaluations such as all those in Tables 1, 2, and 3. In Table 10 in the Supplement, we can see results for predictions in which the embeddings (features before the GAP layer) of the H&E images are modified by subtracting the embeddings of the corresponding nuclear masks. By comparing to Table 1, we see a drop in performance for all methods. However, the drop is largest for our method, with an accuracy below random guessing. This shows that the features computed from H&E-stained images at test time are indeed more similar to features from nuclear masks for our method than for other methods.

Impact of removal of intranuclear texture and colour In Tables 11 and 14 in the Supplement, we consider the performance on modified H&E images, in which intranuclear texture is removed by masking it out with a constant colour (see examples in Figure 2d and 2e). This is of interest due to the observation that intranuclear texture is often different in cancerous nuclei, which can be informative to humans. We can see that if it is replaced by a colour similar to the colour of nuclei, we for our method obtain a performance (Table 11) very similar to the performance with original H&E images (Table 1). On the other hand, changing the colour to white seems to reduce the performance notably (Table 14). This is possibly due to the creation of images with outlier statistics. A more likely explanation is that it is common for H&E stains to have small holes or gaps of white background colour in the stroma, which usually are not discriminative information but rather shear stress artefacts from the tissue cutting process. Therefore, masking nuclei with white masks may effectively remove discriminative information about nuclei. This domain-specific observation may explain the asymmetry in behaviour when masking nuclei with black versus white.

Impact of removal of extranuclear information Table 23 in the Supplement shows results on data where all the non-nuclear background is set to white (see example in Figure 2c). These images can be viewed as an inside-out inverted case of the images evaluated to give the results in Table 11. The common information in both sets of images is the morphology and organisation of nuclei. The performance for our proposed method remains high on these images (Table 23), being close to the best result on original H&E data (Table 1). The experiments in Tables 23 and 11 demonstrate the strong generalisability of focusing on nuclear morphology and organisation in out-of-domain settings.

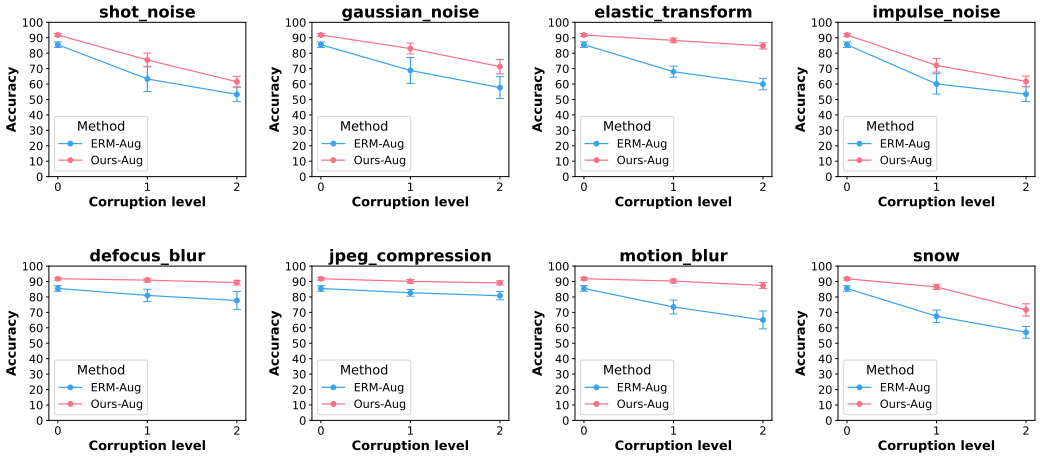


Figure 3: Robustness to added noise described in [61].

Impact of dilution of nuclear shapes We expand nuclei masks by a classic morphological dilation and then blacken the dilated regions in the H&E images. The nuclear shape information in these images is thus progressively reduced compared to the images where only the nuclei are filled with black. Across all methods, we observe a drop in accuracy with an increase in dilation (Tables 11, 12, and 13 in the Supplement), highlighting the critical role of shape in this domain. The proposed method is more robust to moderate shape dilution with a mask size of 5 than the baseline methods. A similar but stronger trend appears in Tables 14, 15, and 16 in the Supplement for whitened nuclei.

Impact of removing nuclei We dilate the nuclear mask image with a kernel size of 5 to encapsulate remnants of the boundary of nuclei and then use the dilated mask to remove nuclei by inpainting [63]. The accuracy with the resulting images (see example in Figure 2g) drops to random guessing for our method (Tables 17 and 18 in the Supplement). Essentially no tiles are classified as tumour, giving a nearly zero recall and low precision (Tables 19, 20, 21, and 22 in the Supplement). Also, this supports that models trained using our method focus on nuclear morphology and organisation, and shows that the models reasonably associate the absence of nuclei with no tumour.

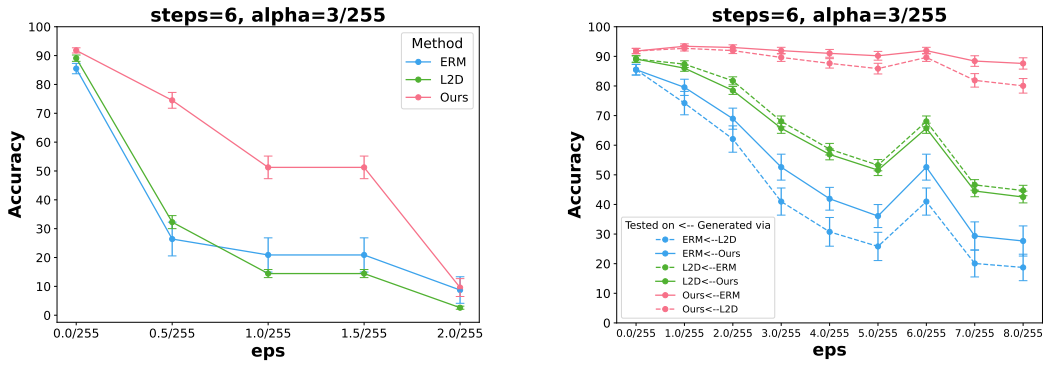
Saliency maps via Integrated Gradients To further demonstrate that our method steers models to focus on nuclei, we generate saliency (pixel attribution) maps using Integrated Gradients [64] and show some randomly selected examples in Figures 6,7 in the Supplement. The saliency maps also indicate that a model trained using our method focuses on nuclei.

Evaluation of L2D and RSC combined with the proposed method Tables 26, 27, and 28 in the Supplement show the results of combining L2D and RSC with the proposed method. Combining the proposed method, which regularises, with L2D, which diversifies, yields mixed results, likely due to the opposing effects of these two interventions. Combining it with RSC results in a small gain over using our method alone. Overall, this demonstrates the effectiveness of the method proposed.

Evaluation on segmentation mask data For the sake of completeness, we show in Tables 24 and 25 in the Supplement that our method also performs well when tested on nuclear masks (as exemplified in Figure 2b) and their inversions (see example in Figure 2f).

Evaluation of robustness to image corruptions Figure 3 shows that the proposed method has notably higher robustness to image corruptions for most experiments in eight types of corruptions described in [61]. Samples of corrupted images are shown in Figure 5 in the Supplement.

Evaluation of robustness against adversarial attacks We evaluated the robustness of models against adversarial attacks [65] using the Projected Gradient Descent (PGD) attack [66]. Figure 4a demonstrates that models trained using our method have significantly higher robustness than ERM and L2D, the latter being the second-best performing method in Tables 1, 2, and 3. Additionally, we



(a) (b)

Figure 4: **(a)** PGD attack on models. **(b)** Cross-model PGD attacks where adversarial images are generated using a model from a method but the accuracy for those images is tested on models from other methods. Results are for the validation subset of each centre in CAMELYON17.

conduct cross-model attacks by generating adversarial images using models trained with one method and evaluating them on models trained with other methods. The results indicate that our models exhibit minimal performance degradation when exposed to adversarial images generated by models from other methods (Figure 4b). In contrast, the accuracy of models trained with ERM and L2D drops substantially. This further demonstrates the superior robustness of models from our method.

A preliminary evaluation on a transformer architecture We perform a comparison using a fine-tuned ViT-Tiny [60] model. The results are shown in Tables 29, 30 and 31 in the Supplement. These results show that our approach obtains superior out-of-domain performance for the CAMELYON17 dataset. The results are more mixed for the other datasets. In particular, it seems that models trained on one of the five centres (Center-4) in CAMELYON17 do not generalise well to other cancer types and are actually also performing sub-optimally in CAMELYON17. For models trained on each of the other four centres in CAMELYON17, the performance with our approach is, on average, better than with other approaches, but the performance increase is lower than for ResNet-50. However, in the same tissue type (CAMELYON17 data), the performance gain is similar for both ViT-Tiny and ResNet-50. Our interpretation of all these results is that our approach can improve out-of-domain performance also for ViT-Tiny, in particular across centres and scanners for the same tissue type, but that it might also fail for a minority of the training datasets. This experiment is preliminary because we took the same hyperparameters as used for ResNet-50, including the same learning rate and $\lambda = 1$, both of which might not be optimal. Also, we note that ViT-Tiny has much fewer parameters than ResNet-50. Experiments with larger transformers might obtain bigger differences, as seen in [67].

Limitations of this study As a limitation, we identify that we have performed these experiments for only one classification task. For medical practitioners, it would be of interest to measure the impact for other tasks, such as tumour grading and survival prediction, when evaluated in an out-of-domain generalisation setup. However, this would require access to multi-centre datasets with relevant labels available. Secondly, we ran the full set of experiments only on one base network, ResNet-50, because we preferred to run a larger set of ablation experiments to understand what actually has been learned when using our method. While we expect results to be qualitatively similar for other CNNs, transformer networks might have different learning dynamics, and results for those with a larger capacity than the ViT-Tiny are of interest in future work. Finally, an extension to other cancer types, such as prostate or colon cancer, would also be of interest.

6 Conclusion

We have shown a simple method to enforce the learning of shape features at training time, which uses unmodified input images at inference time. It shows very good out-of-domain performance and can be combined as a plugin with other methods to enhance out-of-domain generalisation. Aside from out-of-domain accuracy, the proposed method gives improved robustness to image alterations.

Acknowledgments and Disclosure of Funding

The computations were performed on resources provided by Sigma2—the National Infrastructure for High-Performance Computing and Data Storage in Norway—through Project NN8104K. Additionally, we acknowledge Sigma2 for access to the LUMI supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CSC (Finland) and the LUMI consortium through Sigma2, Norway, Project 465000262. This work was supported by the authors' institutions and a grant from the Research Council of Norway (project number 309439). This research / project is supported by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Infocomm Media Development Authority.

References

[1] Marc Macenko, Marc Niethammer, James S Marron, David Borland, John T Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E Thomas. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE international symposium on biomedical imaging: from nano to macro*, pages 1107–1110. IEEE, 2009.

[2] David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen Van Der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical image analysis*, 58:101544, 2019.

[3] Mostafa Jahanifar, Manahil Raza, Kesi Xu, Trinh Vuong, Rob Jewsbury, Adam Shephard, Neda Zamanitajeddin, Jin Tae Kwak, Shan E Ahmed Raza, Fayyaz Minhas, et al. Domain generalization in computational pathology: survey and guidelines. *arXiv preprint arXiv:2310.19656*, 2023.

[4] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2018.

[5] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J Kellman. Deep convolutional networks do not classify based on global object shape. *PLoS computational biology*, 14(12): e1006613, 2018.

[6] Christopher W Elston and Ian O Ellis. Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, 19(5):403–410, 1991.

[7] Daniele Zink, Andrew H Fischer, and Jeffrey A Nickerson. Nuclear structure in cancer cells. *Nature reviews cancer*, 4(9):677–687, 2004.

[8] Edgar G Fischer. Nuclear morphology and the biology of cancer cells. *Acta cytologica*, 64(6): 511–519, 2020.

[9] William H Wolberg, W Nick Street, and Olvi L Mangasarian. Importance of nuclear morphology in breast cancer prognosis. *Clinical Cancer Research*, 5(11):3542–3548, 1999.

[10] Shivang Naik, Scott Doyle, Shannon Agner, Anant Madabhushi, Michael Feldman, and John Tomaszewski. Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. In *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 284–287, 2008. doi: 10.1109/ISBI.2008.4540988.

[11] Jean-Romain Dalle, Wee Kheng Leow, Daniel Racoceanu, Adina Eunice Tutac, and Thomas C Putti. Automatic breast cancer grading of histopathological images. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3052–3055. IEEE, 2008.

[12] Laura E. Boucheron, B. S. Manjunath, and Neal R. Harvey. Use of imperfectly segmented nuclei in the classification of histopathology images of breast cancer. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 666–669, 2010. doi: 10.1109/ICASSP.2010.5495124.

[13] Mitko Veta, Robert Kornegoor, André Huisman, Anoek H J Verschuur-Maes, Max A Viergever, Josien P W Pluim, and Paul J van Diest. Prognostic value of automatically extracted nuclear morphometric features in whole slide images of male breast cancer. *Modern Pathology*, 25(12): 1559–1565, 2012. ISSN 0893-3952. doi: https://doi.org/10.1038/modpathol.2012.126. URL `https://www.sciencedirect.com/science/article/pii/S089339522202943X`.

[14] Paweł Filipczuk, Marek Kowal, and Andrzej Obuchowicz. Automatic breast cancer diagnosis based on k-means clustering and adaptive thresholding hybrid segmentation. In *Image processing and communications challenges 3*, pages 295–302. Springer, 2011.

[15] Hela Masmoudi, Stephen M. Hewitt, Nicholas Petrick, Kyle J. Myers, and Marios A. Gavrielides. Automated quantitative assessment of her-2/neu immunohistochemical expression in breast cancer. *IEEE Transactions on Medical Imaging*, 28(6):916–925, 2009. doi: 10.1109/TMI.2009.2012901.

[16] L Sing Cheong, Angela Jean, Tsu Soo Tan, Waiming Kong, and Soo Yong Tan. Automated segmentation and measurement for cancer classification of her2/neu status in breast carcinomas. In *Biotechno 2011: The Third International Conference on Bioinformatics*. Citeseer, 2011.

[17] Serge Beucher and Christian Lantuéjoul. Use of watersheds in contour detection. workshop published, September 1979. URL `http://cmm.ensmp.fr/~beucher/publi/watershed.pdf`.

[18] Stanley J. Osher and James A. Sethian. Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulations. *Journal of Computational Physics*, 79: 12–49, 1988. URL `https://api.semanticscholar.org/CorpusID:205007680`.

[19] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, Jan 1988. ISSN 1573-1405. doi: 10.1007/BF00133570. URL `https://doi.org/10.1007/BF00133570`.

[20] Akira Hasegawa, Kevin J. Cullen M.D., and Seong Ki Mun. Segmentation and analysis of breast cancer pathological images by an adaptive-sized hybrid neural network. In Murray H. Loew and Kenneth M. Hanson, editors, *Medical Imaging 1996: Image Processing*, volume 2710, pages 752 – 762. International Society for Optics and Photonics, SPIE, 1996. doi: 10.1117/12.237980. URL `https://doi.org/10.1117/12.237980`.

[21] Kyoung-Mi Lee and W.N. Street. An adaptive resource-allocating network for automated detection, segmentation, and classification of breast cancer nuclei topic area: image processing and recognition. *IEEE Transactions on Neural Networks*, 14(3):680–687, 2003. doi: 10.1109/TNN.2003.810615.

[22] Abhishek Vahadane, Tingying Peng, Amit Sethi, Shadi Albarqouni, Lichao Wang, Maximilian Baust, Katja Steiger, Anna Melissa Schlitter, Irene Esposito, and Nassir Navab. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE transactions on medical imaging*, 35(8):1962–1971, 2016.

[23] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001.

[24] Yiqing Shen, Yulin Luo, Dinggang Shen, and Jing Ke. Randstainna: Learning stain-agnostic features from histology slides by bridging stain augmentation and normalization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 212–221. Springer, 2022.

[25] Md Ziaul Hoque, Anja Keskinarkaus, Pia Nyberg, and Tapio Seppänen. Stain normalization methods for histopathology image analysis: A comprehensive review and experimental comparison. *Information Fusion*, page 101997, 2023.

[26] Khrystyna Faryna, Jeroen Van der Laak, and Geert Litjens. Tailoring automated data augmentation to h&e-stained histopathology. In *Medical imaging with deep learning*, 2021.

[27] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[28] David Tellez, Maschenka Balkenhol, Nico Karssemeijer, Geert Litjens, Jeroen van der Laak, and Francesco Ciompi. H and e stain augmentation improves generalization of convolutional networks for histopathological mitosis detection. In *Medical Imaging 2018: Digital Pathology*, volume 10581, pages 264–270. SPIE, 2018.

[29] Joona Pohjonen, Carolin Stürenberg, Atte Föhr, Reija Randen-Brady, Lassi Luomala, Jouni Lohi, Esa Pitkänen, Antti Rannikko, and Tuomas Mirtti. Augment like there's no tomorrow: Consistently performing neural networks for medical imaging. *arXiv preprint arXiv:2206.15274*, 2022.

[30] Niccolò Marini, Sebastian Otalora, Marek Wodzinski, Selene Tomassini, Aldo Franco Dragoni, Stephane Marchand-Maillet, Juan Pedro Dominguez Morales, Lourdes Duran-Lopez, Simona Vatrano, Henning Müller, et al. Data-driven color augmentation for h&e stained images in computational pathology. *Journal of Pathology Informatics*, 14:100183, 2023.

[31] Khrystyna Faryna, Jeroen van der Laak, and Geert Litjens. Automatic data augmentation to improve generalization of deep learning in h&e stained histopathology. *Computers in Biology and Medicine*, 170:108018, 2024.

[32] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part II 16*, pages 124–140. Springer, 2020.

[33] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018.

[34] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12556–12565, 2020.

[35] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.

[36] Lei Li, Ke Gao, Juan Cao, Ziyao Huang, Yepeng Weng, Xiaoyue Mi, Zhengze Yu, Xiaoya Li, and Boyang Xia. Progressive domain expansion network for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 224–233, 2021.

[37] Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh. Learning to diversify for single domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 834–843, 2021.

[38] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[39] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[40] M Tarek Shaban, Christoph Baur, Nassir Navab, and Shadi Albarqouni. Staingan: Stain style transfer for digital histological images. In *2019 Ieee 16th international symposium on biomedical imaging (Isbi 2019)*, pages 953–956. IEEE, 2019.

[41] Thomas de Bel, John-Melle Bokhorst, Jeroen van der Laak, and Geert Litjens. Residual cyclegan for robust domain transformation of histopathological tissue slides. *Medical Image Analysis*, 70:102004, 2021.

[42] Niyun Zhou, De Cai, Xiao Han, and Jianhua Yao. Enhanced cycle-consistent generative adversarial network for color normalization of h&e stained images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*, pages 694–702. Springer, 2019.

[43] Harshal Nishar, Nikhil Chavanke, and Nitin Singhal. Histopathological stain transfer using style transfer network with adversarial loss. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, pages 330–340. Springer, 2020.

[44] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[45] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016.

[46] Edwin Yuan and Junkyo Suh. Neural stain normalization and unsupervised classification of cell nuclei in histopathological breast cancer images. *arXiv preprint arXiv:1811.03815*, 2018.

[47] Hyungjoo Cho, Sungbin Lim, Gunho Choi, and Hyunseok Min. Neural stain-style transfer learning using gan for histopathological images. *arXiv preprint arXiv:1710.08543*, 2017.

[48] Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundström. Measuring domain shift for deep learning in histopathology. *IEEE journal of biomedical and health informatics*, 25(2): 325–336, 2020.

[49] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[50] TorchVision maintainers and contributors. Torchvision: Pytorch's computer vision library. `https://github.com/pytorch/vision`, 2016.

[51] 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.

[52] Mohamed Amgad, Habiba Elfandy, Hagar Hussein, Lamees A Atteya, Mai AT Elsebaie, Lamia S Abo Elnasr, Rokia A Sakr, Hazem SE Salem, Ahmed F Ismail, Anas M Saad, et al. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics*, 35(18): 3461–3467, 2019.

[53] Jeongun Ryu, Aaron Valero Puche, JaeWoong Shin, Seonwook Park, Biagio Brattoli, Jinhee Lee, Wonkyung Jung, Soo Ick Cho, Kyunghyun Paeng, Chan-Young Ock, et al. Ocelot: Overlapped cell on tissue dataset for histopathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23902–23912, 2023.

[54] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021.

[55] Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, Sara Beery, Etienne David, Ian Stavness, Wei Guo, Jure Leskovec, Kate Saenko, Tatsunori Hashimoto, Sergey Levine, Chelsea Finn, and Percy Liang. Extending the wilds benchmark for unsupervised adaptation. In *International Conference on Learning Representations (ICLR)*, 2022.

[56] Dhananjay Tomar. Replication Data for: Are nuclear masks all you need for improved out-of-domain generalization? A closer look at cancer classification in histopathology, 2024. URL `https://doi.org/10.18710/NXPLFL`.

[57] Andreas Kleppe, Ole-Johan Skrede, Sepp De Raedt, Knut Liestøl, David J Kerr, and Håvard E Danielsen. Designing deep learning studies in cancer diagnostics. *Nature Reviews Cancer*, 21 (3):199–211, 2021.

[58] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[59] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. HoVer-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical image analysis*, 58:101563, 2019.

[60] 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 *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

[61] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[62] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Siqi Liu, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, Eric Robert, Yi Kan Wang, Jeremy D. Kunz, Matthew C. H. Lee, Jan Bernhard, Ran A. Godrich, Gerard Oakley, Ewan Millar, Matthew Hanna, Juan Retamero, William A. Moye, Razik Yousfi, Christopher Kanan, David Klimstra, Brandon Rothrock, and Thomas J. Fuchs. Virchow: A million-slide digital pathology foundation model, 2024.

[63] Alexandru Telea. An image inpainting technique based on the fast marching method. *Journal of graphics tools*, 9(1):23–34, 2004.

[64] Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *International conference on machine learning*, pages 9269–9278. PMLR, 2020.

[65] C Szegedy. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[66] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[67] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 1204–1213. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01179. URL `https://doi.org/10.1109/CVPR52688.2022.01179`.

Technical Appendix / Supplement

Table 4: Out-of-domain accuracy on Ocelot evaluated only on the organ BLADDER. The column name indicates the centre used to train models. The best accuracy for each column is in **bold face** and the second best in *italics*.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
ERM	65.0 ± 3.1	58.2 ± 3.5	50.4 ± 0.6	46.4 ± 1.8	53.1 ± 6.9	54.6 ± 3.2
Macenko	71.8 ± 1.4	58.7 ± 2.0	66.2 ± 2.1	60.4 ± 3.0	65.6 ± 5.8	64.6 ± 2.9
RSC	70.3 ± 3.0	58.0 ± 2.7	50.2 ± 0.1	47.8 ± 1.4	50.0 ± 0.9	55.3 ± 1.6
L2D	72.0 ± 1.5	59.5 ± 1.4	59.5 ± 2.2	58.2 ± 4.8	58.4 ± 4.1	61.5 ± 2.8
Ours	73.9 ± 2.7	*74.9 ± 1.7*	*74.6 ± 1.8*	*64.3 ± 3.3*	*74.0 ± 1.0*	*72.4 ± 2.1*
ERM-Aug	*76.6 ± 1.9*	68.3 ± 2.8	63.9 ± 3.3	57.0 ± 1.9	63.7 ± 5.0	65.9 ± 3.0
Macenko-Aug	71.9 ± 2.3	65.9 ± 1.1	70.3 ± 2.4	60.0 ± 3.6	$\mathbf{74.5 \pm 2.0}$	68.5 ± 2.3
RSC-Aug	72.7 ± 3.3	67.9 ± 2.9	65.7 ± 3.8	60.7 ± 1.6	66.6 ± 5.1	66.7 ± 3.3
L2D-Aug	75.9 ± 0.7	74.6 ± 0.4	64.0 ± 0.9	63.7 ± 2.5	73.6 ± 2.2	70.4 ± 1.3
Ours-Aug	$\mathbf{77.3 \pm 1.0}$	$\mathbf{78.2 \pm 2.3}$	$\mathbf{75.2 \pm 1.8}$	$\mathbf{75.5 \pm 0.8}$	73.3 ± 1.7	$\mathbf{75.9 \pm 1.5}$

Table 5: Out-of-domain accuracy on Ocelot evaluated only on the organ ENDOMETRIUM. The column name indicates the centre used to train models. The best accuracy for each column is in **bold face** and the second best in *italics*.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
ERM	68.4 ± 2.9	58.7 ± 2.2	52.3 ± 2.1	47.5 ± 0.7	53.9 ± 5.1	56.2 ± 2.6
Macenko	68.9 ± 2.7	57.4 ± 2.1	71.6 ± 2.0	53.7 ± 1.8	68.4 ± 3.7	64.0 ± 2.4
RSC	71.3 ± 2.0	58.1 ± 1.4	50.8 ± 0.8	48.4 ± 1.0	50.5 ± 1.4	55.8 ± 1.3
L2D	76.6 ± 3.0	58.0 ± 0.5	64.5 ± 2.9	57.4 ± 4.7	62.1 ± 4.2	63.7 ± 3.1
Ours	70.6 ± 1.3	$\mathbf{72.6 \pm 1.4}$	63.9 ± 2.2	62.6 ± 1.4	66.9 ± 1.5	67.3 ± 1.6
ERM-Aug	*78.5 ± 1.6*	64.1 ± 2.6	75.1 ± 2.8	56.5 ± 2.6	73.0 ± 3.9	69.4 ± 2.7
Macenko-Aug	73.2 ± 2.8	64.1 ± 2.3	$\mathbf{78.7 \pm 2.0}$	58.5 ± 4.1	74.5 ± 2.5	69.8 ± 2.7
RSC-Aug	76.2 ± 2.7	64.8 ± 3.8	*75.8 ± 2.6*	57.4 ± 1.1	*74.7 ± 5.6*	69.8 ± 3.2
L2D-Aug	$\mathbf{78.5 \pm 0.6}$	*70.6 ± 0.7*	72.0 ± 0.8	*65.1 ± 3.5*	$\mathbf{79.9 \pm 1.7}$	$\mathbf{73.2 \pm 1.5}$
Ours-Aug	73.5 ± 1.8	69.5 ± 3.0	68.8 ± 2.4	$\mathbf{71.4 \pm 1.3}$	69.9 ± 1.9	*70.6 ± 2.1*

Table 6: Out-of-domain accuracy on Ocelot evaluated only on the organ HEAD AND NECK. The column name indicates the centre used to train models. The best accuracy for each column is in **bold face** and the second best in *italics*.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
ERM	61.9 ± 4.0	53.7 ± 1.6	49.5 ± 1.8	44.1 ± 1.5	52.6 ± 2.9	52.4 ± 2.4
Macenko	57.7 ± 2.1	54.1 ± 1.3	51.8 ± 1.2	56.0 ± 2.9	61.2 ± 3.0	56.2 ± 2.1
RSC	63.0 ± 4.8	55.1 ± 1.5	50.1 ± 0.5	44.9 ± 2.1	50.6 ± 1.2	52.7 ± 2.0
L2D	63.1 ± 2.7	59.3 ± 1.2	47.7 ± 1.4	57.4 ± 1.7	60.0 ± 6.3	57.5 ± 2.7
Ours	66.2 ± 3.1	*74.1 ± 1.3*	*64.6 ± 1.0*	59.6 ± 1.4	*69.0 ± 1.0*	66.7 ± 1.5
ERM-Aug	*72.0 ± 3.8*	67.4 ± 2.4	60.6 ± 2.1	58.9 ± 1.8	63.2 ± 3.9	64.4 ± 2.8
Macenko-Aug	67.2 ± 1.8	63.3 ± 1.9	60.4 ± 3.0	57.6 ± 3.6	68.7 ± 2.7	63.4 ± 2.6
RSC-Aug	68.9 ± 2.1	64.3 ± 2.4	61.9 ± 3.7	61.2 ± 2.6	64.7 ± 3.9	64.2 ± 2.9
L2D-Aug	$\mathbf{73.7 \pm 1.7}$	72.7 ± 0.5	60.2 ± 1.1	*64.4 ± 1.9*	68.6 ± 2.3	*67.9 ± 1.5*
Ours-Aug	71.8 ± 1.2	$\mathbf{74.9 \pm 1.8}$	$\mathbf{68.8 \pm 1.3}$	$\mathbf{74.8 \pm 1.0}$	$\mathbf{70.4 \pm 0.9}$	$\mathbf{72.1 \pm 1.3}$

Table 7: Out-of-domain accuracy on Ocelot evaluated only on the organ KIDNEY. The column name indicates the centre used to train models. The best accuracy for each column is in **bold face** and the second best in *italics*.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
ERM	61.5 ± 4.2	50.2 ± 0.4	50.3 ± 0.5	42.5 ± 2.7	54.7 ± 6.2	51.8 ± 2.8
Macenko	58.3 ± 1.9	50.0 ± 0.7	58.0 ± 2.0	52.6 ± 2.1	62.1 ± 4.3	56.2 ± 2.2
RSC	62.1 ± 4.5	50.5 ± 0.7	50.0 ± 0.1	45.4 ± 2.5	50.9 ± 1.4	51.8 ± 1.8
L2D	54.4 ± 1.4	53.2 ± 0.6	47.0 ± 3.4	51.6 ± 2.3	57.2 ± 5.6	52.7 ± 2.7
Ours	54.9 ± 1.1	$\mathit{59.5 \pm 2.0}$	56.5 ± 3.3	53.9 ± 1.5	62.7 ± 2.3	57.5 ± 2.0
ERM-Aug	$\mathit{66.4 \pm 2.1}$	54.7 ± 1.3	61.9 ± 3.0	52.6 ± 2.2	65.4 ± 3.5	60.2 ± 2.4
Macenko-Aug	59.7 ± 2.0	52.5 ± 0.9	$\mathit{63.4 \pm 2.4}$	54.7 ± 2.9	$\mathit{69.0 \pm 1.8}$	59.9 ± 2.0
RSC-Aug	66.0 ± 3.4	51.2 ± 3.1	$\mathbf{65.4 \pm 4.4}$	56.0 ± 2.1	68.7 ± 3.1	$\mathit{61.5 \pm 3.2}$
L2D-Aug	$\mathbf{69.5 \pm 0.6}$	58.6 ± 0.7	59.0 ± 0.9	$\mathit{57.7 \pm 3.4}$	$\mathbf{71.7 \pm 0.6}$	$\mathbf{63.3 \pm 1.2}$
Ours-Aug	58.1 ± 1.1	$\mathbf{65.7 \pm 3.2}$	57.6 ± 2.4	$\mathbf{59.1 \pm 1.1}$	64.0 ± 2.1	60.9 ± 2.0

Table 8: Out-of-domain accuracy on Ocelot evaluated only on the organ PROSTATE. The column name indicates the centre used to train models. The best accuracy for each column is in **bold face** and the second best in *italics*.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
ERM	68.4 ± 1.9	58.7 ± 1.7	57.6 ± 3.0	46.7 ± 1.0	52.1 ± 2.5	56.7 ± 2.0
Macenko	63.2 ± 2.8	51.5 ± 0.6	66.8 ± 2.3	54.0 ± 1.2	68.5 ± 4.3	60.8 ± 2.2
RSC	70.8 ± 2.6	60.9 ± 1.1	54.9 ± 2.4	48.0 ± 1.2	49.9 ± 0.7	56.9 ± 1.6
L2D	73.1 ± 0.8	54.2 ± 0.7	60.9 ± 2.7	55.6 ± 2.2	58.9 ± 2.0	60.5 ± 1.7
Ours	75.4 ± 0.4	$\mathbf{75.1 \pm 1.2}$	$\mathit{75.4 \pm 0.3}$	$\mathit{68.2 \pm 2.9}$	$\mathit{74.7 \pm 0.6}$	$\mathit{73.8 \pm 1.1}$
ERM-Aug	$\mathbf{76.0 \pm 1.1}$	60.7 ± 2.6	74.2 ± 1.4	60.3 ± 1.1	67.8 ± 2.7	67.8 ± 1.8
Macenko-Aug	72.9 ± 1.5	58.6 ± 0.9	74.2 ± 1.0	60.9 ± 2.5	74.4 ± 1.2	68.2 ± 1.4
RSC-Aug	$\mathit{75.7 \pm 1.7}$	60.2 ± 2.3	74.7 ± 1.6	62.0 ± 1.1	70.3 ± 3.1	68.6 ± 2.0
L2D-Aug	75.5 ± 0.5	67.8 ± 0.5	72.1 ± 0.4	64.7 ± 1.9	74.1 ± 1.3	70.9 ± 0.9
Ours-Aug	75.0 ± 0.2	$\mathit{73.7 \pm 2.1}$	$\mathbf{76.3 \pm 0.5}$	$\mathbf{75.9 \pm 0.4}$	$\mathbf{77.0 \pm 0.5}$	$\mathbf{75.6 \pm 0.7}$

Table 9: Out-of-domain accuracy on Ocelot evaluated only on the organ STOMACH. The column name indicates the centre used to train models. The best accuracy for each column is in **bold face** and the second best in *italics*.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
ERM	71.7 ± 2.4	56.9 ± 3.3	50.8 ± 1.6	47.5 ± 0.7	53.0 ± 6.4	56.0 ± 2.9
Macenko	61.6 ± 3.5	54.2 ± 1.9	62.9 ± 3.2	51.4 ± 3.1	65.5 ± 3.6	59.1 ± 3.1
RSC	72.1 ± 1.8	52.1 ± 0.7	50.1 ± 0.6	48.6 ± 0.9	49.6 ± 1.7	54.5 ± 1.1
L2D	74.1 ± 1.4	63.9 ± 1.2	57.9 ± 1.9	57.9 ± 4.5	68.0 ± 3.8	64.4 ± 2.5
Ours	74.4 ± 1.4	$\mathbf{77.2 \pm 0.7}$	$\mathit{73.4 \pm 0.7}$	$\mathit{70.9 \pm 1.6}$	73.9 ± 1.0	$\mathit{74.0 \pm 1.1}$
ERM-Aug	76.6 ± 1.5	68.0 ± 2.0	70.6 ± 2.2	53.8 ± 1.6	72.2 ± 2.4	68.2 ± 2.0
Macenko-Aug	70.9 ± 3.1	65.8 ± 2.7	72.4 ± 1.9	54.3 ± 2.4	73.5 ± 2.0	67.4 ± 2.4
RSC-Aug	73.3 ± 2.8	68.5 ± 2.5	70.6 ± 1.8	55.3 ± 2.0	74.7 ± 2.0	68.5 ± 2.2
L2D-Aug	$\mathit{77.0 \pm 0.5}$	72.0 ± 0.6	67.1 ± 1.0	61.7 ± 2.9	$\mathit{76.4 \pm 0.7}$	70.8 ± 1.1
Ours-Aug	$\mathbf{77.5 \pm 0.7}$	$\mathit{76.5 \pm 3.5}$	$\mathbf{75.9 \pm 1.0}$	$\mathbf{74.6 \pm 0.8}$	$\mathbf{76.4 \pm 0.9}$	$\mathbf{76.2 \pm 1.4}$

Table 10: Out-of-domain accuracy on CAMELYON17 where embeddings for segmentation masks were subtracted from the embeddings of the original image to see if the accuracy drops. The column name indicates the centre used to train models. The best accuracy for each column is in **bold face** and the second best in *italics*.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
ERM	63.0 ± 7.4	64.2 ± 8.3	58.2 ± 3.6	52.3 ± 3.1	52.6 ± 3.1	58.1 ± 5.1
Macenko	63.3 ± 13.2	58.4 ± 11.5	54.5 ± 3.9	59.8 ± 2.8	52.1 ± 5.5	57.6 ± 7.4
RSC	54.6 ± 3.5	52.7 ± 3.1	64.0 ± 7.6	51.5 ± 5.0	54.1 ± 6.9	55.4 ± 5.2
L2D	$\mathbf{86.4 \pm 1.9}$	48.7 ± 5.0	70.1 ± 8.2	$\mathbf{71.5 \pm 4.8}$	$\mathbf{73.4 \pm 7.9}$	$\mathit{70.0 \pm 5.6}$
Ours	36.6 ± 4.6	51.4 ± 6.5	38.3 ± 4.9	37.5 ± 5.2	56.6 ± 8.8	44.1 ± 6.0
ERM-Aug	75.7 ± 13.3	61.2 ± 9.2	72.1 ± 12.5	65.1 ± 10.2	64.8 ± 12.3	67.8 ± 11.5
Macenko-Aug	76.0 ± 11.8	$\mathbf{68.7 \pm 8.0}$	60.1 ± 7.1	62.7 ± 8.5	62.9 ± 10.5	66.1 ± 9.2
RSC-Aug	69.4 ± 9.0	64.4 ± 8.4	$\mathit{82.1 \pm 7.9}$	56.0 ± 5.9	$\mathit{68.5 \pm 11.4}$	68.1 ± 8.5
L2D-Aug	$\mathit{82.2 \pm 2.7}$	$\mathit{67.3 \pm 2.1}$	$\mathbf{82.2 \pm 3.5}$	$\mathit{70.4 \pm 6.9}$	61.7 ± 4.2	$\mathbf{72.8 \pm 3.9}$
Ours-Aug	47.3 ± 6.6	41.3 ± 4.3	45.4 ± 6.5	39.6 ± 3.9	48.4 ± 2.9	44.4 ± 4.8

Table 11: Out-of-domain accuracy on CAMELYON17 where nuclei are blackened out. The column name indicates the centre used to train models. The best accuracy for each column is in **bold face** and the second best in *italics*.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
ERM	61.2 ± 7.3	53.1 ± 2.7	55.7 ± 11.7	50.1 ± 0.1	51.2 ± 3.6	54.3 ± 5.1
Macenko	50.3 ± 0.9	57.5 ± 12.6	50.5 ± 1.3	56.1 ± 7.0	51.7 ± 3.3	53.2 ± 5.0
RSC	62.5 ± 6.5	52.8 ± 2.4	52.0 ± 5.9	50.4 ± 0.8	49.1 ± 3.5	53.4 ± 3.8
L2D	83.7 ± 4.6	89.4 ± 3.4	83.1 ± 7.5	70.6 ± 8.8	63.7 ± 12.1	78.1 ± 7.3
Ours	91.2 ± 1.0	$\mathbf{92.4 \pm 0.3}$	$\mathit{91.6 \pm 1.4}$	$\mathit{89.1 \pm 2.1}$	$\mathit{86.0 \pm 1.6}$	$\mathit{90.1 \pm 1.3}$
ERM-Aug	79.4 ± 8.7	57.6 ± 8.3	85.8 ± 4.4	58.4 ± 8.1	50.4 ± 0.8	66.3 ± 6.1
Macenko-Aug	75.6 ± 11.6	87.9 ± 6.2	64.2 ± 12.4	59.8 ± 8.4	71.5 ± 12.7	71.8 ± 10.3
RSC-Aug	74.5 ± 10.7	63.2 ± 13.6	86.8 ± 7.9	57.3 ± 5.7	50.7 ± 0.8	66.5 ± 7.7
L2D-Aug	$\mathit{91.9 \pm 0.3}$	87.3 ± 0.8	77.0 ± 3.3	87.7 ± 2.6	79.5 ± 7.8	84.7 ± 3.0
Ours-Aug	$\mathbf{92.3 \pm 0.5}$	$\mathit{91.7 \pm 1.4}$	$\mathbf{92.1 \pm 1.2}$	$\mathbf{93.4 \pm 0.3}$	$\mathbf{86.2 \pm 3.2}$	$\mathbf{91.2 \pm 1.3}$

Table 12: Out-of-domain accuracy on CAMELYON17 where nuclei are blackened out after being expanded with filter size 5, i.e., the blackened out part covers nuclei and some region around nuclei. The column name indicates the centre used to train models. The best accuracy for each column is in **bold face** and the second best in *italics*.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
ERM	56.7 ± 6.1	52.4 ± 2.7	55.7 ± 12.2	50.1 ± 0.2	50.0 ± 3.3	53.0 ± 4.9
Macenko	50.3 ± 1.1	54.8 ± 9.3	50.0 ± 0.1	53.7 ± 7.2	50.2 ± 0.4	51.8 ± 3.6
RSC	57.2 ± 5.9	51.1 ± 2.4	50.8 ± 2.3	50.3 ± 0.5	49.7 ± 0.6	51.8 ± 2.4
L2D	78.0 ± 8.4	$\mathbf{86.0 \pm 8.7}$	83.1 ± 8.3	62.4 ± 9.0	62.3 ± 13.0	74.4 ± 9.5
Ours	89.9 ± 0.4	$\mathit{80.3 \pm 3.5}$	$\mathit{90.9 \pm 1.5}$	$\mathit{90.7 \pm 1.0}$	$\mathit{86.3 \pm 2.0}$	$\mathbf{87.6 \pm 1.7}$
ERM-Aug	72.3 ± 11.2	51.6 ± 2.0	78.6 ± 10.0	50.9 ± 2.3	50.1 ± 0.2	60.7 ± 5.1
Macenko-Aug	68.7 ± 12.3	79.7 ± 11.5	57.8 ± 10.1	53.2 ± 7.3	64.7 ± 13.1	64.8 ± 10.9
RSC-Aug	69.5 ± 11.8	59.8 ± 12.4	85.7 ± 7.0	50.2 ± 0.5	50.1 ± 0.2	63.1 ± 6.4
L2D-Aug	$\mathbf{90.3 \pm 0.6}$	78.6 ± 2.3	67.5 ± 6.1	81.7 ± 5.5	78.8 ± 10.1	79.4 ± 4.9
Ours-Aug	$\mathit{89.9 \pm 0.4}$	75.0 ± 2.6	$\mathbf{91.4 \pm 1.1}$	$\mathbf{91.5 \pm 0.4}$	$\mathbf{86.3 \pm 3.6}$	$\mathit{86.8 \pm 1.6}$

Table 13: Out-of-domain accuracy on CAMELYON17 where nuclei are blackened out after being expanded with filter size 9, i.e., the blackened out part covers nuclei and some region around nuclei. The column name indicates the centre used to train models. The best accuracy for each column is in **bold face** and the second best in *italics*.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
ERM	54.3 ± 4.9	51.9 ± 2.5	55.7 ± 12.2	50.1 ± 0.2	51.1 ± 3.2	52.6 ± 4.6
Macenko	50.3 ± 0.9	53.4 ± 5.8	50.1 ± 0.1	54.5 ± 7.7	50.1 ± 0.1	51.7 ± 2.9
RSC	54.0 ± 4.6	50.4 ± 2.3	50.3 ± 1.0	50.2 ± 0.4	49.8 ± 0.4	50.9 ± 1.7
L2D	70.9 ± 10.9	$\mathbf{81.5 \pm 10.6}$	*76.0 ± 12.3*	58.1 ± 8.6	57.8 ± 9.7	68.9 ± 10.4
Ours	*76.6 ± 3.3*	67.0 ± 5.0	66.9 ± 5.0	$\mathbf{88.6 \pm 0.8}$	$\mathbf{83.0 \pm 1.6}$	$\mathbf{76.4 \pm 3.2}$
ERM-Aug	66.6 ± 11.3	50.1 ± 0.3	71.7 ± 11.4	50.2 ± 0.4	50.0 ± 0.2	57.7 ± 4.7
Macenko-Aug	61.9 ± 10.8	71.6 ± 10.6	54.9 ± 8.1	51.6 ± 4.7	58.7 ± 9.2	59.7 ± 8.7
RSC-Aug	65.3 ± 11.8	56.3 ± 9.7	$\mathbf{77.7 \pm 9.5}$	50.1 ± 0.0	50.1 ± 0.1	59.9 ± 6.2
L2D-Aug	$\mathbf{87.1 \pm 0.5}$	*71.8 ± 4.5*	54.3 ± 3.0	71.9 ± 8.3	76.3 ± 11.3	72.3 ± 5.5
Ours-Aug	75.1 ± 1.3	62.1 ± 4.7	69.7 ± 3.9	*85.1 ± 2.2*	*80.4 ± 5.3*	*74.5 ± 3.5*

Table 14: Out-of-domain accuracy on CAMELYON17 where nuclei are whitened out. The column name indicates the centre used to train models. The best accuracy for each column is in **bold face** and the second best in *italics*.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
ERM	61.3 ± 2.8	50.0 ± 1.4	50.0 ± 0.0	50.1 ± 0.1	48.3 ± 2.4	51.9 ± 1.3
Macenko	50.2 ± 0.5	50.0 ± 0.0	50.0 ± 0.1	50.5 ± 0.9	50.0 ± 0.1	50.1 ± 0.3
RSC	65.6 ± 3.8	50.7 ± 1.8	50.0 ± 0.0	50.3 ± 0.3	50.4 ± 1.4	53.4 ± 1.5
L2D	72.9 ± 9.7	71.1 ± 6.1	54.3 ± 8.1	*72.9 ± 4.1*	53.3 ± 4.1	64.9 ± 6.4
Ours	*79.6 ± 6.0*	*87.8 ± 2.1*	$\mathbf{72.1 \pm 6.5}$	$\mathbf{75.9 \pm 5.2}$	*60.7 ± 5.2*	$\mathbf{75.2 \pm 5.0}$
ERM-Aug	51.3 ± 2.8	46.6 ± 6.2	64.4 ± 11.0	49.1 ± 2.6	50.9 ± 2.5	52.5 ± 5.0
Macenko-Aug	49.6 ± 1.2	48.7 ± 4.5	55.7 ± 10.5	50.4 ± 0.5	50.0 ± 0.0	50.9 ± 3.3
RSC-Aug	52.9 ± 4.5	50.2 ± 0.5	52.7 ± 2.8	50.0 ± 0.0	53.5 ± 6.1	51.9 ± 2.8
L2D-Aug	$\mathbf{85.5 \pm 1.1}$	55.8 ± 3.8	50.5 ± 0.2	59.4 ± 7.1	54.5 ± 4.5	61.1 ± 3.3
Ours-Aug	67.4 ± 7.6	$\mathbf{92.1 \pm 0.6}$	*67.3 ± 11.8*	59.0 ± 7.2	$\mathbf{85.5 \pm 4.7}$	*74.3 ± 6.4*

Table 15: Out-of-domain accuracy on CAMELYON17 where nuclei are whitened out after being expanded with filter size 5, i.e., the whitened out part covers nuclei and some region around nuclei. The column name indicates the centre used to train models. The best accuracy for each column is in **bold face** and the second best in *italics*.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
ERM	61.3 ± 3.0	50.7 ± 1.8	50.0 ± 0.0	50.1 ± 0.3	49.6 ± 1.3	52.3 ± 1.3
Macenko	50.5 ± 0.9	50.0 ± 0.0	50.0 ± 0.1	50.3 ± 0.6	50.0 ± 0.1	50.2 ± 0.3
RSC	63.4 ± 4.8	50.6 ± 1.7	50.0 ± 0.0	50.3 ± 0.3	50.3 ± 0.8	52.9 ± 1.5
L2D	*69.0 ± 7.5*	64.2 ± 6.2	53.8 ± 7.0	$\mathbf{70.4 \pm 5.6}$	52.5 ± 8.4	62.0 ± 6.9
Ours	63.9 ± 9.3	*74.6 ± 2.9*	65.5 ± 7.1	*65.6 ± 5.9*	51.4 ± 0.8	*64.2 ± 5.2*
ERM-Aug	50.9 ± 2.6	49.9 ± 0.2	*69.3 ± 9.3*	50.1 ± 0.2	51.2 ± 1.3	54.3 ± 2.7
Macenko-Aug	51.2 ± 1.3	50.2 ± 0.3	61.9 ± 8.5	52.5 ± 2.6	50.4 ± 1.4	53.2 ± 2.8
RSC-Aug	52.6 ± 4.0	50.4 ± 0.8	55.8 ± 6.2	50.1 ± 0.0	54.6 ± 6.6	52.7 ± 3.5
L2D-Aug	$\mathbf{80.4 \pm 3.2}$	51.0 ± 1.3	50.5 ± 0.2	59.4 ± 6.8	*61.4 ± 7.6*	60.5 ± 3.8
Ours-Aug	53.8 ± 3.2	$\mathbf{82.9 \pm 2.9}$	$\mathbf{71.2 \pm 11.4}$	55.8 ± 6.0	$\mathbf{71.2 \pm 9.9}$	$\mathbf{67.0 \pm 6.7}$

Table 16: Out-of-domain accuracy on CAMELYON17 where nuclei are whitened out after being expanded with filter size 9, i.e., the whitened out part covers nuclei and some region around nuclei. The column name indicates the centre used to train models. The best accuracy for each column is in **bold face** and the second best in *italics*.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
ERM	61.8 ± 3.5	50.5 ± 1.6	50.0 ± 0.0	50.1 ± 0.2	49.9 ± 1.2	52.5 ± 1.3
Macenko	50.5 ± 0.9	50.0 ± 0.0	50.0 ± 0.1	50.1 ± 0.3	50.0 ± 0.2	50.1 ± 0.3
RSC	59.6 ± 5.0	50.2 ± 1.5	50.0 ± 0.0	50.3 ± 0.3	50.1 ± 0.5	52.0 ± 1.5
L2D	$\mathit{63.5 \pm 4.9}$	55.7 ± 3.9	53.0 ± 5.2	$\mathbf{63.9 \pm 5.0}$	48.0 ± 9.6	56.8 ± 5.7
Ours	55.9 ± 6.9	$\mathit{67.3 \pm 3.1}$	61.4 ± 6.8	$\mathit{58.1 \pm 4.7}$	50.9 ± 0.7	$\mathit{58.7 \pm 4.4}$
ERM-Aug	51.0 ± 1.6	50.0 ± 0.0	$\mathbf{68.5 \pm 10.5}$	50.2 ± 0.3	47.3 ± 3.2	53.4 ± 3.1
Macenko-Aug	51.3 ± 1.0	50.0 ± 0.1	62.5 ± 9.4	51.8 ± 1.8	47.7 ± 3.0	52.7 ± 3.1
RSC-Aug	51.3 ± 2.0	50.1 ± 0.3	53.8 ± 6.2	50.1 ± 0.0	50.4 ± 4.4	51.1 ± 2.6
L2D-Aug	$\mathbf{68.2 \pm 4.0}$	49.7 ± 0.1	50.2 ± 0.1	58.0 ± 5.7	$\mathit{51.6 \pm 8.1}$	55.5 ± 3.6
Ours-Aug	50.3 ± 0.3	$\mathbf{76.9 \pm 4.3}$	$\mathit{67.6 \pm 9.7}$	53.1 ± 3.9	$\mathbf{58.9 \pm 8.1}$	$\mathbf{61.3 \pm 5.3}$

Table 17: Out-of-domain accuracy on where nuclei are *inpainted* after being expanded with filter size 5. The column name indicates the centre used to train models. The best accuracy for each column is in **bold face** and the second best in *italics*.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
ERM	54.7 ± 3.3	$\mathit{62.4 \pm 3.7}$	50.5 ± 1.9	49.9 ± 0.2	48.6 ± 7.1	53.2 ± 3.2
Macenko	54.3 ± 7.6	54.5 ± 3.4	50.0 ± 0.0	49.9 ± 0.5	49.6 ± 6.2	51.7 ± 3.5
RSC	59.0 ± 2.2	56.7 ± 2.5	49.9 ± 0.1	50.0 ± 0.1	51.9 ± 4.8	53.5 ± 1.9
L2D	$\mathit{68.0 \pm 3.9}$	$\mathbf{66.1 \pm 0.9}$	50.2 ± 0.4	$\mathit{61.7 \pm 3.7}$	$\mathbf{52.9 \pm 5.8}$	$\mathit{59.8 \pm 2.9}$
Ours	50.0 ± 0.0	50.5 ± 2.1	49.2 ± 0.6	50.0 ± 0.0	49.0 ± 0.9	49.7 ± 0.7
ERM-Aug	51.8 ± 3.1	59.1 ± 4.8	$\mathit{57.5 \pm 5.9}$	58.5 ± 3.1	46.1 ± 3.1	54.6 ± 4.0
Macenko-Aug	55.5 ± 7.5	54.8 ± 3.8	$\mathbf{61.9 \pm 4.3}$	55.3 ± 3.9	52.0 ± 8.5	55.9 ± 5.6
RSC-Aug	48.6 ± 3.5	60.0 ± 5.6	56.0 ± 6.2	60.2 ± 2.4	43.2 ± 5.7	53.6 ± 4.7
L2D-Aug	$\mathbf{69.9 \pm 1.4}$	60.0 ± 1.0	52.1 ± 1.5	$\mathbf{67.9 \pm 2.2}$	$\mathit{52.5 \pm 8.8}$	$\mathbf{60.5 \pm 3.0}$
Ours-Aug	50.0 ± 0.0	48.9 ± 1.6	50.3 ± 1.4	50.0 ± 0.0	49.8 ± 0.1	49.8 ± 0.6

Table 18: In domain accuracy on CAMELYON17 where nuclei are *inpainted* after being expanded with filter size 5. The column name indicates the centre used to train models. The best accuracy for each column is in **bold face** and the second best in *italics*.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
ERM	58.1 ± 4.0	$\mathit{63.6 \pm 8.4}$	47.9 ± 0.3	48.8 ± 0.2	45.2 ± 5.8	52.7 ± 3.7
Macenko	53.4 ± 9.3	53.7 ± 0.2	48.2 ± 0.0	48.3 ± 0.5	52.5 ± 4.3	51.2 ± 2.8
RSC	58.2 ± 4.8	59.5 ± 12.4	48.1 ± 0.1	48.5 ± 0.1	49.9 ± 6.4	52.8 ± 4.8
L2D	66.1 ± 8.5	$\mathbf{70.9 \pm 2.5}$	48.1 ± 0.1	53.6 ± 5.2	51.4 ± 11.3	58.0 ± 5.5
Ours	49.6 ± 0.2	53.5 ± 0.7	47.9 ± 0.4	48.3 ± 0.0	45.2 ± 1.0	48.9 ± 0.5
ERM-Aug	56.3 ± 6.8	57.2 ± 3.0	62.6 ± 11.5	82.5 ± 5.4	44.6 ± 3.3	60.6 ± 6.0
Macenko-Aug	$\mathit{66.2 \pm 5.3}$	54.4 ± 1.3	66.4 ± 14.4	59.4 ± 6.8	$\mathit{59.0 \pm 10.2}$	61.1 ± 7.6
RSC-Aug	65.4 ± 8.8	58.3 ± 10.6	$\mathbf{72.7 \pm 8.6}$	$\mathit{88.2 \pm 2.6}$	44.5 ± 3.2	$\mathit{65.8 \pm 6.8}$
L2D-Aug	$\mathbf{72.8 \pm 0.7}$	37.6 ± 2.2	$\mathit{68.2 \pm 3.1}$	$\mathbf{88.9 \pm 1.3}$	$\mathbf{62.5 \pm 6.6}$	$\mathbf{66.0 \pm 2.8}$
Ours-Aug	49.9 ± 0.0	52.3 ± 2.7	48.0 ± 0.1	48.1 ± 0.2	45.0 ± 1.1	48.6 ± 0.8

Table 19: OUT-OF-DOMAIN recall on CAMELYON17 where nuclei are *inpainted* after being expanded with filter size 5. The column name indicates the centre used to train models. The best accuracy for each column is in **bold face** and the second best in *italics*.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
ERM	$\mathbf{91.4 \pm 11.5}$	60.0 ± 8.9	1.7 ± 4.6	0.1 ± 0.1	21.9 ± 20.3	35.0 ± 9.1
Macenko	11.1 ± 21.4	12.4 ± 10.3	0.0 ± 0.0	0.9 ± 2.0	20.7 ± 13.6	9.0 ± 9.5
RSC	$\mathit{89.9 \pm 8.3}$	$\mathbf{67.3 \pm 4.3}$	0.0 ± 0.1	0.3 ± 0.2	21.8 ± 17.2	$\mathit{35.9 \pm 6.0}$
L2D	72.8 ± 19.8	44.3 ± 2.8	1.0 ± 1.6	$\mathit{32.6 \pm 12.9}$	19.9 ± 22.5	34.1 ± 11.9
Ours	0.0 ± 0.0	2.3 ± 7.0	0.2 ± 0.2	0.0 ± 0.0	0.3 ± 0.2	0.6 ± 1.5
ERM-Aug	10.1 ± 11.2	23.8 ± 9.0	19.6 ± 17.5	24.8 ± 9.7	4.8 ± 3.8	16.6 ± 10.2
Macenko-Aug	46.6 ± 21.4	13.4 ± 8.9	$\mathbf{31.4 \pm 19.9}$	11.2 ± 8.3	$\mathbf{44.0 \pm 21.1}$	29.3 ± 15.9
RSC-Aug	23.4 ± 16.5	30.4 ± 14.5	$\mathit{20.6 \pm 12.0}$	22.3 ± 5.8	3.1 ± 2.8	20.0 ± 10.3
L2D-Aug	74.5 ± 3.3	$\mathit{61.7 \pm 4.7}$	12.5 ± 2.6	$\mathbf{43.6 \pm 7.5}$	$\mathit{36.2 \pm 7.8}$	$\mathbf{45.7 \pm 5.2}$
Ours-Aug	0.1 ± 0.0	2.2 ± 2.8	2.0 ± 3.7	0.0 ± 0.0	0.0 ± 0.0	0.8 ± 1.3

Table 20: IN DOMAIN recall on CAMELYON17 where nuclei are *inpainted* after being expanded with filter size 5. The column name indicates the centre used to train models. The best accuracy for each column is in **bold face** and the second best in *italics*.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
ERM	97.1 ± 4.3	35.0 ± 18.5	0.0 ± 0.0	1.0 ± 0.3	50.6 ± 15.3	36.8 ± 7.7
Macenko	8.7 ± 21.7	0.5 ± 0.7	0.0 ± 0.0	0.9 ± 0.2	17.6 ± 11.6	5.5 ± 6.8
RSC	$\mathbf{99.4 \pm 0.5}$	$\mathbf{47.9 \pm 25.3}$	0.0 ± 0.0	0.4 ± 0.2	$\mathit{50.8 \pm 17.3}$	39.7 ± 8.7
L2D	88.5 ± 17.8	43.6 ± 7.3	0.0 ± 0.0	11.3 ± 10.7	38.4 ± 29.2	36.4 ± 13.0
Ours	0.1 ± 0.1	0.4 ± 1.2	0.0 ± 0.0	0.0 ± 0.0	0.1 ± 0.1	0.1 ± 0.3
ERM-Aug	29.7 ± 26.6	9.5 ± 7.3	28.8 ± 23.2	84.3 ± 13.4	5.4 ± 3.8	31.6 ± 14.8
Macenko-Aug	71.9 ± 19.3	4.0 ± 2.6	37.5 ± 31.0	23.7 ± 15.0	$\mathbf{55.1 \pm 24.0}$	38.5 ± 18.4
RSC-Aug	68.9 ± 31.9	17.8 ± 19.8	$\mathbf{51.2 \pm 20.4}$	$\mathit{86.7 \pm 5.7}$	3.5 ± 2.5	$\mathit{45.6 \pm 16.1}$
L2D-Aug	$\mathit{99.0 \pm 0.1}$	$\mathit{46.7 \pm 4.5}$	$\mathit{39.8 \pm 6.5}$	$\mathbf{93.0 \pm 2.0}$	48.7 ± 12.3	$\mathbf{65.4 \pm 5.1}$
Ours-Aug	0.0 ± 0.0	0.3 ± 0.4	0.1 ± 0.1	0.1 ± 0.1	0.3 ± 0.3	0.2 ± 0.2

Table 21: OUT-OF-DOMAIN precision on CAMELYON17 where nuclei are *inpainted* after being expanded with filter size 5. The column name indicates the centre used to train models. The best accuracy for each column is in **bold face** and the second best in *italics*.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
ERM	52.7 ± 2.2	64.4 ± 6.5	39.1 ± 22.0	53.4 ± 20.5	39.2 ± 21.3	49.8 ± 14.5
Macenko	$\mathbf{88.2 \pm 8.1}$	$\mathbf{83.0 \pm 7.6}$	52.8 ± 45.6	42.6 ± 19.3	48.3 ± 16.9	63.0 ± 19.5
RSC	55.7 ± 2.0	55.7 ± 2.4	15.7 ± 11.9	70.6 ± 17.8	50.3 ± 14.6	49.6 ± 9.7
L2D	69.2 ± 9.2	78.8 ± 2.6	60.7 ± 13.1	79.6 ± 5.5	$\mathit{54.2 \pm 15.3}$	$\mathit{68.5 \pm 9.1}$
Ours	23.2 ± 10.2	20.2 ± 18.3	10.2 ± 3.4	53.2 ± 21.3	11.2 ± 2.0	23.6 ± 11.0
ERM-Aug	61.0 ± 13.0	$\mathit{80.5 \pm 13.3}$	$\mathit{85.0 \pm 8.1}$	76.6 ± 6.5	30.8 ± 17.0	66.8 ± 11.6
Macenko-Aug	57.4 ± 11.6	77.5 ± 10.2	$\mathbf{87.8 \pm 9.6}$	$\mathbf{96.1 \pm 2.1}$	50.3 ± 11.2	$\mathbf{73.8 \pm 9.0}$
RSC-Aug	47.3 ± 15.2	75.4 ± 8.3	72.3 ± 19.8	$\mathit{92.1 \pm 2.6}$	14.9 ± 7.6	60.4 ± 10.7
L2D-Aug	68.3 ± 2.0	59.8 ± 1.5	62.1 ± 8.1	85.4 ± 3.4	$\mathbf{55.3 \pm 13.0}$	66.2 ± 5.6
Ours-Aug	$\mathit{69.8 \pm 10.1}$	34.7 ± 5.6	38.3 ± 22.6	38.6 ± 16.7	2.4 ± 1.2	36.8 ± 11.3

Table 22: IN DOMAIN precision on CAMELYON17 where nuclei are *inpainted* after being expanded with filter size 5. The column name indicates the centre used to train models. The best accuracy for each column is in **bold face** and the second best in *italics*.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
ERM	54.8 ± 2.5	71.2 ± 13.8	0.0 ± 0.0	$\mathbf{99.4 \pm 0.5}$	47.7 ± 7.4	54.6 ± 4.8
Macenko	64.2 ± 25.8	29.0 ± 17.0	0.0 ± 0.0	62.6 ± 21.6	$\mathbf{75.0 \pm 18.7}$	46.2 ± 16.6
RSC	54.7 ± 2.9	56.9 ± 18.6	1.2 ± 4.0	$\mathit{98.4 \pm 1.0}$	51.3 ± 9.9	52.5 ± 7.3
L2D	64.0 ± 10.9	$\mathbf{87.0 \pm 2.6}$	0.8 ± 1.8	90.1 ± 5.9	48.1 ± 15.3	58.0 ± 7.3
Ours	11.1 ± 5.8	14.2 ± 21.4	1.6 ± 2.6	42.9 ± 47.5	0.8 ± 0.8	14.1 ± 15.6
ERM-Aug	58.8 ± 12.9	$\mathit{79.8 \pm 12.3}$	$\mathbf{97.6 \pm 1.6}$	83.0 ± 5.6	38.1 ± 21.3	$\mathit{71.5 \pm 10.7}$
Macenko-Aug	$\mathit{64.7 \pm 2.8}$	61.6 ± 14.2	94.7 ± 4.7	92.5 ± 3.0	60.3 ± 10.5	$\mathbf{74.8 \pm 7.0}$
RSC-Aug	61.9 ± 7.6	69.9 ± 20.5	94.7 ± 5.4	90.3 ± 3.8	27.6 ± 18.1	68.9 ± 11.1
L2D-Aug	$\mathbf{65.1 \pm 0.6}$	36.3 ± 1.2	$\mathit{97.3 \pm 1.3}$	86.6 ± 2.7	$\mathit{69.9 \pm 7.0}$	71.0 ± 2.6
Ours-Aug	25.9 ± 40.5	15.5 ± 12.5	20.9 ± 13.8	18.6 ± 15.0	4.7 ± 2.1	17.1 ± 16.8

Table 23: Out-of-domain accuracy on CAMELYON17 with nuclei on white background. The column name indicates the centre used to train models. The best accuracy for each column is in **bold face** and the second best in *italics*.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
ERM	73.1 ± 7.9	66.4 ± 12.7	50.0 ± 0.0	46.3 ± 2.9	49.5 ± 4.0	57.1 ± 5.5
Macenko	44.9 ± 1.9	63.5 ± 10.4	49.8 ± 0.6	45.4 ± 2.0	50.1 ± 0.2	50.7 ± 3.0
RSC	55.0 ± 7.4	61.4 ± 10.0	50.0 ± 0.0	48.0 ± 2.9	50.3 ± 0.5	52.9 ± 4.2
L2D	53.6 ± 10.1	84.1 ± 1.8	49.9 ± 4.4	41.6 ± 8.9	78.9 ± 7.5	61.6 ± 6.5
Ours	$\mathit{90.1 \pm 1.0}$	$\mathbf{92.9 \pm 0.4}$	$\mathit{89.9 \pm 1.9}$	$\mathbf{92.9 \pm 0.3}$	89.7 ± 1.1	$\mathbf{91.1 \pm 1.0}$
ERM-Aug	58.6 ± 8.1	68.1 ± 9.5	56.9 ± 6.9	52.0 ± 5.1	54.8 ± 6.6	58.1 ± 7.2
Macenko-Aug	52.0 ± 7.1	82.7 ± 9.1	51.2 ± 11.3	51.2 ± 5.7	73.3 ± 9.4	62.1 ± 8.5
RSC-Aug	54.5 ± 8.7	75.9 ± 2.8	48.7 ± 1.6	51.0 ± 5.0	60.4 ± 9.6	58.1 ± 5.5
L2D-Aug	84.1 ± 0.9	89.0 ± 0.6	45.7 ± 0.9	67.5 ± 7.6	$\mathbf{90.3 \pm 1.5}$	75.3 ± 2.3
Ours-Aug	$\mathbf{90.2 \pm 1.1}$	$\mathit{92.2 \pm 0.9}$	$\mathbf{91.0 \pm 2.2}$	$\mathit{91.7 \pm 0.7}$	$\mathit{89.8 \pm 1.3}$	$\mathit{91.0 \pm 1.2}$

Table 24: Out-of-domain accuracy on CAMELYON17 evaluated on binary nuclei segmentation masks. The column name indicates the centre used to train models. The best accuracy for each column is in **bold face** and the second best in *italics*.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
ERM	48.7 ± 4.1	50.0 ± 0.0	50.0 ± 0.0	50.0 ± 0.0	50.8 ± 3.7	49.9 ± 1.6
Macenko	50.0 ± 0.0	49.0 ± 3.1	49.6 ± 1.1	49.4 ± 1.4	50.0 ± 0.0	49.6 ± 1.1
RSC	50.0 ± 0.0	50.0 ± 0.0	49.6 ± 1.3	50.0 ± 0.0	50.0 ± 0.0	49.9 ± 0.3
L2D	59.4 ± 10.5	58.2 ± 6.3	52.7 ± 5.0	50.0 ± 0.0	50.2 ± 0.7	54.1 ± 4.5
Ours	$\mathit{90.8 \pm 1.0}$	$\mathbf{92.9 \pm 0.3}$	$\mathit{92.4 \pm 0.6}$	$\mathit{91.3 \pm 1.2}$	$\mathit{90.8 \pm 0.8}$	$\mathit{91.6 \pm 0.8}$
ERM-Aug	49.3 ± 1.9	49.9 ± 0.3	50.0 ± 0.0	49.0 ± 1.8	49.7 ± 0.6	49.6 ± 0.9
Macenko-Aug	49.4 ± 1.1	50.3 ± 1.0	49.9 ± 1.8	48.2 ± 1.1	50.0 ± 0.0	49.6 ± 1.0
RSC-Aug	49.8 ± 0.5	50.0 ± 0.0	53.1 ± 5.3	50.0 ± 0.0	50.0 ± 0.0	50.6 ± 1.2
L2D-Aug	49.9 ± 1.0	58.4 ± 5.8	50.0 ± 0.0	50.8 ± 2.4	52.4 ± 4.2	52.3 ± 2.7
Ours-Aug	$\mathbf{91.6 \pm 0.5}$	$\mathit{92.4 \pm 0.9}$	$\mathbf{92.5 \pm 1.0}$	$\mathbf{92.4 \pm 0.4}$	$\mathbf{91.4 \pm 0.4}$	$\mathbf{92.1 \pm 0.7}$

Table 25: Out-of-domain accuracy on CAMELYON17 evaluated on inverted nuclei segmentation masks. The column name indicates the centre used to train models. The best accuracy for each column is in **bold face** and the second best in *italics*.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
ERM	47.7 ± 2.5	50.8 ± 2.5	50.0 ± 0.0	47.3 ± 1.9	51.1 ± 3.6	49.4 ± 2.1
Macenko	45.3 ± 1.6	54.6 ± 8.6	49.9 ± 0.4	42.7 ± 8.0	50.0 ± 0.0	48.5 ± 3.7
RSC	48.7 ± 1.9	50.0 ± 0.0	50.0 ± 0.0	48.7 ± 1.9	50.0 ± 0.0	49.5 ± 0.8
L2D	43.5 ± 4.8	72.8 ± 4.7	57.1 ± 9.2	39.6 ± 2.7	48.6 ± 5.1	52.3 ± 5.3
Ours	$\mathit{89.0 \pm 1.0}$	$\mathbf{92.3 \pm 0.4}$	$\mathit{90.6 \pm 2.0}$	$\mathit{88.2 \pm 2.6}$	$\mathit{86.5 \pm 1.8}$	$\mathit{89.3 \pm 1.6}$
ERM-Aug	48.2 ± 2.0	52.4 ± 3.7	53.2 ± 8.9	44.9 ± 1.0	49.4 ± 0.7	49.6 ± 3.3
Macenko-Aug	45.2 ± 3.8	76.0 ± 13.4	49.6 ± 9.1	46.1 ± 0.9	50.0 ± 0.1	53.4 ± 5.5
RSC-Aug	47.6 ± 2.9	51.8 ± 10.6	67.5 ± 13.1	44.9 ± 1.7	49.9 ± 0.4	52.3 ± 5.7
L2D-Aug	64.1 ± 5.0	87.5 ± 1.2	43.2 ± 3.4	58.7 ± 11.0	64.9 ± 9.0	63.7 ± 5.9
Ours-Aug	$\mathbf{91.7 \pm 0.5}$	$\mathit{91.7 \pm 1.0}$	$\mathbf{91.7 \pm 1.4}$	$\mathbf{92.7 \pm 0.3}$	$\mathbf{86.6 \pm 2.3}$	$\mathbf{90.9 \pm 1.1}$

Table 26: Accuracy when combining L2D or RSC with the proposed method on CAMELYON17. The best accuracy for each column is in **bold face** and the second best in *italics*. "+Ours" results are an average of five models for each combination of medical centre and method instead of an average of ten models. All results are using photometric augmentations described in 4.3 even though "-Aug" is omitted from the name in the table.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
L2D	$\mathbf{94.3 \pm 0.1}$	87.6 ± 0.6	87.7 ± 1.4	83.4 ± 2.6	$\mathbf{92.3 \pm 0.9}$	89.1 ± 1.1
L2D+Ours	92.2 ± 0.6	$\mathit{92.8 \pm 0.1}$	$\mathit{92.5 \pm 0.5}$	84.8 ± 1.5	88.9 ± 1.4	90.3 ± 0.8
RSC	$\mathit{93.1 \pm 0.8}$	78.2 ± 2.0	89.3 ± 3.4	77.9 ± 2.2	91.0 ± 1.7	85.9 ± 2.0
RSC+Ours	92.1 ± 0.8	$\mathbf{93.6 \pm 0.4}$	$\mathit{92.5 \pm 1.2}$	$\mathbf{91.3 \pm 0.7}$	$\mathit{91.9 \pm 0.4}$	$\mathbf{92.3 \pm 0.7}$
Ours	91.8 ± 0.7	92.2 ± 1.6	$\mathbf{92.9 \pm 0.7}$	$\mathit{90.4 \pm 1.1}$	91.7 ± 0.5	$\mathit{91.8 \pm 0.9}$

Table 27: Accuracy when combining L2D or RSC with the proposed method on BCSS. The best accuracy for each column is in **bold face** and the second best in *italics*. "+Ours" results are an average of five models for each combination of medical centre and method instead of an average of ten models. All results are using photometric augmentations described in 4.3 even though "-Aug" is omitted from the name in the table.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
L2D	81.9 ± 0.3	74.7 ± 0.6	74.2 ± 0.6	67.5 ± 2.9	$\mathbf{77.2 \pm 2.2}$	75.1 ± 1.3
L2D+Ours	81.8 ± 2.0	$\mathit{82.2 \pm 0.5}$	73.7 ± 1.3	65.7 ± 2.6	73.8 ± 2.2	75.5 ± 1.7
RSC	79.6 ± 1.5	72.1 ± 2.9	71.6 ± 1.7	63.2 ± 1.6	74.1 ± 4.3	72.1 ± 2.4
RSC+Ours	$\mathit{82.0 \pm 1.2}$	$\mathbf{82.5 \pm 1.7}$	$\mathbf{76.7 \pm 2.9}$	$\mathit{77.5 \pm 2.0}$	$\mathit{75.8 \pm 1.6}$	$\mathbf{78.9 \pm 1.9}$
Ours	$\mathbf{82.3 \pm 0.8}$	81.9 ± 2.3	$\mathit{75.7 \pm 2.2}$	$\mathbf{79.6 \pm 1.0}$	74.8 ± 1.9	$\mathit{78.8 \pm 1.6}$

Table 28: Accuracy when combining L2D or RSC with the proposed method on Ocelot. The best accuracy for each column is in **bold face** and the second best in *italics*. "+Ours" results are an average of five models for each combination of medical centre and method instead of an average of ten models. All results are using photometric augmentations described in 4.3 even though "-Aug" is omitted from the name in the table.

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average
L2D	$\mathbf{74.7 \pm 0.6}$	68.3 ± 0.5	65.6 ± 0.7	$\mathit{62.5 \pm 2.7}$	$\mathbf{74.4 \pm 1.2}$	$\mathit{69.1 \pm 1.1}$
L2D+Ours	$\mathit{72.2 \pm 1.6}$	70.8 ± 0.3	$\mathit{69.4 \pm 0.6}$	59.8 ± 1.6	69.9 ± 2.3	68.4 ± 1.3
RSC	71.9 ± 2.2	61.7 ± 2.4	69.2 ± 2.9	58.4 ± 1.4	70.1 ± 3.6	66.3 ± 2.5
RSC+Ours	71.1 ± 0.7	$\mathbf{72.1 \pm 0.6}$	$\mathbf{69.6 \pm 1.7}$	$\mathbf{70.4 \pm 1.7}$	$\mathit{70.9 \pm 1.2}$	$\mathbf{70.8 \pm 1.2}$
Ours	70.8 ± 0.5	$\mathit{72.0 \pm 1.3}$	68.9 ± 1.3	$\mathbf{70.4 \pm 0.7}$	70.7 ± 1.3	$\mathit{70.6 \pm 1.0}$

Table 29: Accuracy when using a ViT-Tiny [60], comparing the baseline against the proposed method on CAMELYON17. The best accuracy for each column is in **bold face** and the second best in *italics*. All results are using photometric augmentations described in 4.3 even though "-Aug" is omitted from the name in the table. 'AwoC4' is 'Average without Centre-4'. 'RSNA' refers to RandStainNA [24].

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average	AwoC4
ERM	94.6 ± 0.5	78.0 ± 4.0	90.0 ± 1.6	84.9 ± 2.9	87.7 ± 2.5	87.0 ± 2.3	86.9 ± 2.3
RSNA	94.6 ± 0.4	$\mathit{86.8 \pm 2.0}$	88.5 ± 2.8	84.9 ± 2.3	$\mathit{90.6 \pm 2.9}$	89.1 ± 2.1	88.7 ± 1.9
DDCA	$\mathit{95.0 \pm 0.5}$	77.9 ± 3.7	91.2 ± 1.6	81.7 ± 4.1	86.2 ± 3.0	86.4 ± 2.6	86.5 ± 2.5
L2D	$\mathbf{95.1 \pm 0.0}$	81.0 ± 0.1	$\mathit{91.8 \pm 0.8}$	$\mathit{87.3 \pm 0.8}$	$\mathbf{91.8 \pm 0.1}$	$\mathit{89.4 \pm 0.4}$	$\mathit{88.8 \pm 0.4}$
Ours	94.5 ± 0.2	$\mathbf{92.5 \pm 0.5}$	$\mathbf{94.0 \pm 0.7}$	$\mathbf{91.8 \pm 1.0}$	86.7 ± 2.1	$\mathbf{91.9 \pm 0.9}$	$\mathbf{93.2 \pm 0.6}$

Table 30: Accuracy when using a ViT-Tiny [60], comparing the baseline against the proposed method on BCSS. The best accuracy for each column is in **bold face** and the second best in *italics*. All results are using photometric augmentations described in 4.3 even though "-Aug" is omitted from the name in the table. 'AwoC4' is 'Average without Centre-4'. 'RSNA' refers to RandStainNA [24].

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average	AwoC4
ERM	79.2 ± 3.3	70.7 ± 4.6	68.6 ± 2.4	66.9 ± 2.2	69.5 ± 5.1	71.0 ± 3.5	71.4 ± 3.1
RSNA	$\mathit{79.9 \pm 1.9}$	$\mathit{77.9 \pm 2.4}$	$\mathbf{74.4 \pm 4.3}$	72.1 ± 2.4	$\mathit{75.2 \pm 5.4}$	$\mathit{75.9 \pm 3.3}$	$\mathit{76.1 \pm 2.8}$
DDCA	78.1 ± 2.7	70.8 ± 3.2	69.9 ± 4.5	66.3 ± 4.4	65.8 ± 4.6	70.2 ± 3.9	71.3 ± 3.7
L2D	$\mathbf{80.8 \pm 0.1}$	75.3 ± 0.1	71.6 ± 4.1	$\mathbf{75.9 \pm 0.3}$	$\mathbf{79.1 \pm 0.3}$	$\mathbf{76.5 \pm 1.0}$	75.9 ± 1.2
Ours	78.6 ± 1.1	$\mathbf{79.5 \pm 1.2}$	$\mathit{74.3 \pm 2.4}$	$\mathit{75.4 \pm 1.3}$	61.9 ± 3.6	74.0 ± 1.9	$\mathbf{77.0 \pm 1.5}$

Table 31: Accuracy when using a ViT-Tiny [60], comparing the baseline against the proposed method on Ocelot. The best accuracy for each column is in **bold face** and the second best in *italics*. All results are using photometric augmentations described in 4.3 even though "-Aug" is omitted from the name in the table. 'AwoC4' is 'Average without Centre-4'. 'RSNA' refers to RandStainNA [24].

Method	Centre-0	Centre-1	Centre-2	Centre-3	Centre-4	Average	AwoC4
ERM	$\mathit{71.7 \pm 1.9}$	61.0 ± 3.5	67.1 ± 3.1	62.6 ± 3.5	67.3 ± 4.3	65.9 ± 3.2	65.6 ± 3.0
RSNA	71.2 ± 1.8	$\mathit{68.6 \pm 2.8}$	67.2 ± 2.5	68.3 ± 1.7	$\mathit{69.2 \pm 3.6}$	$\mathit{68.9 \pm 2.5}$	68.8 ± 2.2
DDCA	$\mathbf{73.0 \pm 1.8}$	61.2 ± 2.0	67.2 ± 0.9	61.6 ± 4.2	62.6 ± 4.1	65.1 ± 2.6	65.7 ± 2.2
L2D	70.8 ± 0.1	65.9 ± 0.1	$\mathit{68.9 \pm 1.0}$	$\mathbf{71.8 \pm 0.7}$	$\mathbf{71.5 \pm 0.3}$	$\mathbf{69.8 \pm 0.4}$	$\mathit{69.3 \pm 0.5}$
Ours	69.3 ± 0.8	$\mathbf{68.8 \pm 0.8}$	$\mathbf{70.0 \pm 0.6}$	$\mathit{70.4 \pm 1.0}$	64.5 ± 2.3	68.6 ± 1.1	$\mathbf{69.6 \pm 0.8}$

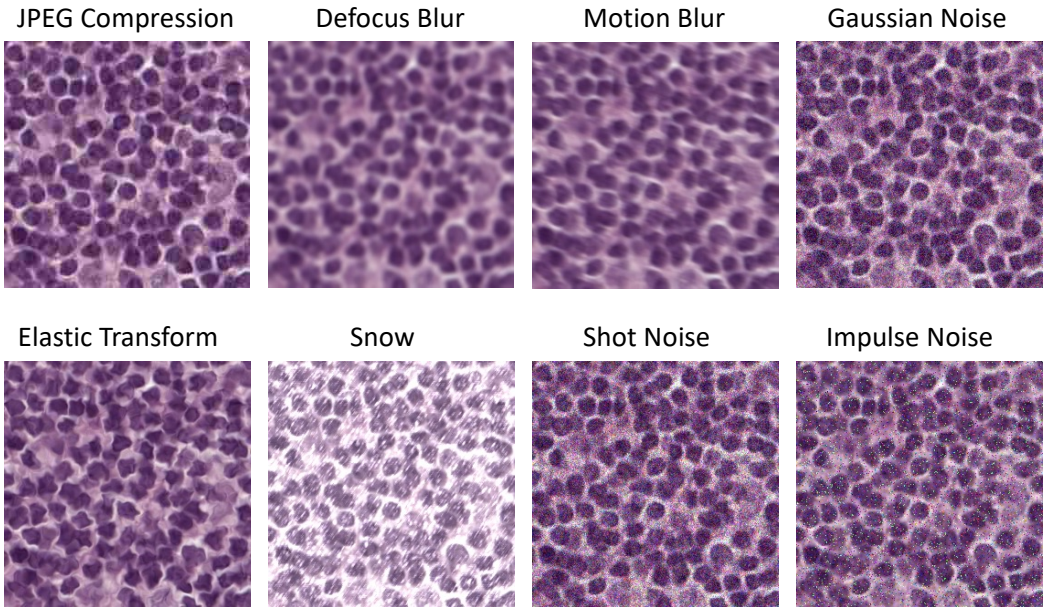


Figure 5: Exemplary image corruptions from [61] applied to an input image used in this study.

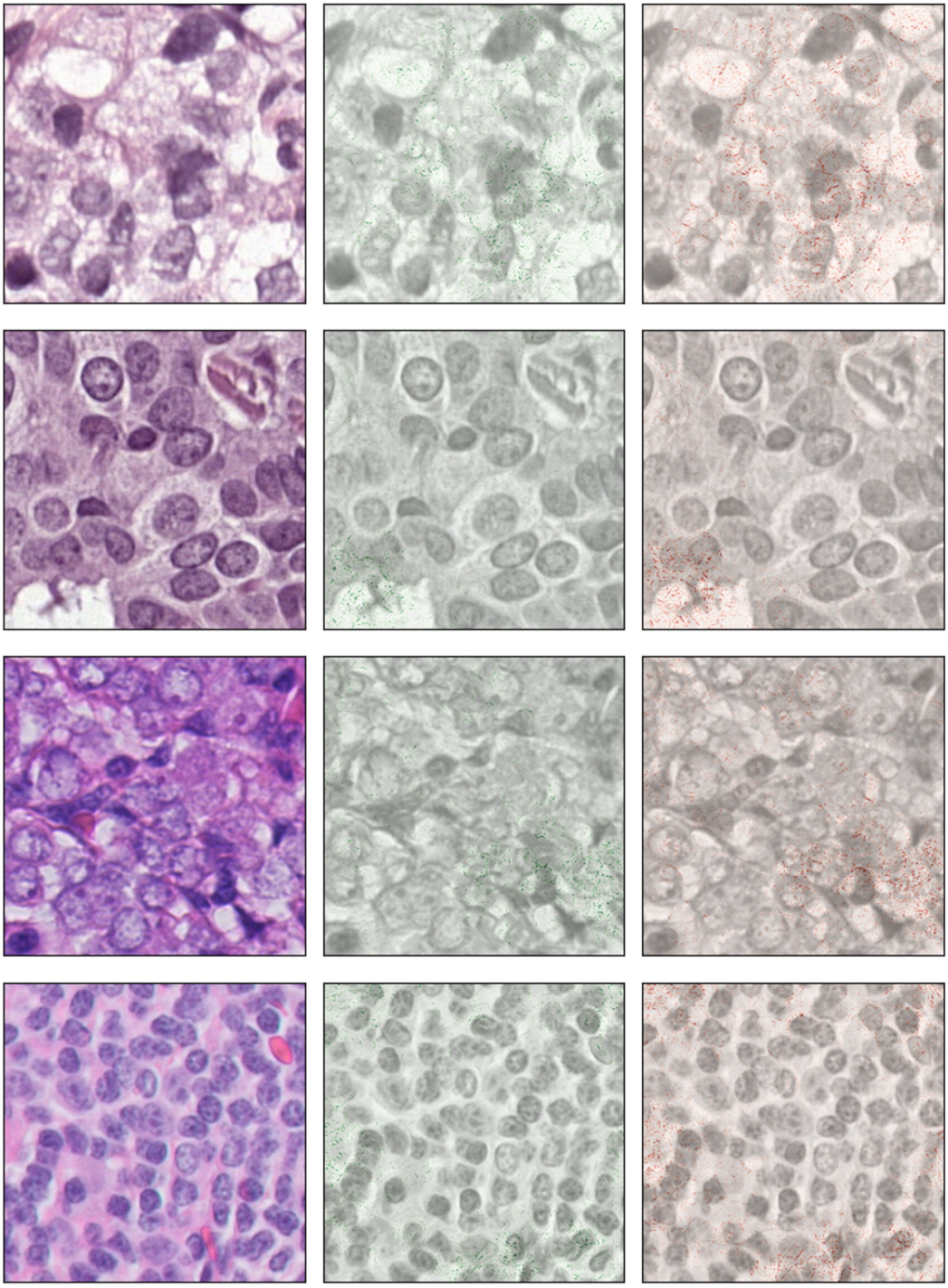


Figure 6: Saliency maps for four randomly selected tiles using Integrated Gradients [64] for a model trained via ERM-Aug. The green-coloured map indicates the contribution towards the positive class (tumour), and the red one towards the negative class (non-tumour).

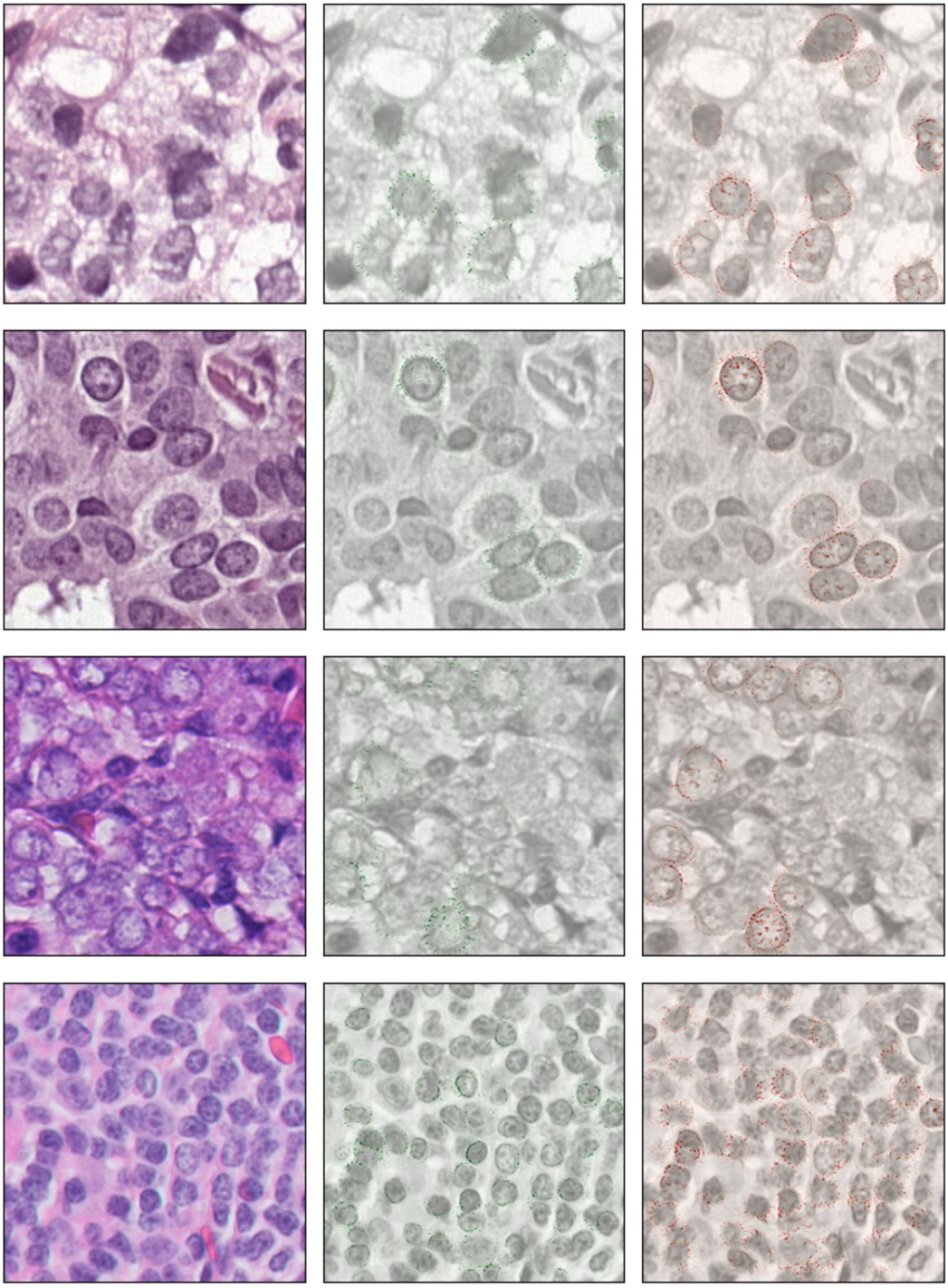


Figure 7: Saliency maps for four randomly selected tiles using Integrated Gradients [64] for a model trained via Ours-Aug. The green-coloured map indicates the contribution towards the positive class (tumour), and the red one towards the negative class (non-tumour).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claim made in the abstract and introduction are supported by the results in the main Tables 1, 2, and 3, and for robustness in Figures 3 and 4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalise to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: see the end of the Section 5 Ablation Study and Discussion for limitations

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The method is described in Section Proposed method 3. The experiments are described in Section Experiments 4, and, there in particular in Subsection 4.3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The datasets are created and available via their owners. Code is available, shareable and presentable to external parties. It is available at `https://github.com/undercutspiky/SFL/`

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experiments are described in Section Experiments 4, and, in particular, in Subsection 4.3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We reported variances of the performance measures over the runs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We disclosed the GPU types, the GPU RAM and training times, see Section 4.3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have checked against the Neurips ethics code. No privacy, security, safety impact expected. "Disclosure of essential elements for reproducibility" has been done in the paper, and we consider it to be sufficient for reproduciblity

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: No societal impact of the work expected

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No direct malicious use is expected from a method to improve out-of-domain generalisation.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The datasets, the model and the deep learning toolbox are all cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We use established datasets [51–53] containing patient data, which were collected by third parties mentioned in the cited publications and which are already published. Applied crowdsourcing is described in those publications.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: We use established datasets [51–53] containing patient data, which were collected by third parties mentioned in the cited publications and which are already published. As a consequence, IRB approval was not required for our study.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.