Samba: Severity-aware Recurrent Modeling for Cross-domain Medical Image Grading

Qi Bi[1]*, **Jingjun Yi**[2], **Hao Zheng**[2]✉, **Wei Ji**[3], **Haolan Zhan**[4],
Yawen Huang[2], **Yuexiang Li**[5]✉, **Yefeng Zheng**[1]✉
[1]Westlake University, China, [2]Jarvis Research Center, Tencent Youtu Lab, China
[3]Yale University, United States, [4]Monash University, Australia,
[5]Guangxi Medical University, China
howzheng@tencent.com, yuexiang.li@ieee.org
zhengyefeng@westlake.edu.cn

Abstract

Disease grading is a crucial task in medical image analysis. Due to the continuous progression of diseases, *i.e.*, the variability within the same level and the similarity between adjacent stages, accurate grading is highly challenging. Furthermore, in real-world scenarios, models trained on limited source domain datasets should also be capable of handling data from unseen target domains. Due to the cross-domain variants, the feature distribution between source and unseen target domains can be dramatically different, leading to a substantial decrease in model performance. To address these challenges in cross-domain disease grading, we propose a Severity-aware Recurrent Modeling (Samba) method in this paper. As the core objective of most staging tasks is to identify the most severe lesions, which may only occupy a small portion of the image, we propose to encode image patches in a sequential and recurrent manner. Specifically, a state space model is tailored to store and transport the severity information by hidden states. Moreover, to mitigate the impact of cross-domain variants, an Expectation-Maximization (EM) based state recalibration mechanism is designed to map the patch embeddings into a more compact space. We model the feature distributions of different lesions through the Gaussian Mixture Model (GMM) and reconstruct the intermediate features based on learnable severity bases. Extensive experiments show the proposed Samba outperforms the VMamba baseline by an average accuracy of 23.5%, 5.6% and 4.1% on the cross-domain grading of fatigue fracture, breast cancer and diabetic retinopathy, respectively. Source code is available at `https://github.com/BiQiWHU/Samba`.

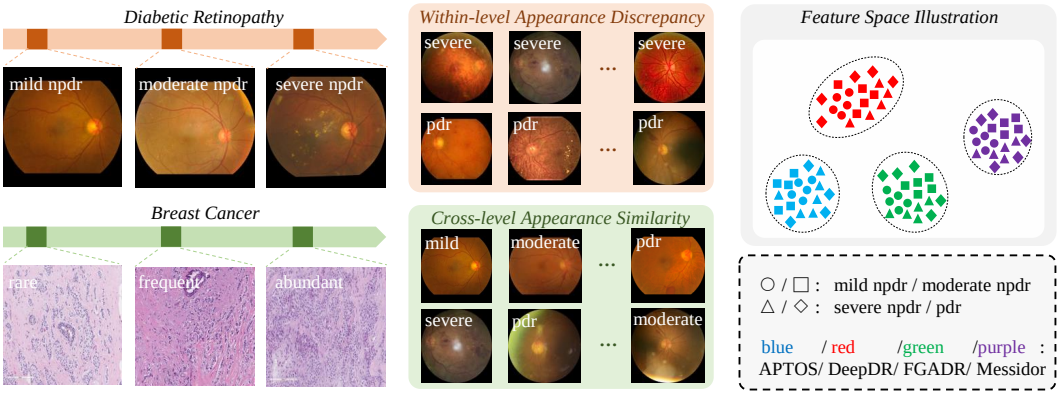
1 Introduction

Disease grading aims to assess the severity level of a disease or a pathological region from a medical image [46, 31, 52, 50, 6]. It is more challenging than conventional deterministic classification with distinctive categories (*e.g.*, cat *vs.* dog), owing to the inherent severity ambiguity within and between levels. This ambiguity arises because the progression of a certain disease or a pathological region is a transitional, continuous and time-growing process (illustrated in Fig. 1a). On the one hand, different medical images within a same severity level can have rather different disease or pathological developments (shown in Fig. 1b). On the other hand, medical images among different severity levels can share similar patterns, as low-level lesions may persist throughout the disease's progression.

The past decade has witnessed the rapid development of disease grading methods [30, 34, 45] owing to the deep learning techniques [25, 24, 28, 26, 57]. However, most of these methods were developed

*This research was conducted under the support from Westlake University and Tencent Youtu Lab.

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



(a) Continuous disease development (b) Within-/cross-level ambiguity (c) Cross-domain clustering challenge

Figure 1: (a) The development of a disease or a pathological region is a continuous progress; (b) The continuous development apart from individual differences and style variation poses both within-level discrepancy (top) and cross-level similarity (bottom) on the medical image appearance; (c) These challenges can make the medical images from the same unseen domain, instead of those from the same grade level, to be clustered in the feature space.

by experts from a specific clinic field (ophthalmology, gynecology, *etc*.). Furthermore, these models usually assume that the medical images used for training and inference are independently and identically distributed (i.i.d.). In practical clinical scenarios, a grading model trained on a number of medical images (source domain) is often required to handle images it has not encountered before (unseen target domains). Due to variations between patients, scanners, imaging parameters, clinic centers, *etc*., the feature distributions of the source domain and unseen target domains can be dramatically different [32, 68, 23, 8]. When the severity level of a disease is measured by the distribution of lesions, this cross-domain variance can lead to the misdetection of crucial lesions, resulting in grading errors [4, 11]. Especially when the appearance of lesions is significantly affected by the style change, it may be observed that medical images from the same *domain* instead of from the same *grade* are clustered in the feature space (illustrated in Fig. 1c). This suggests that the model has learned features with limited generalization ability.

Domain generalized disease grading learns models from only a source domain, but is expected to be applicable to unseen target domains. The key to addressing this problem is accurately identifying the lesions that have a decisive impact on grading [49, 45]. As the disease progresses, multiple lesions may coexist in the image, and the critical aspect of grading is identifying the most severe one among them. However, the most severe lesion may be localized in a small region in the image, exhibiting variable shapes, and being influenced by cross-domain style changes [4, 11]. To overcome these challenges, this paper proposes a severity-aware recurrent modeling method (Samba). Samba encodes image patches in a recurrent manner and recalibrates the state distributions based on learnable bases.

In many disease grading scenarios, the decisive lesions only occupy a small portion of the total area. For instance, in retinal photographs, the affected blood vessels may only involve a small section at the distal. Similarly, in computed tomography (CT) or magnetic resonance imaging (MRI) scans, malignant tumors can also present as small lesions with a diameter less than 3 mm. These small lesions are easily influenced by style variations, which can lead to incorrect grading. Therefore, the model needs to pay sufficient attention to these detailed patches to classify them accurately. To address this issue, we treat the image patches as sequential data and encode them in a recurrent manner. This approach allows the information of decisive lesions to be stored in the hidden states and propagated to subsequent sequences. Furthermore, we adopt bidirectional encoding, enabling critical local information to influence the overall representation. More specifically, we incorporate a bidirectional Mamba [17] layer into the Samba, which supports sequence-to-sequence transformation and efficiently selects data in an input-dependent manner.

The Mamba model achieves its selection mechanism by parameterizing the State Space Model (SSM) based on the input. While this selection mechanism [17, 69, 35] aids in identifying decisive lesions and propagating critical information, these input-dependent parameters are also vulnerable to the influence of image style transformations. When the feature distribution is affected by cross-domain

variations, both the update of hidden states and the gating mechanism are disrupted. To resolve this problem, we utilize learnable tokens to capture the lesion representations, which are then used as bases to map the feature embeddings into a more compact space. To preserve the semantic information within this process, we further employ the Expectation-Maximum (EM) algorithm [14] initialized by these bases to estimate the lesion feature distribution for each image and reconstruct the features accordingly. We refer to this process as EM-based state recalibration in this paper.

Our contributions can be summarized as follows.

- We develop a Severity-aware Recurrent Modeling, dubbed as Samba, for general disease grading within- and cross-domain medical images.
- We propose to encode the image patches in a recurrent manner to accurately capture the decisive lesions and transport the critical information from local to global.
- An EM-based state recalibration mechanism is designed to reduce the impacts of cross-domain variants by mapping the feature embeddings into a compact space.
- Extensive experiments on three cross-domain disease grading benchmarks show the effectiveness of the Samba against the baseline.

2 Related Work

Domain generalization aims to learn a model that can be well generalized to unseen target domains when only trained by the source domain, where the cross-domain feature distribution is usually not identical [65]. In the past few years, a variety of machine learning techniques (*e.g.*, discrepancy minimization [47, 13], knowledge ensemble [12], uncertainty quantification [39, 53], optimal transport [16, 60], self-learning [51, 43], frequency decoupling [59, 9, 10] and casual inference [37, 38]) have been proposed. In the medical imaging community, the effort of bridging the domain gap between training data and unseen inference data is so far mainly focused on medical image segmentation [32, 68, 23, 8, 58] and classification [66, 54]. These methods usually rely on either learning shape-invariant representation or reaching pixel-wise consensus among the source domains. However, they are not especially devised to tackle the key challenge in cross-domain medical image grading, where the medical images from the same severity level instead of the same domain tend to cluster together.

Medical grading has also been extensively studied. For Diabetic Retinopathy (DR) grading, many works aim to highlight the subtle local pathological regions to better discern different severity levels [30, 34, 44, 6, 45, 7]. Similarly, grading models have also been developed for pulmonary nodules [46], fatigue fracture [31], glioma [52], acne vulgaris [50], *etc*. However, most of the existing grading methods are task-specific and assume the training and inference medical images are i.i.d., which is far from reality. Practically, a medical grading model is supposed to show reliable inference on unseen target domains that have different feature distribution from the source domain. *To the best of our knowledge*, only [4] and [11] made an initial investigation on learning domain generalized DR grading.

State Space Model (SSM) [27] contributes to a variety of fields such as robotics, navigation, and control theory, which is a foundational scientific model. In the past few years, SSM has been adapted in the context of deep representation learning, and has shown great success in sequence modeling [19, 20]. More advanced SSM, exemplified by Mamba [17], not only shows stronger representation ability in long sequence modeling, but also exhibits linear scaling ability for long-sequence data. Built upon this, multiple Mamba variations (*e.g.*, Vim [35] and VMamba [69]) have shown effectiveness in the computer vision field. However, these methods mainly focus on enhancing the context representation from the image by exploiting the long-range dependencies. Instead, how to model the cross-level severity development from the medical image by SSM remains unexplored.

3 Methodology

3.1 Problem Definition & Framework Overview

For a given disease grading task, assume we have a number of medical images $\boldsymbol{x}$ and the corresponding severity-level labels $\boldsymbol{y}$ from K different domains, which is denoted as $\mathcal{D}_1 = \{(x_n^{(1)}, y_n^{(1)})\}_{n=1}^{N_1}$,

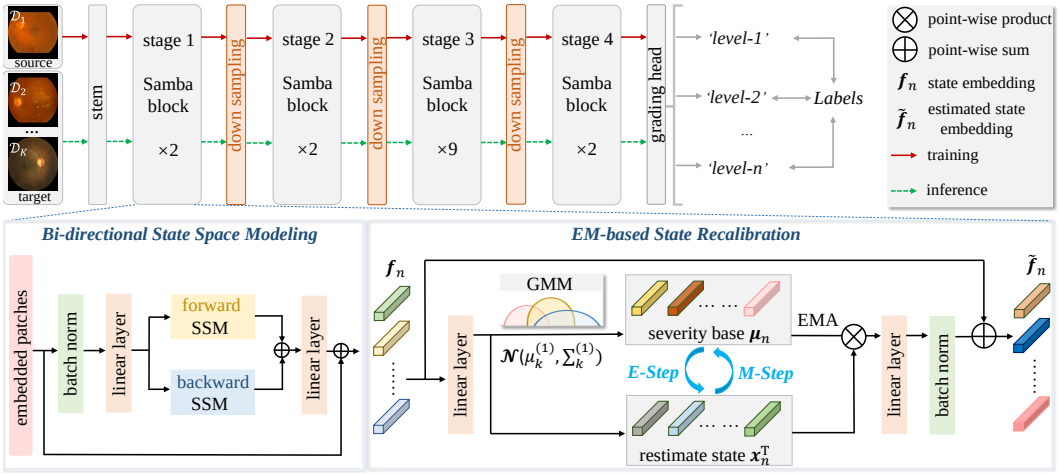


Figure 2: Framework of the proposed Severity-aware Recurrent Modeling (Samba) method. The patch embeddings pass through four encoding stages consisting of different number of severity-aware recurrent layers. Within each Samba block, the embeddings are first input to the bidirectional Mamba layers to store and transport the information about decisive lesions. After that, an EM-based state recalibration module models the feature distribution of lesions via a Gaussian Mixture Model with learnable severity bases. Moreover, the bases are re-estimated by the EM algorithm for each image and reconstruct the features finally.

$\mathcal{D}_2 = \{(x_n^{(2)}, y_n^{(2)})\}_{n=1}^{N_2}, \cdots, \mathcal{D}_K = \{(x_n^{(K)}, y_n^{(K)})\}_{n=1}^{N_K}$. Here N_k denotes the number of images in domain k. For the cross-domain disease grading problem, the objective is to learn a grading model $F_\theta : x \to y$ using images only from a source domain $\mathcal{D}_1$, which is supposed to generalize well on other unseen target domains $\mathcal{D}_2, \cdots, \mathcal{D}_K$. Following prior domain generalization works, each dataset is regarded as a domain $\mathcal{D}_k$, as the samples in a certain dataset are usually collected from the same clinical center by the same scanners and therefore share more similar feature distribution.

The overview of the proposed method is illustrated in Fig. 2. The input image is first encoded into patch embeddings through a stem unit with 4×4 convolutional kernels, where the stem unit partitions the input image into patches. The patch embeddings further pass through four encoding stages. Each Samba block involves a certain number of severity-aware recurrent layers and there are downsampling layers between two consecutive blocks. Finally, a grading head consisting of an average pooling layer and a linear layer generates the final prediction. Within the Samba block, the patch embeddings are first input to the bidirectional Mamba layers to extract the information about decisive lesions. After that, EM-based state recalibration is applied to map the lesion representation into a compact space by learnable bases.

3.2 Recurrent Patch Modeling by State Space Model

The core issue in most medical image disease grading scenarios is to identify the most severe lesion. However, due to the presence of lesions from different stages of the disease in the image, accurately capturing the most severe lesion is highly challenging. When the lesion occupies a large proportion of the image, the model only needs to extract stage-related features. In contrast, when the area of the critical lesion is small, the model needs to simultaneously locate the lesion and extract relevant features. This places higher demands on the model's ability to handle local information. To address this issue, in this paper, we propose to encode the image patches in a recurrent manner. Specifically, the state space model is used to process the sequential patch embeddings.

State Space Model. Let $x(t)$ denote a 1-D input signal. SSM maps it to the 1-D output signal $y(t)$ by an intermediate N-dimensional latent state $u(t)$, given by

$$u'(t) = \boldsymbol{A}u(t) + \boldsymbol{B}x(t), \quad y(t) = \boldsymbol{C}u(t) + \boldsymbol{D}x(t), \tag{1}$$

where $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ denotes the state matrix, while $\boldsymbol{B} \in \mathbb{R}^{N\times 1}$, $\boldsymbol{C} \in \mathbb{R}^{N\times 1}$ and $\boldsymbol{D} \in \mathbb{R}^{N\times 1}$ denote the projection parameters. For deep sequential modeling, $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ and $\boldsymbol{D}$ are parameters that can be learned by gradient descent. The parameter $\boldsymbol{D}$ is omitted for exposition (*i.e.*, $\boldsymbol{D} = 0$) as $\boldsymbol{D}x(t)$ can be regarded as a skip connection and is easy to compute [19, 20].

Discretization. The structured state space [20] and Mamba [17] discretize the above continuous system so as to be tailored for deep representation learning. There are usually two ways for discretization, namely, linear recurrence and discrete convolution. For linear recurrence, instead of a continuous function $x(t)$, a discrete sequence $(x_0, x_1, \cdots)$ is taken as input. Conceptually, we have $x_k = x(k\Delta)$. The state matrix $\boldsymbol{A}$ is approximated as $\overline{\boldsymbol{A}}$ by the zero-order hold rule. The discrete SSM is a sequence-to-sequence map $x_k \mapsto y_k$, given by

$$\begin{aligned} u_k &= \overline{\boldsymbol{A}} u_{k-1} + \overline{\boldsymbol{B}} x_k, \qquad & \overline{\boldsymbol{A}} &= e^{\Delta \boldsymbol{A}}, \\ y_k &= \overline{\boldsymbol{C}} u_k, \qquad & \overline{\boldsymbol{B}} &= \Delta \boldsymbol{B}, \qquad \overline{\boldsymbol{C}} = \boldsymbol{C}. \end{aligned} \tag{2}$$

Selective Scan Mechanism. Prior SSM methods usually focus on the linear time-invariant scenario. Instead, the selective scan mechanism [17], which is the core of SSM operator in Mamba, learns the dynamism of weights from the input and is more aware of the context information.

The Mamba model is a suitable structure that aligns with our needs. When encoding the image patches as sequential data, once important lesion information is discovered, it can be stored in hidden states and propagated to subsequent sequences. Specifically, after sliding the image $\boldsymbol{x} \in \mathbb{R}^{H \times W \times 3}$ into a variety of patches, the input is formed as a sequences of 2-D patches, each of which has a spatial position of $H/4 \times W/4$. Then, in each Samba block, the bi-directional state space modeling module has both feedforward and backforward SSM, where the selective scan mechanism allows to handle the patches in a recurrent manner. The input patches are traversed along two different scanning paths (horizontal and vertical), and each sequence is independently processed by the SSM. Subsequently, the results are merged to construct a 2D feature map as the final output.

By a bidirectional design, the severity information can be transported to each patch. The local-to-global transportation of severity information plays a vital role in the whole process, especially in the selective mechanism. With the guidance of global severity awareness, the update of hidden states can selectively ignore information about low-level lesions, primarily preserving information about the most severe lesions. Specifically, to encode the 2D images, we follow the design of vision Mamba [69] which processes the input features in the forward and backward directions. As illustrated in Fig. 2, the outputs are gated and added together, while there is a skip connection before input to the EM-based state recalibration module.

3.3 EM-based State Recalibration

Another core issue in cross-domain disease grading is the domain generalization ability of the model. Both the intermediate features and the input-dependent parameters in Mamba are affected by the cross-domain variance. To reduce the impact of domain shift, we aim to map the features into a more compact space by feature recalibration. Specifically, the feature distribution of background and grading-related lesions is modeled by a Gaussian Mixture Model (GMM) [42], given by

$$p(\boldsymbol{f}_n) = \sum_{k=1}^{K} z_{nk} \mathcal{N}(\boldsymbol{f}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{3}$$

where K is the total number of the Gaussian models, $\boldsymbol{f}_n$ is the feature embedding of the n-th patch in image $\boldsymbol{x}$, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ denote the mean and covariance of the k-th Gaussian basis, respectively. For simplicity, we set $\boldsymbol{\Sigma}_k$ as the identity matrix $\boldsymbol{I}$. For easy computation, the mixing coefficients of GMM are left out and the exponential inner dot kernel is used. After that, each Gaussian basis is represented by $\boldsymbol{\mu}_k$, which is called severity base in this paper. These bases are learnable parameters to capture the representation of lesions. In the recalibration process, instead of directly reconstructing the features based on the bases, we estimate the lesion distribution of each image which is initialized by the severity bases. This is to prevent the loss of useful information during the compression. Concretely, we adopt the EM algorithm to estimate the GMM of each image.

Within each iteration, we first estimate z_{nk} in the E-step, which denotes the responsibility of the k-th basis to $\boldsymbol{f}_n$. Here we have $1 \leq n \leq N$ and $1 \leq k \leq K$. The posterior probability of $\boldsymbol{f}_n$ given $\boldsymbol{\mu}_k$ can be formulated as $p(\boldsymbol{f}_n | \boldsymbol{\mu}_k) = \mathcal{K}(\boldsymbol{f}_n, \boldsymbol{\mu}_k)$ by a kernel function $\mathcal{K}$. Consequently, estimating the responsibility can be re-formulated into a more general form, given by

$$z_{nk} = \frac{\mathcal{K}(\boldsymbol{f}_n, \boldsymbol{\mu}_k)}{\sum_{i=1}^{K} \mathcal{K}(\boldsymbol{f}_n, \boldsymbol{\mu}_i)}, \tag{4}$$

where for simplicity we directly use the exponential inner dot $\exp(\boldsymbol{f}^{\mathrm{T}}\boldsymbol{\mu})$ as the kernel function.

Given the estimated $\mathbf{Z}^t$, the severity base likelihood maximization, functioning as the M-step, is realized by updating $\boldsymbol{\mu}$. As the bases are supposed to be aligned to the embedding space of each image, the weighted sum is used to update the bases, given by

$$\boldsymbol{\mu}_k^{t+1} = \frac{1}{\sum_{m=1}^{N_p} z_{mk}^t} \sum_{n=1}^{N_p} z_{nk}^t \boldsymbol{f}_n, \tag{5}$$

where t refers to the t-th iteration and N_p denotes the number of patch embeddings.

Assume that the E-step and M-step execute alternately for T_c iterations and the convergence criterion has been reached [14]. The final $\boldsymbol{\mu}^{T_c}$ and $\mathbf{Z}^{T_c}$ are used to recalibrate the image feature $\boldsymbol{F}$, resulting in $\tilde{\boldsymbol{F}}$. Here $\boldsymbol{\mu}^{T_c} = \{\boldsymbol{\mu}_k^{T_c}\}$ and $\mathbf{Z}^{T_c} = \{z_{nk}^{T_c}\}$ refer to the Gaussian basis and the responsibilities of all the patch embeddings from a sample, respectively. This process is mathematically computed as

$$\tilde{\mathbf{F}} = \mathbf{Z}^{T_c}\boldsymbol{\mu}^{T_c}. \tag{6}$$

Then, the recalibrated feature $\tilde{\mathbf{F}}$ is fed into the next Samba module. During this process, grading-related features are mapped to a more compact space, while style differences introduced by image sources are partially removed. Consequently, the critical information transportation within the Mamba model can be more stable in unseen target domains.

To alleviate the potential unstable issue, moving averaging is adapted to update the bases $\boldsymbol{\mu}^0$ during the training process. After the T-th iteration, the generated $\boldsymbol{\mu}^T$ is first averaged over a mini-batch to get $\overline{\boldsymbol{\mu}}^T$. Then, the update of $\boldsymbol{\mu}^0$ with momentum $\alpha \in [0, 1]$ is given by

$$\boldsymbol{\mu}^0 \longleftarrow \alpha\boldsymbol{\mu}^0 + (1-\alpha)\overline{\boldsymbol{\mu}}^T. \tag{7}$$

3.4 Theoretical Analysis

Consider the source domain $\mathcal{D}_1 \sim P(\tilde{\mathbf{X}}^{(1)})$ and a certain unseen target domain $\mathcal{D}_k \sim P(\tilde{\mathbf{X}}^{(k)})$, where $k = 2, \cdots, K$. Given a hypothesis $h \in \mathcal{H}$, according to the domain adaptation/generalization bound theory [5, 2], the relation between the target risk $\epsilon_{\mathcal{D}_k}(h)$ and the source risk $\epsilon_{\mathcal{D}_1}(h)$ can be modeled by a relation inequality, given by

$$\epsilon_{\mathcal{D}_k}(h) \leq \epsilon_{\mathcal{D}_1}(h) + d_{\mathcal{H}\Delta\mathcal{H}}\left(P(\tilde{\mathbf{X}}^{(1)}), P(\tilde{\mathbf{X}}^{(k)})\right) + \min_{P(\tilde{\mathbf{X}})\in P(\tilde{\mathbf{X}}^{(1)}), P(\tilde{\mathbf{X}}^{(k)})} \mathbb{E}\left[|h_{\mathcal{D}_1}(x) - h_{\mathcal{D}_k}(x)|\right], \tag{8}$$

where $d_{\mathcal{H}\Delta\mathcal{H}}\left(P(\tilde{\mathbf{X}}^{(1)}), P(\tilde{\mathbf{X}}^{(k)})\right)$ denotes the distribution gap between the source domain and an unseen target domain, and the right-most term refers to minimal total risk over both domains. In other words, the risk of the proposed Samba on the target domain is bounded by the source domain.

4 Experiments

4.1 Datasets & Evaluation Protocols

Cross-domain Fatigue Fracture Grading Benchmark [31] consists of a total number of 1,785 normal X-ray images and 940 X-ray images with fatigue fracture. They are collected from two hospitals with different types of sensors, which we denote as Domain-1 and Domain-2, respectively. These fatigue fracture images were graded into four stages by three physicians according to the severity level. For simplicity, we denote the grades (including the normal grade) from level-1 to level-5.

Cross-domain Breast Cancer Grading Benchmark consists of a total of 3644 H&E stained breast invasive ductal carcinoma pathological images from two domains.[2] The first domain contains 2,486 images under the 20× magnification (denoted as Domain-1). The second domain contains 1,158 images under the 40× magnification (denoted as Domain-2). Different magnifications make the image appearance dramatically different. For each experiment setting, one is used as the source

[2] `https://github.com/YANRUI121/Breast-cancer-grading`

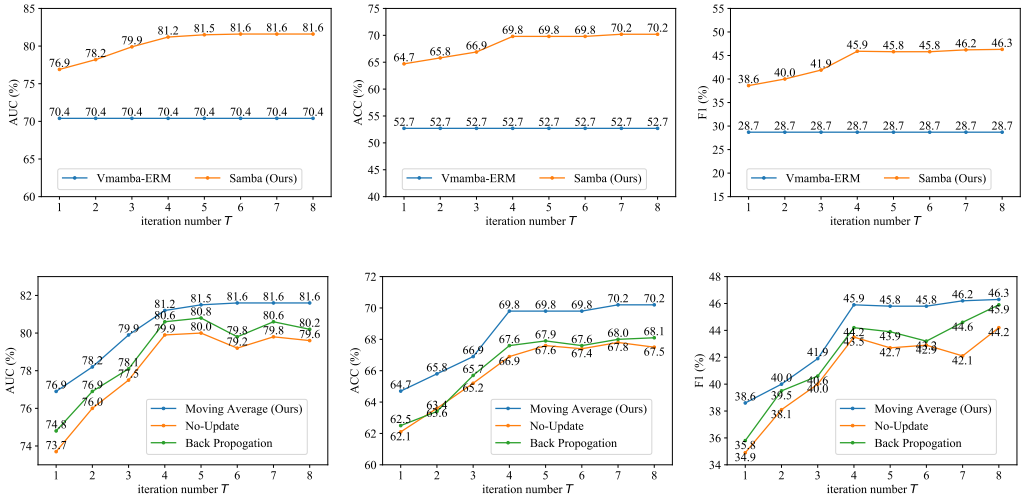


Figure 3: Impact of iteration number T (top row) and severity base (bottom row) updating approaches on generalized medical grading performance. Evaluation metrics AUC, ACC and F1 are presented in percentage (%), from the first to the third column. Domain-1 and Domain-2 in the Fatigue Fracture Grading Benchmark are used as the source and unseen target domain, respectively.

domain and the other is used as the unseen target domain. According to the severity of breast invasive ductal carcinoma, three grades, namely, rare, frequent and abundant, are annotated. For simplicity, we denote them from level-1 to level-3, respectively.

Cross-domain Diabetic Retinopathy Grading Benchmark consists of six DR retinal image datasets, namely, DeepDR [33], Messidor [1], IDRID [40], APTOS [3], FGADR [67], and RLDR [49]. Following recent work [11], the single-domain generalization protocol is adapted. Specifically, one of the above six datasets is used as the source domain, and all the rest datasets are used as unseen target domains. Following [11], two extra large-scale datasets, DDR [29] and EyePACS [15] are used to enrich the source domain for each experiment setting. The development of DR is graded into five levels according to the severity, namely, normal, mild nonproliferative diabetic retinopathy (npdr), moderate npdr, severe npdr and pdr. For simplicity, we denote them from level-1 to level-5.

Under the single-domain generalization protocol, three most common evaluation metrics for grading are used, namely, area under the curve (AUC), accuracy (ACC), and F1-score (F1).

4.2 Results on Fatigue Fracture Grading Benchmark

We conduct extensive ablation experiments to study the impact of iteration number T and optimization of severity basis $\boldsymbol{\mu}$ on the unseen target domain. Images from the first clinical center (Domain-1) are used as the source domain, while images from the second clinical center (Domain-2) are used as the unseen target domain for testing. The vanilla Mamba [35] under the Empirical Risk Minimization (ERM) is the baseline.

Iteration Number T. In the EM algorithm, the iteration number T plays an important role, because it implements the approximation by iteratively conduct the E and M step. Keeping other hyper-parameters and module designs the same, we report the results when the iteration number T of the EM algorithm varies from 1 to 8. The top row of Fig. 3 shows how T impacts the AUC, ACC and F1-score on the unseen target domain. Notice that, the VMamba-ERM baseline does not have EM-based state recalibration. Therefore, the performance of Vmamba-ERM is consistent to T. A too-small T does not reach the convergence criterion, and reduces the effectiveness of feature recalibration. Therefore, a clear performance decline on all the metrics is observed. On the other hand, when T is too large, the representation ability saturates, resulting in little performance improvement, while wasting computation resources.

Severity Base Update. We further study how different optimization approaches of the severity base $\boldsymbol{\mu}_k$ impact the generalization performance on unseen target domain. We study three different settings, namely, no update, only back propogation, and moving average (Eq. 7). The bottom row of Fig. 3

Table 1: Effectiveness of the proposed Samba on recurrent patch modeling. Domain-1 and Domain-2 in the Fatigue Fracture Grading Benchmark are used as the source and unseen target domain, respectively. Metrics presented in percentage (%).

Method	ACC ↑	AUC ↑	F1 ↑
LSTM [22]	39.8	50.2	18.6
UR-LSTM [18]	43.3	61.8	20.9
UR-GRU [18]	45.7	65.1	22.4
ViT [48]	50.0	69.3	26.5
VMamba [69]	52.7	70.4	28.7
Samba	**76.2**	**81.5**	**45.8**

Table 2: Ablation study on each component. BSSM: Bi-directional State Space Modeling; ESR: EM-based State Recalibration. Experiments on the Fatigue Fracture Grading Benchmark. Domain-1 (×20)/Domain-2 (×40) is used as source/target domain. Metrics in percentage (%).

Components			Evaluation Metric		
VMamba	BSSM	ESR	ACC	AUC	F1
✓	✗	✗	52.7	70.4	28.7
✓	✓	✗	57.9	72.1	33.6
✓	✓	✓	**76.2**	**81.5**	**45.8**

Table 3: Category-wise performance and computational cost comparison between VMamba-ERM and the proposed Samba. Experiments are conducted on the DG Breast Cancer Grading Benchmark. Domain-1 (×20)/Domain-2 (×40) is used as source/target domain. Metrics in percentage (%).

Method	Backbone	Computation		Domain-1 as Source			
		GFLOPs	Para.	level-1	level-2	level-3	avg.
ERM	VMama-T	3.7	32.7	22.1	51.5	36.1	40.4
Samba		5.5	32.7	**40.5**	**70.7**	**42.0**	**54.8**
ERM	VMama-S	7.9	63.4	26.7	60.6	38.1	50.1
Samba		11.3	63.4	**47.1**	**71.5**	**43.7**	**56.1**
ERM	VMama-B	14.0	112.4	27.8	75.4	38.2	54.9
Samba		19.6	112.4	**44.8**	**82.5**	**45.2**	**60.5**

Table 4: Impact of the number of components K in GMM. Experiments are conducted on the DG Breast Cancer Grading Benchmark. Domain-1 (×20)/Domain-2 (×40) is used as source/target domain. Metrics presented in percentage (%).

K value	ACC ↑	AUC ↑	F1 ↑
16	58.6	70.0	56.0
32	59.2	71.1	57.2
48	60.4	72.0	58.9
64	**60.5**	**72.3**	**59.1**
96	60.4	72.2	58.8
128	59.5	71.0	57.9

shows the results of the above three settings under a variety of iteration number T. Using moving average to update the severity base $\boldsymbol{\mu}_k$ is able to improve the performance substantially. It may be explained that the proposed state recalibration is differentiable, thereby enabling the application of back-propagation to update $\boldsymbol{\mu}_0$. However, the stability of the update cannot be guaranteed due to the EM iterations. Moving average can update $\boldsymbol{\mu}_0$ to avoid collapse.

Effectiveness on Recurrent Patch Modeling. The proposed Samba realizes the recurrent patch modeling by harnessing the selective state space model. To demonstrate its effectiveness compared with other recurrent or long-context based representation learning methods, we compare the proposed Samba with vanilla VMamba [69], Vision Transformer [48], LSTM [22] and UR-LSTM [18]. Table 1 shows that the proposed Samba has a stronger generalization performance on the unseen target domain, noticeably outperforming the second-best by 23.5%, 11.1% and 17.1% in terms of accuracy, AUC and F1-score, respectively.

Ablation Studies on Each Component. On top of the VMamba baseline, two key components, namely, Bi-directional State Space Modeling (BSSM) and EM-based State Recalibration (ESR), are evaluated. The experiments are conducted on the DG Fatigue Fracture Grading Benchmark. Domain-1/Domain-2 is used as the source/target domain, respectively. The results are reported in Table 2. It is observed that BSSM contributes to an ACC, AUC and F1 improvement of 5.2%, 1.7% and 4.9%, respectively. ESR contributes to an ACC, AUC and F1 improvement of 18.3%, 9.4% and 12.2%, respectively.

4.3 Results on Breast Cancer Grading Benchmark

Grade-wise Improvement Analysis. We provide a break-down analysis on the grade-wise performance of the proposed Samba and the baseline, *i.e.*, VMamba under the empirical risk minimization (ERM). Table 3 reports the performance. The proposed Samba shows a significant performance improvement on each grade level. Especially, on level-1, level-2 and level-3, the accuracy improvement over the ERM baseline is 17.0%, 7.1% and 7.0%, respectively. Compared to VMamba-ERM baseline, the EM-based State Recalibration in Samba models the feature distribution of lesions via Gaussian Mixture Models with learnable severity bases, and re-estimates by E-M algorithm. The grading features are mapped to a more compact space, and are more stable on unseen target domains.

Ablation Studies on the number of Gaussian components K. We study how the number of components K impacts the generalization ability on the unseen target domain. By default K is set

to 64, and we further test the performance when K is set to 16, 32, 48 and 96, respectively. The results are reported in Table 4. When K is set to 64, the proposed Samba achieves the best grading performance, *i.e*., 60.5%, 72.3% and 59.1% in terms of ACC, AUC and F1-Score, respectively.

Scalability to Clinical Computational Pathology. The domain shift in the Cross-domain Breast Cancer Grading Benchmark is from a machine learning perspective, and only handles the magnification difference. However, from a clinical perspective, the computational pathology has to handle the domain shift from not only the magnification difference, but also the staining procedure. However, most existing clinical computational pathology datasets only support the classification task, *i.e*., separating *tumor* category from *normal* category, which is not strongly relevant to our grading task. Appendix A.4 studies the performance of the proposed Samba along with the vanilla VMamba baseline on the CAMELYON17 dataset [3] for a clinical sanity check.

4.4 Results on Diabetic Retinopathy Grading Benchmark

Comparison with State-of-the-art. We compare the proposed Samba with methods from three primary categories: 1) generic domain generalization methods, including Mixup [61], MixStyle [64], DDAIG [63], ATS [55], Fishr [41], and MDLT [56]; 2) state-of-the-art DR grading methods, which focus on DR grading without explicitly addressing domain generalization, including GREEN [34], CABNet [21], Swin-Transformer [36] and MIL-ViT [7]; 3) domain-generalized DR grading methods, including DRGen [4] and GDRNet [11]. Additionally, the vanilla Mamba [35] results under Empirical Risk Minimization (ERM) are provided as a baseline reference. By default, the results are cited directly from [11].

Table 5 presents a comparison between Samba and existing methods within the context of single-domain generalization. Notably, for DG grading tasks, the metric of accuracy and F1-score are more meaningful than AUC, as the AUC can be made artificially high due to the large amount of negative samples belonging to other stages. Therefore, we only involve ACC and F1 for comparison.

The proposed Samba achieves a substantial improvement over state-of-the-art domain generalized DR grading methods. Especially, on APTOS, DeepDR, FGADR, Messidor and RLDR, it outperforms the second-best in terms of ACC and F1 by 5.2% and 2.2%, 27.2% and 5.7%, 60.7% and 31.3%, 6.7% and 1.3%, 28.3% and 4.7%, respectively. The significant improvement on FGDAR may be explained that it has a different severity-level sample distribution than other datasets. The samples without DR (level-1) only occupy only 5.5% among all the training samples, which are far less than others (*e.g*., level-1 samples occupy 49.3% in APTOS). Therefore, existing methods may overfit other severity levels and underfit level-1. In contrast, the selective scan mechanism of the Vmamba-ERM and Samba is robust to this severity distribution shift. The EM state re-calibration in Samba makes the feature space more compact, and improves the generalization.

Additionally, Samba shows a marked improvement over the baseline VMamba model under Empirical Risk Minimization. Especially, on APTOS, DeepDR, FGADR, IDRID, Messidor and RLDR, it outperforms the VMamba-ERM baseline in terms of ACC and F1 by 1.3% and 1.7%, 2.2% and 2.1%, 3.0% and 1.6%, 3.7% and 2.6%, 7.3% and 2.7%, 7.4% and 3.4%, respectively. These results demonstrate its effectiveness in handling domain gap.

Understanding Recurrent Patch Modeling. We validate if the proposed recurrent patch modeling can store and transport the lesion information. An intuitive way is to inspect the correlation between the patch embeddings before and after the recurrent patch modeling. Therefore, we extract the patch embeddings before and after the fourth block. We compute the correlation matrix between the patch-wise embeddings and visualize the results in Fig. 4. After processed by the Recurrent Patch Modeling module, more regions in the correlation matrix have higher responses. Specifically, after passing through certain high-response positions, the relevant information is transmitted to the subsequent patches in the forward direction. The high-response patches have grade-related lesions and the information is transported in the recurrent process.

t-SNE Visualization. To assess the generalization capacity of the proposed Samba, we analyze the feature distribution across the source and unseen target domains using t-SNE visualization in Fig. 5, which compares the t-SNE plots of the ERM baseline (left) and the proposed Samba (right). The results indicate that, after the EM-based state feature calibration, the proposed Samba enables

[3] `https://camelyon17.grand-challenge.org/`

Table 5: Performance comparison of the proposed Samba and existing domain generalized DR grading methods under the single-domain generalization protocol. Evaluation metrics include ACC and F1 (in percentage %). Top three results are highlighted as **best**, second and third, respectively.

Method	APTOS		DeepDR		FGADR		IDRID		Messidor		RLDR		Average	
	ACC↑	F1↑	ACC↑	F1↑	ACC↑	F1↑	ACC↑	F1↑	ACC↑	F1↑	ACC↑	F1↑	ACC↑	F1↑
ResNet-50 based:														
Mixup [61]	49.4	30.2	49.7	33.3	5.8	7.4	64.0	32.6	63.0	32.6	27.7	27.0	43.3	27.2
MixStyle [64]	48.8	25.0	32.0	14.6	7.0	7.9	53.5	19.4	57.6	16.8	18.3	6.4	36.2	15.0
GREEN [34]	52.6	33.3	44.6	31.1	5.7	6.9	60.7	33.0	54.5	33.1	31.9	27.8	41.7	27.5
CABNet [21]	52.2	30.8	55.4	32.0	6.1	7.5	62.7	31.7	63.8	35.3	23.0	25.4	43.8	27.2
DDAIG [63]	48.7	31.6	38.5	29.7	5.0	5.5	60.2	33.4	69.1	35.6	25.4	23.5	41.2	26.7
ATS [55]	51.7	32.4	52.4	33.5	5.3	5.7	66.6	30.6	64.8	32.4	24.2	23.9	44.2	26.4
Fishr [41]	61.7	31.0	61.0	30.1	6.0	7.2	48.0	30.6	52.0	33.8	19.3	21.3	41.3	25.7
MDLT [56]	53.3	32.4	50.2	33.7	7.1	7.8	61.7	32.4	58.9	34.1	29.0	30.0	43.4	28.4
DRGen [4]	60.7	35.7	39.4	31.6	6.8	8.4	67.7	30.6	64.5	37.4	19.0	21.2	43.0	27.5
GDRNet [11]	52.8	35.2	40.0	35.0	7.5	9.2	70.0	35.1	65.7	40.5	44.3	37.9	46.7	32.2
ViT based:														
MIL-ViT [7]	61.8	36.8	38.2	36.3	8.7	9.3	68.6	31.1	67.7	40.7	28.1	34.5	45.5	31.5
Swin-T [36]	64.0	36.7	31.0	32.7	6.0	7.8	**70.4**	38.1	65.6	39.8	27.5	34.5	44.1	31.6
VMamba based:														
ERM	64.6	36.2	65.0	38.6	65.2	38.9	65.2	39.1	65.1	39.1	65.2	39.2	65.1	38.5
Samba (Ours)	**65.9**	**37.9**	**67.2**	**40.7**	**68.2**	**40.5**	68.9	**41.7**	**72.4**	**41.8**	**72.6**	**42.6**	**69.2**	**40.9**

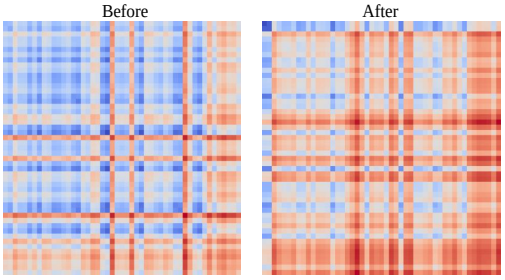


Figure 4: The correlation matrix of each patch embedding before and after processed by the recurrent patch modeling in the forward direction, denoted as 'Before' and 'After' respectively. The higher correlation, the more red a cell is.

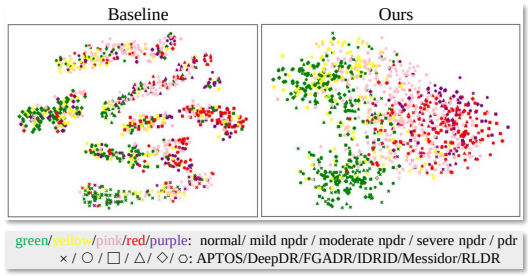


Figure 5: T-SNE visualization of the feature space from the ERM baseline (left), and the proposed Samba (right). APTOS is chosen as the source domain and the rest datasets are used for as target domains.

feature embeddings from different domains to achieve a more uniform distribution, thereby reducing the domain gap. This improved uniformity in the feature space suggests that Samba can enhance generalization, contributing to better performance on unseen domains.

5 Conclusion

In this paper, we aimed to tackle a practical but challenging task, learning domain generalized medical image grading. We mainly focused on two issues: the identification of decisive lesions and the impact caused by inter-domain differences. The proposed severity-aware recurrent modeling adopts a state space model to store and transport the severity information from local to global. To further mitigate the impact of cross-domain variants, an EM-based state recalibration was designed to map the patch embeddings into a compact space. The proposed method can be used in a variety of disease grading scenarios, providing an effective tool for automatic medical image analysis.

Limitation Discussion & Broader Societal Impact. The feature distribution of lesions is modeled by the Gaussian mixture model and estimated by the Expectation-Maximization algorithm. However, when the training source domain has severe class imbalance, the estimated probability distribution by the proposed Samba may not necessarily reflect the domain-agnostic lesion distribution. Nevertheless, the proposed method can be combined with other techniques specifically designed for addressing class imbalance. The proposed method advances the versatility of automatic disease diagnosis, which benefits the human well beings. We do not envision negative societal impact.

Acknowledgments and Disclosure of Funding. This work was supported by the Science and Technology Major Project of Guangxi (AA22096030 and AA22096032), and National Key R&D Program of China under Grant (2020AAA0109500 and 2020AAA0109501).

References

[1] Michael D. Abràmoff, Yiyue Lou, Ali Erginay, and et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Investigative Ophthalmology Visual Science*, 57(13):5200–5206, 2016.

[2] Isabela Albuquerque, João Monteiro, Mohammad Darvishi, Tiago H Falk, and Ioannis Mitliagkas. Generalizing to unseen domains via distribution matching. *arXiv preprint arXiv:1911.00804*, 2019.

[3] APTOS. APTOS 2019 blindness detection, 2019. URL `https://www.kaggle.com/c/aptos2019-blindness-detection`. Accessed on February 20, 2024.

[4] Mostafa Atwany and Muhammad Yaqub. DRGen: Domain generalization in diabetic retinopathy classification. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 635–644. Springer, 2022.

[5] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

[6] Qi Bi, Shuang Yu, Wei Ji, Cheng Bian, Lijun Gong, Hanruo Liu, Kai Ma, and Yefeng Zheng. Local-global dual perception based deep multiple instance learning for retinal disease classification. In *Medical Image Computing and Computer Assisted Intervention*, pages 55–64, 2021.

[7] Qi Bi, Xu Sun, Shuang Yu, Kai Ma, Cheng Bian, Munan Ning, Nanjun He, Yawen Huang, Yuexiang Li, Hanruo Liu, et al. MIL-ViT: A multiple instance vision Transformer for fundus image classification. *Journal of Visual Communication and Image Representation*, 97:103956, 2023.

[8] Qi Bi, Jingjun Yi, Hao Zheng, Wei Ji, Yawen Huang, Yuexiang Li, and Yefeng Zheng. Learning generalized medical image segmentation from decoupled feature queries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 810–818, 2024.

[9] Qi Bi, Shaodi You, and Theo Gevers. Learning generalized segmentation for foggy-scenes by bi-directional wavelet guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 801–809, 2024.

[10] Qi Bi, Shaodi You, and Theo Gevers. Generalized foggy-scene semantic segmentation by frequency decoupling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1389–1399, 2024.

[11] Haoxuan Che, Yuhan Cheng, Haibo Jin, and Hao Chen. Towards generalizable diabetic retinopathy grading in unseen domains. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 430–440. Springer, 2023.

[12] Yimeng Chen, Tianyang Hu, Fengwei Zhou, Zhenguo Li, and Zhi-Ming Ma. Explore and exploit the diverse knowledge in model zoo for domain generalization. In *International Conference on Machine Learning*, pages 4623–4640, 2023.

[13] Rui Dai, Yonggang Zhang, Zhen Fang, Bo Han, and Xinmei Tian. Moderately distributional exploration for domain generalization. In *International Conference on Machine Learning*, pages 6786–6817, 2023.

[14] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[15] EyePACS. Kaggle EyePACS dataset, Accessed 20 February 2024. URL `https://paperswithcode.com/dataset/kaggle-eyepacs`.

[16] Milena Gazdieva, Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Extremal domain translation with neural optimal transport. *Advances in Neural Information Processing Systems*, 36, 2023.

[17] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[18] Albert Gu, Caglar Gulcehre, Thomas Paine, Matt Hoffman, and Razvan Pascanu. Improving the gating mechanism of recurrent neural networks. In *International Conference on Machine Learning*, pages 3800–3809, 2020.

[19] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in Neural Information Processing Systems*, 34:572–585, 2021.

[20] Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

[21] An He, Tianyi Li, Ning Li, Ke Wang, and Huazhu Fu. CABNet: Category attention block for imbalanced diabetic retinopathy grading. *IEEE Transactions on Medical Imaging*, 40(1):143–153, 2020.

[22] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[23] Shishuai Hu, Zehui Liao, Jianpeng Zhang, and Yong Xia. Domain and content adaptive convolution based multi-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(1):233–244, 2023.

[24] Wei Ji, Jingjing Li, Shuang Yu, Miao Zhang, Yongri Piao, Shunyu Yao, Qi Bi, Kai Ma, Yefeng Zheng, Huchuan Lu, et al. Calibrated rgb-d salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9471–9481, 2021.

[25] Wei Ji, Shuang Yu, Junde Wu, Kai Ma, Cheng Bian, Qi Bi, Jingjing Li, Hanruo Liu, Li Cheng, and Yefeng Zheng. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12341–12351, 2021.

[26] Wei Ji, Jingjing Li, Qi Bi, Tingwei Liu, Wenbo Li, and Li Cheng. Segment anything is not always perfect: An investigation of sam on different real-world applications. *Machine Intelligence Research*, 21:617–630, 2024.

[27] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960.

[28] Jingjing Li, Wei Ji, Qi Bi, Cheng Yan, Miao Zhang, Yongri Piao, Huchuan Lu, et al. Joint semantic mining for weakly supervised rgb-d salient object detection. *Advances in Neural Information Processing Systems*, 34:11945–11959, 2021.

[29] Tianyi Li, Yang Gao, Kai Wang, Shuai Guo, Hong Liu, and Hong Kang. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences*, 501:511–522, 2019.

[30] Xiaomeng Li, Xiaowei Hu, Lequan Yu, Lei Zhu, Chi-Wing Fu, and Pheng-Ann Heng. CANet: cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading. *IEEE Transactions on Medical Imaging*, 39(5):1483–1493, 2019.

[31] Yuexiang Li, Yanping Wang, Guang Lin, Yawen Huang, Jingxin Liu, Yi Lin, Dong Wei, Qirui Zhang, Kai Ma, Zhiqiang Zhang, et al. Triplet-branch network with contrastive prior-knowledge embedding for disease grading. *Artificial Intelligence in Medicine*, page 102801, 2024.

[32] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. FedDG: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1013–1023, 2021.

[33] Rui Liu, Xintao Wang, Qi Wu, Lin Dai, Xiaohui Fang, and et al. DeepDRID: Diabetic retinopathy—grading and image quality estimation challenge. *Patterns*, 3(6):100512, 2022.

[34] Shujian Liu, Lihui Gong, Kai Ma, and Yalin Zheng. GREEN: A graph residual re-ranking network for grading diabetic retinopathy. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 585–594, 2020.

[35] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024.

[36] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[37] Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality inspired representation learning for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8046–8056, 2022.

[38] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International Conference on Machine Learning*, pages 7313–7324, 2021.

[39] Xi Peng, Fengchun Qiao, and Long Zhao. Out-of-domain generalization from a single source: An uncertainty quantification approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46 (3):1775–1787, 2022.

[40] Piyush Porwal, Sanket Pachade, Rakesh Kamble, Milind Kokare, Ganesh Deshmukh, and et al. Indian diabetic retinopathy image dataset (IDRiD): A database for diabetic retinopathy screening research. *Data*, 3(3):25, 2018.

[41] Ali Rame, Cedric Dancette, and Matthieu Cord. FishR: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, pages 18347–18377, 2022.

[42] Sylvia Richardson and Peter J Green. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 59(4):731–792, 1997.

[43] Lianghe Shi and Weiwei Liu. Adversarial self-training improves robustness and generalization for gradual domain adaptation. *Advances in Neural Information Processing Systems*, 36, 2023.

[44] Ruoxian Song, Peng Cao, Jinzhu Yang, Dazhe Zhao, and Osmar R Zaiane. A domain adaptation multi-instance learning for diabetic retinopathy grading on retinal images. In *IEEE International Conference on Bioinformatics and Biomedicine*, pages 743–750, 2020.

[45] Rui Sun, Yihao Li, Tianzhu Zhang, Zhendong Mao, Feng Wu, and Yongdong Zhang. Lesion-aware Transformers for diabetic retinopathy grading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10938–10947, 2021.

[46] Wenqiang Tang, Zhouwang Yang, and Yanzhi Song. Disease-grading networks with ordinal regularization for medical imaging. *Neurocomputing*, 545:126245, 2023.

[47] Peifeng Tong, Wu Su, He Li, Jialin Ding, Zhan Haoxiang, and Song Xi Chen. Distribution free domain generalization. In *International Conference on Machine Learning*, pages 34369–34378, 2023.

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[49] Qiang Wei, Xiang Li, Weiyang Yu, Xiang Zhang, Yu Zhang, Bin Hu, Biao Mo, Dawei Gong, Nan Chen, Dian Ding, and et al. Learn to segment retinal lesions and beyond. In *International Conference on Pattern Recognition*, pages 7403–7410. IEEE, 2021.

[50] Xiaoping Wu, Ni Wen, Jie Liang, Yu-Kun Lai, Dongyu She, Ming-Ming Cheng, and Jufeng Yang. Joint acne image grading and counting via label distribution learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10642–10651, 2019.

[51] Zhenbang Wu, Huaxiu Yao, David Liebovitz, and Jimeng Sun. An iterative self-learning framework for medical domain generalization. *Advances in Neural Information Processing Systems*, 36, 2023.

[52] Anqi Xiao, Biluo Shen, Xiaojing Shi, Zhe Zhang, Zeyu Zhang, Jie Tian, Nan Ji, and Zhenhua Hu. Intraoperative glioma grading using neural architecture search and multi-modal imaging. *IEEE Transactions on Medical Imaging*, 41(10):2570–2581, 2022.

[53] Mixue Xie, Shuang Li, Rui Zhang, and Chi Harold Liu. Dirichlet-based uncertainty calibration for active domain adaptation. In *International Conference on Learning Representations*, 2023.

[54] Geng-Xin Xu, Chen Liu, Jun Liu, Zhongxiang Ding, Feng Shi, Man Guo, Wei Zhao, Xiaoming Li, Ying Wei, Yaozong Gao, et al. Cross-site severity assessment of COVID-19 from CT images via domain adaptation. *IEEE Transactions on Medical Imaging*, 41(1):88–102, 2021.

[55] Fan-E Yang, Yu-Chiang Cheng, Zhan-Yuan Shiau, and Yu-Chiang Frank Wang. Adversarial teacher-student representation learning for domain generalization. In *Advances in Neural Information Processing Systems*, volume 34, pages 19448–19460, 2021.

[56] Yue Yang, Hao Wang, and Dina Katabi. On multi-domain long-tailed recognition, imbalanced domain generalization and beyond. In *European Conference on Computer Vision*, pages 57–75. Springer, 2022.

[57] Qinghao Ye, Xiyue Shen, Yuan Gao, Zirui Wang, Qi Bi, Ping Li, and Guang Yang. Temporal cue guided video highlight detection with low-rank audio-visual fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7950–7959, 2021.

[58] Jingjun Yi, Qi Bi, Hao Zheng, Haolan Zhan, Wei Ji, Yawen Huang, Shaoxin Li, Yuexiang Li, Yefeng Zheng, and Feiyue Huang. Hallucinated style distillation for single domain generalization in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 438–448, 2024.

[59] Jingjun Yi, Qi Bi, Hao Zheng, Haolan Zhan, Wei Ji, Yawen Huang, Yuexiang Li, and Yefeng Zheng. Learning spectral-decomposited tokens for domain generalized semantic segmentation. In *ACM Multimedia 2024*, 2024.

[60] Zhongqi Yue, Qianru Sun, and Hanwang Zhang. Make the U in UDA matter: Invariant consistency learning for unsupervised domain adaptation. *Advances in Neural Information Processing Systems*, 36, 2023.

[61] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[62] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[63] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13025–13032, 2020.

[64] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with MixStyle. In *International Conference on Learning Representations*, 2021.

[65] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2022.

[66] Yi Zhou, Lei Huang, Tao Zhou, and Ling Shao. CCT-net: category-invariant cross-domain transfer for medical single-to-multiple disease diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8260–8270, 2021.

[67] Yufan Zhou, Bin Wang, Lei Huang, Shuguang Cui, and Ling Shao. A benchmark for studying diabetic retinopathy: Segmentation, grading, and transferability. *IEEE Transactions on Medical Imaging*, 40(3): 818–828, 2020.

[68] Ziqi Zhou, Lei Qi, and Yinghuan Shi. Generalizable medical image segmentation via random amplitude mixup and domain-specific image restoration. In *European Conference on Computer Vision*, pages 420–436, 2022.

[69] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision Mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.

A Appendix / supplemental material

A.1 Severity Base Normalization

During the iteration steps $t = 1, \cdots, T$, the severity base $\boldsymbol{\mu}_k^{(1),t}$ may not deviate too much from each other, which otherwise can lead to collapse when back propagation. We study the scenarios when no normalization, L-1 normalization and L-2 normalization are used on these severity basis. Fig. 6 shows the results of the above three settings when under a variety of iteration number T on the unseen target domain. L-2 normalization achieves the best performance on all the metrics, especially when T becomes larger.

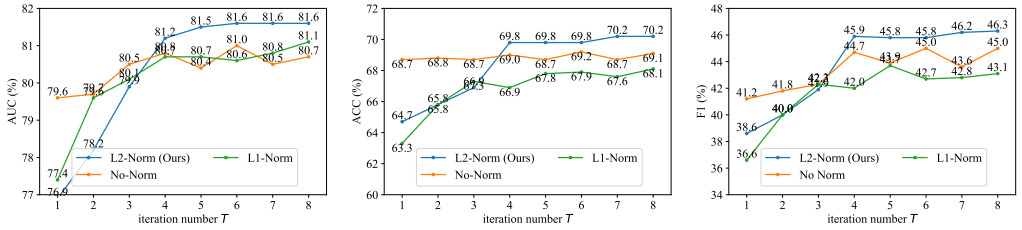


Figure 6: Impact of severity base normalization on generalized medical grading performance. Evaluation metrics AUC, ACC and F1 are presented in percentage (%). Domain-1 and Domain-2 in the Fatigue Fracture Grading Benchmark are used as the source and unseen target domain, respectively. Better zoom in to view.

A.2 Understanding Recurrent Patch Modeling

Fig. 4 in the main text only visualizes the correlation between the patch embeddings before and after the recurrent patch modeling when trained on the APTOS dataset. Here we further demonstrate the results from FGADR, IDRID, Messidor and RLDR. They are visualized in Fig. 7a, b, c and d, respectively. On all these datasets, we can observe a common pattern. After processed by the Recurrent Patch Modeling module, more cells in the correlation matrix have higher response. Usually, a handful of the patches inside the image have grade-related lesions. After the processing of our module, the information of these grade-related lesions is transported to other patches. It allows the model to perceive a more global-wise representation. Consequently, more patches that contain the grade-related lesion information are activated, and more cells are highly responded in the correlation matrix.

A.3 Visualize and Understand the Severity Level

We model the relation between patch embedding from SSM and severity level by drawing inspiration from the class activation map (CAM) mechanism [62]. Specifically, we take the patch embeddings from the last Samba block as input to generate the per-level severity activation patterns. Then, the activated severity patterns are displayed on the original images. We use FGADR as the unseen target domain. The results are shown in Fig. 8, where the activated patches are highlighted in blue boxes. From the first to the fifth row, the samples from level-1 to level-5 are provided accordingly. From the first to the fifth column, the patch activation maps from level-1 to level-5 are displayed. Notice that, as level-1 refers to the normal scenario, each sample has activations on level-1, meaning some patches are normal.

A.4 Application to Computational Pathology Classification

It is important to note that the domain shift in the Breast Cancer Grading Benchmark is technically from a machine learning perspective, and only handles the domain shift from the magnification difference. However, from a clinical perspective, the computational pathology has to handle the domain shift not only from the magnification difference, but also from the staining procedure. Therefore, it is beneficial to test if the hypothesis works in a real-world computational pathology scenario. However, a bottleneck is that, most clinical computational pathology dataset so far conducts the classification task, *i.e.*, separating *tumor* category from *normal* category. Therefore, in this

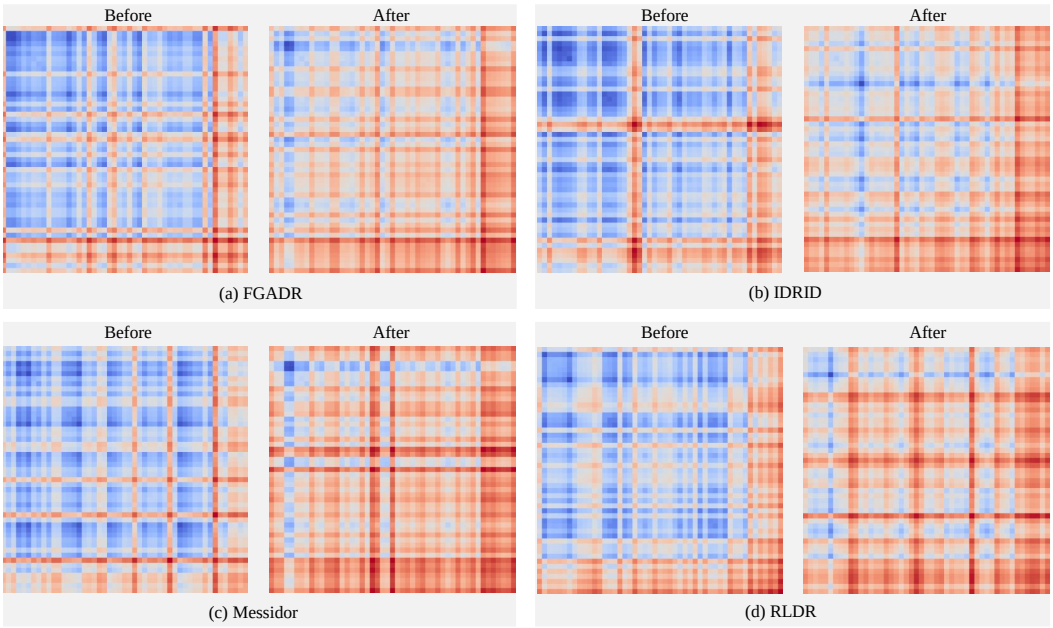


Figure 7: The correlation matrix of each patch embedding before and after processed by the recurrent patch modeling in the forward direction, denoted as 'Before' and 'After', respectively. The higher correlation, the more red a cell is.

Table 6: Impact of the number of components K in GMM on tumor classification performance from unseen target domains. Experiments are conducted on the CAMELYON17. Domain-1 is used as source domain. The rest four are used as unseen target domains. Metrics presented in percentage (%).

K value	Domain-2			Domain-3			Domain-4			Domain-5		
	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
16	81.80	92.09	79.96	79.45	88.96	76.60	83.92	95.38	82.10	75.84	81.99	74.16
32	82.68	93.81	80.95	80.39	91.28	77.97	84.38	96.20	82.96	76.53	82.46	74.92
48	84.06	94.25	81.83	81.84	92.60	78.49	85.87	97.16	83.87	77.69	83.85	75.04
64	**84.59**	**95.67**	**83.10**	**82.48**	**93.32**	**79.85**	**86.50**	**97.82**	**85.13**	**78.64**	**84.70**	**75.32**
96	84.27	95.48	82.53	82.06	92.90	79.44	86.16	97.45	84.80	78.15	84.28	74.86
128	83.65	94.70	81.97	81.50	92.41	78.36	85.47	96.90	84.62	77.38	83.66	74.14

subsection, the proposed Samba along with the vanilla VMamba baseline are benchmarked on the CAMELYON17 dataset [4] for the cross-domain computational pathology classification task.

The first experiment is the impact of the number of components K in GMM, where ACC, AUC and F1 are used as evaluation metric. The results are reported in Table 6. By default K is set to 64, and we further test the performance when K is set to 16, 32, 48 and 96, respectively. When K is set to 64, the proposed Samba achieves the best grading performance. This observation is consistent to the performance on Cross-domain Breast Cancer Grading Benchmark, where a number of 64 Gaussians achieves the optimal performance.

The second experiment is to analyze the trade off between the classification performance and the baseline model. Both the VMamba-ERM and the proposed Samba are tested, where only accuracy is used as the evaluation metric. The results are reported in Table 7. The trend is the same as the trend on the Breast Cancer Grading Benchmark. Using Samba on each type of the VMamba backbone shows a clear performance improvement on unseen domains.

A.5 Attention Maps on Unseen Domains by Samba

Fig. 9 and Fig. 10 show some attention maps of the Samba on unseen retinal images. The proposed Samba is able to model the recurrent relation among patches. Therefore, the activation regions can general cover the lesions and are more robust to the domain shift.

[4] `https://camelyon17.grand-challenge.org/`

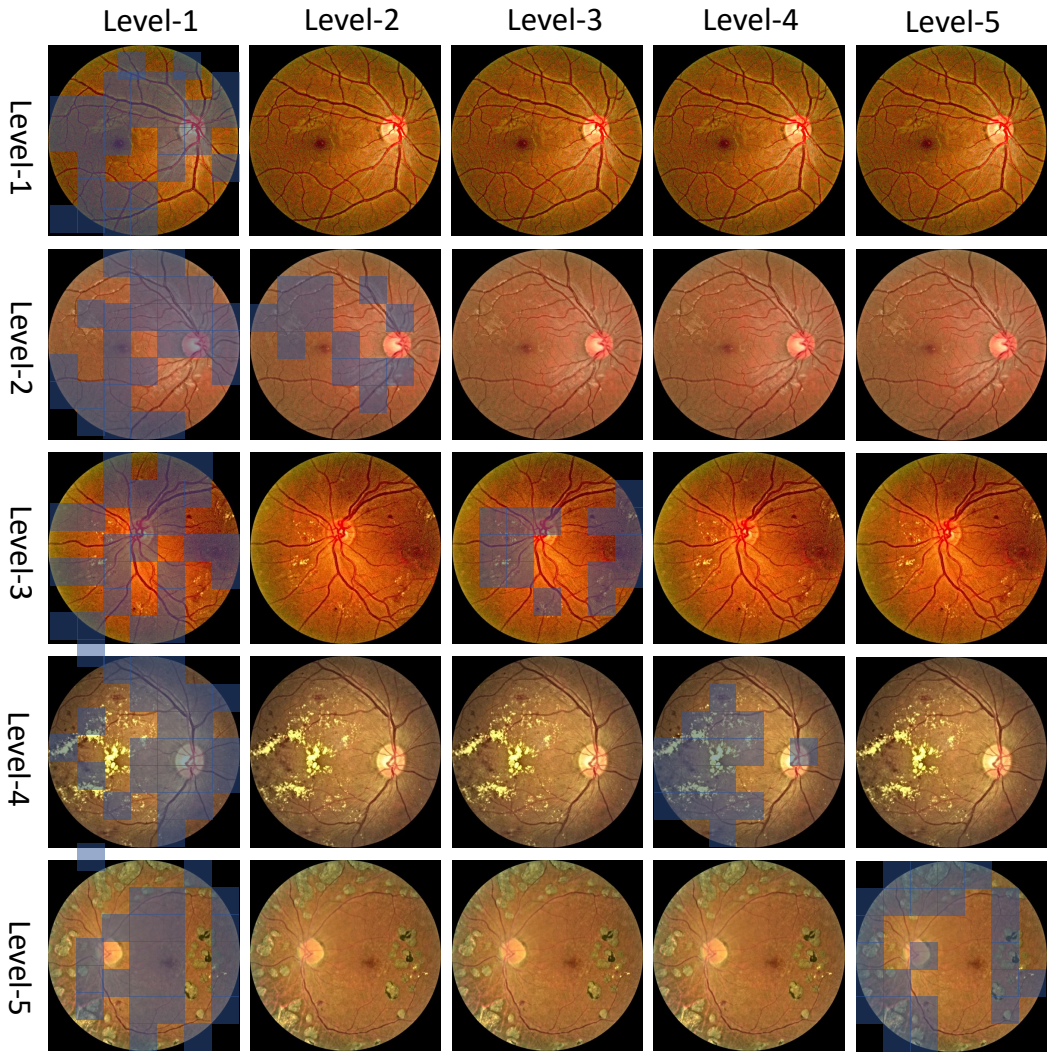


Figure 8: Per-severity activation map of the proposed Samba. From the first to fifth column are the activated patches from level-1 to level-5, highlighted in blue boxes. From the first to the fifth row are the samples with an annotation from level-1 to level-5. FGADR is the unseen target domain.

Table 7: Classification performance comparison between VMamba-ERM and the proposed Samba. Experiments are conducted on the CAMELYON17 dataset for cross-domain tumor classification. Domain-1 is used as the source domain, while the rest four are used as unseen target domains. Metrics in percentage (%).

Method	Backbone	Domain-1 as Source				
		Domain-2	Domain-3	Domain-4	Domain-5	avg.
ERM	VMama-T	70.08	67.29	72.96	63.16	68.37
Samba		**78.74**	**76.15**	**80.06**	**71.05**	**76.50**
ERM	VMama-S	72.86	69.50	75.08	65.72	70.79
Samba		**81.01**	**78.96**	**82.75**	**73.88**	**79.15**
ERM	VMama-B	76.23	74.17	79.53	69.87	74.95
Samba		**84.59**	**82.48**	**86.50**	**78.64**	**83.05**

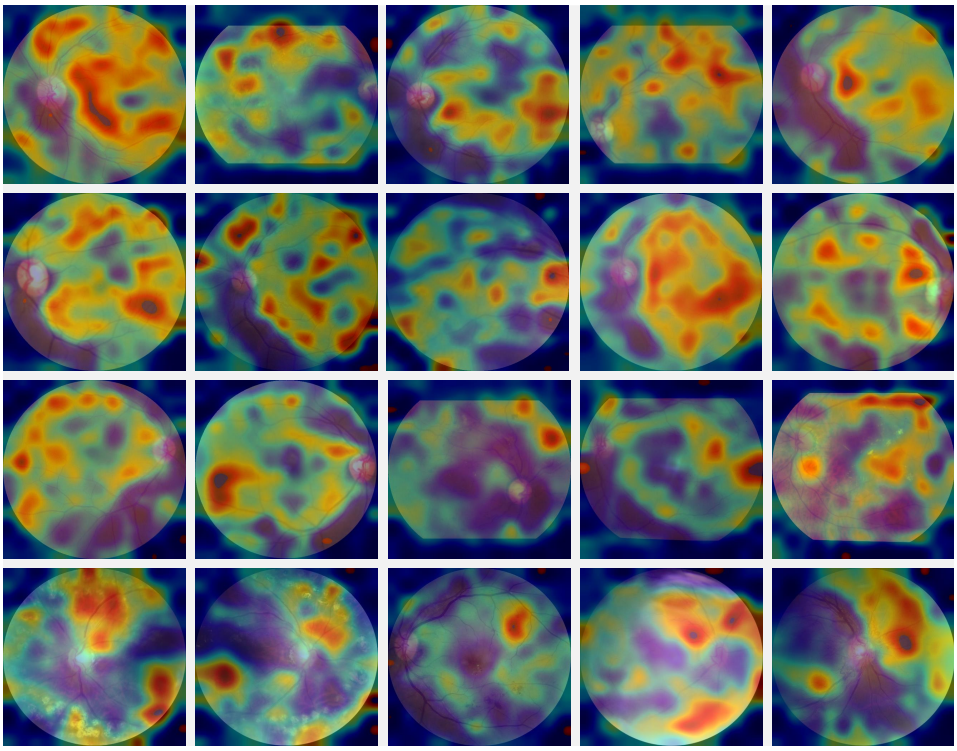


Figure 9: Attention maps of the proposed Samba on retinal images from unseen domains.

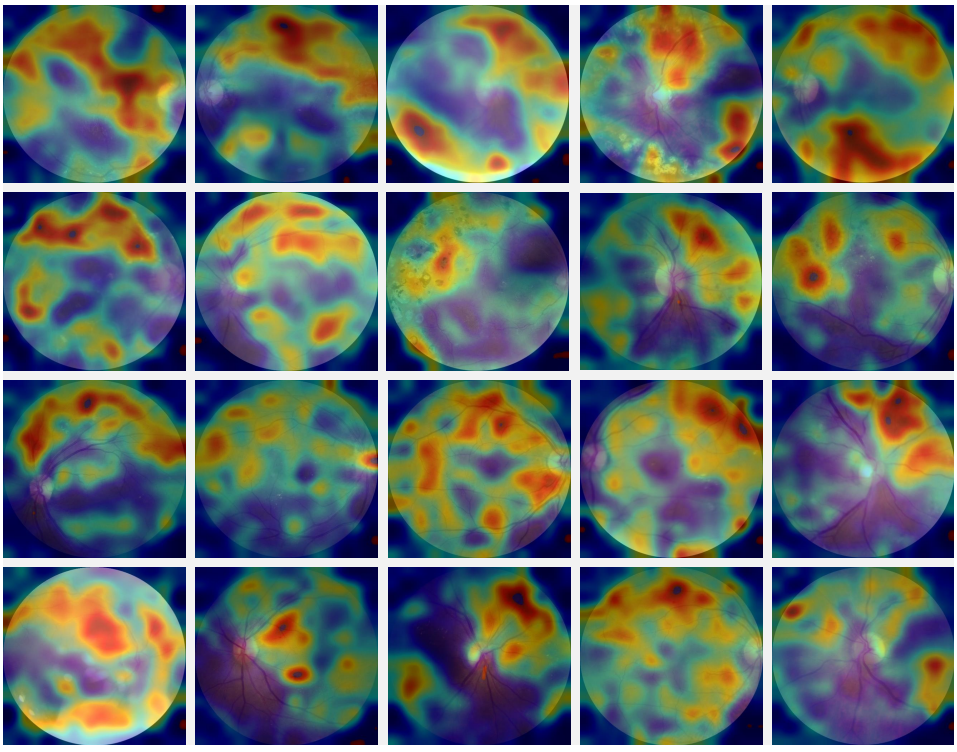


Figure 10: Attention maps of the proposed Samba on retinal images from unseen domains.

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: This paper propose a Severity-aware Recurrent Modeling (Samba) for medical image grading problems on unseen target domains.

   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 generalize 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: At the end of the conclusion section, the limitation of the proposed method has been discussed.

   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: [Yes]

Justification: The theory assumptions are in the methodology section.

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 model realization and implementation details are provided in the experimental section.

   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 public. The source code will be publicly available up on publication.

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, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental settings and details are provided in the experiment section.

   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: [No]

   Justification: The evaluation protocols of these grading datasets do not require report the error bars.

   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: The computation resources and details are discussed in experiment section.

   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: The authors have read the code of ethics. The experiments are all on public datasets without obeying the code of ethics.

   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: [Yes]

    Justification: At the end of the conclusion section, the broader impacts have been discussed.

    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: This work does not have such risks.

    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 asserts used in this paper are all public available for academic researches.

    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: This paper does not release new assets.

    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: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    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: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    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.