No Pains, More Gains:
Recycling Sub-Salient Patches for Efficient High-Resolution Image Recognition

Rong Qin[1], Xin Liu[1], Xingyu Liu[1], Jiaxuan Liu[1], Jinglei Shi[1,5], Liang Lin[2,4], Jufeng Yang[1,2,3]*

[1] VCIP & TMCC & DISSec, College of Computer Science, Nankai University, Tianjin, China.

[2] Pengcheng Laboratory, Shenzhen, China.

[3] Nankai International Advanced Research Institute (SHENZHEN·FUTIAN), Shenzhen, China.

[4] School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou, China.

[5] Key Lab of SCCI, Dalian University of Technology, Dalian, China.

qinrong_nk@mail.nankai.edu.cn, xinliu_0209@163.com

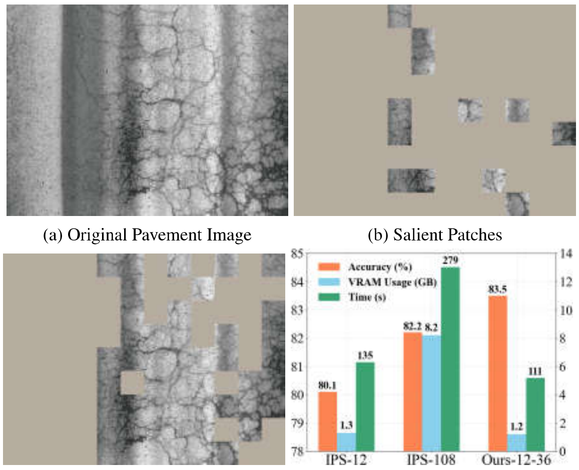
xingyu_liu2002@163.com, jxliu1999@163.com

jinglei.shi@nankai.edu.cn, linliang@ieee.org, yangjufeng@nankai.edu.cn

Abstract

Over the last decade, many notable methods have emerged to tackle the computational resource challenge of the high resolution image recognition (HRIR). They typically focus on identifying and aggregating a few salient regions for classification, discarding sub-salient areas for low training consumption. Nevertheless, many HRIR tasks necessitate the exploration of wider regions to model objects and contexts, which limits their performance in such scenarios. To address this issue, we present a DBPS strategy to enable training with more patches at low consumption. Specifically, in addition to a fundamental buffer that stores the embeddings of most salient patches, DBPS further employs an auxiliary buffer to recycle those sub-salient ones. To reduce the computational cost associated with gradients of sub-salient patches, these patches are primarily used in the forward pass to provide sufficient information for classification. Meanwhile, only the gradients of the salient patches are back-propagated to update the entire network. Moreover, we design a Multiple Instance Learning (MIL) architecture that leverages aggregated information from salient patches to filter out uninformative background within sub-salient patches for better accuracy. Besides, we introduce the random patch drop to accelerate training process and uncover informative regions. Experiment results demonstrate the superiority of our method in terms of both accuracy and training consumption against other advanced methods. ***The code is available in the https://github.com/Qinrong-NKU/DBPS.***

1. Introduction


(a) Original Pavement Image

(b) Salient Patches

(c) Salient and Sub-Salient Patches

(d) Pavement Distress Recognition

Figure 1. In the pavement image (a) of CQU-BPDD dataset [47], a small number of salient patches (b) are insufficient for modeling the distress object and corresponding context. The introduction of sub-salient patches (c) supplements the context for better prediction. As shown in (d), our method utilizes both salient and sub-salient patches for significant accuracy gains while maintaining low training consumption. The 'IPS-M' indicates the IPS-Transformer trained with M patches and 'Ours-M-S' indicates our model optimized with M salient patches and S sub-salient ones.

High resolution images exceeding megapixel have found extensive real-world applications, with HRIR offering significant benefits across various domains [4]. For instance, it aids autopilot systems in accurately identifying distant small traffic signs [44] and assists doctors in making diagnoses based on gigapixel medical images. Despite re-

[*] Corresponding author.

cent advancements in image recognition achieved through deep learning technique, the emphasis has primarily been on downsized images [16, 19]. However, performing recognition tasks on high resolution images presents a greater challenge. On the one hand, directly processing high resolution images with conventional methods often leads to memory overflow. On the other hand, downsizing the original high resolution images to reduce memory consumption results in a loss of discriminative information and thus a notable decline in recognition performance [24, 43].

Although some previous studies have proposed addressing computational resource bottlenecks caused by the high resolution through diverse strategies like non-uniform downsizing, lightweight network design, etc. [35, 36, 40, 42, 46, 50, 53], the strategy of selecting and aggregating salient regions has demonstrated superior performance in terms of both memory consumption and accuracy [15, 17, 37, 38, 54, 59]. For example, [11, 24] leverage attention modules to select salient image patches across downsized images and finalize the classification using the full resolution counterparts of these patches. Furthermore, [9, 26, 51] incorporate this strategy into a multi-stage hierarchical select-and-zoom framework to process gigapixel images. Noting the unreliability of patch selection under downsized images, IPS-Transformer [5] performs both patch selection and aggregation with a learnable query and cross-attention layer. This eliminates the need for training a selection module specifically and allows efficient patch selection on original images in no-gradient mode.

However, the majority of these solutions are founded on the strict assumption that both the objects and context of the scenes can be well modeled by *a small number of image patches*. In fact, for many HRIR scenes, valuable information may be distributed across the entire image, as illustrated by the pavement distress depicted in Fig. 1 (a). This suggests that even with the discriminative details (Fig. 1 (b)), sufficient image patches (Fig. 1 (c)) are still necessary for adequately modeling objects and contexts in these challenging HRIR tasks. Therefore, as shown in Fig. 1 (d), existing methods may fall short in achieving high accuracy. While simply increasing the number of selected patches could potentially improve performance, encoding these additional patches in gradient mode will bring significant increase in training consumption. Another straightforward solution is to employ a limited number of patches for training and utilize more patches as input for inference. However, as demonstrated in Sec. 4.3, this solution hardly improves recognition performance due to the information gap between these two stages. *Thus, it is necessary to find a method to introduce more patches for training without imposing significant cost for HRIR.*

Based on the aforementioned observations, we propose a Dual-Buffer Patch Selection (DBPS) strategy to handle more input information while keeping low consumption during the training stage. Specifically, besides a fundamental buffer to store the embeddings of the patches with the highest saliency scores, DBPS employs an auxiliary buffer to recycle the embeddings of the sub-salient patches, which have slightly lower scores and are typically discarded in previous works. And we endow them with different functionalities in DBPS. Both types of embeddings will be aggregated to provide information for classification in the forward pass, but only the gradients of the salient embeddings are back-propagated to optimize the encoder directly. Though sub-salient embeddings could not optimize the encoder through gradient back-propagation, they still indirectly optimize the encoder by contributing to the final output with supplemented context. Since the sub-salient patch embeddings are all recycled from the patch selection stage and will not back-propagate gradient to the encoder, their participation barely affects the consumption. Considering the hidden uninformative backgrounds inside the sub-salient patches, we further devise a dual-attention MIL architecture to process these two buffers through a progressive aggregation way. Firstly, the most salient patch embeddings are fed into a cross-attention-based transformer to generate the salient query. This query is utilized to further aggregate the embeddings of the sub-salient patches for suppressing the uninformative backgrounds inside sub-salient regions. Moreover, we introduce the random patch drop technique to reduce the patches that need to be traversed and uncover more informative image patches. Results on six challenging datasets demonstrate the superiority of our method in terms of accuracy and training consumption.

Our contributions are three-fold: 1) We propose a Dual-Buffer Patch Selection (DBPS) strategy to ensure sufficient patches for training while maintaining low training consumption. To the best of our knowledge, we are the first to focus on the challenges of scattered distribution in HRIR and suggest introducing sub-salient regions as context supplements in a no-gradient way to address it. 2) To suppress the backgrounds inside sub-salient patches, we devise a dual-attention MIL architecture to generate salient query for the aggregation of sub-salient embeddings. Besides, we introduce the random patch drop to accelerate training and uncover informative regions. 3) We conduct experiments on six HRIR datasets to demonstrate the superiority of our model in terms of accuracy and training consumption.

2. Related Work

2.1. High-Resolution Image Recognition

HRIR is common and valuable in downstream applications, and has gradually attracted the attention of researchers in recent years [4, 14, 24, 26, 36, 40, 47]. Previous works of HRIR mainly focus on using sparse computation to utilize the spatial redundancy of high-resolution images to save memory usage [3, 9, 11, 15, 24, 54]. With the image regions of interest identified on the downsized images, the network only needs to process part of the images at high-resolution. This idea can be traced back to early fast image processing methods [12, 25], and can also be associated with human saccadic eye movements, which may be informed by peripheral vision [6]. To process gigapixel images, [26, 38, 51] expand this idea to a multi-stage sampling strategy, successively adopting attention sampling operation at increasing resolution to find the salient image patches. However, the resolution of the downsized image may be too small for finding salient patches, while the increasing of the resolution will bring more memory usage [5]. IPS-Transformer [5] finishes both informative region selection and patch embedding aggregation through a learnable query and a cross-attention layer. Therefore, it does not need to train a selection module specifically, which allows the efficient selection of informative patches on the original images in no-gradient mode. In this paper, we use two embedding buffers to identify the most salient and sub-salient patches, and aggregate the gradient salient patch embeddings and the no-gradient sub-salient ones for final classification and network training. The gradient-free sub-salient embeddings indirectly optimize the encoder for better performance by affecting the output with supplemented context.

2.2. Multiple Instance Learning

The Multiple Instance Learning (MIL) is first introduced into weakly supervised learning [13] and has achieved significant success in gigapixel whole-slide image (WSI) classification task [7, 27, 29, 45, 49, 57]. For instance, Ilse et al. [23] firstly devise an attention-based MIL architecture, which allocates contribution information to each instance through trainable attention weights. And Dual-stream MIL (DS-MIL) [28] calculates the similarity between the most significant instance and the others as the corresponding attention weights. In contrast to the previous approaches that seek to identify salient instances, Tang et al. [48] aim to enhance the capability of MIL models by intentionally masking the most prominent instances to uncover hard instances.

Image patch selection and the aggregation of patch embeddings are the two major parts of HRIR. Therefore, efficient aggregation through MIL architecture and associating MIL with patch selection have attracted the attention of researchers [5, 11, 24, 26, 51]. For instance, [24, 26] propose to utilize the attention scores computed by patch selection module as a MIL operator for embedding aggregation. Differential Patch Selection (DPS) [11] verifies the advantages of self-attention in aggregating patch embeddings. Recently, IPS-Transformer [5] introduces learnable query and cross-attention to unify patch selection and embedding aggregation, and has demonstrated its efficiency. In this work, we devise a dual-attention MIL, which adaptively generates salient query for the aggregation of sub-salient patch embeddings. Moreover, we introduce random patch drop technique to uncover more informative instances.

3. Methodology

3.1. Limitation of Patch Selection Strategy in HRIR

In this sub-section, we take the most advanced HRIR method IPS-Transformer [5] as an example to introduce the widely applied single buffer patch selection strategy in HRIR methods and explain its limitation. Specifically, in IPS-Transformer, given a high resolution image, it is initially divided into N patches. Then each patch $\boldsymbol{x}$ is embedded into an embedding $\boldsymbol{e} \in \mathbb{R}^D$ by an encoder as:

$$\boldsymbol{e} = \mathcal{F}(\boldsymbol{x}), \tag{1}$$

where $\mathcal{F}(\cdot)$ denotes the mapping function of the encoder. With a learnable query token $\boldsymbol{q} \in \mathbb{R}^D$, the attention score a of $\boldsymbol{x}$ can be computed through a cross-attention layer:

$$a = \frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{D'}}, \tag{2}$$

where query $\boldsymbol{Q} = \boldsymbol{q}\boldsymbol{W}^q$ and key $\boldsymbol{K} = \boldsymbol{e}\boldsymbol{W}^k$ are respectively the linear transforms of $\boldsymbol{q} \in \mathbb{R}^D$ and $\boldsymbol{e}$ through $\boldsymbol{W}^q \in \mathbb{R}^{D\times D'}$ and $\boldsymbol{W}^k \in \mathbb{R}^{D\times D'}$. And a can serve as a metric for the saliency of each patch. Because computing the saliency scores of all patches in parallel can result in GPU memory overflow, the most salient patch embeddings can be selected iteratively with the assistance of a buffer P_M^t. Here t represents a specific update step, and M denotes the capacity of the buffer. And the buffer concept here represents a preset space in GPU for storing tensors. Let us note that all these operations occur in no-gradient mode.

The final selected salient embeddings will be re-embedded in gradient mode and then aggregated for classification and optimization. Besides the ideal scenarios like recognizing traffic signs where objects are concentrated within a few patches, many HRIR tasks necessitate the model to select more patches for the exploration of a broader perspective. However, encoding more patches in gradient mode can easily exceed the memory limit [5, 11, 24, 26]. Conversely, using a small number of patches for training and more for test, as shown in Tab. 1, barely improves the accuracy. *This dilemma necessitates a resource-efficient patch utilization method to introduce more patches for training without imposing significant cost.*

3.2. Dual-Buffer Patch Selection

In this sub-section, we further introduce our proposed DBPS strategy for addressing the above mentioned limitation. Specifically, in addition to the aforementioned fundamental buffer P_M^t, DBPS further employs an auxiliary

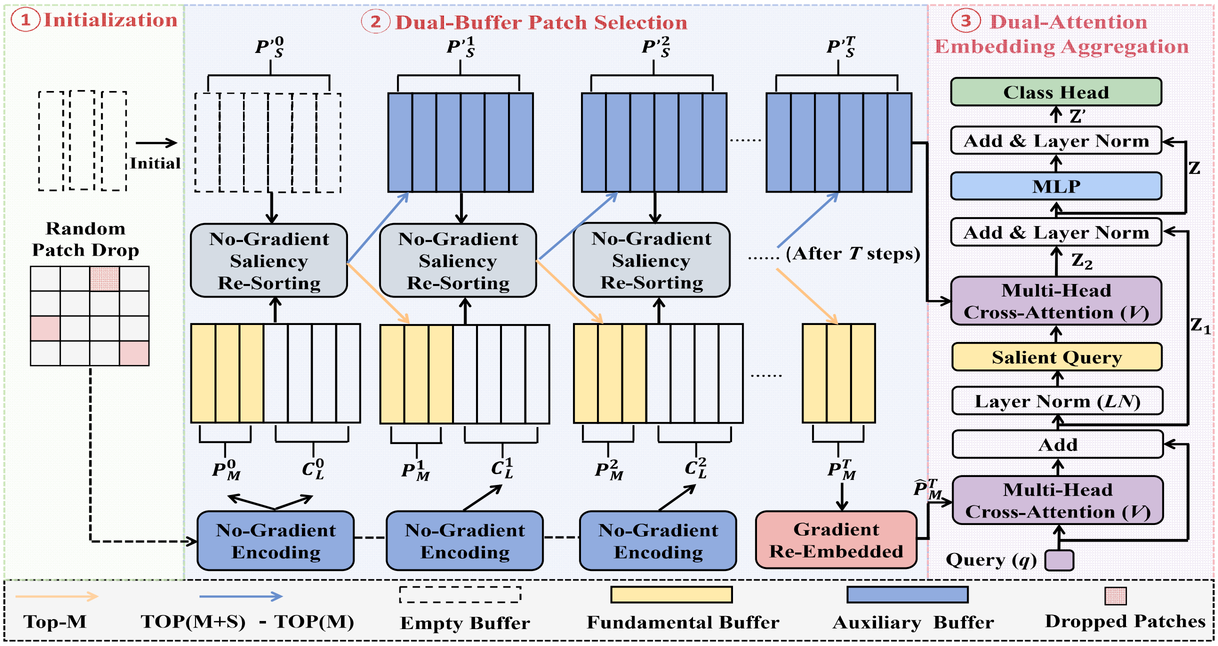


Figure 2. The pipeline of our proposed method. We firstly divide the high-resolution image into image patches and randomly drop some of them. Next, we initial two buffers P^0_M and P'^0_S, and iteratively update them in no-gradient mode to find the most salient patches and the sub-salient patches. After patch selection, P^T_M is embedded again in gradient mode, and then input into the cross-attention layer to be aggregated. The aggregated embedding can be denoted as the salient query for the better aggregation of no-gradient P'^t_S.

buffer P'^t_S to store and recycle the embeddings of the sub-salient S patches that are originally discarded in the single patch selection. P'^t_S has larger capacity than P^t_M (i.e., $S > M$). As shown in Fig. 2, P'^t_S is initialized as an empty buffer as $P'^0_S = \{\}$, while P^t_M is initialized with the embeddings of the first M patches ($P^0_M = \{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_i, \cdots, \boldsymbol{e}_M\}$). During the traversal of the rest $(N - M)$ patches, a collection of L patch embeddings C_L is considered at each step. After obtaining the scores of the existing $(M + S + L)$ embeddings in P^{t-1}_M, P'^{t-1}_S and C^{t-1}_L, P^{t-1}_M and P'^{t-1}_S will be updated to be P^t_M and P'^t_S through selecting the new M most salient embeddings and the new S sub-salient ones, respectively:

$$P^t_M = Top\text{-}M\{P^{t-1}_M \cup P'^{t-1}_S \cup C^{t-1}_L | \boldsymbol{A}_{M+S+L}\}, \quad (3)$$

$$P'^t_S = Top\text{-}S\{(P^{t-1}_M \cup P'^{t-1}_S \cup C^{t-1}_L) - P^t_M | \boldsymbol{A}_{M+S+L}\}, \quad (4)$$

where $\boldsymbol{A}_{M+S+L}$ is the score vector of $(M+S+L)$ embeddings, and the maximal traversal step is $T = \lceil (N-M)/L \rceil$.

As explained earlier, re-embedding the two types of patch embeddings in gradient mode will lead to GPU memory overflow. To ensure low computational resource increasing, we suggest to only re-embed P^T_M in gradient mode and add their aggregation $\boldsymbol{Z}_1$ with the aggregation $\boldsymbol{Z}_2$ of no-gradient P'^T_S for the final classification. Compared to the original single buffer patch selection, our DBPS only recycles the originally discarded sub-salient patch embeddings, hence just increasing ultra-low cost of storing P'^T_S. The addition of no-gradient sub-salient patch embeddings can directly optimize aggregation module and indirectly optimize the encoder by providing necessary context information that significantly affects final classification. More concisely, P'^T_S will not introduce additional gradients to optimize the encoder, but only improves the prediction and adjusts the gradients belonging to the re-embedded salient patch embeddings $\hat{P}^T_M$ for better optimization of the encoder. The pseudo code of DBPS is available in the Algorithm 1.

3.3. Dual-Attention Embedding Aggregation

Although there is considerable context information among P'^T_S, introducing P'^T_S for final classification may also bring uninformative backgrounds, which will limit the accuracy gains or even degrade model performance [5, 11, 28, 33] . To address this challenge, inspired by [28], we propose a dual-attention MIL architecture as our aggregation module (as shown in the right part of Fig. 2), which processes the $\hat{P}^T_M$ and P'^T_S with a progressive aggregation way:

1). Classic HRIR works [5, 11, 24, 26] have demonstrated the superiority of attention-based MIL operations in patch embedding aggregation, while the cross-attention transformer is an efficient MIL architecture. Therefore, we employ cross-attention transformer as the basic architecture of our aggregation module. $\hat{P}^T_M$ with high salient scores and a learnable query q are firstly fed into the cross-attention layer to get the aggregation $\boldsymbol{Z}_1$ of the $\hat{P}^T_M$:

$$\boldsymbol{Z}_1 = \mathcal{V}(\boldsymbol{q}, \hat{P}^T_M) + \boldsymbol{q}, \quad (5)$$

where $\mathcal{V}(\cdot)$ denotes a cross-attention layer based on Eq. 2.

Note that the scores of the embeddings in $\hat{P}_M^T$ will be normalized in $\mathcal{V}(\cdot)$ through $Softmax(\cdot)$ for valid aggregation.

2). Since the embeddings in $\hat{P}_M^T$ are relatively salient, $\boldsymbol{Z}_1$ can be considered to contain relatively little backgrounds. Therefore, we replace the input-irrelevant $\boldsymbol{q}$ with $\boldsymbol{Z}_1$ as the adaptive salient query for the aggregation of $P_S'^T$:

$$\boldsymbol{Z}_2 = \mathcal{V}(LN(\boldsymbol{Z}_1), P_S'^T) = \mathcal{V}(LN(\mathcal{V}(\boldsymbol{q}, \hat{P}_M^T) + \boldsymbol{q}), P_S'^T), \quad (6)$$

$$\boldsymbol{Z} = LN(\mathcal{V}(\boldsymbol{q}, \hat{P}_M^T) + \boldsymbol{q} + \mathcal{V}(LN(\mathcal{V}(\boldsymbol{q}, \hat{P}_M^T) + \boldsymbol{q}), P_S'^T)), \quad (7)$$

where $LN(\cdot)$ represents layer normalization [2] and $\boldsymbol{Z}$ represents the total aggregation. According to the Eq. 6, the attention scores of $P_S'^T$ are not only related to $\boldsymbol{q}$ but are also their correlation with the salient embeddings in $\hat{P}_M^T$.

Finally, we adopt a 3-layer fully connected layers MLP$(\cdot)$ to get the ultimate embedding $\boldsymbol{Z}'$ for classification:

$$\boldsymbol{Z}' = LN(\text{MLP}(\boldsymbol{Z}) + \boldsymbol{Z}). \quad (8)$$

The dual-attention MIL aims to allocate greater attention scores to sub-salient embeddings that are more correlated to the $\hat{P}_M^T$, and thus suppress uninformative backgrounds.

3.4. Boosting HRIR with Random Patch Drop

In recent years, the random patch drop technique has achieved impressive success in self-supervised pre-training tasks [1, 20, 52, 58]. Based on the natures of HRIR, we suggest that the random patch drop and the HRIR task are also a perfect match. Firstly, the HRIR methods are mainly based on the patch-selection-aggregation architecture, which allows the random patch drop to be directly applied for pruning the training input without additional designs. Besides, there is a significant spatial redundancy in high-resolution images, which saves the random patch drop from causing severe information loss and disturbing model training.

Therefore, we introduce the random patch drop technique to boost the training process of HRIR models in terms of effectiveness and efficiency. Precisely, given a preset drop ratio $r \in [0, 1)$ and total N patches of an image, only $\lceil (1-r) \times (N-M) + M \rceil$ patches are randomly kept for patch selection phase. Albeit simple, random patch drop can reduce the number of patches to be traversed, thereby accelerating training and saving more GPU memory. Besides, it could also prevent the model from overfitting on simple or incorrect regions. Moreover, compared to competitors such as random erasing that employs a single rectangle region, the randomly dropped patches are uniformly distributed across the image, which leads to better performance when a large r is adopted (More details in Sec. 4.5).

Algorithm 1: The pseudo code of DBPS

```
with torch.no_grad():
    # Initialization
    M_buf = net.encode(patches[:, :M].cuda())
    idx = torch.arange(N, dtype=torch.int64,
     device=device).unsqueeze(0).expand(B, -1)
    M_idx = idx[:, :M]
    S_buf = torch.zeros((B, S, D)).cuda()
    A_S = torch.zeros((B, S)).cuda() - 1000
    # Patch selection in no-gradient mode
    for i in range(T):
        start = i * L + M
        end = min(start + L, N)
        L_idx = idx[:, start:end]
        L_buf = net.encode(patches[L_idx].cuda())
        M_buf, M_idx, A_ML = net.select1(M_buf,
         L_buf, M_idx, L_idx, M)
        S_buf, A_S = net.select2(M_buf, L_buf,
         S_buf, A_ML, A_S, M, S)
# Re-embedding salient patches in gradient mode
M_patch =net.select3(patches, M_idx).cuda()
M_buf = net.encode(M_patch))
# Aggregating all embeddings in gradient mode
image_emb = net.transf(M_buf, S_buf)
preds = net.get_preds(image_emb)
```

3.5. Efficiency Analysis

Finally, we analyze the efficiency of DBPS strategy. Let $\mathcal{O}(1)$ be the computational cost of encoding and scoring a patch at no-gradient mode while $\hat{\mathcal{O}}(1)$ be that of encoding a patch in gradient mode, Besides, $\overline{\mathcal{O}}(1)$ is the computational cost of aggregating a patch embedding at gradient mode, and $\mathcal{G}(1)$ denotes the memory required to store a patch embedding in GPU. Note that $\hat{\mathcal{O}}(1) > \mathcal{O}(1) > \overline{\mathcal{O}}(1)$. The computational cost of the single buffer patch selection strategy during training could be denoted as $\mathcal{O}(N) + \hat{\mathcal{O}}(M) + \overline{\mathcal{O}}(M)$ and that of the DBPS could be denoted as $\mathcal{O}(N) + \hat{\mathcal{O}}(M) + \overline{\mathcal{O}}(M+S)$. The buffer memory of the single buffer patch selection strategy during training could be denoted as $\mathcal{G}(M+L)$ and that of the DBPS could be denoted as $\mathcal{G}(M+L+S)$. It can be concluded that DBPS strategy introduces S extra patches during training for significant accuracy gains while only incurring the ultra-low additional cost of $\overline{\mathcal{O}}(M+S)$ and $\mathcal{G}(S)$. After applying random patch drop, our cost and buffer memory can be further reduced to $\mathcal{O}(\lceil (1-r) \times (N-M) + M \rceil) + \hat{\mathcal{O}}(M) + \overline{\mathcal{O}}(M+S)$ and $\mathcal{G}(M + \lceil (1-r) \times L \rceil + S)$, which are similar or even less than that of the single buffer patch selection strategy.

4. Experiments

4.1. Datasets

We perform experiments on six HRIR datasets: 1) the CQU-BPDD dataset [47], 2) the Functional Map of the World (fMoW) dataset [10], 3) the Swedish Traffic Signs Recognition dataset [24], 4) the CAMELYON16 dataset [32], 5) the DDR dataset [30], 6) the MAME dataset [39].

4.2. Evaluation Metrics and Settings

Following [5], we use maximum GPU memory usage (VRAM) and batch training runtime (Time) for a batch size of 16 to evaluate computational efficiency, whose corresponding units are GB and ms. We set $\hat{L} = \lceil L \times (1-r) \rceil$

Table 1. Results of different methods and settings on six downstream datasets, in which different solutions are highlighted with different colors. The 'Scale' represents the input image scale. The 'Ratio' represents the patch drop ratio. The 'M' represents the number of salient image patches for training. The 'M'' represents the number of salient image patches for evaluation. The 'S' represents the number of sub-salient image patches for training and evaluation. The 'UN' represents that the corresponding value is unknown. The 'NA' represents corresponding value is not available in the method.

CQU-BPDD (Pavement)									**Functional Map of the World (Satellite)**								
Methods	Scale	Ratio	M	M'	S	ACC	VRAM ↓	Time ↓	Methods	Scale	Ratio	M	M'	S	ACC	VRAM ↓	Time ↓
ResNet-18 [19]	0.3×	NA	NA	NA	NA	75.9	0.8	32	ResNet-18 [19]	0.3×	NA	NA	NA	NA	76.6	0.69	25
	1×	NA	NA	NA	NA	81.3	7.0	212		1×	NA	NA	NA	NA	80.2	5.75	163
KIPRN [41]	0.5×	NA	NA	NA	NA	82.1	UN	UN	Zoom-In [26]	1×	0.0	8	8	NA	72.9	UN	UN
WSPLIN [21]	1×	NA	NA	NA	NA	81.5	UN	UN	Zoom-In+ [26]	1×	0.0	8	8	NA	74.3	UN	UN
IPSformer [5] **(Baseline)**	1×	0.0	108	108	NA	82.2	8.2	279	IPSformer [5] **(Baseline)**	1×	0.0	81	81	NA	80.1	6.16	212
	1×	0.0	36	36	NA	81.5	3.1	194		1×	0.0	36	36	NA	78.3	3.1	171
	1×	0.0	12	108	NA	69.9	1.3	135		1×	0.0	12	81	NA	76.3	1.3	110
	1×	0.0	12	36	NA	78.5	1.3	135		1×	0.0	12	36	NA	76.9	1.3	110
	1×	0.0	12	12	NA	80.1	1.3	135		1×	0.0	12	12	NA	77.1	1.3	110
DBPSformer **(ours)**	1×	0.0	12	12	60	82.5	1.3	136	DBPSformer **(ours)**	1×	0.0	36	36	24	79.9	3.1	172
	1×	0.2	12	12	60	82.6	1.3	117		1×	0.1	36	36	24	80.4	3.0	164
	1×	0.3	12	12	60	82.9	1.2	112		1×	0.2	36	36	24	**80.5**	2.9	156
	1×	0.3	12	12	48	82.9	1.2	111		1×	0.3	12	12	48	78.3	1.2	96
	1×	0.3	12	12	36	**83.5**	1.2	111		1×	0.3	12	12	36	78.3	1.2	96
	1×	0.3	12	12	24	83.2	1.2	111		1×	0.3	12	12	24	78.8	1.2	96
DDR (Retinopathy)									**MAME (Artwork)**								
Methods	Scale	Ratio	M	M'	S	ACC	VRAM ↓	Time ↓	Methods	Scale	Ratio	M	M'	S	ACC	VRAM ↓	Time ↓
ResNet-50 [19]	0.3×	NA	NA	NA	NA	73.4	3.8	98	ResNet-18 [19]	0.3×	NA	NA	NA	NA	80.4	0.8	30
	0.75×	NA	NA	NA	NA	80.3	21.4	568		1×	NA	NA	NA	NA	86.5	7.0	211
CBANet [18]	0.5×	NA	NA	NA	NA	77.7	UN	UN	DPSformer [11]	1×	0.0	30	30	NA	84.0	4.2	403
CANet [31]	0.5×	NA	NA	NA	NA	77.5	UN	UN		1×	0.0	10	10	NA	80.2	2.3	182
IPSformer [5] **(Baseline)**	1×	0.0	18	18	NA	80.1	19.9	833	IPSformer [5] **(Baseline)**	1×	0.0	100	100	NA	86.7	7.5	259
	1×	0.0	9	9	NA	79.4	10.3	586		1×	0.0	30	30	NA	85.5	2.6	170
	1×	0.0	4	18	NA	74.8	4.9	442		1×	0.0	10	100	NA	81.3	1.3	119
	1×	0.0	4	9	NA	75.2	4.9	442		1×	0.0	10	30	NA	83.7	1.3	119
	1×	0.0	4	4	NA	77.0	4.9	442		1×	0.0	10	10	NA	83.9	1.3	119
DBPSformer **(ours)**	1×	0.0	4	4	20	80.1	4.9	443	DBPSformer **(ours)**	1×	0.0	30	30	45	86.1	2.6	170
	1×	0.1	4	4	20	80.5	4.9	394		1×	0.1	30	30	45	86.6	2.6	160
	1×	0.2	4	4	20	**81.0**	4.9	370		1×	0.2	30	30	45	**86.8**	2.5	151
	1×	0.3	4	4	20	80.7	4.9	343		1×	0.3	10	10	60	85.6	1.1	96
	1×	0.3	4	4	16	80.7	4.9	342		1×	0.3	10	10	45	85.9	1.1	96
	1×	0.3	4	4	12	80.5	4.9	342		1×	0.3	10	10	30	85.5	1.1	96
Swedish Traffic Signs (Traffic Signs)									**CAMELYON16 (WSI)**								
Methods	Scale	Ratio	M	M'	S	ACC	VRAM ↓	Time ↓	Methods	Scale	Ratio	M	M'	S	AUC	VRAM ↓	Time ↓
RPSformer [5]	1×	0.0	10	10	NA	50.7	1.3	31	DSMIL-LC [28]	1×	0.0	$8k$	$8k$	NA	91.7	UN	UN
	1×	0.0	2	2	NA	28.3	0.9	26	CLAM [34]	1×	0.0	$42k$	$42k$	NA	93.6	UN	UN
DPSformer [11]	1×	0.0	10	10	NA	94.0	3.5	337	CLAM-SB [56]	1×	NA	UN	UN	NA	94.2	UN	UN
	1×	0.0	2	2	NA	97.2	2.5	296	Challenge [55]	1×	NA	NA	NA	NA	92.5	UN	UN
TopMIL [8]	1×	0.0	10	10	NA	97.7	4.5	109	TopMIL [8]	1×	0.0	$5k$	$5k$	NA	71.8	19.8	84
	1×	0.0	2	2	NA	98.2	4.5	96		1×	0.0	$1k$	$1k$	NA	76.2	19.8	76
DeepMIL [23]	1×	0.0	192	192	NA	96.2	14.2	323		1×	0.0	100	100	NA	84.4	19.8	74
DeepMIL+ [23]	1×	0.0	192	192	NA	97.7	14.3	323	DeepMIL [23]	1×	0.0	$70k$	$70k$	NA	94.5	4.5	26
GF-ResNet [22]	1×	NA	NA	NA	NA	89.5	1.9	UN		1×	0.0	$50k$	$50k$	NA	93.8	22.5	110
GF-ResNet+ [22]	1×	NA	NA	NA	NA	91.5	2.8	UN		1×	0.0	$10k$	$10k$	NA	84.1	31.5	150
IPSformer [5] **(Baseline)**	1×	0.0	192	192	NA	98.4	14.3	495	IPSformer [5] **(Baseline)**	1×	0.0	$5k$	$5k$	NA	98.1	4.7	315
	1×	0.0	10	10	NA	98.6	1.6	202		1×	0.0	$1k$	$1k$	NA	97.5	4.1	313
DBPSformer **(ours)**	1×	0.00	10	10	20	99.2	1.6	205	DBPSformer **(ours)**	1×	0.0	$1k$	$1k$	$4k$	98.3	3.9	317
	1×	0.05	10	10	20	**99.3**	1.5	196		1×	0.1	$1k$	$1k$	$4k$	**98.6**	3.7	297
	1×	0.10	10	10	20	**99.3**	1.4	188		1×	0.2	$1k$	$1k$	$4k$	98.4	3.6	275
	1×	0.15	10	10	20	98.6	1.4	182		1×	0.3	$1k$	$1k$	$4k$	98.2	3.4	253

when random patch drop is applied with ratio r. For CQU-BPDD, fMoM, MAME, and Swedish Traffic Signs Recognition, we adopt ResNet-18 [19] with ImageNet-1k weights as encoder for all methods, while ResNet-50 [19] is used for DDR dataset. For CAMELYON16, we adopt ResNet-50 [19] pre-trained on CAMELYON16 as the encoder for our methods and then fixed. The patch size of DDR is 200×200, while that of CAMELYON16 is 256×256. And the patch sizes of the other four datasets are 100×100.

4.3. Validity of Our DBPS strategy

Our work is built on the assumption that for many HRIR tasks, sufficient image patches are necessary to adequately model objects and contexts. To prove this assumption, we firstly evaluate three types of IPS-Transformer [5]. The first two are trained with a small number of patches, but separately uses insufficient and sufficient patches for evaluation, and the last one is always with sufficient patches. The results of six HRIR tasks are shown in Tab. 1. We can firstly

Table 2. Ablation experiments of our method on CQU-BPDD dataset and Functional Map of the World dataset. 'Dual-Buffer' indicates the use of dual-buffer patch selection strategy. 'Dual-Attention' indicates the use of dual-attention Multiple Instance Learning. 'Patch Drop (r)' indicates the use of random patch drop training strategy with a certain drop ratio r. 'IPS' indicates the IPS-Transformer.

IPS [5] (Baseline)	Dual-Buffer	Dual-Attention	Patch Drop (0.1)	Patch Drop (0.2)	Patch Drop (0.3)	CQU-BPDD			Functional Map		
						Acc	VRAM ↓	Time ↓	Acc	VRAM ↓	Time ↓
✓						80.1	1.3	135	77.1	1.3	110
✓	✓					82.2	1.3	135	77.9	1.3	110
✓	✓	✓				82.8	1.3	135	78.4	1.3	110
✓			✓			80.8	1.3	127	77.3	1.3	103
✓				✓		81.1	1.3	118	77.4	1.2	100
✓					✓	81.2	1.2	110	77.2	1.2	95
✓	✓	✓	✓			82.9	1.3	129	78.6	1.3	106
✓	✓	✓		✓		83.0	1.3	118	**79.0**	1.2	101
✓	✓	✓			✓	**83.2**	1.2	111	78.8	1.2	96

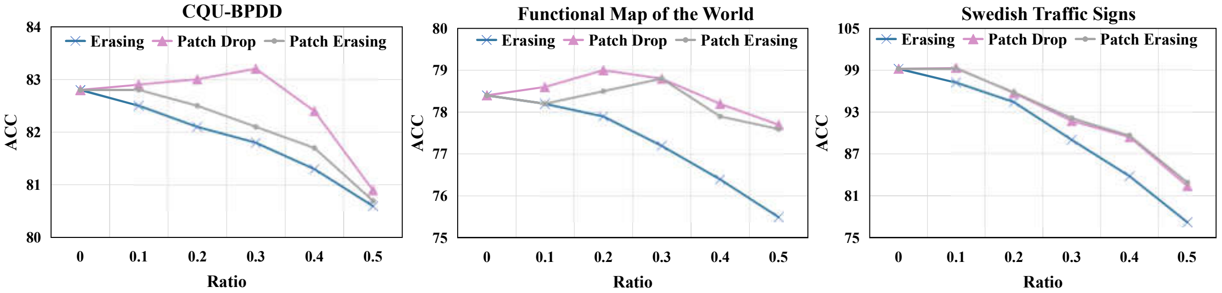


Figure 3. The comparison results between random erasing, random patch erasing, and our random patch drop under different ratio.

Table 3. The inference costs of both IPS-Transformer and our method (both with batch size 16 for inference).

Methods	CQU-BPDD		FMoW	
	Memory(MB) ↓	*Time(ms)* ↓	*Memory(MB)* ↓	*Time(ms)* ↓
IPS-Transformer	1075.5	105	890.1	82
DBPS-Transformer (Ours)	1076.4	108	891.7	83

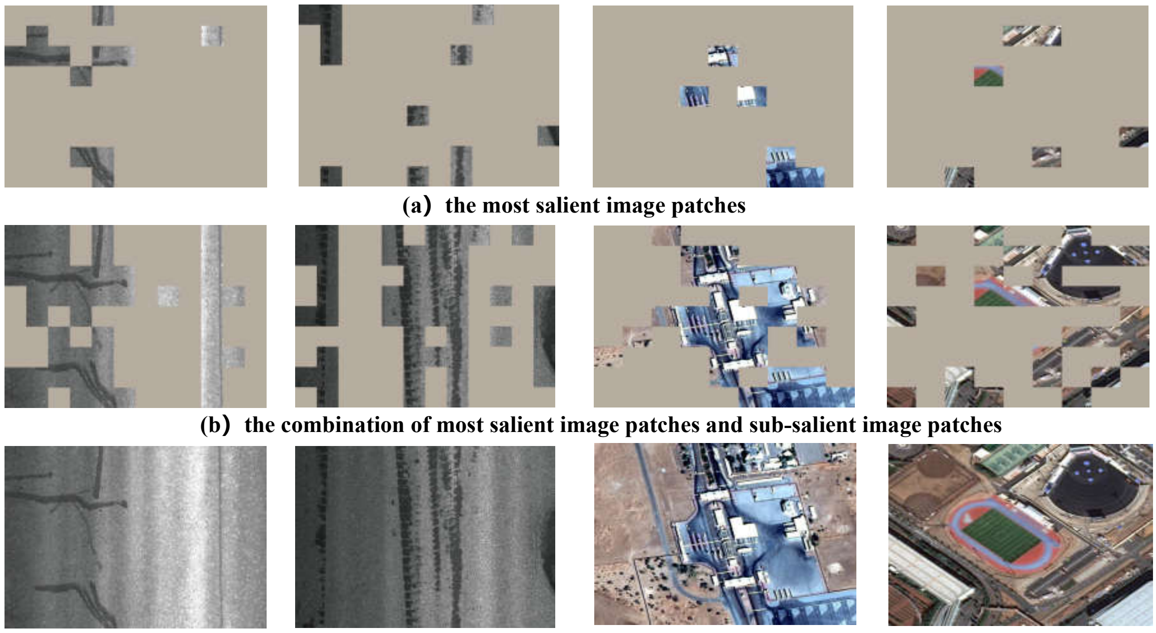
observe that the IPS-Transformer trained with more patches significantly performs better than those with less patches in four challenging HRIR tasks (pavement, satellite, retinopathy, and artwork). Another valuable observation is that the IPS-Transformer trained with a small number of patches but uses more for evaluation achieves the worst accuracy. These observations indicate that introducing enough patches is necessary for training effective HRIR models.

Although simply adding more patches during the training of IPS-Transformer can alleviate this problem, it brings much more GPU memory and time consumption for training. However, this additional consumption could be avoided if we utilize abundant sub-salient patches for training in a no-gradient way. As shown in Tab. 1, our proposed DBPS strategy has the ability to significantly boost the model with the context information contained in sub-salient patches, but almost no additional training costs are added due to the no-gradient encoding. When boosting the DBPS strategy with random patch drop technique, the training consumption of our method could be reduced below that of the standard single buffer selection strategy. Besides, as reported in Tab. 3, the additional inference cost brought by the extra sub-salient patches is also quite low. This is because that both two methods need to encode and score each patch during patch selection stage, which means no cost difference. And the aggregation of the sub-salient patch embeddings is only operated within one cross-attention transformer layer, which also means low computational cost.

4.4. Unanimous Improvements over HRIR Tasks

We conduct extensive experiments on diverse scenes to verify the strong generalization ability of our method in improving HRIR, as shown in Tab. 1. For the HRIR tasks that need sufficient patches (pavement, satellite, retinopathy, and artwork), it is obvious that our method achieves the most advanced performance with low training cost. By suppressing uninformative regions and uncovering informative ones through dual-attention MIL and random patch drop, the accuracy of our method can even exceed that of the IPS-Transformer trained with all patches, but uses much less training consumption. For instance, our method gains about 85% memory reduction, 60% training time reduction, and better performance than IPS-Transformer trained with even all patches in pavement distress recognition.

And when focusing on the HRIR tasks that only need a small part of the image to identify objects (traffic sign and WSI), we are not surprised to see that our proposed method also outperforms all comparison methods. This finding further tells that with our dual attention MIL architecture, the introduced sub-salient patches will not disturb model performance in simple situations, where the objects are concentrated in a few regions. Another experimental conclusion is that even in these scenes, instead of misdirecting model, random patch drop strategy with appropriate ratio could still improve model performance. This conclusion demonstrates the severe spatial redundancy in high-



(a) the most salient image patches

(b) the combination of most salient image patches and sub-salient image patches

(c) the original images

Figure 4. The visualization results on CQU-BPDD dataset and Functional Map of the World dataset.

resolution images, thus random patch drop strategy rarely loses all object information under appropriate drop ratio.

4.5. Ablation study

As shown in Tab. 2, from the first row and the second row, it is obvious the DBPS strategy can significantly enhance the performance of the model. It is because that DBPS strategy could introduce more informative regions into model training to get a wider perspective. From the second and the third rows, we can observe that by generating a salient query to aggregate the sub-salient patch embeddings, the model performance can be further improved. From the last three rows, it can be observed that our method can achieve the best performance with the addition of random patch drop.

And as shown in Fig. 3, it can be observed that random patch drop outperforms random erasing and random patch erasing under all ratios, while random patch erasing performs better than random erasing. This conclusion along with the results from the fourth to sixth rows in Tab. 2 proves that drop pattern and patch pattern are more appropriate than erasing pattern and rectangle pattern in HRIR.

4.6. Visualization

According to Fig. 4, we can firstly observe that the visualization results of salient patches are highly consistent with the most discriminative regions, but provide unclear information due to the small coverage area. Besides, although including some backgrounds, the visualization results of sub-salient patches could completely supplement the rest regions of interest. The combination of them can accurately and completely cover the regions of objects and corresponding context, which demonstrates the validity of our method.

5. Conclusion

In this paper, we propose the dual-buffer patch selection (DBPS) method to increase the number of image patches used in training HRIR models while keeping computational resource consumption at a low level. To suppress the uninformative background information in the sub-salient image patches, we devise a dual-attention MIL architecture to generate a salient query for aggregating sub-salient patch embeddings. Additionally, we introduce an efficient random patch drop training strategy to uncover informative image regions while reducing both the training time and GPU memory usage. Experimental results demonstrate the effectiveness of our approach in terms of accuracy and training consumption across various HRIR tasks and datasets.

6. Acknowledgments

This work was supported by the Natural Science Foundation of Tianjin, China (No.24JCZXJC00040), Shenzhen Science and Technology Program (No. JCYJ20240813114229039), the National Natural Science Foundation of China (No. 623B2056, 624B2072, 62302240), the Funding from SCCI, Dalian University of Technology (No. SCCI2023YB01), the Fundamental Research Funds for the Central Universities, the Supercomputing Center of Nankai University (NKSC).

References

[1] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *European Conference on Computer Vision*, pages 456–473. Springer, 2022. 5

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[3] Arian Bakhtiarnia, Qi Zhang, and Alexandros Iosifidis. Accurate gigapixel crowd counting by iterative zooming and refinement. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6510–6514. IEEE, 2024. 3

[4] Arian Bakhtiarnia, Qi Zhang, and Alexandros Iosifidis. Efficient high-resolution deep learning: A survey. *ACM Computing Surveys*, 56(7):1–35, 2024. 1, 2

[5] Benjamin Bergner, Christoph Lippert, and Aravindh Mahendran. Iterative patch selection for high-resolution image recognition. *arXiv preprint arXiv:2210.13007*, 2022. 2, 3, 4, 5, 6, 7

[6] Peter J Burt. Attention mechanisms for vision in a dynamic world. In *9th international conference on pattern recognition*, pages 977–978. IEEE Computer Society, 1988. 3

[7] Gabriele Campanella, Vitor Werneck Krauss Silva, and Thomas J Fuchs. Terabyte-scale deep multiple instance learning for classification and localization in pathology. *arXiv preprint arXiv:1805.06983*, 2018. 3

[8] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019. 6

[9] Xuanyao Chen, Zhijian Liu, Haotian Tang, Li Yi, Hang Zhao, and Song Han. Sparsevit: Revisiting activation sparsity for efficient high-resolution vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2061–2070, 2023. 2, 3

[10] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018. 5

[11] Jean-Baptiste Cordonnier, Aravindh Mahendran, Alexey Dosovitskiy, Dirk Weissenborn, Jakob Uszkoreit, and Thomas Unterthiner. Differentiable patch selection for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2351–2360, 2021. 2, 3, 4, 6

[12] Sean M Culhane and John K Tsotsos. An attentional prototype for early vision. In *Computer Vision—ECCV'92: Second European Conference on Computer Vision Santa Margherita Ligure, Italy, May 19–22, 1992 Proceedings 2*, pages 551–560. Springer, 1992. 3

[13] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997. 3

[14] Mingyu Ding, Xiaochen Lian, Linjie Yang, Peng Wang, Xiaojie Jin, Zhiwu Lu, and Ping Luo. Hr-nas: Searching efficient high-resolution neural architectures with lightweight transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2982–2992, 2021. 2

[15] Nanqing Dong, Michael Kampffmeyer, Xiaodan Liang, Zeya Wang, Wei Dai, and Eric Xing. Reinforced auto-zoom net: towards accurate and fast breast cancer segmentation in whole-slide images. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 317–325. Springer, 2018. 2, 3

[16] Alexey DOSOVITSKIY. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[17] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4438–4446, 2017. 2

[18] Along He, Tao Li, Ning Li, Kai Wang, and Huazhu Fu. Cabnet: Category attention block for imbalanced diabetic retinopathy grading. *IEEE Transactions on Medical Imaging*, 40(1):143–153, 2020. 6

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2, 6

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 5

[21] Guixin Huang, Sheng Huang, Luwen Huangfu, and Dan Yang. Weakly supervised patch label inference network with image pyramid for pavement diseases recognition in the wild. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7978–7982. IEEE, 2021. 6

[22] Gao Huang, Yulin Wang, Kangchen Lv, Haojun Jiang, Wenhui Huang, Pengfei Qi, and Shiji Song. Glance and focus networks for dynamic visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4605–4621, 2022. 6

[23] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*, pages 2127–2136. PMLR, 2018. 3, 6

[24] Angelos Katharopoulos and François Fleuret. Processing megapixel images with deep attention-sampling models. In *International Conference on Machine Learning*, pages 3282–3291. PMLR, 2019. 2, 3, 4, 5

[25] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology*, 4(4):219–227, 1985. 3

[26] Fanjie Kong and Ricardo Henao. Efficient classification of very large images with tiny objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2384–2394, 2022. 2, 3, 4, 6

[27] Marvin Lerousseau, Maria Vakalopoulou, Marion Classe, Julien Adam, Enzo Battistella, Alexandre Carré, Théo Estienne, Théophraste Henry, Eric Deutsch, and Nikos Paragios. Weakly supervised multiple instance learning histopathological tumor segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, pages 470–479. Springer, 2020. 3

[28] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2021. 3, 4, 6

[29] Honglin Li, Yunlong Zhang, Pingyi Chen, Zhongyi Shui, Chenglu Zhu, and Lin Yang. Rethinking transformer for long contextual histopathology whole slide image analysis. *arXiv preprint arXiv:2410.14195*, 2024. 3

[30] Tao Li, Yingqi Gao, Kai Wang, Song Guo, Hanruo Liu, and Hong Kang. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences*, 501:511–522, 2019. 5

[31] 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. 6

[32] 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. 5

[33] Xin Liu, Rong Qin, Junchi Yan, and Jufeng Yang. Ncmnet: Neighbor consistency mining network for two-view correspondence pruning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 4

[34] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021. 6

[35] Wenqi Lu, Simon Graham, Mohsin Bilal, Nasir Rajpoot, and Fayyaz Minhas. Capturing cellular topology in multi-gigapixel pathology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 260–261, 2020. 2

[36] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. A full-image full-resolution end-to-end-trainable cnn framework for image forgery detection. *IEEE Access*, 8:133488–133502, 2020. 2

[37] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. *Advances in neural information processing systems*, 27, 2014. 2

[38] Athanasios Papadopoulos, Pawel Korus, and Nasir Memon. Hard-attention for scalable image classification. *Advances in Neural Information Processing Systems*, 34:14694–14707, 2021. 2, 3

[39] Ferran Parés, Anna Arias-Duart, Dario Garcia-Gasulla, Gema Campo-Francés, Nina Viladrich, Eduard Ayguadé, and Jesús Labarta. The mame dataset: on the relevance of high resolution and variable shape image properties. *Applied Intelligence*, 52(10):11703–11724, 2022. 5

[40] Hans Pinckaers, Bram Van Ginneken, and Geert Litjens. Streaming convolutional neural networks for end-to-end learning with multi-megapixel images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1581–1590, 2020. 2

[41] Rong Qin, Luwen Huangfu, Devon Hood, James Ma, and Sheng Huang. Kernel inversed pyramidal resizing network for efficient pavement distress recognition. In *International Conference on Neural Information Processing*, pages 302–312. Springer, 2022. 6

[42] Adria Recasens, Petr Kellnhofer, Simon Stent, Wojciech Matusik, and Antonio Torralba. Learning to zoom: a saliency-based sampling layer for neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 51–66, 2018. 2

[43] Carl F Sabottke and Bradley M Spieler. The effect of image resolution on deep learning in radiography. *Radiology: Artificial Intelligence*, 2(1):e190015, 2020. 2

[44] Furkan E Sahin. Long-range, high-resolution camera optical design for assisted and autonomous driving. In *photonics*, page 73. MDPI, 2019. 1

[45] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34:2136–2147, 2021. 3

[46] Yuehao Song, Xinggang Wang, Jingfeng Yao, Wenyu Liu, Jinglin Zhang, and Xiangmin Xu. Vitgaze: gaze following with interaction features in vision transformers. *Visual Intelligence*, 2(1):1–15, 2024. 2

[47] Wenhao Tang, Sheng Huang, Qiming Zhao, Ren Li, and Luwen Huangfu. An iteratively optimized patch label inference network for automatic pavement distress detection. *IEEE Transactions on Intelligent Transportation Systems*, 23 (7):8652–8661, 2021. 1, 2, 5

[48] Wenhao Tang, Sheng Huang, Xiaoxian Zhang, Fengtao Zhou, Yi Zhang, and Bo Liu. Multiple instance learning framework with masked hard instance mining for whole slide image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4078–4087, 2023. 3

[49] Wenhao Tang, Fengtao Zhou, Sheng Huang, Xiang Zhu, Yi Zhang, and Bo Liu. Feature re-embedding: Towards foundation model-level performance in computational pathology. *arXiv preprint arXiv:2402.17228*, 2024. 3

[50] David Tellez, Geert Litjens, Jeroen van der Laak, and Francesco Ciompi. Neural image compression for gigapixel histopathology image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):567–578, 2019. 2

[51] Kevin Thandiackal, Boqi Chen, Pushpak Pati, Guillaume Jaume, Drew FK Williamson, Maria Gabrani, and Orcun Goksel. Differentiable zooming for multiple instance learning on whole-slide images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 699–715. Springer, 2022. 2, 3

[52] Keyu Tian, Yi Jiang, Qishuai Diao, Chen Lin, Liwei Wang, and Zehuan Yuan. Designing bert for convolutional networks: Sparse and hierarchical masked modeling. *arXiv preprint arXiv:2301.03580*, 2023. 5

[53] Maria Tzelepi and Anastasios Tefas. Improving the performance of lightweight cnns for binary classification using quadratic mutual information regularization. *Pattern Recognition*, 106:107407, 2020. 2

[54] Burak Uzkent and Stefano Ermon. Learning when and where to zoom with deep reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12345–12354, 2020. 2, 3

[55] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*, 2016. 6

[56] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, 81:102559, 2022. 6

[57] Xuenian Wang, Shanshan Shi, Renao Yan, Qiehe Sun, Lianghui Zhu, Tian Guan, and Yonghong He. Task-oriented embedding counts: Heuristic clustering-driven feature finetuning for whole slide image classification. *arXiv preprint arXiv:2406.00672*, 2024. 3

[58] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, pages 1–8, 2024. 5

[59] Wenqing Zhao and Lijiao Xu. Weakly supervised target detection based on spatial attention. *Visual Intelligence*, 2(1): 2, 2024. 2