PyraE2E: Enhancing End-to-End WSI Analysis via Cross-Scale Super-Resolution

Yuechuan Lin[1], Yujian Liu[1], Weipeng Zhang[1], Yanyu Fan[1], Zikang Wang[1], Dongxu Shen[2], Liqin Fei[1], Xiaoli Liu[3], and Shidang Xu[1(✉)]

[1] South China University of Technology, Guangzhou, China
xsud@scut.edu.cn
[2] Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
[3] AiShiWeiLai AI Research, Beijing, China

Abstract. Offline feature extraction with pretrained encoders followed by multiple instance learning (MIL) aggregation is still the dominant paradigm for whole slide image (WSI) analysis, but the domain gap between natural and histopathology images limits representation quality. End-to-end optimization can reduce this gap by training the encoder with slide-level objectives, but it often suffers from (i) cost-driven fine-grained and morphological information loss caused by low-magnification small inputs and random patch sampling, and (ii) sparse supervision from slide labels alone. We propose PyraE2E, an end-to-end framework that turns the intrinsic multi-resolution WSI pyramid into a dense self-supervision signal via super-resolution (SR). For each sampled low-resolution (LR) patch at low magnification, we use its spatially aligned high-resolution (HR) patch at higher magnification as a reconstruction target, providing cross-scale pixel-level supervision while keeping computation bounded by operating on LR inputs. PyraE2E combines an embedded Cluster-Score Sampling module to select informative LR regions with a shared Global–Local Partitioned (GLP) encoder whose HR-informed features are jointly optimized by an SR reconstruction head and a slide-level prediction head. Joint training couples dense cross-scale supervision with slide-level objectives, improving end-to-end representations and downstream WSI prediction under controlled computational cost.

Keywords: Whole Slide Image · End-to-End Learning · Super-Resolution

1 Introduction

Whole slide images (WSIs) are gigapixel-scale, yet training supervision is often limited to a single slide-level label [5, 35, 59]. This severe imbalance between data size and label granularity makes WSI recognition challenging and motivates patch-based learning under weak supervision [23, 31, 32].

Yuechuan Lin, Yujian Liu, and Weipeng Zhang—Equal contribution.
(✉) Corresponding author.

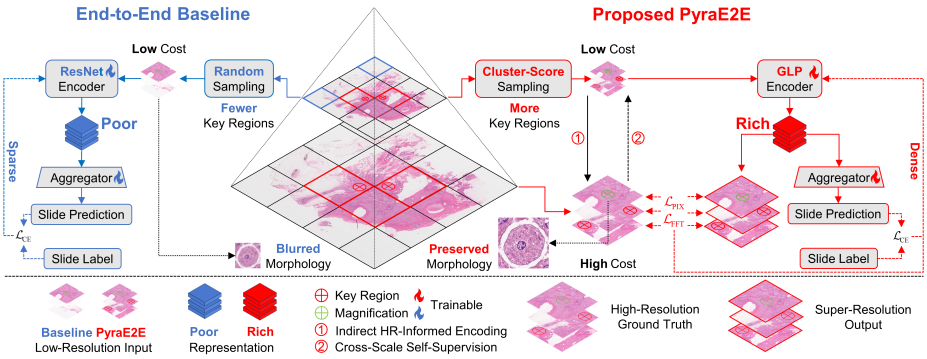


Fig. 1: Motivation of the proposed PyraE2E and comparison with common end-to-end baselines. Left: End-to-end baselines can yield suboptimal representation quality due to (i) fine-grained and morphological information loss from LR patches and random sampling, both used to reduce computational cost; (ii) sparse supervision from slide labels. Middle: A WSI pyramid and our motivation: LR patches are cheap but lose fine-grained morphology, so we encode LR inputs into HR-informed representations and use spatially aligned HR patches as supervision without incurring HR-input cost. Right: PyraE2E improves representation quality by (i) preserving fine-grained morphology under low-cost LR inputs via cluster-score sampling and HR-informed encoding, and (ii) injecting dense cross-scale reconstruction supervision.

A widely adopted solution is the two-stage weakly supervised multiple instance learning (MIL) paradigm: a pretrained encoder extracts offline patch embeddings, and MIL aggregates them for slide-level prediction [24, 40]. While this makes gigapixel processing feasible with only slide-level labels, it introduces a persistent domain gap: encoders pretrained on natural images are not optimized for histopathology morphology, yielding suboptimal slide representations and a performance bottleneck [28, 38, 47].

End-to-end learning is typically optimized with slide-level supervision and offers a principled remedy by jointly optimizing the encoder and slide-level prediction, allowing representations to adapt to pathology tasks and reducing the domain gap [53]. However, in practice, current end-to-end approaches may lead to relatively poor representations (Fig. 1, left). First, to reduce computation, practical end-to-end training commonly relies on low-magnification small patches [9,57] and random sampling [6, 55], which in turn cause fine-grained morphological information loss. Low-magnification small patches blur fine-grained cellular details, weaken high-frequency textures, and compress continuous morphological units (*e.g.*, cells, glands, tumor boundaries, and stroma) [61, 65]. In addition, cost-driven random sampling can miss morphologically critical regions, causing key evidence to be underrepresented during training [1, 29, 36]. Second, supervision from slide labels alone is sparse for a long encoder–aggregator chain, making it difficult to train intermediate representations and optimize the model [4]. Existing self-supervised auxiliaries [16, 26] remain difficult to use effectively in end-to-end WSI training under realistic budgets. Contrastive and distillation signals

are typically too coarse, as they mainly provide feature-level or instance-level supervision rather than dense morphological guidance [25, 58]. Reconstruction provides pixel-level supervision, but its cost becomes prohibitive at WSI scale, making it difficult to embed into end-to-end pipelines [2, 60]. Therefore, the sparsity bottleneck remains largely unresolved. Fundamentally, the bottleneck comes from the computational cost constraint due to the gigapixel scale of WSIs and the need to backpropagate through the encoder in end-to-end optimization, where blindly using low-magnification small patches overlooks the fine-grained morphological information inherent in the WSI itself. These observations raise a key question: under realistic budgets, can we train end-to-end models with low-magnification small inputs while still obtaining representations that encode high-magnification, large-patch morphological information, and leverage a proper self-supervised learning method to obtain dense supervision beyond slide labels?

Our answer comes from a simple but often overlooked fact that WSIs are stored as multi-resolution pyramids [12,19,27,64]. A low-resolution (LR) patch at low magnification is inexpensive to process but discards fine-grained morphology, whereas its spatially aligned high-resolution (HR) patch at higher magnification preserves cellular detail but is costly to encode (Fig. 1, middle). Because of computational constraints, end-to-end pipelines are often forced to operate on LR inputs. Under this constraint, the challenge is to recover fine-grained morphology from LR inputs. This suggests a natural strategy: instead of paying the full cost of directly encoding HR patches, we seek a lightweight mechanism that injects HR information while keeping LR patches as the encoder input. The WSI pyramid makes this learnable: for each sampled LR location, the corresponding spatially aligned HR patch naturally provides dense cross-scale supervision through pixel-level reconstruction without requiring additional annotation. As illustrated in Fig. 1 (right), this aligns directly with super-resolution (SR) [33, 34, 44, 67], in which the model takes an LR patch at low magnification as input, predicts the corresponding HR patch at higher magnification, and is trained using reconstruction losses against the HR target. In end-to-end WSI learning, this SR objective serves as a well-suited auxiliary task because it encourages the encoder to preserve fine-grained morphology, provides dense supervision beyond slide labels, and keeps computation controllable by operating on LR inputs [15, 52].

Based on this insight, we propose PyraE2E, a pyramid-driven end-to-end framework for whole slide image analysis via super-resolution. PyraE2E selects low-cost LR patches using an embedded Cluster-Score Sampling module that efficiently leverages internal features and attention cues. The selected patches are processed by a shared Global–Local Partitioned (GLP) encoder that produces HR-informed representations. These shared features are then routed to two heads for joint optimization: (i) an SR branch that reconstructs the aligned HR patch at a higher magnification level, supervised by pixel-domain and frequency-domain losses to densify training signals beyond slide labels; and (ii) a slide-level prediction branch for the downstream task. By optimizing both branches, PyraE2E improves representation quality and boosts slide-level prediction performance under controlled computational cost. In summary, our contributions are:

1. **Embedded Cluster-Score Sampling Module.** We introduce an embedded sampling module that exploits internal features and attention cues within the end-to-end framework to efficiently select informative patches.
2. **SR Supervision for End-to-End WSI Learning.** We introduce SR supervision into end-to-end WSI training, where spatially aligned HR patches provide dense pixel- and frequency-domain signals to the lightweight Global–Local Partitioned (GLP) encoder beyond slide-level labels.
3. **PyraE2E Framework.** We propose PyraE2E, a pyramid-driven end-to-end framework based on SR, which turns the intrinsic WSI pyramid into a self-supervised signal while maintaining controllable computational cost.

2 Related Work

2.1 Two-Stage MIL for WSI Analysis

Whole slide images (WSIs) provide high-resolution tissue morphology for computational pathology tasks [10,11,51], but their gigapixel scale makes direct processing computationally prohibitive. Therefore, most existing methods follow a two-stage multiple instance learning (MIL) paradigm: an ImageNet-pretrained encoder [13] extracts patch-level features offline, and a separate aggregator produces slide-level predictions. Attention-based MIL [24,38] learns instance weights to perform weighted pooling, yielding compact slide representations while suppressing noise from non-diagnostic regions. Building on this idea, prior work improves aggregation through sparse attention-based instance selection [63, 68], long-range dependency modeling [30, 62], and structure-aware reasoning [32, 69] under weak supervision. The use of fixed offline features in two-stage pipelines limits task adaptation and exacerbates domain gaps, motivating slide-level supervised end-to-end learning.

2.2 End-to-end learning for WSI Analysis

Weakly Supervised End-to-End Learning. End-to-end learning jointly optimizes feature extraction and prediction, reducing the domain gap induced by fixed pretrained encoders. Current end-to-end computational pathology methods adopt either instance-level or slide-level supervision. Instance-level approaches [39, 45, 46] train encoders using patch-wise pseudo-labels to ease optimization under gigapixel constraints, but they decouple from the global slide context, serving as a pragmatic approximation rather than true end-to-end modeling. Currently, the mainstream end-to-end WSI paradigm is slide-level supervised learning [9, 14, 43, 49, 53, 55], which jointly optimizes the encoder and MIL aggregator using slide labels. Yet its representation quality often suffers due to compute-driven fine-grained and morphological information loss from (i) low-magnification small patches [9, 49] and cost-driven random sampling [6, 53], and (ii) sparse slide-level supervision [14, 43] over a long encoder–aggregator chain.

Self-Supervised End-to-End Learning. To address the sparsity of slide-level supervision in end-to-end WSI learning, recent studies [2, 16, 25, 26, 58, 60]

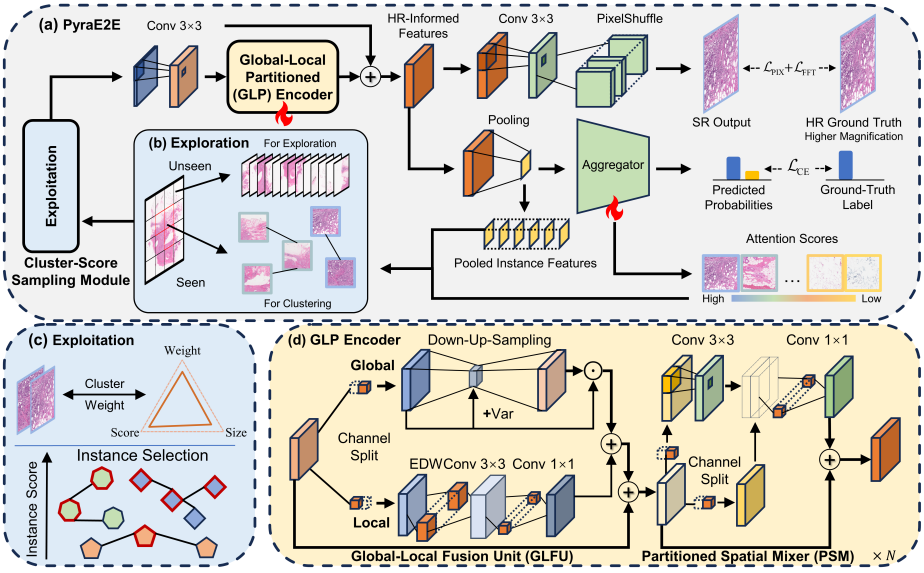


Fig. 2: Overview of the proposed PyraE2E framework. (a) Starting from low-resolution inputs sampled by the Cluster-Score Sampling module, patches are processed by the Global–Local Partitioned (GLP) encoder to produce features with high-resolution information. These features are then split into a super-resolution branch and a slide-level prediction branch. (b-c) The Cluster-Score Sampling module consists of two components: exploration and exploitation, both guided by network features and attention scores. (d) The GLP encoder is composed of N stacked Global–Local Fusion Units and Partitioned Spatial Mixers.

introduce self-supervised objectives as auxiliary training signals. These methods can be broadly categorized into contrastive learning-based approaches and reconstruction-based approaches. Contrastive learning–based methods [16, 25, 58] introduce extra-cost contrastive objectives at the instance or slide level to enhance representation discrimination under weak supervision, but the signal is coarse-grained and remains indirect compared to pixel-level supervision. Reconstruction-based methods [2, 60] employ generative reconstruction losses to provide denser training signals during end-to-end optimization. However, the gigapixel scale of WSIs makes training prohibitively expensive, so both categories still largely rely on pretrained feature extractors rather than fully end-to-end optimization from raw WSI patches. As a result, achieving dense, pixel-level supervision in fully end-to-end WSI training remains challenging.

3 Method

End-to-end WSI learning is commonly constrained by two factors: (i) fine-grained and morphological information loss caused by low-magnification small inputs and

cost-driven random patch sampling, and (ii) sparse supervision when the encoder and aggregator are optimized only with slide labels. PyraE2E (Fig. 2) addresses these limitations by (1) selecting informative low-resolution (LR) patches efficiently under a fixed budget and (2) injecting dense cross-scale supervision via super-resolution (SR) reconstruction.

PyraE2E leverages the intrinsic pyramid structure of WSIs: an LR patch at low magnification (*e.g.*, $5\times$, 512×512) corresponds to the same physical tissue region as a higher-resolution (HR) patch at higher magnification (*e.g.*, $20\times$, 2048×2048). Using this cross-scale correspondence, PyraE2E combines (a) a Cluster-Score Sampling module to select LR patches and (b) a lightweight Global–Local Partitioned (GLP) encoder that learns HR-informed representations from LR inputs. The shared features are jointly optimized by two objectives: (i) an SR branch that reconstructs the paired HR patch under pixel- and frequency-domain losses, and (ii) a slide-level prediction branch using attention-based MIL. Importantly, MIL attention and instance embeddings are fed back to update sampling states, forming a closed-loop end-to-end training pipeline.

3.1 Cluster-Score Sampling Module

Cost-driven random sampling can overlook diagnostically critical tissue, while purely attention-based sampling tends to over-focus on a small set of high-score instances and reduce morphological diversity. Inspired by the exploration–exploitation trade-off, our Cluster-Score Sampling module balances coverage and utility by jointly operating on instance features and attention scores.

Exploration. We aim to expose the input to more instances rather than letting it be dominated by a small subset in the exploration stage. As shown in Fig. 2(b), in each epoch, we sample K instances per slide. We reserve K_e instances for exploration by uniformly sampling from the unseen set $\mathcal{U}$, and allocate the remaining $K_r = K - K_e$ instances to exploitation from the seen set $\mathcal{S}$,

$$\mathcal{U} = \{i \mid s_i = 0\}, \qquad \mathcal{S} = \{i \mid s_i \geq 1\}. \tag{1}$$

Here, s_i denotes how many times instance i has been sampled so far. To enable stable online sampling, for each instance i, we maintain and update an online state $(a_i, \mathbf{f}_i, s_i, c_i)$: attention score a_i, instance feature $\mathbf{f}_i$, seen count s_i, and cluster label c_i. All states are initialized as a_i=0, $\mathbf{f}_i$=$\mathbf{0}$, s_i=0, and c_i=-1. In the first epoch, since all instances satisfy $s_i = 0$, we perform exploration-only sampling to bootstrap these states: each sampled instance yields a feature $\mathbf{f}_i$ from the encoder and an attention score a_i from the MIL aggregator (Sec. 3.3). Then, we run k-means every T_c epochs on the collection of instance features to assign cluster labels c_i. The number of clusters is chosen adaptively based on the number of seen instances (details in Sec. 4.1). Therefore, each explored instance obtains its online state $(a_i, \mathbf{f}_i, s_i, c_i)$ via a single forward pass. In particular, its seen count is updated from s_i=0 to 1, and it is thus moved into the seen set $\mathcal{S}$.

Exploitation. Exploitation aims to prioritize informative tissue while maintaining diversity across tissue types in the seen set $\mathcal{S}$. Let n_c be the size of cluster c

and $\bar{a}_c$ its mean attention. We compute a stage-wise cluster weight (Fig. 2(c)):

$$w_c \propto \begin{cases} \mathrm{Norm}(n_c), & t < T_w, \\ (1-\lambda)\,\mathrm{Norm}(n_c) + \lambda\,\mathrm{Norm}(\bar{a}_c), & t \geq T_w, \end{cases} \tag{2}$$

where T_w is a warm-up epoch to avoid early noise from unstable attention, $\mathrm{Norm}(\cdot)$ denotes min–max normalization, and λ balances coverage and attention-based importance. We allocate quotas q_c to clusters using a largest-remainder scheme:

$$q_c = \left\lfloor K_r \cdot \frac{w_c}{\sum_{c'} w_{c'}} \right\rfloor, \tag{3}$$

and distribute the remaining $K_r - \sum_c q_c$ slots to clusters with the largest fractional remainders.

Within each cluster, when $t < T_w$ we sample uniformly to avoid early noise from unstable attention. After T_w, we rank instances with a seen-penalized score:

$$r_i = \mathrm{Norm}(a_i) - \beta\,\mathrm{Norm}(s_i), \tag{4}$$

and select the top-q_c instances according to r_i within each cluster. They are then forwarded through the model and we use refreshed cluster labels to update their states $(a_i, \mathbf{f}_i, s_i, c_i)$. We update a_i and $\mathbf{f}_i$ with EMA:

$$\mathbf{f}_i \leftarrow \alpha_2 \mathbf{f}_i + (1-\alpha_2)\hat{\mathbf{f}}_i, \qquad a_i \leftarrow \alpha_1 a_i + (1-\alpha_1)\hat{a}_i, \tag{5}$$

where $\hat{\mathbf{f}}_i$ and $\hat{a}_i$ are from the current forward pass, and $\alpha_1, \alpha_2 \in [0, 1)$.

3.2 Global–Local Partitioned (GLP) Encoder

Most end-to-end WSI pipelines rely on low magnification and small patches to meet computational constraints, which may sacrifice subtle morphology. PyraE2E employs the Global–Local Partitioned (GLP) encoder (Fig. 2(d)), an efficient shared backbone that learns HR-informed representations from LR inputs. We emphasize that "HR-informed" refers to learning under HR-supervised SR losses (Sec. 3.3), which means HR patches are used only as reconstruction targets and are never fed into the encoder at training or inference.

Given an LR patch $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$, we first extract shallow features and then refine them with N identical blocks. Each block contains (i) a Global–Local Fusion Unit (GLFU) and (ii) a Partitioned Spatial Mixer (PSM). GLFU injects coarse contextual cues while preserving local details; PSM performs spatial mixing on only a subset of channels for efficiency. Within each block, we apply GLFU and add it back via a residual connection, followed by PSM with a second residual.

Global–Local Fusion Unit (GLFU). Inspired by global–local representation modeling [20, 41, 67], GLFU fuses a global stream and a local stream in a lightweight manner. Let $\mathbf{F}_{t-1} \in \mathbb{R}^{C\times H\times W}$ denote the input to block t (with $\mathbf{F}_0$

as shallow features extracted by a 3×3 convolution). We apply a 1×1 projection and obtain two feature streams via channel partition:

$$[\mathbf{X}, \mathbf{Y}] = \mathrm{Split}\big(\mathrm{Conv}_{1\times1}^{C\to 2C}(\mathbf{F}_{t-1})\big)\,, \tag{6}$$

where $\mathrm{Split}(\cdot)$ divides channels into two equal parts, and $\mathrm{Conv}_{1\times1}^{a\to b}(\cdot)$ denotes a 1×1 convolution that maps channels from a to b while preserving resolution.

The global unit downsamples the global stream $\mathbf{X}$ with strided max pooling (factor s) and applies a depthwise 3×3 convolution:

$$\mathbf{X}_s = \mathrm{DWConv}_{3\times3}(\mathrm{MaxPool}_s(\mathbf{X}))\,, \tag{7}$$

and produces a gated global feature:

$$\mathbf{X}_g = \mathbf{X} \odot \mathrm{Interp}(\boldsymbol{\alpha} \odot \mathbf{X}_s + \boldsymbol{\beta} \odot \mathrm{Var}(\mathbf{X}))\,, \tag{8}$$

where $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^{C\times1\times1}$ are learnable per-channel weights, $\mathrm{Var}(\cdot)$ is channel-wise variance, and $\mathrm{Interp}(\cdot)$ applies a 1×1 convolution followed by GELU and then upsamples to the original resolution.

The local unit refines local stream $\mathbf{Y}$ using a lightweight local mixer that performs depthwise expansion, pointwise mixing, and projection:

$$\mathbf{Y}_l = \mathcal{L}(\mathbf{Y}). \tag{9}$$

Here, $\mathcal{L}(\cdot)$ is a sequential convolutional module consisting of (i) an expanded depthwise 3×3 convolution that lifts channels from C to $2C$ (implemented with groups $= C$), followed by (ii) a 1×1 pointwise mixing in the $2C$ space with GELU, and (iii) a 1×1 projection back to C. We fuse both streams and project:

$$\mathcal{G}(\mathbf{F}_{t-1}) = \mathrm{Conv}_{1\times1}^{C\to C}(\mathbf{X}_g + \mathbf{Y}_l)\,, \tag{10}$$

where $\mathcal{G}(\mathbf{F}_{t-1}) \in \mathbb{R}^{C\times H\times W}$ denotes the GLFU output given the input $\mathbf{F}_{t-1}$.

Partitioned Spatial Mixer (PSM). PSM mixes spatial information on only a subset of channels. Given $\mathbf{F}'_{t-1} = \mathbf{F}_{t-1} + \big(\mathcal{G}(\mathbf{F}_{t-1})\big) \in \mathbb{R}^{C\times H\times W}$, which serves as the input to the subsequent PSM after GLFU, we first expand channels with a 1×1 convolution with GELU:

$$\mathbf{H} = \mathrm{Conv}_{1\times1}^{C\to 2C}(\mathbf{F}'_{t-1}). \tag{11}$$

We split $\mathbf{H} = [\mathbf{H}_p, \mathbf{H}_r]$ along channels, where $\mathbf{H}_p$ contains a fraction p of channels. Only $\mathbf{H}_p$ undergoes 3×3 spatial mixing with GELU, followed by concatenation and projection back to C:

$$\mathcal{M}(\mathbf{F}'_{t-1}) = \mathrm{Conv}_{1\times1}^{2C\to C}\Big([\mathrm{Conv}_{3\times3}(\mathbf{H}_p), \mathbf{H}_r]\Big), \tag{12}$$

where $[\cdot, \cdot]$ denotes channel concatenation, and $\mathcal{M}(\mathbf{F}'_{t-1}) \in \mathbb{R}^{C\times H\times W}$ is the PSM output given the input $\mathbf{F}'_{t-1}$.

3.3 End-to-End Joint Optimization

End-to-end WSI baselines supervise the encoder only via slide-level loss, which is often too sparse for long pipelines. We therefore adopt joint optimization: the GLP encoder produces a shared feature map $\mathbf{z} = \mathbf{F}_0 + \mathbf{F}_N$, which is routed to (i) an SR branch supervised by pixel and frequency reconstruction losses using cross-scale paired HR crops obtained via pyramid coordinate mapping (same tissue region), and (ii) a slide-level prediction branch for classification.

SR Branch. Given the HR-informed feature map $\mathbf{z} \in \mathbb{R}^{C\times H\times W}$, we reconstruct an HR (super-resolved) RGB patch via a 3×3 convolution followed by a PixelShuffle layer [50] with upscaling factor r:

$$\hat{\mathbf{x}}^{\mathrm{SR}} = \mathrm{PixelShuffle}_r\Big(\mathrm{Conv}_{3\times 3}^{C\to 3r^2}(\mathbf{z})\Big) \in \mathbb{R}^{3\times(rH)\times(rW)}. \tag{13}$$

Here, $\mathrm{PixelShuffle}_r(\cdot)$ upsamples features by rearranging channel groups of size r^2 into an $r \times r$ spatial neighborhood, converting a tensor of shape $(3r^2) \times H \times W$ into an RGB image of shape $3 \times (rH) \times (rW)$. We supervise $\hat{\mathbf{x}}^{\mathrm{SR}}$ using the paired HR patch $\mathbf{x}^{\mathrm{HR}}$ cropped from the WSI pyramid, with a pixel-domain and a frequency-domain loss:

$$\mathcal{L}_{\mathrm{PIX}} = \big\|\hat{\mathbf{x}}^{\mathrm{SR}} - \mathbf{x}^{\mathrm{HR}}\big\|_1, \qquad \mathcal{L}_{\mathrm{FFT}} = \big\|\mathcal{F}\big(\hat{\mathbf{x}}^{\mathrm{SR}}\big) - \mathcal{F}\big(\mathbf{x}^{\mathrm{HR}}\big)\big\|_1, \tag{14}$$

where $\|\cdot\|_1$ denotes the element-wise ℓ_1 norm, and $\mathcal{F}(\cdot)$ denotes the 2D fast Fourier transform (FFT) applied per channel. The FFT term encourages preservation of high-frequency morphology that may be smoothed by pixel losses alone.

Slide-Level Prediction Branch. We obtain an instance embedding for each patch via attention pooling over spatial locations:

$$\hat{\mathbf{f}} = \mathrm{AttnPool}(\mathbf{z}) \in \mathbb{R}^D. \tag{15}$$

For each WSI, the instance embeddings $\{\hat{\mathbf{f}}_i\}_{i=1}^K$ are fed into an attention-based aggregator [24, 53], which computes the attention scores $\{\hat{a}_i\}_{i=1}^K$ and slide-level predicted probabilities $\hat{\mathbf{y}}$ as

$$\begin{aligned} \hat{a}_i &= \frac{\exp\Big(\mathbf{v}^\top\big(\tanh(\mathbf{V}\hat{\mathbf{f}}_i^\top)\odot\sigma(\mathbf{U}\hat{\mathbf{f}}_i^\top)\big)\Big)}{\sum_{j=1}^K \exp\Big(\mathbf{v}^\top\big(\tanh(\mathbf{V}\hat{\mathbf{f}}_j^\top)\odot\sigma(\mathbf{U}\hat{\mathbf{f}}_j^\top)\big)\Big)}, \\ \hat{\mathbf{y}} &= \mathrm{Aggregator}\Big(\{\hat{\mathbf{f}}_i\}_{i=1}^K, \{\hat{a}_i\}_{i=1}^K\Big), \end{aligned} \tag{16}$$

where $\odot$ denotes element-wise multiplication and $\sigma(\cdot)$ is the sigmoid gate. We return $\{\hat{a}_i\}_{i=1}^K$ as sampling scores and $\{\hat{\mathbf{f}}_i\}_{i=1}^K$ as features for clustering in the Cluster-Score Sampling module (Sec. 3.1).

Overall Objective. We jointly optimize SR reconstruction and WSI classification with a weighted sum of losses:

$$\mathcal{L}_{\mathrm{total}} = \alpha\,\mathcal{L}_{\mathrm{PIX}} + \beta\,\mathcal{L}_{\mathrm{FFT}} + \gamma\,\mathcal{L}_{\mathrm{CE}} + \delta\,\mathcal{L}_{\mathrm{KLD}}. \tag{17}$$

Here, $\mathcal{L}_{\mathrm{CE}}$ is the slide-level cross-entropy loss, $\mathcal{L}_{\mathrm{KLD}}$ is a cluster-wise KL regularizer that encourages the attention mass to be more evenly distributed among instances within the same cluster, mitigating attention collapse to a few patches.

In summary, our end-to-end training loop proceeds as follows: (i) The Cluster-Score Sampling module selects K instances per WSI from LR pyramid level. (ii) GLP encodes each selected LR patch and outputs a shared feature map. (iii) The SR head reconstructs the paired HR patch, while the MIL head outputs $\hat{\mathbf{f}}_i$ and $\hat{a}_i$. (iv) We backpropagate $\mathcal{L}_{\mathrm{total}}$ to update the shared encoder and both heads. (v) We update the online state $(a_i, \mathbf{f}_i, s_i, c_i)$, where a_i and $\mathbf{f}_i$ are updated by EMA, s_i is incremented upon sampling, and c_i is refreshed by running k-means every T_c epochs, closing the loop between sampling and representation learning.

4 Experiments

4.1 Experiment Settings

Datasets. To evaluate PyraE2E, we conduct experiments on three public WSI benchmarks: TCGA-RCC, TCGA-NSCLC, and TCGA-BRCA. TCGA-RCC is a three-class classification task, including 940 slides from clear cell (TCGA-KIRC, 519), papillary (TCGA-KIRP, 300), and chromophobe (TCGA-KICH, 121) renal cell carcinoma. TCGA-NSCLC is a binary lung cancer subtyping task between LUAD and LUSC, containing 1,053 slides (541 LUAD and 512 LUSC). TCGA-BRCA is a binary breast cancer subtyping task between IDC and ILC, containing 985 slides (787 IDC and 198 ILC). All datasets provide slide-level labels.

Evaluation Metrics. We evaluate classification performance using accuracy (ACC) and the area under the ROC curve (AUC). For the quality of super-resolution (SR), we report Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS).

Compared Methods. We compare PyraE2E with representative and state-of-the-art WSI classification paradigms on the above three datasets, covering two-stage MIL, foundation-model-based, self-supervised, and end-to-end pipelines. Specifically, for two-stage MIL baselines, we adopt a ResNet [21] encoder and evaluate MeanMIL, MaxMIL, ABMIL [24], CLAM [38], TransMIL [48], RRT-MIL [54], FOCUS [18], SMMILE [17], and 2D Mamba [66]. For foundation models, we evaluate UNI [7] and CONCH [37]. For self-supervised learning, we include SimCLR [8] and CDSR [36] in the main comparison, and further report ICMIL [56] in the efficiency comparison. For end-to-end methods, we compare against C2C [49], Streaming [42], the ABMILX end-to-end pipeline [53]. To ensure a fair comparison, we use ABMIL and ABMILX as aggregators for feature-based pipelines based on foundation-model and self-supervised features, while end-to-end baselines use their original aggregation heads.

Implementation Details. We use $5\times$ RGB patches of 512×512 as inputs. The Cluster-Score Sampling module selects K=160 instances per WSI from LR pyramid level, with a warm-up of T_w=30 epochs. Every T_c=10 epochs, we adaptively

Table 1: Comparison with state-of-the-art methods for WSI classification on TCGA-RCC, TCGA-NSCLC, and TCGA-BRCA (mean±std). "E2E" denotes whether the model is trained in an end-to-end manner. "–L" denotes the method with ABMIL aggregator, while "–X" denotes the method with ABMILX aggregator.

Encoder	Method	E2E	TCGA-RCC		TCGA-NSCLC		TCGA-BRCA	
			ACC	AUC	ACC	AUC	ACC	AUC
ResNet	MeanMIL	✗	$85.4_{\pm 1.4}$	$95.2_{\pm 0.3}$	$80.8_{\pm 1.3}$	$89.6_{\pm 0.5}$	$83.8_{\pm 1.0}$	$87.0_{\pm 2.3}$
	MaxMIL	✗	$88.2_{\pm 2.0}$	$97.6_{\pm 0.4}$	$85.2_{\pm 2.0}$	$93.6_{\pm 1.7}$	$83.8_{\pm 4.2}$	$83.6_{\pm 0.8}$
	ABMIL	✗	$89.1_{\pm 0.7}$	$98.4_{\pm 0.4}$	$88.3_{\pm 2.1}$	$95.8_{\pm 1.4}$	$88.9_{\pm 1.5}$	$92.0_{\pm 3.3}$
	CLAM	✗	$90.5_{\pm 0.9}$	$98.7_{\pm 0.1}$	$89.9_{\pm 2.1}$	$95.5_{\pm 2.2}$	$88.7_{\pm 1.2}$	$93.5_{\pm 1.7}$
	TransMIL	✗	$90.2_{\pm 0.9}$	$97.9_{\pm 0.5}$	$85.7_{\pm 1.9}$	$93.7_{\pm 1.3}$	$87.9_{\pm 1.8}$	$91.9_{\pm 0.8}$
	RRT-MIL	✗	$90.3_{\pm 0.8}$	$99.1_{\pm 0.1}$	$87.0_{\pm 2.5}$	$95.6_{\pm 1.7}$	$87.8_{\pm 2.3}$	$93.8_{\pm 2.6}$
	FOCUS	✗	$91.6_{\pm 2.3}$	$98.9_{\pm 0.7}$	$89.5_{\pm 1.7}$	$96.0_{\pm 1.7}$	$87.7_{\pm 1.7}$	$93.0_{\pm 2.5}$
	SMMILE	✗	$91.2_{\pm 1.2}$	$98.9_{\pm 0.3}$	$86.8_{\pm 2.5}$	$94.7_{\pm 1.8}$	$85.0_{\pm 4.8}$	$92.9_{\pm 2.0}$
	2D Mamba	✗	$90.6_{\pm 1.5}$	$98.3_{\pm 0.9}$	$89.6_{\pm 1.7}$	$96.2_{\pm 1.3}$	$90.1_{\pm 1.7}$	$94.0_{\pm 1.4}$
UNI	ABMIL	✗	$92.0_{\pm 1.0}$	$99.1_{\pm 0.1}$	$90.7_{\pm 0.7}$	$97.1_{\pm 0.5}$	$91.2_{\pm 0.5}$	$94.7_{\pm 0.4}$
	ABMILX	✗	$91.5_{\pm 0.9}$	$99.0_{\pm 0.2}$	$90.1_{\pm 2.6}$	$96.1_{\pm 2.1}$	$\mathbf{93.2}_{\pm 0.8}$	$\mathbf{95.1}_{\pm 0.5}$
CONCH	ABMIL	✗	$91.6_{\pm 0.5}$	$98.1_{\pm 0.2}$	$91.0_{\pm 0.8}$	$97.0_{\pm 0.5}$	$90.6_{\pm 0.6}$	$95.0_{\pm 1.0}$
	ABMILX	✗	$92.8_{\pm 0.6}$	$\mathbf{99.3}_{\pm 0.3}$	$91.8_{\pm 1.0}$	$97.0_{\pm 0.6}$	$92.0_{\pm 0.3}$	$94.7_{\pm 0.9}$
ResNet	SimCLR-L	✗	$90.1_{\pm 1.2}$	$98.0_{\pm 0.4}$	$89.7_{\pm 2.3}$	$96.6_{\pm 1.2}$	$88.3_{\pm 3.7}$	$92.2_{\pm 1.2}$
	SimCLR-X	✗	$90.2_{\pm 1.7}$	$97.7_{\pm 0.3}$	$90.1_{\pm 2.2}$	$96.7_{\pm 1.3}$	$89.2_{\pm 2.3}$	$92.9_{\pm 1.6}$
CDSR	ABMIL	✗	$89.9_{\pm 2.4}$	$97.8_{\pm 0.5}$	$88.3_{\pm 2.1}$	$95.7_{\pm 0.5}$	$87.3_{\pm 2.4}$	$90.9_{\pm 3.6}$
	ABMILX	✗	$90.9_{\pm 1.2}$	$98.2_{\pm 0.3}$	$89.0_{\pm 1.3}$	$95.9_{\pm 0.5}$	$87.4_{\pm 1.6}$	$91.5_{\pm 3.5}$
ResNet	C2C	✓	$90.3_{\pm 0.7}$	$98.0_{\pm 0.2}$	$88.8_{\pm 1.7}$	$94.8_{\pm 1.1}$	$88.0_{\pm 1.9}$	$88.6_{\pm 0.8}$
	Streaming	✓	$90.2_{\pm 1.1}$	$97.9_{\pm 0.2}$	$90.2_{\pm 1.1}$	$96.2_{\pm 0.4}$	$88.0_{\pm 1.1}$	$88.0_{\pm 2.9}$
	ABMILX	✓	$91.3_{\pm 1.3}$	$98.1_{\pm 0.1}$	$90.2_{\pm 1.0}$	$96.3_{\pm 0.3}$	$87.0_{\pm 3.9}$	$90.1_{\pm 4.5}$
GLP	PyraE2E-L	✓	$90.6_{\pm 2.0}$	$97.8_{\pm 0.6}$	$91.4_{\pm 0.8}$	$97.5_{\pm 0.7}$	$90.0_{\pm 0.7}$	$92.5_{\pm 1.0}$
	PyraE2E-X	✓	$\mathbf{93.1}_{\pm 1.2}$	$98.3_{\pm 0.2}$	$\mathbf{92.0}_{\pm 1.4}$	$\mathbf{97.6}_{\pm 0.4}$	$90.5_{\pm 1.3}$	$93.5_{\pm 0.7}$

set the number of clusters $c \in \{2, 4, 6, 8\}$ based on the predefined seen-instance count thresholds $\{500, 1500, 3500\}$. The shared GLP encoder uses N=8 identical blocks to produce HR-informed features. In the SR branch, we reconstruct a $\times 4$ super-resolved RGB patch with r=4 (to $20\times$, 2048×2048), supervised by the spatially corresponding HR patch only during training; inference requires no HR ground truth. We jointly optimize the SR and slide-level prediction branches with loss weights $(\alpha, \beta, \gamma, \delta) = (1.0, 0.05, 0.5, 0.5)$ for $\mathcal{L}_{\text{PIX}}, \mathcal{L}_{\text{FFT}}, \mathcal{L}_{\text{CE}}, \mathcal{L}_{\text{KLD}}$, respectively. Please refer to Supplementary Material for more details.

4.2 Model Performance Analysis

Table 1 presents the quantitative experimental results of multiple frameworks on three datasets. Overall, PyraE2E demonstrates strong competitiveness across datasets. Compared with non-foundation two-stage baselines, PyraE2E achieves consistently strong performance. Moreover, PyraE2E also surpasses standard end-to-end pipelines, suggesting that the gain does not come from end-to-end training alone. Instead, it stems from turning the intrinsic WSI pyramid into an

Table 2: Ablation studies on TCGA-NSCLC with ABMILX aggregator. (a) Ablation study on the sampling strategy, where "Attn" refers to Attention, and "Emb" indicates whether the strategy is implemented inside the end-to-end training loop without external preprocessing. "Time" reports the additional training time per epoch (min) relative to random sampling. (b) Loss sensitivity study. The total loss is $\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{PIX}} + \beta\mathcal{L}_{\text{FFT}} + \gamma\mathcal{L}_{\text{CE}} + \delta\mathcal{L}_{\text{KLD}}$, where $\mathcal{L}_{\text{PIX}}$ and $\mathcal{L}_{\text{FFT}}$ are super-resolution losses in the pixel and frequency domains, respectively.

(a) Sampling strategy.

Cluster	Attn.	Emb.	ACC	AUC	Time
✗	✗	✓	$89.9_{\pm 0.8}$	$96.3_{\pm 1.3}$	0
✓	✗	✓	$90.7_{\pm 0.9}$	$97.0_{\pm 0.7}$	**+3.3m**
✓	✗	✗	$90.5_{\pm 0.7}$	$96.9_{\pm 0.4}$	+8.1m
✗	✓	✗	$91.1_{\pm 1.0}$	$97.1_{\pm 0.6}$	+9.6m
✓	✓	✓	$\mathbf{92.0}_{\pm 1.4}$	$\mathbf{97.6}_{\pm 0.4}$	+4.2m

(b) Loss sensitivity.

Variant	α	β	γ	δ	ACC	AUC
0.5× SR	.5	.025	.5	.5	$91.2_{\pm 1.3}$	$97.2_{\pm 0.6}$
default	1.0	.05	.5	.5	$\mathbf{92.0}_{\pm 1.4}$	$\mathbf{97.6}_{\pm 0.4}$
2× SR	2.0	.10	.5	.5	$91.8_{\pm 1.2}$	$97.5_{\pm 0.4}$
w/o FFT	1.0	0	.5	.5	$86.2_{\pm 1.8}$	$93.6_{\pm 0.9}$
w/o KLD	1.0	.05	.5	0	$90.5_{\pm 1.4}$	$96.2_{\pm 0.6}$

SR-based self-supervised signal that injects cross-scale dense supervision, alleviating the morphological information loss induced by low-magnification small patches and random sampling, which improves end-to-end representations. To further assess generalization beyond the TCGA cohorts, we also report additional external validation results on Camelyon16 in the Supplementary Material.

PyraE2E achieves a strong balance between accuracy and efficiency. Compared to self-supervised baselines, PyraE2E achieves better performance while requiring lower computational cost in practice. Moreover, PyraE2E achieves performance comparable to foundation models on several datasets, indicating that our framework can approach the performance of foundation models pretrained on large data at lower computational cost. Computational cost is further analyzed in Sec. 4.4.

4.3 Ablation Study

In this section, we conduct ablation studies to verify the effectiveness of the sampling strategy and loss design in PyraE2E. We report ACC and AUC on TCGA-NSCLC.

Ablation on the Sampling Strategy. We evaluate the impact of different sampling strategies in our model, as shown in Table 2a. The combination of cluster and attention achieves the best results, highlighting that our Cluster-Score Sampling module, with both components, offers more efficient sampling for the model. Meanwhile, compared with its embedded counterpart, the external clustering strategy [49] and the external attention strategy [36] incur higher overhead while offering no additional performance gain. This supports our embedded design, which is faster and performs better.

Loss Sensitivity Analysis. We further conducted a loss sensitivity analysis. As shown in Table 2b, PyraE2E remains stable across 0.5×–2× SR weights, with the default setting achieving the best ACC and AUC. This stability suggests that SR does not dominate optimization toward non-diagnostic high-frequency

Table 3: Efficiency comparison on TCGA-NSCLC with ABMILX aggregator. "E2E" indicates end-to-end training, "SSL" indicates self-supervised learning, "FM" indicates foundation-model encoders and "Params" denotes the number of model parameters used for feature extraction and slide-level prediction. All measurements are conducted on a single NVIDIA A800 GPU unless specified. For UNI/CONCH, pretraining cost and training time are taken from their official Hugging Face model cards (multi-A100 setups), where ">" indicates the time would be longer on a single A800. Inference time (s/slide) for UNI/CONCH is measured by us on the same A800 under the same strategy as other methods.

Method	E2E	SSL	FM	Pretrain Cost	Training Time	Inference Time	Params	AUC
SimCLR	✗	✓	✗	80G	6.4d	11.3s	28M	96.7
ICMIL	✗	✓	✗	80G	2.3d	8.2s	9M	95.3
CDSR	✗	✓	✗	80G	2.1d	12.6s	128M	95.9
UNI	✗	✗	✓	32×80G	>1.3d	17.0s	307M	96.1
CONCH	✗	✗	✓	8×80G	>0.9d	7.8s	90M	97.0
C2C	✓	✗	✗	–	**0.8d**	3.2s	24M	94.8
Streaming	✓	✗	✗	–	1.8d	**2.1s**	21M	96.2
Ours	✓	✓	✗	–	1.1d	3.1s	**2M**	**97.6**

noise, but instead serves as auxiliary dense supervision for the shared encoder. This behavior is further supported by the design that SR is constrained by both sampling and joint optimization: Cluster-Score Sampling applies SR supervision to diagnostic regions, as shown in Fig. 4, while $\mathcal{L}_{\mathrm{CE}}$ keeps features aligned with slide labels and $\mathcal{L}_{\mathrm{KLD}}$ avoids attention collapse. Moreover, drops without FFT or KLD indicate the importance of frequency-domain loss and attention regularization. Specifically, $\mathcal{L}_{\mathrm{FFT}}$ complements pixel-domain reconstruction by constraining high-frequency morphological cues, while $\mathcal{L}_{\mathrm{KLD}}$ stabilizes cluster-wise attention and promotes more consistent slide-level aggregation.

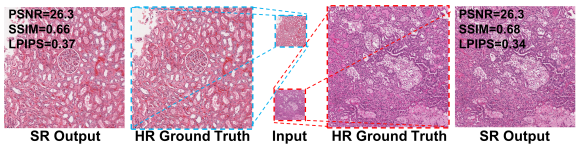
4.4 Computational Cost Analysis

Table 3 summarizes the efficiency comparison in terms of pretrain GPU memory cost, training time, inference time, parameters, and AUC on TCGA-NSCLC. Overall, PyraE2E is more efficient than the self-supervised and foundation-model pipelines in Table 3, with lower training/inference time under comparable evaluation settings, while remaining competitive in AUC. Compared with end-to-end baselines, PyraE2E achieves comparable efficiency with improved performance.

This advantage mainly comes from three factors. First, PyraE2E operates on low-magnification patches with improved sampling, which directly reduces the input cost and shortens training. Second, we adopt a lightweight encoder built on depthwise–pointwise mixing and partitioned spatial interaction, where spatial mixing is applied to only a subset of channels. This design avoids heavy attention blocks and wide feature widths, keeping the parameter count and optimization overhead low. Third, by tackling key optimization challenges in end-to-end training, our framework better realizes two core advantages of end-to-end WSI

Dataset	PSNR	SSIM	LPIPS
TCGA-NSCLC	$25.1_{\pm 4.8}$	$0.67_{\pm 0.17}$	$0.38_{\pm 0.07}$
TCGA-RCC	$24.5_{\pm 2.6}$	$0.64_{\pm 0.14}$	$0.39_{\pm 0.05}$
TCGA-BRCA	$23.7_{\pm 4.5}$	$0.62_{\pm 0.17}$	$0.40_{\pm 0.10}$

(a) Super-resolution metrics.



(b) Visual comparison of super-resolution quality.

Fig. 3: Super-resolution quality and visualization.

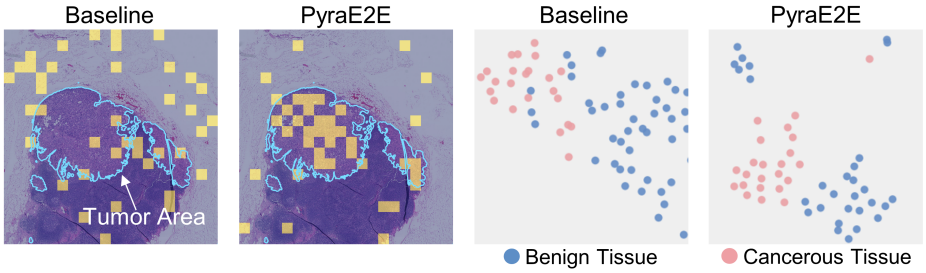


Fig. 4: Feature and sampling visualization. The features and sampling strategy of "Baseline" are taken from the end-to-end baseline [49]. All results are reported on the Camelyon16 dataset [3], which provides patch-level annotations.

learning: (i) it performs feature extraction and slide-level prediction in a single pass, eliminating pretraining cost and enabling faster inference than both self-supervised and foundation-model pipelines; and (ii) it learns task-focused representations optimized directly for the downstream objective, leading to higher performance than self-supervised pipelines.

4.5 Visualization Analysis

Super-Resolution Quality Analysis. To evaluate the effectiveness of the SR branch, we perform $\times 4$ super-resolution by taking 512×512 LR patches as input and reconstructing their spatially aligned 2048×2048 HR counterparts. Given the complexity and multi-scale nature of WSI data, the quantitative results in Fig. 3(a) indicate that the model achieves reasonable fidelity in this challenging cross-scale setting. As shown in Fig. 3(b), the SR reconstructions are visually consistent with the HR ground truth, suggesting that the learned representations capture informative HR cues.

Feature and Sampling Visualization. To verify the effectiveness of our sampling module and instance features, we conduct a visualization analysis, as shown in Fig. 4. The left two figures show the visualization results of the model's sampling. It can be seen that random sampling covers fewer diagnostic regions, while the regions sampled by the Cluster-Score Sampling module better cover the tumor and exhibit diversity, providing better inputs for the model. The right two figures embed the features into a two-dimensional space using UMAP [22]: the

features extracted by the end-to-end baseline [49] are relatively dispersed in the feature space, with obvious overlap between tumor and normal instances. In contrast, our features can aggregate instances of the same class more densely while enlarging the gap between different classes, indicating improved representations.

5 Conclusion

End-to-end learning often suffers from (i) fine-grained and morphological information loss caused by low-magnification inputs and cost-driven patch sampling, and (ii) sparse supervision from slide labels alone. We propose PyraE2E, an end-to-end framework that turns the intrinsic multi-resolution Whole Slide Image (WSI) pyramid into dense self-supervision via super-resolution (SR). PyraE2E mitigates information loss through an embedded Cluster-Score Sampling module to efficiently select informative low-resolution regions. These regions are processed by a shared Global–Local Partitioned (GLP) encoder, which reconstructs spatially aligned high-resolution patches and is jointly optimized for both SR and slide-level prediction. On three TCGA datasets, PyraE2E achieves strong performance, showing SR as an effective auxiliary task for capturing the large-scale and fine-grained morphology of WSIs. Future work includes extending SR to more downstream tasks, designing SR encoders better tailored to WSI morphology, and leveraging SR as a lightweight objective to encode larger regions (even whole slides) with fewer patch partitions to reduce structure fragmentation.

Acknowledgements. This work was funded by the National Natural Science Foundation of China (52373136), the Guangdong Basic and Applied Basic Research Foundation (2024B1515020048), the Science and Technology Projects in Guangzhou (2024A04J3809), and the Science and Technology Projects in Guangzhou (2024D03J0004).

References

1. Afonso, M., Bhawsar, P.M., Saha, M., Almeida, J.S., Oliveira, A.L.: Multiple instance learning for wsi: A comparative analysis of attention-based approaches. Journal of Pathology Informatics **15**, 100403 (2024)
2. An, J., Bai, Y., Chen, H., Gao, Z., Litjens, G.: Masked autoencoders pre-training in multiple instance learning for whole slide image classification. In: Medical imaging with deep learning (2022)
3. Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama **318**(22), 2199–2210 (2017)
4. Campanella, G., Fluder, E., Zeng, J., Vanderbilt, C., Fuchs, T.J.: Beyond multiple instance learning: full resolution all-in-memory end-to-end pathology slide modeling. arXiv preprint arXiv:2403.04865 (2024)
5. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine **25**(8), 1301–1309 (2019)

6. Cao, L., Wang, J., Zhang, Y., Rong, Z., Wang, M., Wang, L., Ji, J., Qian, Y., Zhang, L., Wu, H., Song, J., Liu, Z., Wang, W., Li, S., Wang, P., Xu, Z., Zhang, J., Zhao, L., Wang, H., Sun, M., Huang, X., Yin, R., Lu, Y., Liu, Z., Deng, K., Wang, G., Qiu, M., Li, K., Wang, J., Hou, Y.: E2efp-mil: End-to-end and high-generalizability weakly supervised deep convolutional network for lung cancer classification from whole slide image. Medical Image Analysis **88**, 102837 (2023)
7. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature medicine **30**(3), 850–862 (2024)
8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709 (2020)
9. Chen, Y., Liu, J., Zuo, Z., Jiang, P., Jin, Y., Wu, G.: Classifying pathological images based on multi-instance learning and end-to-end attention pooling. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
10. Cifci, D., Veldhuizen, G.P., Foersch, S., Kather, J.N.: Ai in computational pathology of cancer: improving diagnostic workflows and clinical outcomes? Annual Review of Cancer Biology **7**(1), 57–71 (2023)
11. Cui, M., Zhang, D.Y.: Artificial intelligence and computational pathology. Laboratory Investigation **101**(4), 412–422 (2021)
12. D'Amato, M., Szostak, P., Torben-Nielsen, B.: A comparison between single-and multi-scale approaches for classification of histopathology images. Frontiers in public health **10**, 892658 (2022)
13. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
14. Dooper, S., Pinckaers, H., Aswolinskiy, W., Hebeda, K., Jarkman, S., van der Laak, J., Litjens, G.: Gigapixel end-to-end training using streaming and attention. Medical Image Analysis **88**, 102881 (2023)
15. El-Shafai, W., Ali, A.M., El-Nabi, S.A., El-Rabaie, E.S.M., Abd El-Samie, F.E.: Single image super-resolution approaches in medical images based-deep learning: a survey. Multimedia Tools and Applications **83**(10), 30467–30503 (2024)
16. Fashi, P.A., Hemati, S., Babaie, M., Gonzalez, R., Tizhoosh, H.: A self-supervised contrastive learning approach for whole slide image representation in digital pathology. Journal of Pathology Informatics **13**, 100133 (2022)
17. Gao, Z., Mao, A., Dong, Y., Clayton, H., Wu, J., Liu, J., Wang, C., He, K., Gong, T., Li, C., et al.: Smmile enables accurate spatial quantification in digital pathology using multiple-instance learning. Nature Cancer pp. 1–17 (2025)
18. Guo, Z., Xiong, C., Ma, J., Sun, Q., Feng, L., Wang, J., Chen, H.: Focus: Knowledge-enhanced adaptive visual compression for few-shot whole slide image classification. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15590–15600 (2025)
19. Guo, Z., Zhao, W., Wang, S., Yu, L.: Higt: Hierarchical interaction graph-transformer for whole slide image analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 755–764. Springer (2023)
20. Hatamizadeh, A., Yin, H., Heinrich, G., Kautz, J., Molchanov, P.: Global context vision transformers. In: International conference on machine learning. pp. 12633–12646. PMLR (2023)

21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
22. Healy, J., McInnes, L.: Uniform manifold approximation and projection. Nature Reviews Methods Primers **4**(1), 82 (2024)
23. Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H.: Patch-based convolutional neural network for whole slide tissue image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2424–2433 (2016)
24. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 2127–2136. PMLR (10–15 Jul 2018)
25. Jin, J., Liu, X., Gao, T., Shi, Z., Liang, Y., Zheng, R., Kuang, H., Zeng, M., Kan, S.: Dynamic residual encoding with slide-level contrastive learning for end-to-end whole slide image representation. In: Proceedings of the 33rd ACM International Conference on Multimedia. p. 8389–8398. MM '25, Association for Computing Machinery, New York, NY, USA (2025)
26. Juyal, D., Shingi, S., Javed, S.A., Padigela, H., Shah, C., Sampat, A., Khosla, A., Abel, J., Taylor-Weiner, A.: Sc-mil: Supervised contrastive multiple instance learning for imbalanced classification in pathology. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 7946–7955 (January 2024)
27. Khened, M., Kori, A., Rajkumar, H., Krishnamurthi, G., Srinivasan, B.: A generalized deep learning framework for whole-slide image segmentation and analysis. Scientific reports **11**(1), 11579 (2021)
28. Konz, N., Mazurowski, M.A.: The effect of intrinsic dataset properties on generalization: Unraveling learning differences between natural and medical images. arXiv preprint arXiv:2401.08865 (2024)
29. Lerousseau, M., Vakalopoulou, M., Deutsch, E., Paragios, N.: Sparseconvmil: sparse convolutional context-aware multiple instance learning for whole slide image classification. In: MICCAI Workshop on Computational Pathology. pp. 129–139. PMLR (2021)
30. Li, H., Zhang, Y., Chen, P., Shui, Z., Zhu, C., Yang, L.: Rethinking transformer for long contextual histopathology whole slide image analysis. Advances in Neural Information Processing Systems **37**, 101498–101528 (2024)
31. Li, H., Zhang, Y., Zhu, C., Cai, J., Zheng, S., Yang, L.: Long-mil: scaling long contextual multiple instance learning for histopathology whole slide image analysis. arXiv preprint arXiv:2311.12885 (2023)
32. Li, J., Chen, Y., Chu, H., Sun, Q., Guan, T., Han, A., He, Y.: Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11323–11332 (2024)
33. Li, J., Pei, Z., Li, W., Gao, G., Wang, L., Wang, Y., Zeng, T.: A systematic survey of deep learning-based single-image super-resolution. ACM Computing Surveys **56**(10), 1–40 (2024)
34. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1833–1844 (2021)

35. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)
36. Liu, Y., Lin, Y., Shen, D., Li, H., Wang, Y., Liu, X., Xu, S.: Minimal high-resolution patches are sufficient for whole slide image representation via cascaded dual-scale reconstruction. arXiv preprint arXiv:2508.01641 (2025)
37. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**, 863–874 (2024)
38. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)
39. Luo, X., Qu, L., Guo, Q., Song, Z., Wang, M.: Negative instance guided self-distillation framework for whole slide image analysis. IEEE Journal of Biomedical and Health Informatics **28**(2), 964–975 (2024)
40. Mercan, C., Mercan, E., Aksoy, S., Shapiro, L.G., Weaver, D.L., Elmore, J.G.: Multi-instance multi-label learning for whole slide breast histopathology. In: Medical Imaging 2016: Digital Pathology. vol. 9791, pp. 48–58. SPIE (2016)
41. Peng, Z., Huang, W., Gu, S., Xie, L., Wang, Y., Jiao, J., Ye, Q.: Conformer: Local features coupling global representations for visual recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 367–376 (2021)
42. Pinckaers, H., Bulten, W., Van der Laak, J., Litjens, G.: Detection of prostate cancer in whole-slide images through end-to-end training with image-level labels. IEEE Transactions on medical imaging **40**(7), 1817–1826 (2021)
43. Pinckaers, H., Van Ginneken, B., Litjens, G.: 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)
44. Qiu, D., Cheng, Y., Wang, X.: Medical image super-resolution reconstruction algorithms based on deep learning: A survey. Computer Methods and Programs in Biomedicine **238**, 107590 (2023)
45. Qu, L., Ma, Y., Luo, X., Guo, Q., Wang, M., Song, Z.: Rethinking multiple instance learning for whole slide image classification: A good instance classifier is all you need. IEEE Transactions on Circuits and Systems for Video Technology **34**(10), 9732–9744 (2024)
46. Qu, L., Wang, M., Song, Z., et al.: Bi-directional weakly supervised knowledge distillation for whole slide image classification. Advances in Neural Information Processing Systems **35**, 15368–15381 (2022)
47. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: Understanding transfer learning for medical imaging. Advances in neural information processing systems **32** (2019)
48. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)
49. Sharma, Y., Shrivastava, A., Ehsan, L., Moskaluk, C.A., Syed, S., Brown, D.: Cluster-to-conquer: A framework for end-to-end multi-instance learning for whole slide image classification. In: Medical imaging with deep learning. pp. 682–698. PMLR (2021)
50. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1874–1883 (2016)

51. Song, A.H., Jaume, G., Williamson, D.F., Lu, M.Y., Vaidya, A., Miller, T.R., Mahmood, F.: Artificial intelligence for digital and computational pathology. Nature Reviews Bioengineering **1**(12), 930–949 (2023)
52. Sun, K., Gao, Y., Xie, T., Wang, X., Yang, Q., Chen, L., Wang, K., Yu, G.: Multiscale super-resolution generation of low-resolution scanned pathological images. arXiv preprint arXiv:2105.07200 (2021)
53. Tang, W., Qin, R., Fang, H., Zhou, F., Chen, H., Li, X., Cheng, M.M.: Revisiting end-to-end learning with slide-level supervision in computational pathology. arXiv preprint arXiv:2506.02408 (2025)
54. Tang, W., Zhou, F., Huang, S., Zhu, X., Zhang, Y., Liu, B.: Feature re-embedding: Towards foundation model-level performance in computational pathology. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11343–11352 (2024)
55. Tang, Y., He, Y., Nath, V., Guo, P., Deng, R., Yao, T., Liu, Q., Cui, C., Yin, M., Xu, Z., et al.: Holohisto: End-to-end gigapixel wsi segmentation with 4k resolution sequential tokenization. arXiv preprint arXiv:2407.03307 (2024)
56. Wang, H., Luo, L., Wang, F., Tong, R., Chen, Y.W., Hu, H., Lin, L., Chen, H.: Iteratively coupled multiple instance learning from instance to bag classifier for whole slide image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 467–476. Springer (2023)
57. Wang, J., Mao, Y., Guan, N., Xue, C.J.: Advances in multiple instance learning for whole slide image analysis: Techniques, challenges, and future directions. arXiv preprint arXiv:2408.09476 (2024)
58. Wang, W., Ma, S., Xu, H., Usuyama, N., Ding, J., Poon, H., Wei, F.: When an image is worth 1,024 x 1,024 words: A case study in computational pathology (2023)
59. Wang, X., Xiang, J., Zhang, J., Yang, S., Yang, Z., Wang, M.H., Zhang, J., Yang, W., Huang, J., Han, X.: Scl-wc: Cross-slide contrastive learning for weakly-supervised whole-slide image classification. Advances in neural information processing systems **35**, 18009–18021 (2022)
60. Wu, K., Zheng, Y., Shi, J., Xie, F., Jiang, Z.: Position-aware masked autoencoder for histopathology wsi representation learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 714–724. Springer (2023)
61. Xu, J., Shi, L., Zhang, Y., Zhao, G., Gao, Y.: A joint magnification and attention sampling based cascade network for brca mutation classification from histopathology images. In: International Conference on Advanced Data Mining and Applications. pp. 50–65. Springer (2025)
62. Yang, S., Wang, Y., Chen, H.: Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. In: International conference on medical image computing and computer-assisted intervention. pp. 296–306. Springer (2024)
63. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y.: Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18802–18812 (2022)
64. Zhang, J., Hao, F., Liu, X., Yao, S., Wu, Y., Li, M., Zheng, W.: Multi-scale multi-instance contrastive learning for whole slide image classification. Engineering Applications of Artificial Intelligence **138**, 109300 (2024)

65. Zhang, J., Ma, K., Van Arnam, J., Gupta, R., Saltz, J., Vakalopoulou, M., Samaras, D.: A joint spatial and magnification based attention framework for large scale histopathology classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3776–3784 (2021)
66. Zhang, J., Nguyen, A.T., Han, X., Trinh, V.Q.H., Qin, H., Samaras, D., Hosseini, M.S.: 2dmamba: Efficient state space model for image representation with applications on giga-pixel whole slide image classification. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3583–3592 (2025)
67. Zheng, M., Sun, L., Dong, J., Pan, J.: Smfanet: A lightweight self-modulation feature aggregation network for efficient image super-resolution. In: European conference on computer vision. pp. 359–375. Springer (2024)
68. Zheng, T., Jiang, K., Yao, H.: Dynamic policy-driven adaptive multi-instance learning for whole slide image classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8028–8037 (2024)
69. Zheng, Y., Sharma, H., Betke, M., Cherry, J.D., Mez, J.B., Beane, J.E., Kolachalama, V.B.: Fouriermil: Fourier filtering-based multiple instance learning for whole slide image analysis: Y. zheng et al. International Journal of Computer Vision **134**(1), 26 (2026)